1 Today’s Agenda

1. Market Efficiency
2. Variance Ratio Tests
3. Forecastability of Returns: Univariate and Bivariate Systems
4. Long-Horizon Forecasts of Returns
5. Multivariate Systems (VAR): Reduced forms
6. A Basic Structural model: Consumption-based model
7. Risk Return Trade-Off (ICAPM)
8. ARCH
9. GMM

Warm-up question: Is FED policy risk priced? Can you answer this question within the framework of the APT?
2 Market Efficiency (Chapter 2)

- Earlier Models of Asset Pricing:
  - Fair Game: Cardano’s (c.1565) *Liber de Ludo Aleae* (CLM, 1997)

  The most fundamental principle of all in gambling is simply equal conditions, e.g. of opponents, of bystanders, of money, of situation, of the dice box, and of the die itself. To the extent to which you depart from that equality, if it is in your opponent’s favor, you are a fool, and if in your own, you are unjust.

- We have already seen the market efficiency condition in class 1, but now we will go deeper and provide a test based on the price process.

  \[ E(P_{t+1} | I_t) = P_t \]

  which implies that

  \[ P_{t+1} = P_t + \varepsilon_{t+1} \]

  with \[ E(\varepsilon_{t+1} | I_t) = 0 \]

- A martingale is a special case of the above. If \[ I_t = \{ P_t, P_{t-1}, \ldots \} \], so that

  \[ E(P_{t+1} | P_t, P_{t-1}, \ldots) = P_t \]
• Note: This is an AR(1) model with $\phi = 1$ (non-stationary).

• The martingale condition (also known as a “Random Walk”) places restrictions only on the first moments on the process. However, it is silent on second, third, fourth, etc. moments.
  – Specifically, we know that there has got to be a trade-off between risk and return.

• LeRoy (1973), Lucas (1978): The Martingale condition is neither a necessary nor a sufficient condition for rational expectations models of asset prices.

• However, the martingale is an important starting point.
• There are three versions of the Random Walk model.

• The variations place various restrictions on the innovations (surprise, news) process, $\varepsilon_t$ in

$$P_t = \mu + P_{t-1} + \varepsilon_t$$

– $\varepsilon_t$ are i.i.d $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$.

– If we place some restrictions, we can drop the "identically distributed" condition, but keep the "independent" condition. I.e. $\varepsilon_t$ are independent.

– $\varepsilon_t$ are uncorrelated (but they might be dependent).

• Note that the above process has a "stochastic trend". To see why, notice that, if we solve recursively the above process, we have

$$P_t = \mu t + \sum_{i=1}^{t} \varepsilon_i + P_0$$

– $E(P_t) = \mu t$

– $Var(P_t) = \sigma^2 t$

– Note that all three forms have the same two moments.

• Review Q: Do you see why $P_t$ is not (covariance) stationary?
- Is the process $P_t = \mu + P_{t-1} + \varepsilon_t$ adequate to model stock prices? Here is a hint:
We will use \( p_t = \log P_t \) (or \( P_t = e^{p_t} \)) and

\[
p_t = \mu + p_{t-1} + \varepsilon_t
\]

so that \( p_t = \mu t + \sum_{i=1}^{t} \varepsilon_t + p_0 \), or to get back to prices, \( P_t = e^{P_0} e^{\mu t} e^{\sum_{i=1}^{t} \varepsilon_t} \).

- Note that \( p_t - p_{t-1} = \mu + \varepsilon = r_t \) (in our previous notation), is the continuously compounded return.

- If we assume that \( \varepsilon_t \) (and hence \( r_t \)) is normally distributed, then we are back to the case where \( r_t \) is normally distributed and \( 1 + R_t = e^{r_t} \) is lognormally distributed.

- The process \( p_t = \mu + p_{t-1} + \varepsilon_t \) is a discrete (arithmetic) Brownian motion

- The process \( P_t = e^{p_t} \) is a discrete geometric Brownian motion.
• Note: An immediate observation of the above specification for prices is that returns are not forecastable by past returns or by any other conditioning variable, since

\[ r_t = \mu + \varepsilon_t \]

and

• \( \varepsilon_t \) are either uncorrelated or independent.
• \( \mu \) is the mean and it does not vary with time.
3 Variance Ratio Tests

- The variance ratio test is a neat and “natural” way to test market efficiency.

- We cannot work with prices $p_t$ or $P_t$, because those series are non-stationary. So far, we know how to work only with stationary time-series.

- Recall: DO NOT work with non-stationary time series, unless you know what you are doing.

- The non-stationary processes of interest to us ($p_t$) can be stationarized....How?

- Therefore, we will have a preference to work with ....
• As we saw above, an implication of the Random Walk model of prices is that:
  – Continuously compounded (or log) returns are unforecastable
  – The increments (log returns) are uncorrelated (or independent, or iid)
• Here is a useful observation about the variance of a two period (log) return \( r_{t,t+2} \)
  \[
  Var (r_{t,t+2}) = Var (r_{t+1} + r_{t+2}) \\
  = Var (r_{t+1}) + Var (r_{t+2}) \\
  = 2\sigma_r^2
  \]
• In general, for the \( q \)-th period (long-horizon) return \( r_{t,t+q} \):
  \[
  Var (r_{t,t+q}) = Var (r_{t+1} + \ldots + r_{t+q}) \\
  = q\sigma_r^2
  \]
• Those implications were derived under the hypothesis (the null) that returns are unforecastable.
• So, we have a “natural” test for the Random Walk model (or equivalently, for unforecastable returns):

\[ \text{Var}(r_{t,t+q}) \underset{\text{Under the Null of a Random Walk}}{=} q \text{Var}(r_{t+1}) \]

or

\[ VR(q) = \frac{\text{Var}(r_{t,t+q})}{q \text{Var}(r_{t+1})} = 1 \]

• This is a simple test.

• Compute the variance of the \( q \)-period returns.

• Compute the variance of the 1-period returns.

• It must be the case that the ratio of \( \frac{\text{Var}(r_{t,t+q})}{q \text{Var}(r_{t+1})} \) must be close to 1.

• Q: How close? We should not forget that the test \( VR(q) \) is a random variable with a corresponding density, etc.

• Q: For what values of \( VR(q) \) is the model a RW and for what values it is not?

• ASIDE: We can relate the \( VR(q) \) test to autocorrelations (see CLM, 1997).

• ASIDE: What \( q \) to choose?
• So, under the null hypothesis that \( r_t \) are i.i.d., we can use the CLT to show that

\[
\sqrt{Tq} \left( \frac{VR(q) - 1}{Tq} \right) \sim^{a} \mathcal{N}(0, 2(q - 1))
\]

(recall last lecture and the results about \( \overline{X} \) and \( \hat{\beta} \))

or

\[
\left( \frac{VR(q) - 1}{Tq} \right) \sim^{a} \mathcal{N} \left( 0, \frac{2(q - 1)}{Tq} \right)
\]

• Therefore, we can conduct testing in the usual way: Form a statistic

\[
Z = \frac{VR(q) - 1}{\sqrt{\frac{2(q - 1)}{Tq}}} \sim^{a} \mathcal{N}(0, 1)
\]

• So, if \( Z \) is greater than 1.96, we reject the null of IID returns at the 5% level.
• Note: We must be careful when computing the long-horizon returns \( r_{t,t+q} \) from overlapping observations.

• Q: Why?

• A: Because, under the null hypothesis, \( r_{t,t+q} \) are i.i.d. If we use overlapping observations, they will NOT be i.i.d. by construction.

• Some people advocate the use of overlapping observations and correcting for correlation from the overlap as:

\[
\sqrt{Tq} \left( \frac{1}{V R_{\text{overlap}} (q)} - 1 \right) \sim a N \left( 0, \frac{2 (2q - 1) (q - 1)}{3q} \right)
\]

• The correction of the variance is supposed to correct for the overlap. At the same time we have more observations, so the test might be “better”, or more powerful.

• People are deluding themselves!

• If we correct exactly for the overlap, whether we use overlap or not would make no difference. But the uncertainty introduced around how to deal with the overlap makes the second test less desirable.

• At the end of the day, we have the same amount of information (returns), not more.
• NOTE: As part of your homework, you will have to program the $VR$ test using non-overlapping and overlapping returns and for different portfolios.

• Compare your results to those in CLM (chapter 2).

• Answer the question: Do you believe the RW hypothesis?
4 Forecastability of Returns: Univariate and Bivariate Systems (CLM, Chapters 2 and 7)

• Thus far, we defined returns \( r_t = \log (1 + R_t) \) as

\[
\begin{align*}
t_r &= p_t - p_t-1 \\
   &= \log \frac{P_t}{P_{t-1}}
\end{align*}
\]

• Of course, if \( r_t = \mu + \varepsilon_t \) is true, we can also run simple regression tests:
  – We can write \( r_t = \mu + \phi r_{t-1} + \varepsilon_t \), where under the null hypothesis, \( \phi = 0 \).
  – Run: \( r_t \) on \( r_{t-1} \) (this is an AR(1) model).
  – Test that the parameter \( \phi \) is equal to zero.
  – Those tests are perfectly valid and can be related to the \( VR \) test.
• ASIDE: There is some evidence that, at short horizons, \( r_t \) is positively correlated for some stocks. That implies that if returns of certain stocks are higher than average at \( t - 1 \), they would tend to be higher than average at \( t \). (Jagadeesh and Titman)

• This is called MOMENTUM and it is not uncontroversial.

• Momentum disappears at longer horizon and even “reverses” itself.
  – Q: If agents are rational and markets are efficient, why don’t people take advantage of momentum?
  – Q: Can we find “rational” explanations of momentum?

• There is no agreed upon consensus on momentum.
• If

\[ r_t = p_t - p_{t-1} \]
\[ = \log \frac{P_t}{P_{t-1}} \]

• Then

\[ e^{r_t} = (1 + R_t) = \frac{P_t}{P_{t-1}} \]
\[ R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \]

• Note: This definition has no place for dividends. But dividends must carry important information about the viability of a company.

• Can it be the case that, if we do the calculations correctly and incorporate dividends, then that information might forecast returns?
• The exact definition of a return is

\[ R_{t+1} = \frac{P_{t+1} + D_{t+1} - P_t}{P_t} \]

or

\[ 1 + R_{t+1} = \frac{P_{t+1}}{P_t} + \frac{D_{t+1}}{P_t} \]

• Therefore, the increase in returns can come either from an increase in the stock price or from a dividend payment.

• Here, we will use a trick that has proven useful.

• Take logs

\[ \log (1 + R_{t+1}) = \log (P_{t+1} + D_{t+1}) - \log (P_t) \]

• Now, I will do a trick using the Taylor’s expansion (see CLM, p. 261)

\[ r_{t+1} \approx k + \rho p_{t+1} - p_t + (1 - \rho) d_{t+1} \]

where

- \( d_{t+1} = \log (D_{t+1}) \) is the log dividend
- \( k \) is a constant term
- \( \rho \) is a constant of linearization that we will treat as a parameter.
• The above transformation is “cool” because now returns, prices, and dividends are related by a linear relation.

• But there is a problem: This is not exactly a regression, because prices and dividends are not stationary....Remember.

• But, we can write

\[ p_t \approx k + \rho p_{t+1} - r_{t+1} + (1 - \rho) d_{t+1} \]

• Solve FOREWARD (not backward as before)

\[ p_{t+1} \approx k + \rho p_{t+2} - r_{t+2} + (1 - \rho) d_{t+2} \]

or

\[ p_t \approx k + \rho [k + \rho p_{t+2} - r_{t+2} + (1 - \rho) d_{t+2}] - r_{t+1} + (1 - \rho) d_{t+1} \]

\[ \approx K + \rho^2 p_{t+2} - \rho r_{t+2} - r_{t+1} + \rho (1 - \rho) d_{t+2} + (1 - \rho) d_{t+1} \]

\[ \approx \ldots \]

\[ \approx K + \rho^j p_{t+j} - \sum_{i=0}^{j} \rho^i r_{t+1+i} + \sum_{i=0}^{j} \rho^i (1 - \rho) d_{t+1+i} \]

• Now, we have to assume a no-bubbles condition (interpretation):

\[ \lim_{j \to \infty} \rho^j p_{t+j} = 0 \]
Therefore:
\[ p_t \approx K - \sum_{i=0}^{\infty} \rho^i r_{t+1+i} + \sum_{i=0}^{\infty} \rho^i (1 - \rho) d_{t+1+i} \]

- This is quite intuitive: If price is “high”, it means that either
  - dividends in the future would be high
  - returns in the future would be low
  - This presumes a no-bubbles condition. Note a difference between fundamentals and the price can exist, but it cannot last indefinitely.
• The last step is to subtract the above expression from \(d_t\) to obtain:

\[
d_t - p_t = K - \sum_{i=0}^{\infty} \rho^i (1 - \rho) \Delta d_{t+1+i} + \sum_{i=0}^{\infty} \rho^i r_{t+1+i}
\]

• This is a very, very useful relation, because
  – The log dividend price ratio \((d_t - p)\) is stationary
  – \(\Delta d_{t+1+i}\) is stationary
  – \(r_{t+1+i}\) is stationary.
  – The log dividend price ratio must forecast either future dividend growth or future returns
  – This is the relation we had obtained a few slides ago, but now it is in a “regression” form.
  – We can use regressions to test this relation.

• The last observation is that we are dealing with future values. But this has never stopped us.

\[
d_t - p_t = K - E_t \left\{ \sum_{i=0}^{\infty} \rho^i (1 - \rho) \Delta d_{t+1+i} + \sum_{i=0}^{\infty} \rho^i r_{t+1+i} \right\}
\]
Before going further, let’s take a step back and think of what we have done so far:

• In the absence of dividends, we argued that returns are unforecastable

• Now, we argue that dividends can help us forecast returns.

• Isn’t that inconsistent?
Testing the above relation:

- Regress: $r_{t+1}$ on $d_t - p_t$
- Regress: $\Delta d_{t+1}$ on $d_t - p_t$
- Which relationship is stronger?
- Note: This is a forecasting relationship: the dividend price ratio today helps us forecast tomorrow’s (expected) returns
- It is different from the CAPM, APT, where we were trying to explain ex-post variations in returns.
- Here, we are doing an ex-ante regression.
5 Long-Horizon Regressions:

People have noticed that the relation
\[ d_t - p_t = K - \sum_{i=0}^{\infty} \rho^i (1 - \rho) \Delta d_{t+1+i} + \sum_{i=0}^{\infty} \rho^i r_{t+1+i} \]
might not hold in the short run, but it has got to hold in the long run.

• The above observation has prompted people to run regressions:
  \[ -r_{t+1} + \ldots + r_{t+k} = \alpha_1 + \beta_1 (d_t - p_t) + u_{t+k} \]
  \[ -\Delta d_{t+1} + \ldots + \Delta d_{t+k} = \alpha_2 + \beta_2 (d_t - p_t) + e_{t+k} \]

• But then, it must be the case that
  \[ \beta_1 + \beta_2 = 1 \]
  by definition. We have not used anything, but the definition of returns, no-bubbles condition, and a log-linearization.

• The above restriction is never used in practice. How can it be implemented?
• Remark: As in the variance ratio tests, it is important to run the long horizon regressions using non-overlapping returns.

• If we use overlapping returns, we get spurious results.

• Correcting for the overlap is difficult and introduces another layer of complication.
Vector Autoregressions (VAR)

- First, we should not confuse VAR (vector autoregression) with VaR (Value at Risk).
- The VAR is a natural generalization of the autoregressive process, for multivariate series.
- Suppose we are interested in the joint, dynamic interaction between a few series, say returns and volatility.
  - We suppose that higher volatility must lead to higher returns
  - We also think that there might be some feedback effect from returns to volatility Campbell and Hentschel (1992).
  - Ultimately, we are not sure about the dynamic relationship between those two series
Then, we write

\[ Y_t = \begin{bmatrix} r_t \\ \sigma_t \end{bmatrix} \]

where \( r_t \) is the return and \( \sigma_t \) is its volatility (standard deviation) at time \( t \). We want to write the following dynamic relationship

\[
\begin{align*}
    r_t &= \alpha_1 + \beta_{11} r_{t-1} + \beta_{12} \sigma_{t-1} + \varepsilon_{1,t} \\
    \sigma_t &= \alpha_2 + \beta_{21} r_{t-1} + \beta_{22} \sigma_{t-1} + \varepsilon_{2,t}
\end{align*}
\]

- Both series depend on their own lagged values (as in the AR process)
- Both series depend on the lagged realization of the other process
- The residuals \([\varepsilon_{1t} \ \varepsilon_{2t}]\) have a covariance matrix \( \Sigma \).
- This system can be written in a more elegant form as:

\[
Y_t = \alpha + \Phi Y_{t-1} + \varepsilon_t
\]

where \( \Phi = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \) and \( \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \).
The beauty of VARs

- Very easy to estimate: Run regressions of variables on their lagged values and on the lagged values of all other variables.

- Run the regressions, equation by equation, and then stack them together.

- We don’t need a priori theory to know what variable causes what variable, etc.

- The VARs help us make the notion of statistical “causality” very precise.
Statistical (Granger) causality:
- Example: Martian observes that when people bring umbrellas, it starts raining. She concludes that people bringing umbrellas causes rain.
- But it is the other way around...The anticipation of rain makes people bring their umbrellas.
- In this example, the statistical causality runs from “umbrellas” to “rain”
- The structural causality runs from “rain” to “umbrellas”.
- Weather and orange juice futures example. – argument for market efficiency.
- Problem: In empirical work, if we don’t have a model (or more information about the system), we cannot distinguish between the two alternatives.
Causality in a bivariate VAR:

\[ r_t = \alpha_1 + \beta_{11} r_{t-1} + \beta_{12} \sigma_{t-1} + \epsilon_{1,t} \]
\[ \sigma_t = \alpha_2 + \beta_{21} r_{t-1} + \beta_{22} \sigma_{t-1} + \epsilon_{2,t} \]

- Suppose that \( \beta_{12} = 0 \). Then \( r_t = \alpha_1 + \beta_{11} r_{t-1} + \epsilon_{1,t} \), or \( r_t \) is not affected by \( \sigma_{t-1} \). In such a case, we say that \( \sigma_{t-1} \) does not Granger-cause (or just cause) \( r_t \).

- However, since \( \beta_{21} \neq 0 \), \( r_{t-1} \) does Granger-cause \( \sigma_t \).

- It is important to understand that Granger-causality gives us the timing (umbrella, then rain) but not the economic story.

- This is a very common mistake in academia, in practice, and in everyday life.
Test for Granger-causality:

- We run the two regressions.
- For the hypothesis: \( \sigma_{t-1} \text{ Granger-causes } r_t \), we test \( \beta_{12} = 0 \)
- For the hypothesis: \( r_{t-1} \text{ Granger-causes } \sigma_t \), we test \( \beta_{21} = 0 \)
- It is always a good idea to start an empirical work with some background Granger-causality tests, be it only to get a feel for the data.
- But it is not a good idea to start weaving economic stories based on that evidence alone.
- We will come back to VARs.
Here is a real research problem (Goto and Valkanov (2001)):

- Fact: The effect of the FED actions on the stock market is negative. In other words, a contractionary monetary policy (aimed at curbing inflation) will result in lower returns for some time in the future.

- Question: Why is this so? There are two possibilities
  - Fed has a better forecast of the state of the economy, but its policy has no real effect on stock fundamentals (the umbrella). This is only Granger-causation without structural effect.
  - Fed has an impact on the economy and by contracting the economy, cash flows go down, returns decrease (the cloud). This is Granger-causation and a structural effect.

- Q: Is there something about the stock market that would help us distinguish between the two alternatives?
6 A Basic Structural model: Consumption-based model (CCAPM)

- So far, we have estimated the APT and the CAPM
- The CAPM and the APT capture risk and return, but are their related to our more fundamental needs: consumption of goods.
- Some of you have asked me to clarify: What do you mean by “The equity premium puzzle is too high”?
- We will work out a “simple” model where assets are priced explicitly relative to our utility from consumption.
- This explicit model will generate a familiar stochastic discount factor pricing relation.
• First, we have to model the behavior of a representative investor

• Think of this investor as the average person in the economy.

• The investor invests primarily so he will consume goods (bread, cheese, or Ferraris) tomorrow.

• The utility function of this investor, today and tomorrow is:

\[ U(c_t, c_{t+1}) = u(c_t) + \beta u(c_{t+1}) \]

where \( c_t \) denotes consumption at date \( t \)

• \( \beta \) is a subjective discount factor, usually \( \beta < 1 \).

• Note: Some people have argued that \( \beta \geq 1 \) (Behavioral finance–fitness club experiments).

• Ultimately, we won’t be estimating \( \beta \). Set \( \beta = 0.995 \).
• At the beginning, we will work with \( u(c_t) \), where
  – \( u(.) \) is concave, reflecting a decreasing marginal
    value of consumption
  – \( u(.) \) is increasing, reflecting a insatiable desire for
    more consumption
  – the curvature generates aversion to risk and to
    inter-temporal substitution: The investor prefers a
    consumption stream that is steady over time and
    across states of nature.

• But to make the model operational (estimable), we
  have to give a functional form to \( u(.) \).

• We will assume that

\[
    u(c_t) = \frac{1}{1 - \gamma} c_t^{1-\gamma}
\]

where \( \gamma \) captures the curvature.
Here is the game:
- We want to value an asset with an uncertain payoff $x_{t+1}$
- It is a two-period problem (today/tomorrow or young/old, etc.)
- The problem is:
  \[
  \max_{\zeta} u(c_t) + E_t [\beta u(c_{t+1})] \\
  \text{such that} \quad c_t = e_t - p_t \zeta \\
  c_{t+1} = e_{t+1} + x_{t+1} \zeta
  \]
- $e_t$ is the original endowment of the individual (cash he inherited from his parents)
- At time $t$, he has an endowment $e_t$.
- He decided to purchase $\zeta$ shares of the asset at price $p_t$
- Whatever is left over after the purchase of the asset $e_t - p_t \zeta$, is used for consumption.
- At time $t + 1$, we has an endowment $e_{t+1}$ but also the payoff from the asset $x_t$ (think $p_{t+1} + d_{t+1}$) time the number of shares.
- Since the individual “dies” at $t + 1$, he must consume everything, so $c_{t+1} = e_{t+1} + x_{t+1} \zeta$.
- No bequest—Can’t analyze inheritance here.
• To recapitulate: We want to find the number of shares that would be bought at price $p_t$, given that the payoff from this investment is $x_{t+1}$.

• Implicitly, we will find the price of the asset as a function of the payoff and everything else.

• This is called an equilibrium model.

• This is also a structural model.

• Solving the maximization problem yields the FOC:
  
  $$p_t u'(c_t) = E_t [\beta u'(c_{t+1}) x_{t+1}]$$

  or

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

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

is quite interesting.

1. Note that if we write \( m_{t+1} = \beta \frac{u'(c_{t+1})}{u'(c_t)} \) and interpret \( m_{t+1} \) as the stochastic discount factor, then we get the familiar pricing equation, where \( m_{t+1} \) is a function of consumption. Hence, this is called a consumption-based pricing model.

\[ p_t = E_t [m_{t+1} x_{t+1}] \]

2. Note that we cannot go beyond this point without specifying a functional form for \( u() \), and hence, for \( u'(c) \). In the above parameterization, \( u'(c) = c^{-\gamma} \) and the FOC is:

\[ p_t = E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} x_{t+1} \right] \]

3. Now, we can rewrite the model as:

\[ 1 = E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \frac{x_{t+1}}{p_t} \right] = E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} (1 + R_{t+1}) \right] \]

4. If we have data on \( R_{t+1} \) and on \( c_t \), we can think of a way to estimate the parameters \( \gamma \) and \( \beta \).

5. Problem: The above relationship is nonlinear...we only know how to run linear regressions....
• Here is a good insight: The equation

\[ 1 = E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} (1 + R_{t+1}) \right] \]

must hold for any asset. Indeed, we did not specify a particular asset when deriving the above equations.

– Let’s consider a particular asset, the risk free asset with return \( R^f \) (and forget uncertainty and expectations for a moment)

– The pricing equation can be rewritten as

\[ (1 + R^f) = \frac{1}{\beta} \left( \frac{c_{t+1}}{c_t} \right) ^\gamma \]

– Take logs:

\[ \ln(1 + R^f) = r^f = - \ln \beta + \gamma (\ln c_{t+1} - \ln c_t) \]

– Note that \( \ln c_{t+1} - \ln c_t \) is nothing but the growth in consumption between \( t + 1 \) and \( t \).

– We can estimate the above equation if we have data on \( r^f \) and \( c_t \).

– What is the interpretation of \( \gamma \)?
• Now, we have to take care of the uncertainly. For that we will use log-normality

• If $X$ is conditionally lognormally distributed, it has the convenient property

$$\ln E_t (X) = E_t (\ln(X)) + \frac{1}{2} \text{Var}_t (\ln(X))$$

• Recall our discussion of: $g(E(X)) \neq E(g(X))$. In the above example, $g() = \ln()$.

• In addition, we will assume that $\text{Var}_t (\ln(X)) = \text{Var} (\ln(X)) = \sigma_x^2$. (What is this assumption called?)
• Recall

\[ 1 = E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} (1 + R_{t+1}) \right] \]

• Taking logs

\[ 0 = \ln E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} (1 + R_{t+1}) \right] \]

\[ = E_t \left[ \ln \left( \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} (1 + R_{t+1}) \right) \right] + \]

\[ \frac{1}{2} Var \left( \ln \left( \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} (1 + R_{t+1}) \right) \right) \]

\[ = E_t \left[ \ln (1 + R_{t+1}) + \ln \beta - \gamma (\ln c_{t+1} - \ln c_t) \right] + \]

\[ \frac{1}{2} Var \left( \ln (1 + R_{t+1}) + \ln \beta - \gamma (\ln c_{t+1} - \ln c_t) \right) \]

\[ = E_t r_{t+1} + \ln \beta - \gamma E_t \left[ (\ln c_{t+1} - \ln c_t) \right] + \]

\[ \frac{1}{2} \left[ \sigma_r^2 + \gamma^2 \sigma_{\Delta c}^2 - 2 \gamma \sigma_r \Delta c \right] \]
- The equation

\[ 0 = \mathbb{E}_t r_{t+1} + \ln \beta - \gamma \mathbb{E}_t [(\ln c_{t+1} - \ln c_t)] + \frac{1}{2} \left[ \sigma_r^2 + \gamma^2 \sigma_{\Delta c}^2 - 2 \gamma \sigma_{r,\Delta c} \right] \]

is valid for any asset. It holds for the risk free asset, where \( \sigma_{rf}^2 = \sigma_{rf,\Delta c} = 0 \). Therefore, we can write

\[ r_f = -\ln \beta + \gamma \mathbb{E}_t [(\ln c_{t+1} - \ln c_t)] - \frac{1}{2} \gamma^2 \sigma_{\Delta c}^2 \]

- Comparing this formula to what we had before, the term \(-\frac{1}{2} \gamma^2 \sigma_{\Delta c}^2\) adjusts for the variance, or uncertainly in consumption, provided that all processes are lognormal.
• The advantage of this formula is that we can handle uncertainly and any other asset.

• From

\[ E_t r_{t+1} = -\ln \beta + \gamma E_t ((\ln c_{t+1} - \ln c_t)] - \frac{1}{2} [\sigma_r^2 + \gamma^2 \sigma_{\Delta c}^2 - 2\gamma \sigma_r \Delta c] \]

• And

\[ r^f = -\ln \beta + \gamma E_t ((\ln c_{t+1} - \ln c_t)] - \frac{1}{2} \gamma^2 \sigma_{\Delta c}^2 \]

we obtain the cool expression

\[ E_r r_{t+1} - r^f = -\gamma \sigma_r \Delta c - \frac{\sigma_r^2}{2} \]

\[ = -\gamma Cov \left( r_{t+1}, (\ln c_{t+1} - \ln c_t) \right) - \frac{Var \left( r_{t+1} \right)}{2} \]

• In words, the excess return is equal to the covariance of the asset return with consumption growth.

• What is this result reminiscent of?

• This is called the Consumption CAPM (or CCAPM)?

• But then:
  – The CAPM does not hold. Does CCAPM hold?
  – Why is this model better? (before \( \beta \), now \( \gamma \))
• The CCAPM is as successful as the CAPM, or even less, but
  – The coefficient $\gamma$ has a very nice interpretation: It measures our aversion to risk.
  – We have a consumption variable in the pricing kernel.
  – To test the CAPM we needed the market portfolio (Roll’s critique). Similarly, now we need consumption.
  – The CCAPM (with the added log-linearity restrictions) is easy to test using regressions.
• Note that we have two regressions that we can run in order to estimate $\gamma$

• First, using the riskless rate

$$r^f = -\ln \beta + \gamma E_t [(\ln c_{t+1} - \ln c_t)] - \frac{1}{2} \gamma^2 \sigma^2_{\Delta c}$$

• Second, using the risky rate

$$E_r r_{t+1} - r^f = -\gamma \sigma_{r, \Delta c} - \frac{\sigma^2_r}{2}$$

• Note that both equations must give us the same result (statistically speaking, at least).

• The trouble is that the estimates of $\gamma$ in those regressions are in total disagreement.

• Either the risky rate has been “too high” or the riskless rate has been “too low” to reconcile the model with the data.

• What’s next: The assumption that the risk premium $E_r r_{t+1} - r^f$ does not vary with time has been seen, lately, as being particularly bothersome.

• A few models have tried to relax this assumption, while keeping the economic story of the model.
7 Risk Return Trade-Off (ICAPM)

- One of the main tenets of modern finance is that we have to be compensated with higher returns if we are exposed to higher risk.
- We measure risk with the variance.
- So far, we have assumed that $\text{Var}(r_t) = \sigma^2$.
- This is clearly an untenable assumption.
- There is no reason why the variance of return will stay constant.
- Why does the variance fluctuate (Schwert (1990))?
  - Macroeconomic factors
  - Microeconomic factors
  - Who knows why?
- Bottom line: $\text{Var}(r_t) = \sigma^2_t$ (What does this imply for returns?)
- Modelling $\sigma_t$ will be a separate discussion.
• Suppose we have an estimate of $\sigma_t$.

• We can run the regression

$$r_t = \alpha + \beta \sigma_{t-1} + \varepsilon_t$$

• Under the null hypothesis that there is a risk-return trade-off, we expect $\beta > 0$.

• If the relationship does not hold, there are several possibilities:
  – The relationship is not well specified (perhaps see VAR)
  – The volatility is not well estimated
  – There are other variables that must enter into a VAR (investment opportunity set is time-varying).
  – A combination of the above

• Surprisingly, thus far, the evidence for $\hat{\beta} > 0$ is non-existent.

• Several papers find $\hat{\beta} < 0$. 
A naive way to estimate $\sigma_t$ (it is unobservable)

- Note that before we can estimate the conditional second moment, we have to estimate the conditional first moment. A lot of people forget to specify the conditional first moment and get garbage. From the definition, $\text{Var}_t(r_{t+1}) = E_t \left( (r_{t+1} - E_t r_{t+1})^2 \right) = E_t \left( r_{t+1}^2 \right) - \left( E_t (r_{t+1}) \right)^2$

- Here are the steps.
  - Run daily returns on lagged daily returns and lagged dividend yields. The residuals from this regression will be
    $$\hat{e}_t = r_t - \hat{E}_{t-1}(r_t)$$
    where $\hat{e}$ denotes the residual from the OLS regression.
  - Using the daily data, we can form an estimate for the monthly volatility as in:
    $$\hat{\sigma}_t^2 = \left( \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 \right) \times 22$$
    where $i = 1, \ldots, n$ are the daily observations within month $t$.
  - This is a particularly simple way to estimate the monthly $\sigma_t^2$. It is non-parametric (no parameters to estimate, just $\sigma_t$).
• If we plot \( \hat{\sigma}_t^2 \) over time, it becomes immediately clear that \( \hat{\sigma}_t^2 \) is not white noise, but follows a certain process.

• Recall, how we characterized the levels \( Y_t = \phi Y_{t-1} + \varepsilon_t \).

• We want to find a similar model for the volatility, something like \( \hat{\sigma}_t^2 = \phi \hat{\sigma}_{t-1}^2 + \varepsilon_t \).

• But a simple AR model would not work. Why?
8 Autoregressive Conditional Heteroskedasticity (ARCH)

• Suppose we have the returns process $\{r_t\}_{t=1}^T$.

• First, we model the conditional mean, for example as

\[ r_t = \mu + \phi r_{t-1} + u_t \]

• We know that the unconditional first moments are

\[ E( r_t ) = \frac{\mu}{1-\phi} \text{ and } Var( r_t ) = \frac{\sigma_u^2}{1-\phi^2} \]

• We also know that the conditional first moment

\[ E_{t-1}( r_t ) = \mu + \phi r_t \] is time varying, even though the unconditional moment is not!

• Q: Can we also have the same situation for the second conditional moment, i.e. to have a time-varying conditional second moment, although the unconditional second moment constant over time?

• A: Yes.

• Note: The unconditional second moment of $u_t$ is $\sigma_u^2$. 
• Intuitively, we want to model $u_t^2$ to follow an AR process just as we had $r_t$ follow an AR process.

• We can write such a process as

$$u_t^2 = \zeta + \alpha u_{t-1}^2 + w_t$$

where $w_t$ is a white noise process with $E(w_t^2) = \sigma_w^2$.

• Note that $E_{t-1}(u_t^2) = \zeta + \alpha u_{t-1}^2$, so that the conditional moment is time-varying although $E(u_t^2) = \sigma_u^2$.

• The above process is called an autoregressive conditional heteroskedasticity model of order 1, or ARCH(1).

• We can generalize to an ARCH(p) model as:

$$u_t^2 = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_p u_{t-p}^2 + w_t$$
• Note that a variance cannot be negative. We need to place certain restrictions on
\[ u_t^2 = \zeta + \alpha u_{t-1}^2 + w_t \]
in order to insure that \( u_t^2 \) is always positive.

• We need:
  – \( w_t \) to be bound from below by \( -\zeta \), where \( \zeta > 0 \).
  – \( \alpha \geq 0 \).
  – For covariance stationarity of \( u_t^2 \), we also need \( \alpha < 1 \) as in the other AR model.

• With all those conditions, we can see that
\[ Var \left( u_t \right) = E \left( u_t^2 \right) = \sigma_u^2 = \frac{\zeta}{1 - \alpha} \]
• This is the ARCH(1) model and its historical relation to what we have done before

• It is more convenient, but less intuitive, to present the ARCH(1) model as:

\[ u_t = \sqrt{h_t} v_t \]

where \( v_t \) is iid with mean 0, and \( E \left( v_t^2 \right) = 1 \).

• Suppose that

\[ h_t = \zeta + au_{t-1}^2 \]

then combining the above equations, we obtain:

\[ u_t^2 = h_t v_t^2 \]

• Now, since \( v_t \) is iid then

\[
E_{t-1} \left( u_t^2 \right) = E_{t-1} \left( h_t^2 v_t^2 \right) \\
= E_{t-1} \left( h_t^2 \right) E_{t-1} \left( v_t^2 \right) \\
= \zeta + au_{t-1}^2
\]

as before.
• Reconciling the two definitions.
• From one side, we have \( u_t^2 = h_t v_t^2 \)
• From another side, we have \( u_t^2 = h_t + w_t \). Therefore
  \[
  h_t v_t^2 = h_t + w_t
  \]
or
  \[
  w_t = h_t (v_t^2 - 1)
  \]
• From here, we can see that \( E_{t-1}(w_t^2) \) is time varying, whereas \( E(w_t^2) = \sigma_w^2 \)
• Note: The unconditional second moment of \( w_t \) (the unconditional fourth moment of \( u_t \)) does not always exist for an ARCH model. Not a big deal, but might be annoying if we want to look at conditional kurtosis.
• The ARCH process gives us conditional heteroskedasticity, but it turns out that $\sigma_u^2$ is a very persistent process.

• We can capture such a process with an ARCH(p) process $u_t^2 = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_p u_{t-p}^2 + w_t$ where $p$ is very large.

• This solution is inefficient. There are too many parameters to estimate!

• What to do? GARCH

• The GARCH, Generalized ARCH allows us to capture the persistence of conditional volatility in a parsimonious way.
• Recall that we could write:
  \[ u_t = \sqrt{h_t} v_t \]
  where \( h_t = \zeta + au_{t-1}^2 \) for an ARCH process.

• GARCH: Suppose, we specify \( h_t \) as
  \[ h_t = \zeta + \delta h_{t-1} + au_{t-1}^2 \]

• The direct link between \( h_t \) and \( h_{t-1} \) is exactly what is needed to capture the dependence between \( \sigma_t^2 \) and \( \sigma_{t-1}^2 \).

• A process with \( h_t = \zeta + \delta h_{t-1} + au_{t-1}^2 \) is called a GARCH(1,1)

• Of course, we can generalize to a GARCH(p,q) as:
  \[ u_t^2 = \zeta + \delta_1 h_{t-1} + \delta_2 h_{t-2} + \ldots + \delta_p h_{t-p} + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_q u_{t-q}^2 \]
• For most practical purposes a GARCH(1,1) is GREAT.
  – There is a trade-off. You introduce more parameters to capture the accurate dynamics, but there are more parameters to estimate
  – Those parameters have restrictions. The estimation is tricky.
  – Bottom line, for 99% of the applications, GARCH(1,1) does a great job.

• GARCH is successful, because it can capture the persistence in $\sigma_t^2$, which is the most significant feature that needs to be captured.
• Another useful model to estimate is the IGARCH model, or integrated GARCH
• The IGARCH(1,1) is a GARCH(1,1) where
  \[ \delta + \alpha = 1 \]
• If this condition is satisfied, it can be shown that the conditional variance of \( u_t \) is infinite.
• The processes \( u_t \) and \( u_t^2 \) are not covariance stationary.
• However, the process \( u_t \) is stationary (i.e. its conditional density does not depend on \( t \)).
• The IGARCH is important because it captures the important case of a strong dependence that leads to non-stationarity.
Asymmetricities: We can go crazy with an autoregressive specification. For instance, Glosten, Jagannathan, and Runkle (1993) use a specification of the following kind:

\[ \sigma_t^2 = \kappa + \beta \sigma_{t-1}^2 + \alpha u_{t-1}^2 + \eta u_{t-1}^2 1_{\{u_{t-1}>0\}} \]

- This type of an effect is suitable to test the hypothesis that negative surprises increase volatility more than positive surprises. If this hypothesis is true, we expect

\[ \eta < 0 \]
The GARCH literature has gone crazy chasing after the perfect conditional heteroskedasticity model. Some of the models we have are:

- ARCH in Means
- Exponential GARCH
- Nonlinear GARCH
- Asymmetric GARCH
- Fractionally Integrated GARCH (FIGARCH)
- ABS.ASYMM.FIGARCH ????
• The interest in forecasting conditional volatility using past volatility has been greater in the applied field more so than in academia.

• The GARCH models are unsatisfactory, from an economic perspective, because
  – Explaining vol with past vol tells us nothing about the underlying economic factors that cause the volatility to move.
  – If a structural break occurs, a recursive model will fail miserably...a structural model might not.
  – In general, the GARCH models are difficult to generalize to a multivariate setting.
• Asymmetricities: We can go crazy with an autoregressive specification. For instance, Glosten, Jagannathan, and Runkle (1993) use a specification of the following kind:

\[ \sigma_t^2 = \kappa + \beta \sigma_{t-1}^2 + \alpha u_{t-1}^2 + \eta u_{t-1}^2 1_{\{u_{t-1} > 0\}} \]

• This type of an effect is suitable to test the hypothesis that negative surprises increase volatility more than positive surprises. If this hypothesis is true, we expect

\[ \eta < 0 \]
• Fact: Volatility is very persistent. It might be more persistent than what is allowed by an exponentially decaying GARCH-type model.

• There is another type of GARCH that plays an increasingly important role in finance. It is called Fractionally integrated GARCH.
Other methods for modelling volatility:

- French, Schwert and Stambaugh (1987): The idea is to use higher frequency data to estimate the variance as:

\[
\sigma_t^2 = \frac{1}{k} \sum_{d=1}^{k} \varepsilon_{t+d}^2
\]

- where \( \varepsilon_t \) are measured in days, and we estimate monthly variance.

- This produces a monthly sequence \( \{\hat{\sigma}_t^2\} \) of estimated variances.

- There is nothing wrong with this scheme.

- Another method: AR model for volatility:

\[
|\varepsilon_t| = \eta + \gamma |\varepsilon_{t-1}| + \nu_t
\]

- where the \( \varepsilon_t \) are estimated from a first step procedure.

- There is nothing wrong with this method. It provides another model for stochastic volatility.

- Since we don’t observe true volatility, we can’t really say which method is the best at capturing it.
• GARCH in Mean or GARCH-M models (Engle, Lilien, and Robins (1987)):

\[ r_t = a + bx_{t-1} + c\sigma_t^2 + \varepsilon_t \]

\[ \varepsilon_t = z_t\sigma_t \]

\[ \sigma_t^2 = \kappa + \beta(L)\sigma_{t-1}^2 + \alpha(L)\varepsilon_{t-1}^2 \]

• The difference from the previous models is that the volatility enters also in the mean of the return.

• This is exactly what Merton’s (1973, 1980) ICAPM produces–risk-return tradeoff.

• It must be the case that \( b > 0 \).

• The GARCH-M is estimated with ML or QML.

• The evidence on the risk-return tradeoff is not good.

• French, Schwert and Stambaugh conduct similar tests, but their method is a two-step procedure (inefficient, and potentially problematic.)
• Current Research: MIDAS estimators—Mixed Data Sampling estimators (Ghysels, Santa-Clara, Valkanov (2002a,b))

• Idea: Use data at different frequencies to estimate the risk-return tradeoff

\[ R_{t+1} = \alpha + \beta \left( \sum_{d=1}^{D} w_d r_{t-d}^2 \right) + \varepsilon_t \]

• where \( R_t \) is at monthly frequency and \( r_t \) is at daily.

• The weights \( w_d \) sum up to one.

• Given that vol is persistent, there might be many weights to estimate, which would result in inefficient estimators.

• Hence, we parameterize \( w_d(\theta) \) and estimate the shape of the weights.

• There are several advantages:
  – Higher frequency data, i.e. better estimates of vol.
  – Joint estimation of \( \theta, \alpha, \beta \)
  – Flexibility of weights
  – Easy to implement other variables, asymmetries.
• Stochastic volatility models:

\[ r_t = a + br_{t-1} + \varepsilon_t \]
\[ \varepsilon_t = z_t \sigma_t \]
\[ \sigma_t = \kappa + \beta \sigma_{t-1} + \nu_t \]

• The difference here is that the shocks that govern the volatility are not necessarily \( \varepsilon_t^2 \)’s.

• This is really a discretization of a continuous-time model, where the mean and the variance follow two OU processes.

• Stochastic vol models can be estimated by MLE or other methods.
We can also think of modelling the entire variance covariance matrix (see homework).
- Bollerslev (1990) provides a particularly elegant model with constant correlations, but time-varying covariances.

Here is a new and good way of modelling the entire variance-covariance matrix.
- To keep it simple, we focus on 2 assets ($R_{1,t}$ and $R_{2,t}$) and 1 exogenous variable, $X_{t-1}$ (think D/P ratio). First, we model the conditional means as:

$$R_{1,t+1} = \mu_1 + k_1 X_t + Y_{1,t+1}$$
$$R_{2,t+1} = \mu_2 + k_2 X_t + Y_{2,t+1}$$

Let $Y_{t+1} = (Y_{1,t+1} Y_{2,t+1})$, and $\Sigma_t = E (Y'_{t+1} Y_{t+1} | X_t) = E\begin{pmatrix} Y_{1,t+1}^2 & Y_{1,t+1} Y_{2,t+1} \\ Y_{1,t+1} Y_{2,t+1} & Y_{2,t+1}^2 \end{pmatrix} | X_t$.

- **NOTE:** Usually, we handle conditional expectations with projections (regressions), but we cannot regress $Y_{1,t+1}^2$, $Y_{1,t+1} Y_{2,t+1}$, and $Y_{2,t+1}^2$ on $X_t$ because $\Sigma_t$ must be positive semi-definite.

**LET:**

$$\Sigma_t = \hat{\Sigma}_t + \begin{pmatrix} \epsilon_{11,t+1} & \epsilon_{12,t+1} \\ \epsilon_{21,t+1} & \epsilon_{22,t+1} \end{pmatrix}$$

Then, we can use any triangular decomposition
(say Cholesky) to write:

\[ \hat{\Sigma}_t = U_t'U_t \]

where

\[ U_t = \begin{pmatrix} U_{11,t} & U_{12,t} \\ 0 & U_{22,t} \end{pmatrix} \]

\[ = \begin{pmatrix} \alpha_{11} + \beta_{11}X_t + \gamma_{11}U_{11,t-1} & \alpha_{12} + \beta_{12}X_t + \gamma_{12}U_{12,t-1} \\ 0 & \alpha_{22} + \beta_{22}X_t + \gamma_{22}U_{22,t-1} \end{pmatrix} \]

Then, we can write

\[ Y_{1,t+1}^2 = \hat{\Sigma}_{11,t} + \varepsilon_{11,t+1} = U_{11,t}^2 + \varepsilon_{11,t+1} \]
\[ Y_{1,t+1}Y_{2,t+1} = \hat{\Sigma}_{12,t} + \varepsilon_{12,t+1} = U_{11}U_{12} + \varepsilon_{12,t+1} \]
\[ Y_{2,t+1}^2 = \hat{\Sigma}_{22,t} + \varepsilon_{22,t+1} = U_{12,t}^2 + U_{22,t}^2 + \varepsilon_{22,t+1} \]

– **NOTE:** The positive definiteness restrictions are insured by the Cholesky decomposition.

– **NOTE:** Estimating the \( \alpha' \)'s, \( \beta' \)'s, and \( \gamma' \)'s is done with non-linear least squares (GMM with \( W=I \)).

• **NOTE:** 2 assets, 5 explanatory variables—5 seconds.
• Comments:
  – Easy to generalize to N assets and M exogenous variables.
  – Note that we can write:

\[
U_{11,t} = \kappa_{11} + \beta_{11} \sum_{k=0}^{\infty} \gamma_{11}^k X_{t-k-1}
\]
• This is great, but how do we estimate a GARCH model.
• OLS clearly does not work
• We need a method that would allow us to do non-linear estimation
• We also need a general method that we can apply to any nonlinear problem, with minimal assumptions.
• We do not want to have assumptions on the distribution of the residuals.
• Therefore, we need GMM (Generalized Method of Moments)
9 Simple Introduction to GMM

• Recall that any variable $x_t$ has a distribution $F_x(x)$. If $x$ has moments $E(x^j)$, $j = 1, \ldots$ then those moments can be used to retrieve $F_x(x)$.

• Caution: Some variables do not have moments (Cauchy distribution case).

• Suppose we have random variables $x_t, y_t, z_t$.
  – A population moment of those variables is
    $$E\left[g\left(x_t, y_t, z_t\right)\right]$$
  – A sample moment of those variables is
    $$\frac{1}{T} \sum_{t=1}^{T} g\left(x_t, y_t, z_t\right)$$
  – By the ergodicity theorem (or the LLN in cross section, we know that)
    $$\frac{1}{T} \sum_{t=1}^{T} g\left(x_t, y_t, z_t\right) \rightarrow^p E\left[g\left(x_t, y_t, z_t\right)\right]$$
    under some mild conditions on the function $g(.)$.  

• In other words, the distance between the sample and the population moment goes to zero in probability as $T \to \infty$:

$$\left\{ \frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t) - E\left[g(x_t, y_t, z_t)\right] \right\} \to^p 0$$

• Can we use this “insight” to estimate parameters. Suppose that the function $g$ depends not only on the data but also on the unknown parameters, $\theta$.

• We want to choose the parameter $\theta$ in order to minimize the distance between the data and the population moment.

• In a simpler example, let’s concentrate on a univariate case. Then $g(x|\theta) = \mu$, the population mean. In other words, $\theta = \mu$.

• The problem becomes (trivially):

$$\left\{ \frac{1}{T} \sum_{t=1}^{T} x_t - \mu \right\} \to^p 0$$
Here is a more interesting example: OLS as GMM

The FOC in the OLS case could be written as:

\[ E(x_t \varepsilon_t) = 0 \]

This is a moment condition that also depends on parameters. To see that, write

\[ E(x_t (y_t - x_t \beta)) = 0 \]

\[ E(x_t y_t) = \beta E(x_t^2) \]

\[ \beta = \frac{E(x_t y_t)}{E(x_t^2)} \]

Therefore, approximating the population means by their sample analogues, we get

\[ \frac{1}{T} \sum_{t=1}^{T} (x_t (y_t - x_t \beta)) = 0 \]

\[ \frac{1}{T} \sum_{t=1}^{T} x_t y_t = \beta \frac{1}{T} \sum_{t=1}^{T} x_t^2 \]

\[ \hat{\beta} = \frac{\frac{1}{T} \sum_{t=1}^{T} x_t y_t}{\frac{1}{T} \sum_{t=1}^{T} x_t^2} \]
• But we can also write another moment condition:

\[ E \left( x_t^2 \varepsilon_t \right) = 0 \]

• Then, as above

\[
E \left( x_t^2 (y_t - x_t \beta) \right) = 0
\]

\[
\beta = \frac{E \left( x_t^2 y_t \right)}{E \left( x_t^3 \right)}
\]

• Therefore, using sample moments to approximate population moments, we get

\[
\hat{\beta}_2 = \frac{1}{T} \sum_{t=1}^{T} x_t^2 y_t
\]

\[
\frac{1}{T} \sum_{t=1}^{T} x_t^3
\]

• We can also use

\[
E \left( g(x_t) \varepsilon_t \right) = 0
\]

for some function \( g(\cdot) \). Note: You should also be able to show that \( E \left( x_t \varepsilon_t \right) = 0 \) implies \( E \left( g(x_t) \varepsilon_t \right) = 0 \). Then, for a known function \( g(\cdot) \)

\[
\hat{\beta}_g = \frac{1}{T} \sum_{t=1}^{T} g(x_t) y_t
\]

\[
\frac{1}{T} \sum_{t=1}^{T} g(x_t) x_t
\]
• Oupss! Problem. We have one parameter, $\beta$, but three possible estimators
\[
\hat{\beta} = \frac{1}{T} \sum_{t=1}^{T} x_t y_t \xrightarrow{p} \beta
\]
\[
\hat{\beta}_2 = \frac{1}{T} \sum_{t=1}^{T} x_t^2 y_t \xrightarrow{p} \beta
\]
\[
\hat{\beta}_g = \frac{1}{T} \sum_{t=1}^{T} g(x_t) y_t \xrightarrow{p} \beta
\]

• Which one do we choose?

• Result: Under some very restrictive assumptions (i.e. exogeneity of $x_t$, homoskedasticity, uncorrelated $\varepsilon_t$, etc), the OLS is the best linear unbiased estimator (BLUE). In other words, in has the smallest variance.

• However, who knows if those assumptions are satisfied. In all likelihood, they are not.

• Q: Can we stack all the moments in a vector as
\[
E(g(x|\beta)) = E \begin{bmatrix}
x_t \varepsilon \\
x_t^2 \varepsilon \\
g(x_t) \varepsilon_t
\end{bmatrix} = 0
\]
and choose the value of $\beta$ that satisfies the three sample moments?
• A: Off course, not! Three equations, potentially nonlinear, with only one unknown....Who knows how many solutions there are, if any.

• But, we can construct a quadratic function, as:

\[
E \left( g \left( x | \beta \right)^{\prime} W g \left( x | \beta \right) \right) = 0
\]

for some symmetric positive definite matrix \( W \).

• Now, we have the information into the three equations, weighted by the elements of the matrix \( W \).

• Problem: What matrix \( W \) to choose.

• A: Any symmetric positive matrix will give us consistent estimates (i.e. \( \hat{\beta}_W \rightarrow^p \beta \)), but we are concerned with efficiency, or smallest possible standard errors around \( \hat{\beta}_W \).
• Endogeneity, Instrumental Variables (IV) and GMM.

• By construction, we had $E(\varepsilon \mid x) = 0$ implied that $E(\varepsilon x) = 0$. In other words, the residuals and the explanatory variables are uncorrelated.

• However, in structural models, it is often the case that we want to run regressions when this requirement is not satisfied. For example:

$$\text{FirmValue}_t = \alpha + \beta \text{Debt}_t + \varepsilon_t$$

• But it is not reasonable to assume that Debt is an exogenous variable. For example, new (relatively low Firm Value) firms do not have access to debt. Indeed, we might try to run the opposite regression:

$$\text{Debt}_t = \delta + \zeta \text{FirmValue}_t + \nu_t$$

• So, here

$$E(\text{Debt}_t \varepsilon_t) = E((\delta + \zeta \text{FirmValue}_t + \nu_t) \varepsilon_t) \neq 0$$

• Q: If $E(\text{Debt}_t \varepsilon_t) \neq 0$, can we still have $\hat{\beta} \rightarrow^p \beta$?

• Breaking the $E(\text{Debt}_t \varepsilon_t) = 0$ condition is the cardinal sin in empirical work!!!

• Q: What to do?

• Note: We can argue that most equations in finance suffer from this endogeneity problem.
• Well, we can look for a variable, $Z_t$ which is:
  – Correlated with $Debt_t$ (it proxies for $Debt_t$)
  – Is uncorrelated with $\varepsilon_t$.

• If such a variable exists, then we can write the moment condition:

$$E(Z_t \varepsilon_t) = 0$$

• Or, if we look at the sample moments, then

$$\frac{1}{T} \sum_{t=1}^{T} Z_t \varepsilon_t = \frac{1}{T} \sum_{t=1}^{T} Z_t (y_t - \beta x_t)$$

$$\frac{1}{T} \sum_{t=1}^{T} Z_t y_t = \beta \frac{1}{T} \sum_{t=1}^{T} Z_t x_t$$

$$\hat{\beta}_{IV} = \frac{\frac{1}{T} \sum_{t=1}^{T} Z_t y_t}{\frac{1}{T} \sum_{t=1}^{T} Z_t x_t}$$

• Then we can show that $\hat{\beta}_{IV} \rightarrow_p \beta$, whereas $\hat{\beta}_{ols}$ does not.

• This estimator was motivated from GMM.
• GMM is a very powerful way of looking at an estimation problem.
• All we need is a moment condition that holds.
• The problem does not have to be linear
• No distributional assumptions are needed.
• We will use GMM to estimate
  – The non-linearized version of the CCAPM
  – Nonlinear process, such as ARCH, GARCH, etc.
  – Joint estimation of linear models (CAPM).