Today’s Agenda

1. Will finish volatility modelling later
   – Can anything beat a GARCH(1,1)?

2. Spurious regressions—non-stationary processes

3. Aside: Wold’s Decomposition—Stationary Processes

4. GMM—Simple Introduction
   – ARCH estimation
   – Other volatility estimations
   – CCAPM estimation

5. MLE—Simple Introduction
   – GARCH estimation

6. Announcements
1 Spurious Regressions and Non-Stationary Time Series

• Thus far, we have been looking at stationary time series. We have focused on $r_t$, $\sigma^2_t$, assuming that they are stationary. For some series, this assumption is more tenable than for others.

• But suppose you want to work with non-stationary time-series, i.e. prices, volume, number of investors in a particular fund, number of funds, etc. Those processes are inherently non-stationary.

• Let $p_t$ be the log-price. We know that
  
  $p_t = p_{t-1} + \varepsilon_t$

  or $p_t$ is an AR(1) process with an unit-root.

• This process is non-stationary. We cannot apply the CLT.

• But we are still interested in testing the null $\phi = 1$ versus $\phi < 1$.

• Problem. Under the null, the process is non-stationary.

• Under the alternative, the process is stationary.
• It turns out that (Functional CLT)
\[
\frac{1}{\sqrt{T}} p_t = \frac{1}{\sqrt{T}} \sum_{s=1}^{t} \varepsilon_s
\]
\[
= \frac{1}{\sqrt{T}} \sum_{s=1}^{[rT]} \varepsilon_s \Rightarrow W(r)
\]
where \( W(r) \) is a Brownian motion on \([0, 1]\). I.e., \( W(r) \sim N(0, r), 0 < r < 1 \).

• Note that this result is asymptotic. We don’t have to assume that the \( \varepsilon 's \) are normal (hence the Functional CLT, or FCLT).

• Q: Can’t we standardize the non-stationary processes by a power of \( T \) in order for them to converge.

• A: Yes.

• Let’s get a “flavor” of how things work:
Recall that
\[ \hat{\phi} = \frac{\sum p_t p_{t-1}}{\sum p_{t-1}^2} = \frac{\sum p_{t-1} (\phi p_{t-1} + \varepsilon_t)}{\sum p_{t-1}^2} \]
\[ = \phi + \frac{\sum p_{t-1} \varepsilon_t}{\sum p_{t-1}^2} \]

If \( \phi < 1 \), we had
\[ \hat{\phi} = \phi + \frac{1}{T} \sum \frac{\varepsilon_t p_{t-1}}{\sum p_{t-1}^2} \xrightarrow{p} \phi \]
\[ \sqrt{T} \left( \hat{\phi} - \phi \right) \sim N \left( 0, \sigma_{\phi}^2 \right) \]

But if \( \phi = 1 \), the results do not hold. But
\[ \hat{\phi} = \phi + \frac{\sum \varepsilon_t p_{t-1}}{\sum p_{t-1}^2} \]
\[ T \left( \hat{\phi} - \phi \right) \Rightarrow O_p(1) \]

In other words, \( \hat{\phi} \) is super-consistent.

Q: But since we don’t know the distribution of \( \hat{\phi} \), can we use this result for testing?

A: Yes, if we simulate the distribution.
Dickey-Fuller (DF) Test:

- $H_o : \phi = 1$
- $H_a : \phi < 1$

The test is: 

$$t = \frac{\hat{\phi} - 1}{se(\hat{\phi})}$$

Suppose that $\varepsilon_t$ follows an AR(p) process. The distribution of the DF test is influenced by those parameters. Not good.

To get rid of those parameters, we run the following regression:

$$p_t = \phi p_{t-1} + \zeta_1 \Delta p_{t-1} + \zeta_2 \Delta p_{t-2} + \ldots + \zeta_k \Delta p_{t-k} + \nu_t$$

Then, 

$$t = \frac{\hat{\phi} - 1}{se(\hat{\phi})}.$$

This is called the Augmented DF, or ADF test.
• Summary of ADF test–Testing for a unit root:
  – Regress $p_t$ on $p_{t-1}, \Delta p_{t-1}, \Delta p_{t-2}, \ldots, \Delta p_{t=k}$
  – Form: $t = \frac{\hat{\phi} - 1}{se(\hat{\phi})}$
  – Get the critical value from simulations.
So, is working with non-stationary variables that easy?

NO:

Suppose \( p_1^t \) and \( p_2^t \) are the prices of the same asset traded on two markets. Then, it must be the case that

\[
p_1^t = p_2^t
\]

Empirically, this is almost true. We find

\[
p_1^t - p_2^t = \varepsilon_t
\]

where \( \varepsilon_t \) is almost iid, and \( E(\varepsilon_t) = 0 \).

How do we take advantage of this?

Regress:

\[
p_1^t = \gamma p_2^t + \varepsilon_t
\]

If

\[
\varepsilon_t > 0
\]
\[
p_1^t > \gamma p_2^t
\]

The asset is “too expensive” in market 1.

Easy, right?
• Wrong!

• Suppose we have two assets, with prices $p_t$ and $q_t$. One might be tempted to look for arbitrage strategies as in

$$p_t = \gamma q_t + \varepsilon_t$$

• If $\gamma > 0$, there is a relationship, and we can trade.

• No.

• Suppose

$$p_t = p_{t-1} + u_t$$
$$q_t = q_{t-1} + v_t$$
$$\text{cov}(u_t v_t) = 0$$

• Note: The two log-prices represent two independent discrete Brownian motions.
• But:

\[
\hat{\gamma} = \frac{\sum_{t=1}^{T} p_t q_t}{\sum_{t=1}^{T} q_t^2} = \frac{\sum_{t=1}^{T} \left( \sum_{s=1}^{t} v_s \right) \left( \sum_{s=1}^{t} u_s \right)}{\sum_{t=1}^{T} \left( \sum_{s=1}^{t} v_s \right)^2} = \frac{1}{T^2} \sum_{t=1}^{T} \left( \sum_{s=1}^{t} v_s \right) \left( \sum_{s=1}^{t} u_s \right) = O_p(1)
\]

• Similarly:

\[
t = \frac{\hat{\gamma}}{se(\hat{\gamma})}
\]

is not consistent.

• The $R^2$ does not converge to 0, as $T \to \infty$.

• Illustration: spurious.m
Co-integration:

- Suppose that $p_t$ and $q_t$ are unit root (integrated) processes, but there is a linear combination of $p_t$ and $q_t$ that is stationary. That is, there exists a vector $\gamma = \begin{bmatrix} 1 & -\gamma_1 \end{bmatrix}$ such that
  \[ p_t - \gamma q_t = \varepsilon_t \]
  and $\varepsilon_t$ is a stationary process. Then, $p_t$ and $q_t$ are said to be cointegrated.

- There are formal tests for cointegration, but they have low power against the alternative. Why? (Think spurious correlation).

- Cointegration is only between contemporaneous variables. I.e. $p_t - \gamma q_t$ is a cointegrating vector. But $\Delta p_t$ is not. However, the latter is also a way of stationarizing the process.

- Cointegration occurs “naturally” in economics. It is dictated by theory.
• Examples:
  – Prices $p^1_t$ and $p^2_t$ of the same asset traded on two markets.
  – Dividend price ratio is cointegrated, or:
    \[ d_t - p_t \]
    must be stationary
  – The long and the short rate must be cointegrated
  – Consumption and GDP must be cointegrated, etc.
  – Lettau & Ludvigson (2001), the “cay” ratio:
    \[ c_t - \gamma_1 a_t - \gamma_2 y_t \]
2 Wold Decomposition: Stationary Processes

- Q: Isn’t the AR(1) (or ARMA(p,q)) model restrictive?
- No, because of the Wold decomposition result

Wold’s (1938) Theorem: Any zero-mean covariance stationary process \( Y_t \) can be represented in the form

\[
Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t
\]

where \( \psi_0 = 1 \) and \( \sum_{j=0}^{\infty} \psi_j^2 < \infty \) (square summable). The term \( \varepsilon_t \) is white noise and represents the linear projection error of \( Y_t \) on lagged \( Y_t \)'s

\[
\varepsilon_t = Y_t - E(Y_t|Y_{t-1}, Y_{t-2}, \ldots).
\]

The value \( \kappa_t \) is uncorrelated with \( \varepsilon_{t-j} \) for any \( j \) and is a purely deterministic term.

- Can we estimate all \( \psi_j \) in the Wold’s decomposition?
• The stationary ($|\phi| < 1$) AR(1) model can be written as

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

$$(1 - \phi L) Y_t = \varepsilon_t$$

$$Y_t = (1 - \phi L)^{-1} \varepsilon_t$$

$$= \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

or $\psi_j = \phi^j$. This is the restriction for the AR(1) model.

• The stationary ARMA(1,1) model can be written as

$$Y_t = \phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$(1 - \phi L) Y_t = (1 + \theta L) \varepsilon_t$$

$$Y_t = \frac{\varepsilon_t}{(1 - \phi L)} + \frac{\theta \varepsilon_{t-1}}{(1 - \phi L)}$$

$$= \varepsilon_t + \sum_{j=1}^{\infty} \phi^{j-1} (\phi + \theta) \varepsilon_{t-j}$$

or $\psi_j = \phi^{j-1} (\phi + \theta)$.

• And so on.
• Another interesting process: Fractionally differencing

\[ Y_t = (1 - L)^{-d} \varepsilon_t \]

• where \( d \) is a number between 0 and 0.5.

• It can be shown (Granger and Joyeux (1980), and Josking (1981)) that

\[ Y_t = \sum_{j=0}^{\infty} \eta_j \varepsilon_{t-j} \]

\[ \eta_j = \frac{1}{j!} (d + j - 1) (d + j - 2) (d + j - 3) \ldots (d + 1) d \]

\[ \eta_j \approx (j + 1)^{d-1}, \text{ for large } j \]

• Plot of \( \eta_j \) for \( d = 0.25 \) and \( \phi^j \) for \( \phi = 0.5 \) and \( \phi = 0.95 \)
There is a similar representation in the spectral domaine

Spectral Representation Theorem [e.g., Cramer and Leadbetter (1967)]: Any covariance stationary process \( Y_t \) with absolutely summable autocovariances can be represented as

\[
Y_t = \mu + \int_0^\pi \left[ \alpha(\omega) \cos(\omega t) + \delta(\omega) \sin(\omega t) \right] d\omega
\]

where \( \alpha(.) \) and \( \delta(.) \) are zero-mean random variables for any fixed frequency \( \omega \in [0, \pi] \). Also, for any frequencies \( 0 < \omega_1 < \omega_2 < \omega_3 < \omega_4 < \pi \), \( \int_{\omega_1}^{\omega_2} \alpha(\omega) d\omega \) is uncorrelated with \( \int_{\omega_3}^{\omega_4} \alpha(\omega) d\omega \) and the variable \( \int_{\omega_1}^{\omega_2} \delta(\omega) d\omega \) is uncorrelated with \( \int_{\omega_3}^{\omega_4} \delta(\omega) d\omega \).

Different (but equivalent) way of looking at a time-series.
3 Simple Introduction to GMM

- Recall that any variable $x_t$ has a distribution $F_x(x)$. If $x$ has moments $E(x^j)$, $j = 1, ...$ 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 $x_t, y_t, z_t$, and a function $g(.)$.
  - A population moment of those variables is $E[g(x_t, y_t, z_t)]$
  - A sample moment of those variables is $\frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t)$
  - By the ergodicity theorem (or the LLN in cross section, we know that) $\frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t) \rightarrow_p E[g(x_t, y_t, z_t)]$ 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 $T \to \infty$:

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

• 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 model is:

\[ y_t = x_t\beta + \varepsilon_t \]

The FOC in the OLS case could be written as a moment:

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

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  
\[ E \left( x_t^2 (y_t - x_t \beta) \right) = 0 \]
\[ \beta = \frac{E (x_t^2 y_t)}{E (x_t^3)} \]

• Therefore, using sample moments to approximate population moments, we get  
\[ \hat{\beta}_2 = \frac{\frac{1}{T} \sum_{t=1}^{T} x_t^2 y_t}{\frac{1}{T} \sum_{t=1}^{T} x_t^3} \]

• We can also use  
\[ E (g(x_t) \varepsilon_t) = 0 \]
for some function \( g(.) \). Note: You should also be able to show that \( E (x_t \varepsilon_t) = 0 \) implies \( E (g(x_t) \varepsilon_t) = 0 \). Then, for a known function \( g(.) \)
\[ \hat{\beta}_g = \frac{\frac{1}{T} \sum_{t=1}^{T} g(x_t) y_t}{\frac{1}{T} \sum_{t=1}^{T} g(x_t) x_t} \]
Oupss! Problem. We have one parameter, $\beta$, but three possible estimators

$$\hat{\beta} = \frac{1}{T} \sum_{t=1}^{T} x_t y_t \rightarrow^p \beta$$

$$\hat{\beta}_2 = \frac{1}{T} \sum_{t=1}^{T} x_t^2 y_t \rightarrow^p \beta$$

$$\hat{\beta}_g = \frac{1}{T} \sum_{t=1}^{T} g(x_t) y_t \rightarrow^p \beta$$

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 among all linear unbiased estimators.

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:
  \[ E \left( g(x|\beta)^\prime W g(x|\beta) \right) = 0 \]
  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. \( \hat{\beta}_W \to^p \beta \)), 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 $E(\varepsilon|x) = 0$ 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:

$$FirmValue_t = \alpha + \beta Debt_t + \varepsilon_t$$

• 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:

$$Debt_t = \delta + \zeta FirmValue_t + \nu_t$$

• So, here

$$E(Debt_t\varepsilon_t) = E((\delta + \zeta FirmValue_t + \nu_t)\varepsilon_t) \neq 0$$

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

• Breaking the $E(Debt_t\varepsilon_t) = 0$ 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

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

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

\[
\hat{\beta}_{IV} = \frac{\frac{1}{T} \sum_{t=1}^{T} Z_t y_t}{\frac{1}{T} \sum_{t=1}^{T} Z_t x_t} = \frac{\frac{1}{T} \sum_{t=1}^{T} Z_t (\beta x_t + \varepsilon_t)}{\frac{1}{T} \sum_{t=1}^{T} Z_t x_t}
\]

\[ = \beta + \frac{\frac{1}{T} \sum_{t=1}^{T} Z_t \varepsilon_t}{\frac{1}{T} \sum_{t=1}^{T} Z_t x_t} \]

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

• This estimator was motivated from GMM.
• All this is great, but how do we choose the instrument $Z_t$.

• This is usually the big question.

• Usually, $Z_t = Debt_{t-k}$, because

$$E(Z_t \varepsilon_t) = E(Debt_{t-k} \varepsilon_t)$$

$$= E((\delta + \zeta FirmValue_{t-k} + \nu_{t-k}) \varepsilon_t)$$

• Predetermined regressors can be thought of as IV’s. They are generally OK.

• Weak instruments literature: Theoretically, we only need small correlation between the instrument and the variable. However, the bigger the correlation, the better.

• Generalization of IV: Two-Stage Least Squares (TSLS)

• Conclusion: Don’t pile up too many weak instruments. 10,000 weak instruments are no substitute for a strong instrument!
3.1 Estimation of ARCH using GMM

- Recall the model:
  \[ r_t = \mu + \phi r_{t-1} + u_t \]
  \[ u_t^2 = h_t + w_t = \zeta + \alpha u_{t-1}^2 + w_t \]
  \[ E_{t-1} (u_t^2) = \zeta + \alpha u_{t-1}^2 \]

  - First equation: Conditional Mean
  - Second equation: Conditional Variance

- The four moment conditions are:
  \[ E(u_t) = 0 \]  
  \[ E (u_t r_{t-1}) = E ((r_t - \mu - \phi r_{t-1}) r_{t-1}) = 0 \]
  \[ E \left( u_t^2 - \frac{\zeta}{1 - \alpha} \right) = 0 \]
  \[ E (w_t r_{t-1}) = E \left( (u_t^2 - \zeta - \alpha u_{t-1}^2) r_{t-1} \right) = 0 \]

- Note: four moments and four parameters \((\mu, \phi, \zeta, \alpha)\).

- Replace those conditions by their sample analogues:
  \[
  \frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix} (r_t - \mu - \phi r_{t-1}) \\ ((r_t - \mu - \phi r_{t-1}) r_{t-1}) \\ (u_t^2 - \frac{\zeta}{1 - \alpha}) \\ ((u_t^2 - \zeta - \alpha u_{t-1}^2) r_{t-1}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
  \]

- Solve for \((\mu, \phi, \zeta, \alpha)\).
• We need $u_{t-1}$ for the last equation.
• A better way to estimate ARCH/GARCH: maximum likelihood
• 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 can use GMM to estimate
  – The non-linearized version of the Consumption CAPM.
  – Nonlinear process, such as ARCH, GARCH, etc.
  – Interesting interest rate models (Chan et. al (1992)).
- Practical Considerations:
  - We need at least as many conditions as parameters (just-identified case)
  - If there are more moments, they can be used to test the model (J test).
  - Too many moments are not desirable in practice.
  - The conditioning information matters (what variables are included in the moments—as with other estimators).
  - People have raised questions regarding the small sample properties of GMM. Unsubstantiated, perhaps.
3.2 Specification and GMM Estimation of Non-Autoregressive Models of Volatility

- Chan et al. (1992): Estimate a “very” general process for the short rate.

- Q: Why?

- A: Because, many models in finance have to specify a short rate. Usually they specify the short rate process that is most convenient, i.e. the short rate process that allows them to derive nice, simple, closed-form results.

- But one has to realize that the results from those models are conditional upon having the right short-rate process.
The proposed (general) process for the short rate is:

\[ dr = (\alpha + \beta r) \, dt + \sigma r \, dZ \]

The short rate is written in continuous time to relate it to most finance models.

The estimation is done in discrete time.

The discretization is done by taking that \( dt \) is one month.

No other adjustments are made.

Discretization

\[
\begin{align*}
    r_{t+1} - r_t &= \alpha + \beta r_t + \sigma r_t v_t \\
    r_{t+1} &= \alpha + (1 + \beta) r_t + \sigma r_t v_t
\end{align*}
\]
Implementation:


- The short rate is:
  \[
  \Delta r_{t+1} = \alpha + \beta r_t + u_{t+1} \\
  \text{or } r_{t+1} = \alpha + (1 + \beta) r_t + u_{t+1} \\
  E\left(u_{t+1}^2\right) = \sigma^2 r_t^{2\gamma}
  \]

- Note that while the continuous time processes assumed that \(dZ_t\) is normally distributed, this is purely for tractability.

- The shock \(u_{t+1}\) can have any distribution!

- The processes have to be stationary!–This is the only restriction.

- The parameters to be estimated are: \(\alpha, \beta, \gamma, \) and \(\sigma^2\).
• Of course, this setup is very similar to our first (ARCH) setup, but:

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

\[ u_t = \sigma r_{t-1}^{\gamma} v_t \]

\[ E_{t-1}(u_t^2) = \sigma^2 r_{t-1}^{2\gamma} \]

– First equation: Conditional Mean

* We can take the first difference of returns and write (Chan et al., 1992)

\[ r_{t+1} - r_t = \mu + (\phi - 1) r_t + \sigma r_{t-1}^{\gamma} v_t \]

– Second equation: Conditional Variance

– Note: The variance is not auto-regressive—not suitable for forecasting

• The four moment conditions are:

\[ E(u_t) = 0 \]  \hspace{1cm} (3)

\[ E(u_t r_{t-1}) = E((r_t - \mu - \phi r_{t-1}) r_{t-1}) = 0 \]  \hspace{1cm} (4)

\[ E(u_t^2 - \sigma^2 r_{t-1}^{2\gamma}) = 0 \]

\[ E(w_t r_{t-1}) = E\left(\left(u_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) r_{t-1}\right) = 0 \]

• Note: four moments and four parameters \((\mu, \phi, \sigma, \gamma)\).

• Replace those conditions by their sample ana-
logues:

\[
\frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix}
(r_t - \mu - \phi r_{t-1}) \\
((r_t - \mu - \phi r_{t-1}) r_{t-1}) \\
(u_t^2 - \sigma^2 r_{t-1}^{2\gamma}) \\
((u_t^2 - \sigma^2 r_{t-1}^{2\gamma}) r_{t-1})
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\]

- Solve for \((\mu, \phi, \sigma, \gamma)\).
- The same algorithm applies for all problems that can be written as a moment condition.
- Example: stock_vol_illustration2.m
We could also proceed in the following way:

- Regress $r_t$ on $r_{t-1}$ using OLS. Get estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\varepsilon}_t$.

- Estimate $\sigma^2$ and $\gamma$ using non-linear least squares:

$$\min_{\sigma^2, \gamma} \sum_{t=1}^{T} \left( \hat{\varepsilon}_{t+1}^2 - \sigma^2 r_t^{2\gamma} \right)^2$$

- Would we obtain the same answer?
- What method is preferable in theory?
- What method is preferable in practice?
• A good exercise: Redo the Chan et al. (1992) paper with data up to 2001 and with monthly and daily data. Compare the results and compare them to the GARCH homework.
3.3 Example: CCAPM and GMM estimation

(Hansen, EMA 1982; Hansen and Singleton, EMA 1982):

\[ p_t = E_t (m (z_t | \theta) x_{t+1}) \]

- \( m (z_t | \theta) \) is the stochastic discount factor that depends on some state variables and parameters.
- \( x_{t+1} \) is the payoff from the asset.

- Example from last lecture:

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

- We don’t need to log-linearize anything with GMM.
- Non-linearities are not a problem.
- Robustness is a real issue.
- But this is not what people use. Why?
- We can re-write the pricing equation as:

\[ 1 = E_t \left( m(z_t|\theta) \frac{x_{t+1}}{p_t} \right) \]

\[ E_t (m(z_t|\theta) R_{t+1}) - 1 = 0 \]

where \( R_{t+1} = x_{t+1}/p_t \). In the case of stocks, think \( x_{t+1} = p_{t+1} + d_{t+1} \).

- In this case, returns are stationary.

- If \( z_t \) in the stochastic discount factor is also stationary, then the ergoticity theorem conditions will be satisfied (under some other mild conditions).

- Example from last lecture:

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

- We have the data \( \{c_t, R_t\} \).

- Have to estimate one parameter, \( \gamma \) (assuming \( \beta \) is 0.995).
4 Maximum Likelihood Estimation

(Preliminaries for GARCH/Stochastic Volatility & Kalman Filtering)

- Suppose we have the series \( \{Y_1, Y_2, \ldots, Y_T\} \) with a joint density \( f_{Y_T, \ldots, Y_1}(\theta) \) that depends on some parameters \( \theta \) (such as means, variances, etc.)
- We observe a realization of \( Y_t \).
- If we make some functional assumptions on \( f \), we can think of \( f \) as the probability of having observed this particular sample, given the parameters \( \theta \).
- The maximum likelihood estimate (MLE) of \( \theta \) is the value of the parameters \( \theta \) for which this sample is most likely to have been observed.
- In other words, \( \hat{\theta}^{MLE} \) is the value that maximizes \( f_{Y_T, \ldots, Y_1}(\theta) \).
• Q: But, how do we know what \( f \)–the true density of the data–is?

• A: We don’t.

• Usually, we assume that \( f \) is normal, but this is strictly for simplicity. The fact that we have to make distributional assumptions limits the use of MLE in many financial applications.

• Recall that if \( Y_t \) are independent over time, then

\[
f_{Y_T,...,Y_1}(\theta) = f_{Y_T}(\theta_T)f_{Y_{T-1}}(\theta_{T-1})...f_{Y_1}(\theta_1) = \prod_{i=1}^{T} f_{Y_i}(\theta_i)
\]

• Sometimes it is more convenient to take the log of the likelihood function, then

\[
\Lambda(\theta) = \log f_{Y_T,...,Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i}(\theta)
\]
• However, in most time series applications, the independence assumption is untenable. Instead, we use a conditioning trick.

• Recall that
  \[ f_{Y_2Y_1} = f_{Y_2|Y_1} f_{Y_1} \]

• In a similar fashion, we can write
  \[ f_{Y_{T-1}...Y_1}(\theta) = f_{Y_{T-1}|Y_{T-2}...Y_1}(\theta) f_{Y_{T-2}|Y_{T-3}...Y_1}(\theta) ... f_{Y_1}(\theta) \]

• The log likelihood can be expressed as
  \[ \Lambda(\theta) = \log f_{Y_{T-1}...Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i|Y_{i-1}...Y_1}(\theta_i) \]
• Example: The log-likelihood of an AR(1) process
\[ Y_t = c + \phi Y_{t-1} + \varepsilon_t \]

• Suppose that \( \varepsilon_t \) is iid \( N(0, \sigma^2) \)

• Recall that \( E(Y_t) = \frac{c}{1-\phi} \) and \( Var(Y_t) = \frac{\sigma^2}{1-\phi^2} \)

• Since \( Y_t \) is a linear function of the \( \varepsilon_t \)'s, then it is also Normal (sum of normals is a normal).

• Therefore, the density (unconditional) of \( Y_t \) is Normal.

• Result: If \( Y_1 \) and \( Y_2 \) are jointly Normal, then the marginals are also normal.

• Therefore,
\[
    f_{Y_2|Y_1} \text{ is } N \left( (c + \phi y_1), \sigma^2 \right)
\]

or
\[
    f_{Y_2|Y_1} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ \frac{-(y_2 - c - \phi y_1)^2}{2\sigma^2} \right]
\]
Similarly,

\[ f_{Y_3|Y_2} \text{ is } N \left( (c + \phi y_2, \sigma^2) \right) \]

or

\[
f_{Y_3|Y_2} = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ -\frac{(y_3 - c - \phi y_2)^2}{2\sigma^2} \right]
\]
Then, the log likelihood can be written as

\[
\Lambda(\theta) = \log f_{Y_1} + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}
\]

\[
= -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \left(\sigma^2 / (1 - \phi^2)\right) - \frac{\{y_1 - (c / (1 - \phi))\}^2}{2\sigma^2 / (1 - \phi^2)}
\]

\[
- \frac{T - 1}{2} \log (2\pi) - \frac{(T - 1)}{2} \log (\sigma^2)
\]

\[
- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}
\]

- The unknown parameters are collected in \(\theta = (c, \phi, \sigma)\)

- We can maximize \(\Lambda(\theta)\) with respect to all those parameters and find the estimates that maximize the probability of having observed such a sample.

\[
\max_{\theta} \Lambda(\theta)
\]

- Sometimes, we can even put constraints (such as \(|\phi| < 1\))

- Q: Is it necessary to put the constraint \(\sigma^2 > 0\)?
• Note: If we forget the first observation, then we can write (setting $c = 0$) the FOC:

$$- \sum_{t=2}^{T} \frac{\partial}{\partial \phi} \frac{(y_t - \phi y_{t-1})^2}{2\sigma^2} = 0$$

$$\sum_{t=2}^{T} y_{t-1} (y_t - \phi y_{t-1}) = 0$$

$$\hat{\phi} = \frac{\sum_{t=2}^{T} y_{t-1}y_t}{\sum_{t=2}^{T} y_{t-1}^2}$$

• RESULT: In the *univariate linear* regression case, OLS, GMM, MLE are equivalent!!!
• To summarize the maximum likelihood principle:
  (a) Make a distributional assumption about the data
  (b) Use the conditioning to write the joint likelihood function
  (c) For convenience, we work with the log-likelihood function
  (d) Maximize the likelihood function with respect to the parameters

• There are some subtle points.
  – We had to specify the unconditional distribution of the first observation
  – We had to make an assumption about the dependence in the series

• But sometimes, MLE is the only way to go.

• MLE is particularly appealing if we know the distribution of the series. Most other deficiencies can be circumvented.
Now, you will ask: What are the properties of $\hat{\theta}^{MLE}$? More specifically, is it consistent? What is its distribution, where

$$\hat{\theta}^{MLE} = \arg \max \Lambda (\theta)$$

Yes, $\hat{\theta}^{MLE}$ is a consistent estimator of $\theta$.

As you probably expect the asymptotic distribution of $\hat{\theta}^{MLE}$ is normal.

Result:

$$T^{1/2} \left( \hat{\theta}^{MLE} - \theta \right) \sim ^a N (0, V)$$

$$V = \left[ - \frac{\partial^2 \Lambda (\theta)}{\partial \theta \partial \theta'} |_{\theta^{MLE}} \right]^{-1}$$

or

$$V = \sum_{t=1}^{T} l \left( \hat{\theta}^{MLE}, y \right) l \left( \hat{\theta}^{MLE}, y \right)$$

$$l \left( \hat{\theta}^{MLE}, y \right) = \frac{\partial f}{\partial \theta} \left( \hat{\theta}^{MLE}, y \right)$$

But we will not dwell on proving those properties.
Another Example: The log-likelihood of an AR(1)+ARCH(1) process

\[ Y_t = c + \phi Y_{t-1} + u_t \]

- where,
  \[ u_t = \sqrt{h_t} v_t \]

- ARCH(1) is:
  \[ h_t = \zeta + au_{t-1}^2 \]
  where \( v_t \) is iid with mean 0, and \( E(v_t^2) = 1 \).

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

- Recall that \( E(Y_t) = \frac{c}{1-\phi} \) and \( Var(Y_t) = \frac{\sigma^2}{1-\phi^2} \)

- Since \( Y_t \) is a linear function of the \( \varepsilon_t \)'s, then it is also Normal (sum of normals is a normal).

- Therefore, the density (unconditional) of \( Y_t \) is Normal.

- Result: If \( Y_1 \) and \( Y_2 \) are jointly Normal, then the marginals are also normal.

- Therefore,
  \[ f_{Y_2|Y_1} \text{ is } N \left( (c + \phi y_1), h_2 \right) \]
  or for the ARCH(1)
  \[ f_{Y_2|Y_1} = \frac{1}{\sqrt{2\pi (\zeta + au_1^2)}} \exp \left[ -\frac{(y_2 - c - \phi y_1)^2}{2(\zeta + au_1^2)} \right] \]
• Similarly,

\[ f_{Y_3|Y_2} \text{ is } N \left( (c + \phi y_2), h_3 \right) \]

or

\[
f_{Y_3|Y_2} = \frac{1}{\sqrt{2\pi (\zeta + \alpha u_2^2)}} \exp \left[ -\frac{(y_3 - c - \phi y_2)^2}{2 (\zeta + \alpha u_2^2)} \right]
\]
Then, the conditional log likelihood can be written as

$$\Lambda (\theta |y_1) = \sum_{t=2}^{T} \log f_{Y_t | Y_{t-1}}$$

$$= -\frac{(T - 1)}{2} \log (2\pi) - \frac{(1)}{2} \sum_{t=2}^{T} \log \left( \zeta + au_{t-1}^2 \right)$$

$$- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2 \left( \zeta + au_{t-1}^2 \right)}$$

- The unknown parameters are collected in $\theta = (c, \phi, \zeta, \alpha)$

- We can maximize $\Lambda (\theta)$ with respect to all those parameters and find the estimates that maximize the probability of having observed such a sample.

$$\max_{\theta} \Lambda (\theta)$$

- Example: mle_arch.m
• Similarly for GARCH(1,1):

\[ \Lambda (\theta | y_1) = \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}} \]

\[ = -\frac{(T - 1)}{2} \log (2\pi) - \frac{(1)}{2} \sum_{t=2}^{T} \log (h_t) \]

\[ - \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2 (h_t)} \]

• where

\[ h_t = \zeta + \delta h_{t-1} + \alpha u_{t-1}^2 \]