Today’s Agenda

1. Announcement:
   – Brief comments about HW 1 and recent financial events
   – Homework 2 (and dataset) on website tomorrow
   – Proposals due Week 5

2. Conditional Expectation and Linear Regression Analysis (OLS–Ordinary Least Squares)
   – Consistency and asymptotic normality under general conditions
   – Univariate tests

3. Market Efficiency

4. Variance Ratio Tests

5. Forecastability of Returns: Univariate and Bivariate Systems

6. Long-Horizon Forecasts of Returns

7. Multivariate Systems (VAR): Reduced forms
Recall from last time, that we can write any variable $y$ as $y = E(y|x) + \{y - E(y|x)\}$

We denote $\varepsilon = \{y - E(y|x)\}$.

Note that $E(\varepsilon|x) = 0$ by construction.

Recall that we were looking at $E(y|x)$ as a function of $x$.

Now, we assume that $E(y|x)$ is an affine function of $x$, or $E(y|x) = \alpha + \beta x$.

The problem is that we don’t know the parameters $\alpha$ and $\beta$.

But...we can estimate them...and this is the goal of a linear regression analysis.
• Linear Regression analysis, also known as Ordinary Least Squares (or OLS) is by far the most popular model of a conditional mean

• This popularity is due to three factors:
  – The analysis and estimation of a linear function is far, far easier than the estimation of nonlinear functions...but you will find that out for yourselves.
  – Any function $g(x)$ can be approximated (Taylor expansion) with a line around the points of interest.
  – Most finance (and economic) models are linear.
  – Aside: Benefits of non-lineararities and complexity
    * Cost-Benefit Analysis: What are the benefits from modelling higher order terms? If the benefits outweigh the cost of dealing with complex systems, then it is worth doing.
• BIGGER QUESTION: Why do we estimate the relationship between $y_t$ and $x_t$?

• We want to know what is the statistical effect of $x_t$ on $y_t$.

• IMPORTANT: If you have two variables $y_t$ and $x_t$, can we always estimate the linear relationship between them?

• I.e. can we always estimate $\beta$ accurately?

• ANSWER: No. We need the condition $E(\varepsilon_t|x_t) = 0$, or that $\varepsilon_t$ is truly the residual in the regression and that it is not correlated with the explanatory variables $x_t$.

• The variable $x_t$ is called exogenous. It is not related to the shocks $\varepsilon_t$ that affect $y_t$. If it is related to the shocks, we will not be able to distinguish $x_t$ from the shocks, and we will not be able to estimate $\beta$ consistently.
  – If $E(\varepsilon_t|x_t) = 0$ is not satisfied, our estimator of $\beta$ will not be consistent. In other words, it will not be close to $\beta$.

• This is the cardinal sin in empirical work!!!
1 Time Series Linear Regression Models

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

- Given the observations \((y_t, x_t)\), the ordinary least squares estimate of \(\beta\), denoted by \(\hat{\beta}_{ols}\), or simply \(\hat{\beta}\), is the value of \(\beta\) that minimizes the residual sum of squares, or

\[ \hat{\beta}_{ols} = \arg \min_{\beta} \sum_{t=1}^{T} (y_t - x_t' \beta)^2 \]

- The solution is

\[ \hat{\beta} = \left( \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left( \sum_{t=1}^{T} x_t y_t \right) \]

- Note that

\[ \hat{\beta} = \left( \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left( \sum_{t=1}^{T} x_t y_t \right) \]

\[ = \left( \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left( \sum_{t=1}^{T} x_t (x_t \beta + \varepsilon_t) \right) \]

\[ = \left( \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left( \sum_{t=1}^{T} x_t x_t \right) \beta + \left( \sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t \varepsilon_t \]

\[ = \beta + \left( \sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t \varepsilon_t \]
• Note: The solution $\hat{\beta}$ to the minimization problem is very simple.

• But, we can also solve the problem by brute force, by numerically minimizing $\sum_{t=1}^{T} (y_t - x_t'\beta)^2$ w.r.t. $\beta$.

• We should obtain the same solution as $\hat{\beta}$, if our numerical minimization routine is good.

• This is the goal of simpleregression2.m
• Note:

\[ \hat{\beta} - \beta = \left( \sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t \varepsilon_t \]

• If \( \hat{\beta} \) is a good estimate of \( \beta \), it better be the case that \( \left( \sum_{t=1}^{T} x_t' x_t \right)^{-1} \sum_{t=1}^{T} x_t' \varepsilon_t \) is small.

• In fact, we can argue that this piece is “small” and “decreases” as \( T \to \infty \). For that, we need that \( E(x_t \varepsilon_t) = 0 \). Then, we can argue that \( \text{plim}_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t = E(x_t \varepsilon_t) = 0 \) and that \( \frac{1}{T} \sum_{t=1}^{T} x_t x_t' \) converges to a finite number.
Note that $\hat{\beta}$ is a random variable. For different samples $(y_t, x_t)$, we will get different estimates $\hat{\beta}$ of $\beta$.

Therefore, as any random variable, $\hat{\beta}$ has a density.

What is the density of $\hat{\beta}$?

Q: Why do we want to know that?

A: Because we want to know whether $\beta$ is “close” to some hypothesized number $\beta_0$ i.e. we will want to test the null hypothesis: $\beta = \beta_0$.

We can test the hypothesis by seeing how far away $\hat{\beta}$ is from $\beta_0$.

Q: What do we mean by “far away”?

A: To answer this question, we need to know the density of $\hat{\beta}$. 
• Illustration: Suppose in our simulated regression (simpleregression.m), we want to test whether $\beta = 3.1$ and we have $T = 100$.

• Suppose $T = 1000$ and we want to test the same hypothesis. How about for $T = 5000$?

• What do we conclude?

• IMPORTANT: We were able to do this testing because we were able to simulate the distribution of $\hat{\beta}$.

• But when we work with data, we don’t know what is the true data generating process (DGP).

• We need to derive some theoretical results that tell us what is the distribution of $\hat{\beta}$ without relying on simulations.
Before we go any further, we need to (re?) introduce the CLT (Central Limit Theorem).

- **CLT**: Suppose that \( \{x_t\}_{t=1}^{T} \) is a random sample from a distribution with mean \( \mu \) and variance \( \sigma^2 < \infty \). The sample mean, denoted by \( \bar{X} = (1/T) \sum_{t=1}^{T} x_t \), scaled by \( \sqrt{T} \) has an asymptotically normal distribution, or

\[
\sqrt{T} (\bar{X} - \mu) \rightarrow^d N(0, \sigma^2)
\]

- **Interpretation 1**: As we have more and more observations (information), the rescaled mean \( \sqrt{T \bar{X}} \) converges to a normal distribution with mean \( \mu \) and variance \( \sigma^2 \).

- **Interpretation 2**: As we have more and more observations (information), the mean \( \bar{X} \) converges to a normal distribution with mean \( \mu \) and variance \( \frac{\sigma^2}{T} \). This is often written as:

\[
\bar{X} \sim^a N \left( \mu, \frac{\sigma^2}{T} \right)
\]
• Note: The variance of $X$ is $\sigma^2$, but the variance of $\overline{X}$ is $\sigma^2/T$.

• Note: As $T \to \infty$, the variance of $\overline{X}$ decreases, i.e. we gain certainty about the exact location of the population mean $\mu$. 
• Now, let’s go back to the regression analysis, where
\[ \hat{\beta} - \beta = \left( \sum_{t=1}^{T} x'_t x_t \right)^{-1} \sum_{t=1}^{T} x_t \varepsilon_t \]

• First, assuming that \( x_t \) is ergodic for the second moment, then \( \frac{1}{T} \sum_{t=1}^{T} x'_t x_t \to \Sigma_x \), where \( \Sigma_x \) is ??????????

• Second, assume that the product \( x_t \varepsilon_t \) is a random variable, so that (from the CLT)
\[
\left( \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \right) \sim^a N \left( E(x_t \varepsilon_t), \frac{E(x_t \varepsilon_t \varepsilon_t x'_t)}{T} \right)
\]
Here comes a crucial step in regression analysis, a step that we will use over and over again in different "flavors"

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

The above condition comes from the FOC of the minimization problem.

This condition implies that the explanatory variables \( x_t \) and the errors \( \varepsilon_t \) are uncorrelated.

Recall that our model was: \( y_t = \alpha + x'_t\beta + \varepsilon_t \). We want \( E(y|x) = \alpha + x'_t\beta \), so that \( E(\varepsilon_t|x_t) = 0 \).

But we also want \( E(\varepsilon_t) = 0 \). What does this imply for the relationship between \( \varepsilon_t \) and \( x_t \)?

You will have to show for homework that if \( E(\varepsilon_t|x_t) = 0 \), then \( E(\varepsilon_t x_t) = 0 \).

Therefore, the \( E(\varepsilon_t x_t) = 0 \) condition is really an implication of our model.

As mentioned above \( x_t \) is called exogenous.
Given that

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

we have

\[
\left( \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \right) \sim^a N \left( 0, \frac{E(x_t \varepsilon_t \varepsilon_t x_t')}{T} \right)
\]
Now comes another crucial assumption.

- We assume that $E(x_t \varepsilon_t \varepsilon_t' x_t') = E(x_t x_t') E(\varepsilon_t \varepsilon_t) = \sum_x \sigma_{\varepsilon}^2$, where $\sum_x$ is the covariance matrix of $x_t$ and $\sigma_{\varepsilon}^2$ is the variance of the residuals.

- This assumption will be explained later....but it has a lot of sub-assumptions. For example
  - The $\varepsilon_t$ are serially uncorrelated
  - The $\varepsilon_t$ are stationary
  - The variance of $\varepsilon_t$ does not depend on $x_t$.

- The assumption implies that:
  $$\left(\frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \right) \sim^a N \left( 0, \frac{\sum_x \sigma_{\varepsilon}^2}{T} \right)$$
Now we put all the pieces together:
\[
\hat{\beta} - \beta = \left( \sum_{t=1}^{T} x'_t x_t \right)^{-1} \sum_{t=1}^{T} x_t \varepsilon_t
\]
\[
= \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \xrightarrow{p} 0
\]

- Or, \( \text{plim} \hat{\beta} = \beta \). Intuitively, as \( T \) increases, the estimate \( \hat{\beta} \) gets closer and closer to \( \beta \).
- To emphasise once again, if \( E(\varepsilon_t x_t) \neq 0 \), then
  \[
  \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \xrightarrow{p} E(\varepsilon_t x_t) \neq 0,
  \]
  and
  \[
  \hat{\beta} - \beta = \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \xrightarrow{p} \frac{E(\varepsilon_t x_t)}{\sum_x} \neq 0
  \]
- No matter how much data we have, \( \hat{\beta} = \beta + R.V. \)
- We don’t have a good estimate of \( \beta \).
More interestingly, 
\[
\left( \hat{\beta} - \beta \right) = \frac{1}{T} \frac{\sum_{t=1}^{T} x_t \varepsilon_t}{\sum_{t=1}^{T} x_t' x_t} = N \left( 0, \frac{\sum x \sigma^2_{\varepsilon}}{T} \right)
\]
\[
\sim a \frac{\sum x}{\sum x} \sim a N \left( 0, \frac{\sigma^2_{\varepsilon} \Sigma^{-1}_x}{T} \right)
\]

Or, we can also write
\[
\sqrt{T} \left( \hat{\beta} - \beta \right) \rightarrow^d N \left( 0, \sigma^2_{\varepsilon} \Sigma^{-1}_x \right)
\]

- In words, the estimate $\hat{\beta}$ has an approximately normal distribution, centered at $\beta$, and with variance $\frac{\sigma^2_{\varepsilon} \Sigma^{-1}_x}{T}$.

- We can estimate $\sigma^2_{\varepsilon}$ and $\Sigma^{-1}_x$ using their sample moments (and invoking ergoticity).

- Hence, we don’t have to rely on simulations to know the distribution of $\hat{\beta}$. 
• We have made the strong assumption that \( E(x_t \varepsilon_t \varepsilon_t' x_t') = E(x_t x_t') E(\varepsilon_t \varepsilon_t) \). Suppose that this is not the case.

• Instead \( E(x_t \varepsilon_t \varepsilon_t' x_t') = \sum_{x \varepsilon} \).

• Then, (we still have that \( E(x_t \varepsilon_t) = 0 \))

\[
\left( \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \right) \sim^a N \left( 0, \frac{\sum_{x \varepsilon}}{T} \right)
\]

• Therefore,

\[
\left( \hat{\beta} - \beta \right) = \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t \cdot \frac{1}{T} \sum_{t=1}^{T} x_t' x_t \\
\sim^a N \left( 0, \frac{\sum_{x \varepsilon}}{T} \right) \sim^a N \left( 0, \frac{\sum_{x}^{-1} \sum_{x \varepsilon} \sum_{x}^{-1}}{T} \right)
\]

• Now, the question is: How do we estimate \( \sum_{x \varepsilon} \). This will wait.

• The point is that all the results still go through, with the exception that the standard error of \( \hat{\beta} \) is different.

• But the estimator is consistent. (It will not be if \( E(\varepsilon x) \neq 0 \).)
Since we have gotten too deep into econometrics and we really want to go to applications, here is a result that we will not prove.

- You probably know that if we have a random sample \( \{x_t\}_{t=1}^T \), then \( \frac{X - \mu}{s/\sqrt{T}} \sim^a N(0, 1) \) (From CLT).

- Similarly, in a regression, \( \frac{\hat{\beta} - \beta}{s_b/\sqrt{T}} \sim^a N(0, 1) \), where \( s_b = \text{diagonal element of the matrix } \sigma^2 \sum_{x}^{-1} \).

- This statistic is called a \( t \) statistic, because when the sample size is small (say 20 observations), the distribution of \( \frac{\hat{\beta} - \beta}{s_b/\sqrt{T}} \) is better approximated by a \( t \)-distribution. When the sample size increases, we use the normal distribution.

- Usually, we want to know whether \( \beta = 0 \). This is called a null hypothesis. Under this null hypothesis, the statistic becomes

\[
t = \frac{\hat{\beta} - 0}{s_b/\sqrt{T}}
\]

- This test is fundamental in analyzing regressions.
In-Sample Fit

We might be interested in knowing how well the data fits our model. But in order to answer this question, we first have to define a metric that measures the fit of the model.

• Recall that \( \frac{1}{T} \sum_{t=1}^{T} (y - \bar{y})_t^2 \) is the total variance of the data.

• But, note that

\[
y_t = \hat{\alpha} + x_t\hat{\beta} + \hat{\varepsilon}_t \\
= \bar{y} - \bar{x}\hat{\beta} + x_t\hat{\beta} + \hat{\varepsilon}_t \\
y_t - \bar{y} = \hat{\beta} (x_t - \bar{x}) + \hat{\varepsilon}_t
\]

• We can then write

\[
\frac{1}{T} \sum_{t=1}^{T} (y - \bar{y})_t^2 = \frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})^2 \hat{\beta}^2 + \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t^2
\]

• Therefore, dividing by \( \frac{1}{T} \sum_{t=1}^{T} (y - \bar{y})_t^2 \), we obtain

\[
1 = \frac{\frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})^2 \hat{\beta}^2}{\frac{1}{T} \sum_{t=1}^{T} (y - \bar{y})_t^2} + \frac{\frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t^2}{\frac{1}{T} \sum_{t=1}^{T} (y - \bar{y})_t^2}
\]

• The first term is the variance of \( y_t \) explained by the conditional mean. This term is denoted by \( R^2 \). Properties of \( R^2??? \)

• The second term is the fraction of the variance of \( y_t \) that is still unexplained.
Important:

- Always run regressions with an intercept. I.e., include a column of 1's in the $x$ matrix.

- Including an intercept has the effect of immediately demeaning the $y$ and $x$ variables (recall our example during last class of regressing $y_t$ on 1's).

- If the intercept is zero, we can test it. No impact on the estimates of the slope coefficients.

- If it is not zero and we don’t include an intercept, then the estimates of the slope coefficients would be inconsistent!

- If we include an intercept, then the $R^2$ is always between 0 and 1.
Let’s slow down and think of what we have just done:

- We have specified the conditional mean as a linear function that has parameters to be estimated.
- We have estimated those parameters in such a way that the sum of the squared residuals is minimized.
- In this fashion, we have obtained an estimate of the parameters, called $\hat{\beta}$. This estimate has some nice properties:
  - As we increase $T$, the estimate gets closer and closer to the true value $\beta$.
  - We know that $\hat{\beta}$ has a normal distribution with mean $\beta$ and a variance that we can estimate.
  - Therefore, we can construct a statistical test whether the true $\beta$ is zero (or any other value).
Digression: Robust (Quantile) Regression

- So far, we have focused on the conditional mean $E(y_t|x_t)$ and assuming that $E(y_t|x_t) = x_t'\beta$, we decided to minimize the sum of the squared residuals, to obtain

$$\hat{\beta}^{ols} = \arg \min_{\beta} \sum_{t=1}^{T} (y_t - x_t'\beta)^2$$

- But suppose that the conditional distribution of $y_t$ given $x_t$ is ill behaved (outliers, skewness, kurtosis, etc.)

- We want to estimate the entire distribution of $y_t$ given $x_t$.

- We will see later how to do that...But here we can do something else.

- Let the unconditional distribution of $y_t$ be $F_{y_t}(y) = \Pr(y_t \leq y)$. Then, for any quantile $\tau$, $0 < \tau < 1$, we can define the inverse of $F_{y_t}(\cdot)$ as $Q_{y_t}(\tau) = \inf \{ r : F_{y_t}(y) \geq \tau \}$. The function $Q_{y_t}(\cdot)$ is called the unconditional quantile function of $y$. $Q_y(0.5)$ is the 50th quantile, or the median, of $y_t$. 
• The introduction of conditional quantiles is easily understood by making an analogy to the familiar least squares estimation.

• The conditional quantile function $Q_{y|x}(\tau|X = x) = x'_t \beta(\tau)$ can be estimated by solving:

$$\hat{\beta}(\tau) = \arg \min_{\beta} \sum_{t} \rho_{\tau}(y_t - x'_t \beta)$$

where $\rho_{\tau}(.)$ is a piecewise linear “check function,” defined as $\rho_{\tau}(u) = u(\tau - I(u < 0))$ and $I(.)$ is the indicator function.
The function $\rho_\tau(\cdot)$ selects the quantile $\tau$ to be estimated (see Koenker and Hallock (2000)).

For the case $\tau = 0.5$, $\rho_{0.5}(u) = |u|$ and the solution of the above problem, $\hat{\eta}(0.5)$, is equivalent to minimizing the sum of absolute values of the residuals.

From the definition, $\hat{Q}_{y|x}(0.5|x = x_t) = x_t^T \hat{\beta}(0.5)$ represents the estimate of the conditional median of $y_t$.

For different values of $\tau$, the estimate $\hat{\beta}(\tau)$ is the effect of $x_t$ on the $\tau$th quantile of $y_t$.

An estimate of the entire function $\hat{Q}_{y|x}(\tau|y = x_t)$ can be computed from the above relation. For a more detailed introduction to quantile regressions, please refer to Koenker and Hallock (2000), and Koenker (2000).
2 Market Efficiency (Chapter 2)

- Earlier Models of Asset Pricing:
  - Fair Game: Cardano’s (c.1565) *LiberdeLudoAleae* (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

  \[ E (P_{t+1}|P_t, P_{t-1}, \ldots) = P_t \]
Market Efficiency: An efficient market is one in which prices fully reflect available information.

- Levels of Market Efficiency
  
  (a) Weak Form: prices reflect information set comprising past market trading data (i.e. prices, volume, dividends, etc.)

  (b) Semi-Strong Form: prices reflect information set comprising past market trading data plus all other currently available public information

  (c) Strong Form: prices reflect all public and private information relevant to a stock

Notes:

1. The different forms of market efficiency focus on what kind of information is included in the information set $I_t$.

2. This is an economic statement of how people behave and how fast they incorporate relevant information into prices.

3. It is always a good idea to ask “How efficient are markets?” rather than “Are markets efficient?”
• Note: The martingale model of prices is nothing but 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 price (return) 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 \( P_t = e^{pt} \)) 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 \( p_t - p_{t-1} = \mu + \varepsilon_t = r_t \) (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 \( P_t = e^{pt} \) 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 \( r_{t,t+2} \):

\[
\text{Var} (r_{t,t+2}) = \text{Var} (r_{t+1} + r_{t+2}) = \text{Var} (r_{t+1}) + \text{Var} (r_{t+2}) = 2\sigma_r^2
\]

In general, for the q-th period (long-horizon) return \( r_{t,t+q} \):

\[
\text{Var} (r_{t,t+q}) = \text{Var} (r_{t+1} + \ldots + r_{t+q}) = q\sigma_r^2
\]

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):

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

or

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

This is a simple test.

1. Compute the variance of the \( q \)-period returns.
2. Compute the variance of the 1-period returns.
3. It must be the case that the ratio of \( \frac{V \text{ar} (r_{t,t+q})}{q \text{Var} (r_{t+1})} \) 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
\[
\sqrt{Tq} \left( \hat{VR}(q) - 1 \right) \sim^a N \left( 0, 2(q-1) \right)
\]
(recall last lecture and the results about \( \bar{X} \) and \( \hat{\beta} \)) or
\[
\left( \hat{VR}(q) - 1 \right) \sim^a N \left( 0, \frac{2(q-1)}{Tq} \right)
\]
• Therefore, we can conduct testing in the usual way:
Form a statistic
\[
Z = \frac{\hat{VR}(q) - 1}{\sqrt{\frac{2(q-1)}{Tq}}} \sim^a N \left( 0, 1 \right)
\]
• 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:

$$\sqrt{Tq \left( \widehat{VR_{\text{overlap}}} (q) - 1 \right)} \sim^a N \left( 0, \frac{2(2q-1)(q-1)}{3q} \right)$$

• 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

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

- 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
\[ p_t \approx k + \rho \left[ k + \rho p_{t+2} - r_{t+2} + (1 - \rho) d_{t+2} \right] - r_{t+1} + (1 - \rho) d_{t+1} \]
\[ \approx K + \rho^2 p_{t+2} - \rho r_{t+2} - r_{t+1} + \rho (1 - \rho) d_{t+2} + (1 - \rho) d_{t+1} \]
\[ \approx \ldots \]
\[ \approx K + \rho^j p_{t+j} - \sum_{i=0}^{j} \rho^i r_{t+1+i} + \sum_{i=0}^{j} \rho^i (1 - \rho) d_{t+1+i} \]

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 Predictive Regressions

5.1 Econometrics

A predictive regression is

\[ r_{t+1} = \alpha + \beta x_t + \varepsilon_{t+1} \]
\[ x_{t+1} = \mu + \phi x_t + u_{t+1} \]

Notes:
- The timing of \( r_t \) and \( x_t \) is crucial.
- \( x_t \) must be observable (for now).
- \( \varepsilon_t \) and \( u_t \) might be (often are) correlated. If \( v_{t+1} = [\varepsilon_{t+1} \ u_{t+1}]' \), then
  \[ E(v_{t+1}v_{t+1}') = \begin{bmatrix} \sigma^2_{\varepsilon} & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & \sigma^2_u \end{bmatrix} \]
- \( x_t \) might be (often is) serially correlated.
- Since \( E(\varepsilon_{t+1}|x_t) = 0 \), we can run a regression.
- But notice that \( E(\varepsilon_{t+1}|x_{t+1}) \) might not be equal to zero, if \( \varepsilon_{t+1} \) and \( u_{t+1} \) are correlated.
- In small samples, we will have a bias, but asymptotically everything is OK.
- This is the typical predictive regression in finance.
  - Example: \( r_{t+1} \) is predicted by \( d_t - p_t \)
  - Example: \( \Delta d_{t+1} \) is predicted by \( d_t - p_t \)
Bias in predictive regressions:
We can show that (Kendall (1954), Stambaugh (1999)):

\[ E \left( \hat{\beta} - \beta \right) = \frac{\sigma_{\varepsilon u}}{\sigma_u^2} E \left( \hat{\phi} - \phi \right) \]

We can also show that

\[ E \left( \hat{\phi} - \phi \right) = -\frac{(1 + 3\phi)}{T} + O(1/T^2) \]

Then

\[ E \left( \hat{\beta} - \beta \right) = -\frac{\sigma_{\varepsilon u}(1 + 3\phi)}{\sigma_u^2 \frac{T}{T}} \]

In sum, the bias depends on two factors:

- The persistence of the predictor ($\phi$).
- The correlation between the shocks of the predictor and the shocks to returns
Stambaugh (1999): Table 1–Calibrated to US stock market

<table>
<thead>
<tr>
<th></th>
<th>Sample Period</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A. True properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0.07</td>
<td>0.18</td>
<td>0.18</td>
<td>0.42</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.16</td>
<td>0.33</td>
<td>0.27</td>
<td>0.45</td>
</tr>
<tr>
<td>skewness</td>
<td>0.71</td>
<td>0.83</td>
<td>0.98</td>
<td>1.29</td>
</tr>
<tr>
<td>kurtosis</td>
<td>3.84</td>
<td>4.14</td>
<td>4.62</td>
<td>5.83</td>
</tr>
<tr>
<td>p-value for $\beta = 0$</td>
<td>0.17</td>
<td>0.42</td>
<td>0.15</td>
<td>0.64</td>
</tr>
<tr>
<td>B. Properties in the standard regression setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bias</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.14</td>
<td>0.27</td>
<td>0.20</td>
<td>0.30</td>
</tr>
<tr>
<td>skewness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>kurtosis</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>p-value for $\beta = 0$</td>
<td>0.06</td>
<td>0.22</td>
<td>0.02</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Can we get rid of the bias?

Proposition:

\[ \hat{\beta} - Bias \]

\[ \hat{\beta} - \left( -\frac{\sigma_{\varepsilon u}(1 + 3\phi)}{\sigma^2_u T} \right) \]

Problem: We need the true \( \phi \). Estimates are bias!

How about the following adjustment:

\[ \hat{\beta} - \left( -\frac{\sigma_{\varepsilon u}(1 + 3)}{\sigma^2_u T} \right) \]

Problem: Over-corrects for (not a function of) the persistence
5.2 Economics

Recall the decomposition

$$r_{t+1} = E(r_{t+1}|I_t) + \varepsilon_{t+1}$$

- Predictive regressions focus on $E(r_{t+1}|I_t)$ rather than $\varepsilon_{t+1}$.

- The rational for that is the long-horizon decomposition.

- Predictive regressions are run at long-horizons.

- Broader question: Are expected returns time-varying? Do they vary with the business cycle

- Expected returns can vary because
  - Time varying risk aversion (e.g., CAPM)
  - Time varying risk premium (e.g., CAPM)

- Other predictors:
  - Term Spread (aka Term Premium)
  - Default Spread (aka Term Premium)
  - Interest rate
  - Eg: Fama and French (1993)
6 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:
  \[ -r_{t+1} + ... + r_{t+k} = \alpha_1 + \beta_1 (d_t - p_t) + u_{t+k} \]
  \[ -\Delta d_{t+1} + ... + \Delta d_{t+k} = \alpha_2 + \beta_2 (d_t - p_t) + e_{t+k} \]

- 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

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

- 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 \([\epsilon_{1t} \; \epsilon_{2t}]\) have a covariance matrix \( \Sigma \).
- This system can be written in a more elegant form as:

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

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

- Suppose that \( \beta_{12} = 0 \). Then \( r_t = \alpha_1 + \beta_{11} r_{t-1} + \varepsilon_{1,t} \), 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: “$\sigma_{t-1}$ Granger-causes $r_t$”, we test $\beta_{12} = 0$

- For the hypothesis: “$r_{t-1}$ 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?