Introduction to Econometrics: Notes

Daniel Cullen & Travis Cyronek

University of California, Santa Barbara

Updated: January 15, 2018

Contents

1 Introduction

One of the main goals of science is to uncover causal relationships. The problem, for those concerned with social or economic policy, is that we seldom are able to perform controlled experiments similar to those conducted by natural scientists. Instead, we draw our inferences from the analysis of non-experimental data, and that is the function of econometrics. The main tool of econometricians is the regression, a statistical tool for understanding the relationship between different variables. Before learning about the regression we first need to review some math and statistics.

2 Math & Stats Review

2.1 Expected Value

One of the more important features of the distribution of a random variable is its *expected value*. This feature gives us an idea about where a distribution is centered. Some books might call the expected value the "long-run average" of an experiment. That is, it's the average of an experiment if we were to continually repeat the experiment. How we write the expected value depends on whether or not the variable is discrete or continuous (but notice the similarities).

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_i \qquad \text{(Discrete)}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{(Continuous)}$$

There are a few useful properties of expected values that we should be very comfortable with in this class. Let X and Y be random variables; let a and b be scalar constants.

$$\begin{aligned}
&\mathbb{E}[a] = a \\
&\mathbb{E}[aX] = a\mathbb{E}[X] \\
&\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \\
&\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y] \qquad \text{(this simply combines the previous two)}
\end{aligned}$$

This is often called the "linearity of expectations" because scalar multiplication and addition are both linear operators. As a side note, because we like to write things succinctly, we often times will define $\mu \equiv \mathbb{E}[X]$.

2.2 Variance, Standard Deviation, and Covariance

Another important feature of a distribution is its *variance* (and, by extension, its standard deviation). This gives us a measure of the spread, or dispersion, of the outcomes of a random variable. It is defined as follows.

$$\begin{aligned} Var(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \qquad \text{(another way to write it)} \end{aligned}$$

Again, because we love notation, we'll often write $Var(X) = \sigma^2$. Next, the *standard deviation* of the random variable X is simply the square root of its variance. I.e.

$$St.dev(X) = \sqrt{Var(X)} \qquad \text{or} \qquad \sigma = \sqrt{\sigma^2}$$

Last, the covariance of two random variables X and Y can tell us how they move together. That is, if X is high when Y is high, then the covariance is positive. If X is high when Y is low (or vice versa), the covariance is negative. The covariance is defined as

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \equiv \sigma_{XY}$$

Note: $corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$.

Just like with the expected value, there are a lot of useful properties that we should be very comfortable with.

$$\begin{aligned} &Var(a) = 0 \\ &Var(aX) = a^2 Var(X) \\ &Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X, Y) \end{aligned}$$

$$\begin{aligned} &Cov(a + X, b + Y) = Cov(X, Y) \\ &Cov(aX, bY) = ab\, Cov(X, Y) \\ &Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \\ &Cov(X, X) = Var(X) \\ &Cov(X, a) = 0 \end{aligned}$$

2.3 Derivatives and Optimization

Here we'll just do a few examples to refresh our memory on differentiation. (The last one is also useful to remind us how summations work.)

$$f(x,y) = log(xy) \qquad \longrightarrow \qquad \frac{\partial f}{\partial x} = \frac{y}{xy} = \frac{1}{x}, \qquad \frac{\partial f}{\partial y} = \frac{1}{y}$$

$$y = \alpha + \beta x + \delta z + \lambda xz \qquad \longrightarrow \qquad \frac{\partial y}{\partial z} = \delta + \lambda x$$

$$c(x) = \sum_{i=1}^{n} (\beta x + 5m_i) \qquad \longrightarrow \qquad \frac{dc(x)}{dx} = n\beta$$

Now consider a problem of optimization. Suppose we know that $wage = \beta_1 + \beta_2 Age + \beta_3 Age^2$ where $\beta_1 = 200$, $\beta_2 = 800$, and $\beta_3 = -10$. At what age would you earn the highest wage? $\boxed{40}$

2.4 Standard Error & Standard Deviation

The standard error is the standard deviation of a estimator (a function of random variables). The standard error of the sample mean is an estimate of how far the sample mean is likely to be from the population mean, whereas the standard deviation of the sample is the degree to which individuals within the sample differ from the sample mean. If the population standard deviation is finite, the standard error of the mean of the sample will tend to zero with increasing sample size, because the estimate of the population mean will improve, while the standard deviation of the sample will tend to approximate the population standard deviation as the sample size increases.

Random Variable:	Y	$\sigma_Y = \sqrt{Var(Y)}$
Estimator:	$\widehat{\mu} = \frac{1}{n}\sum y_i$	$\sigma_{\widehat{\mu}} = \sqrt{Var(\widehat{\mu})}$

Standard Error of $\widehat{\mu}$:

$$\begin{aligned}
\sigma_{\widehat{\mu}} &= \sqrt{Var(\widehat{\mu})} \\
&= \sqrt{Var\left(\frac{1}{n}\sum y_i\right)} \\
&= \sqrt{\left(\frac{1}{n}\right)^2 Var\left(\sum y_i\right)} \\
&= \sqrt{\left(\frac{1}{n}\right)^2 \left[\sum \underbrace{Var\left(y_i\right)}_{\sigma_Y^2} + \sum_{i\neq j} \underbrace{Cov(y_i, y_j)}_{0}\right]} \\
&= \sqrt{\left(\frac{1}{n}\right)^2 \left[\sum \sigma_Y^2\right]} \\
&= \sqrt{\left(\frac{1}{n}\right)^2 [n\sigma_Y^2]} \\
&= \sqrt{\frac{\sigma_Y^2}{n}} \\
&= \frac{\sigma_Y}{\sqrt{n}}
\end{aligned}$$

Standard Error of $\widehat{\beta}$:

$$
\begin{aligned}
SE\left(\widehat{\beta}\right) = \sigma_{\widehat{\beta}} &= \sqrt{Var(\widehat{\beta})} \\
&= \sqrt{Var\left(\frac{\sum x_i y_i}{\sum x_i^2}\right)} \\
&= \sqrt{Var\left(\frac{\sum x_i(\beta x_i + \varepsilon_i)}{\sum x_i^2}\right)} \\
&= \sqrt{Var\left(\frac{\beta \sum x_i^2}{\sum x_i^2} + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right)} \\
&= \sqrt{Var\left(\beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right)} \\
&= \sqrt{Var\left(\frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right)} \\
&= \sqrt{\left(\frac{1}{\sum x_i^2}\right)^2 Var\left(\sum x_i \varepsilon_i\right)} \\
&= \sqrt{\left(\frac{1}{\sum x_i^2}\right)^2 \left[\sum Var\left(x_i \varepsilon_i\right) + \sum_{i \neq j} x_i x_j Cov(\varepsilon_i, \varepsilon_j)\right]} \\
&= \sqrt{\left(\frac{1}{\sum x_i^2}\right)^2 \left[\sum x_i^2 \underbrace{Var\left(\varepsilon_i\right)}_{\sigma_\varepsilon^2} + \sum_{i \neq j} x_i x_j \underbrace{Cov(\varepsilon_i, \varepsilon_j)}_{0}\right]} \\
&= \sqrt{\left(\frac{1}{\sum x_i^2}\right)^2 \left[\sum x_i^2 \sigma_\varepsilon^2\right]} \\
&= \sqrt{\left(\frac{1}{\sum x_i^2}\right)^2 \left[\sigma_\varepsilon^2 \sum x_i^2\right]} \\
&= \sqrt{\left(\frac{\sigma_\varepsilon^2}{\sum x_i^2}\right)} \\
&= \frac{\sigma_\varepsilon}{\sqrt{\sum x_i^2}}
\end{aligned}
$$

3 Parameters & Estimators

First we need to learn some definitions to fix some ideas. Being able to understand the different "levels" that we will be working with will make other things much more easy to wrap your heads around.

Parameter: true characteristic / feature about something (often unknown); (e.g. $\mu = \mathbb{E}[Y_i]$, $\sigma^2 = Var(Y_i)$, β)

Estimator: a function of a random variable(s) used to obtain an estimate for some parameter of interest; (e.g. $\widehat{\mu} = \frac{1}{n}\sum Y_i$, $\widehat{\sigma^2} = \frac{1}{n-1}\sum(Y_i - \overline{Y})^2$, $\widehat{\beta}$)

Estimate: the value obtained after applying an estimator to actual data; (e.g. m, s^2, b)

3.1 Model of the Sample Mean

Here we will formalize what we already know about the sample mean and situate ourselves in the "big picture" we just saw. First, we'll want to be explicit about the assumptions we are making for this model. This might seem weird at first, but (in due time) this should hopefully make more sense.

	$A1$:	$Y_i = \mu + \varepsilon_i$ is the true DGP and we have $i = 1, \ldots, n$ observations
	$A2$:	$\mathbb{E}[\varepsilon_i] = 0$
	$A3$:	$Var(\varepsilon_i) = \sigma^2 \, \forall i$
		$Cov(\varepsilon_i, \varepsilon_j) = 0 \, \forall i \neq j$
(sometimes)	$A4$:	$\varepsilon_i \sim Normally$

What we are interested in is trying to determine what μ is (the parameter of interest). To get an idea, we can utilize the sample mean estimator, $\widehat{\mu} = \frac{1}{n}\sum_i^n Y_i$, to give us an estimate (presuming, of course, that we have data on y_i).

3.2 Important Properties of Estimators

Unbiasedness: An estimator of β is unbiased if and only if $\mathbb{E}[\widehat{\beta}] = \beta$. This is to say that, on average, the estimates obtained from the estimator are the truth.

Example: consider the sample average as an estimator for μ and show that it is unbiased.

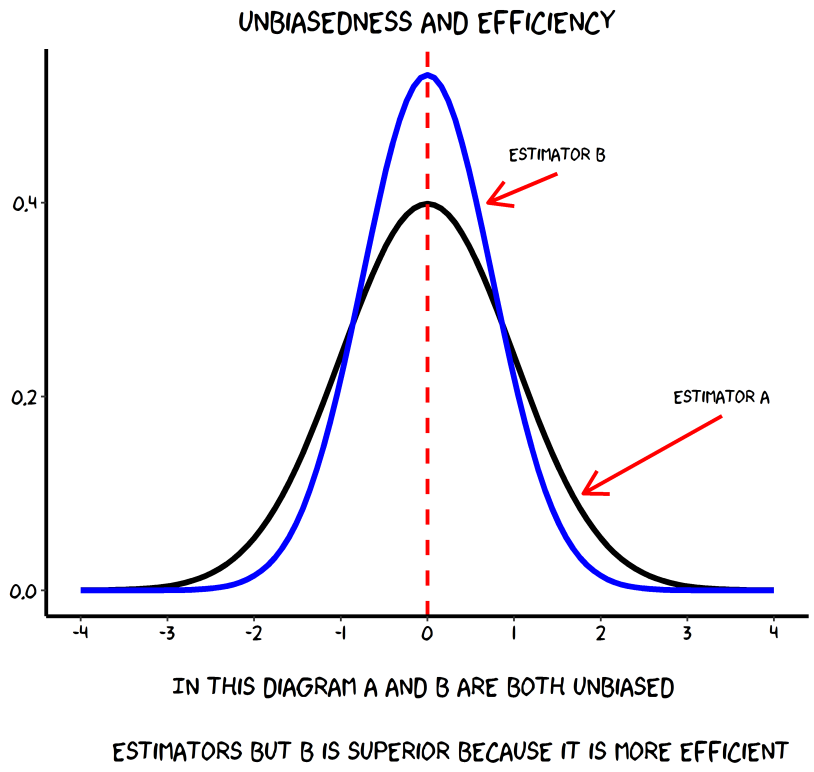
$$
\begin{aligned}
\mathbb{E}[\widehat{\mu}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\mu+\varepsilon_i)\right] && \text{(A1)} \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mu + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right] && \text{(arithmetic)} \\
&= \mu + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\varepsilon_i] && \text{(distributing the expected value)} \\
&= \mu && \text{(A2)}
\end{aligned}
$$

Efficiency: This is a way of getting at how accurate an estimator is. Where as unbiasedness is concerned about the mean of an estimator, efficiency is concerned about the variance of an estimator. If an estimator has low variance, it is said to be efficient. (Note that efficiency is a relative term. That is, one estimator is more efficient than another.) To give us a simple warm-up exercise, let's calculate the variance of an estimator.

Example: consider the sample average as an estimator for μ and find its variance.

$$
\begin{aligned}
Var(\widehat{\mu}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= Var\left(\frac{1}{n}\sum_{i=1}^{n}(\mu+\varepsilon_i)\right) && \text{(A1)} \\
&= Var\left(\mu + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right) && \text{(arithmetic)} \\
&= Var\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right) && \text{(variance of a constant is 0)} \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(\varepsilon_i) + \sum_{i\neq j} Cov(\varepsilon_i,\varepsilon_j)\right] && \text{(generalized variance rule)} \\
&= \frac{1}{n^{\cancel{2}}}\cancel{n}\sigma^2 && \text{(A3)} \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Below is a graph that illustrates both of the above properties.



3.3 (Linear) Regression Estimator

Now let's take this structure to regressions. A *regression* is a statistical process for estimating the relationship between variables. There are many "types" of regressions that we might want to run. We will mostly consider regression models of the following variety (for now we abstract away from an intercept for simplicity of algebra).

$$y_i = \beta x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n \qquad \text{(Model / DGP)}$$

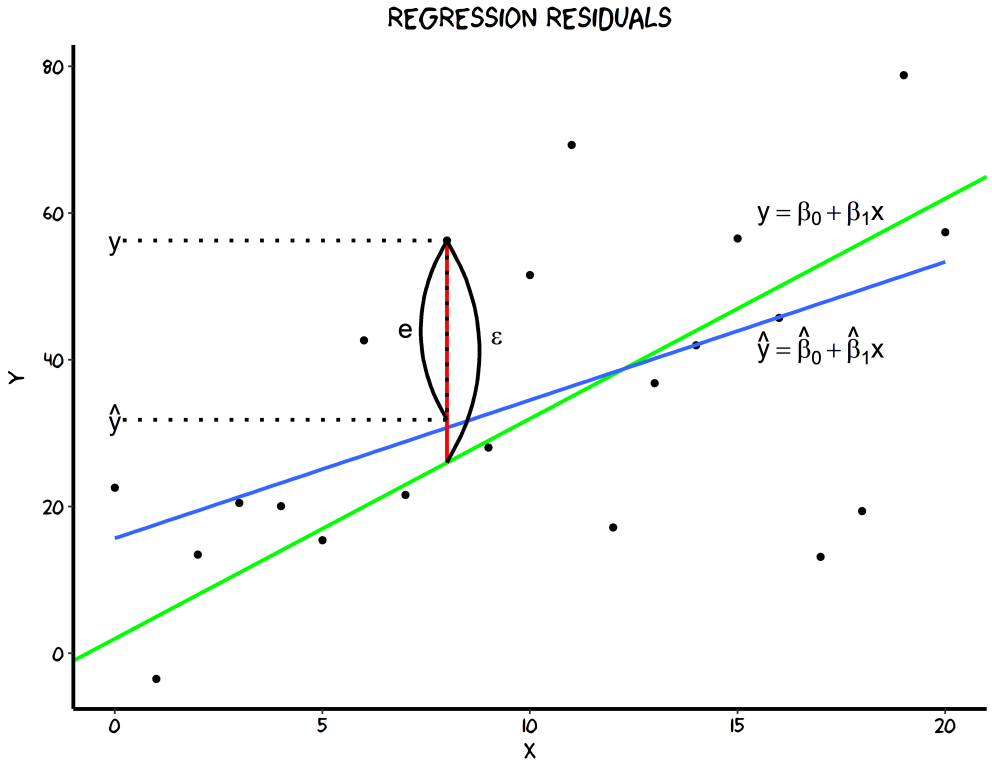
$$y_i = \widehat{\beta} x_i + e_i$$

where y_i is the dependent variable, x_i is the independent variable, β is the slope parameter, ε_i is the error term, $\widehat{\beta}$ is the slope estimator, and e_i is the residual. One of the most important "things" in the regression model is how the error term is structured. The error term (which is unobserved) captures all of the things that impact y_i that are not included as right hand side variables.

The "residual" is the term used for what we can think of as the "estimator of the error term." That is, suppose we had some estimate for β obtained using $\widehat{\beta}$. The residuals would be given by the vertical distance between an observation and the estimated line:

$$e_i = y_i - \widehat{\beta} x_i$$

Let's take a step back and try to visualize these things (remembering the different "levels" that we are operating on from the beginning of today).



The question still remains, though: how do we (formally) draw a line of "best fit"? There are a couple different ways. The first, most natural way might be to minimize the sum of the residuals (the distances between the points and our line). Does anyone see a problem with this? (Indeed, by doing so points above the line might cancel out with points below the line and we would effectively be throwing away important information.) To address this problem of "negative" residuals canceling out positive ones, we will square the residuals, making them all positive. Then we minimize the sum of those. That is, we will draw a line such that we are minimizing the sum of the squared residuals (SSR). This will give us the Ordinary Least Squares estimator of β (sometimes written as $\widehat{\beta}^{OLS}$). Mathematically,

$$\widehat{\beta}^{OLS} \equiv \underset{\widehat{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \widehat{\beta} x_i\right)^2 \qquad \text{(recalling that } e_i = y_i - \widehat{\beta} x_i\text{)}$$

Now we want to actually find the equation for $\widehat{\beta}^{OLS}$. All we have to do is find the critical point that minimizes the function above. To do so, we differentiate w.r.t. $\widehat{\beta}$, set it equal to zero, and solve for $\widehat{\beta}$.

$$
\begin{aligned}
\text{F.O.C.:} \quad & \sum_{i=1}^{n} -2(y_i - \widehat{\beta} x_i) x_i = 0 \\
\iff \quad & 2\sum_{i=1}^{n} y_i x_i - 2\widehat{\beta}\sum_{i=1}^{n} x_i^2 = 0 \\
\iff \quad & \sum_{i=1}^{n} y_i x_i = \widehat{\beta}\sum_{i=1}^{n} x_i^2 \\
\iff \quad & \boxed{\widehat{\beta}^{OLS} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}}
\end{aligned}
$$

3.4 Example

Spring 2015, Midterm 1, Question 3. You have the following data

$$
\begin{aligned}
x_1 &= 1 & y_1 &= 2 \\
x_2 &= 6 & y_2 &= 6 \\
x_3 &= 4 & y_3 &= 8
\end{aligned}
$$

and the true DGP is $y_i = \beta x_i + \varepsilon_i$. Your friend calculated two estimates for β: $\widehat{\beta}_1 = 1$ and $\widehat{\beta}_2 = 2$.

(a) Find the sum of squared residuals for $\widehat{\beta}_1$. (hint: you should get an actual number)

Answer:

$$
\begin{aligned}
SSR_1 &= e_1^2 + e_2^2 + e_3^2 \\
&= (y_1 - \widehat{\beta}_1 x_1)^2 + (y_2 - \widehat{\beta}_1 x_2)^2 + (y_2 - \widehat{\beta}_1 x_2)^2 \\
&= (2 - 1(2))^2 + (6 - 1(6))^2 + (8 - 1(4))^2 \\
&= 17
\end{aligned}
$$

(b) Find the sum of squared residuals for $\widehat{\beta}_2$. (hint: you should get an actual number)

Answer:

$$\begin{aligned} SSR_2 &= e_1^2 + e_2^2 + e_3^2 \\ &= (y_1 - \widehat{\beta}_2 x_1)^2 + (y_2 - \widehat{\beta}_2 x_2)^2 + (y_2 - \widehat{\beta}_2 x_2)^2 \\ &= (2 - 2(2))^2 + (6 - 2(6))^2 + (8 - 2(4))^2 \\ &= 36 \end{aligned}$$

(c) Which one of these two estimates is more likely to be the OLS estimate of β? Why?

Answer: $\widehat{\beta}_1$ is more likely to be the OLS estimate of β since it has the smaller SSR.

(d) Find the estimate of β that has the smallest sum of squared residuals.

Answer: The OLS estimate of β that has the smallest SSR is given by

$$\widehat{\beta}^{OLS} = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{2(1) + 6(6) + 8(4)}{1^2 + 6^2 + 4^2} = 1.32$$

4 Ordinary Least Squares

4.1 OLS and the "Classic" Assumptions

There are some assumptions that we tend to make. These are often called the "classic" OLS assumptions. These assumptions enable us to prove unbiasedness and calculate the variance of the OLS estimator. Note that we did not have to assume any of these prior to our derivation of the OLS estimator.

	$A1$:	$y_i = \beta x_i + \varepsilon_i$ is the true DGP	
	$A2$:	x_i is nonrandom	
	$A3$:	$\mathbb{E}[\varepsilon_i] = 0 \ \forall i$	
	$A4$:	$Var(\varepsilon_i) = \sigma^2 \ \ i = 1, \ldots, n$	("homoskedastic")
		$Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$	("no correlation")
(sometimes)	$A5$:	$\varepsilon_i \sim Normally$	

You might wonder what "sometimes" means. If we make assumption $A5$, then we'll have

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right).$$

If we don't then we'll have (using the Central Limit Theorem)

$$\widehat{\beta} \sim^{A} N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right).$$

The difference is whether or not we think our sample size is large enough for the distribution to converge. If we don't think so, assuming that the errors are normal will do the trick. If we do think so, then we might avoid making the assumption altogether. Now lets calculate the mean and variance of the OLS estimator. First, the mean.

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}^{OLS}\right] &= \mathbb{E}\left[\frac{\sum_i y_i x_i}{\sum_i x_i^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_i (\beta x_i + \varepsilon_i) x_i}{\sum_i x_i^2}\right] && \text{(A1)} \\
&= \mathbb{E}\left[\beta + \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2}\right] \\
&= \beta + \frac{\sum_i x_i \mathbb{E}[\varepsilon_i]}{\sum_i x_i^2} && \text{(A2)} \\
&= \beta && \text{(A3)}
\end{aligned}$$

That is, the OLS estimator is unbiased. Now we find the variance (which is slightly more difficult).

$$
\begin{aligned}
Var\left(\widehat{\beta}^{OLS}\right) &= Var\left(\frac{\sum\limits_i x_i y_i}{\sum\limits_i x_i^2}\right) \\
&= Var\left(\frac{\sum\limits_i x_i(\beta x_i + \varepsilon_i)}{\sum\limits_i x_i^2}\right) && \text{(A1)} \\
&= Var\left(\beta + \frac{\sum\limits_i x_i \varepsilon_i}{\sum\limits_i x_i^2}\right) \\
&= Var\left(\frac{\sum\limits_i x_i \varepsilon_i}{\sum\limits_i x_i^2}\right) && \text{(the variance of a constant is zero)} \\
&= \frac{1}{\left(\sum\limits_i x_i^2\right)^2} Var\left(\sum_i x_i \varepsilon_i\right) && \text{(pull constant out and square)} \\
&= \frac{1}{\left(\sum\limits_i x_i^2\right)^2}\left[\sum_i x_i^2 Var(\varepsilon_i) + \sum_{i \neq j} x_i x_j Cov(\varepsilon_i, \varepsilon_j)\right] \\
& && \text{(apply general formula; A2)} \\
&= \frac{1}{\left(\sum\limits_i x_i^2\right)^2}\left[\sigma^2 \sum_i x_i^2\right] && \text{(A4)} \\
&= \frac{\sigma^2}{\sum\limits_i x_i^2}
\end{aligned}
$$

4.2 Why do we use OLS?

We might ask ourselves, why are we using OLS instead of some other estimator. Restricting ourselves might seem severe. After all, we often have reason to believe that variables are related in *nonlinear* ways. However, we should not fret. That OLS is a *linear* model simply means that it is linear in the coefficients (not the variables). Consider the following:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \varepsilon_i.$$

Many people would say that this is not a linear regression. However, what if we simply called $x_{2i} = x_{1i}^2$ (the computer doesn't really care what we call the variables; all it does it take the

numbers and doesn't interpret whether or not they are nonlinear). Thus we can write our problem as

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.$$

In fact we can do this for any transformation of our variables (*logs*, for instance).

In addition to the above, there is a very important theorem that explains why OLS might be the preferred. Colloquially, if the classic OLS assumptions introduced earlier hold, OLS is the best in the class of linear estimators.

Gauss-Markov Theorem: If $\mathbb{E}[\varepsilon_i] = 0$, $Var(\varepsilon_i) = \sigma^2 < \infty \ \forall \ i$, and $Cov(\varepsilon_i, \varepsilon_j) = 0$ $forall \ i \neq j$, then the Ordinary Least Squares estimator is the Best Linear Unbiased Estimator (BLUE), where best means "lowest variance."

4.3 Functional Forms

Here we will get more in depth about how flexible our linear regression model is. When writing down our typical multiple regression model,

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, \ldots, n$$

we are only requiring that equation is linear *in the coefficients*. This essentially means that the LHS variable is a linear combination of RHS variables. We are free, then, to use RHS variables that are transformations of what we typically use.

The usual transformations are higher ordered (e.g. quadratic) polynomial terms and log transformations. Implementation of these different types of functional forms is fairly straightforward. Given some starting multiple regression model, suppose we are wondering whether or not we want to include a quadratic term for the variable Age_i. What we could do (and something I recommend) is to plot the relationship between the left hand variable, let's call it $Income_i$, and Age_i. Does it look linear? Perhaps something else, like a parabola? If so we might want to add in an Age_i^2 term.

Here's some tips you might want to remember.

- If you are including a cubic term, you should always include a quadratic and a linear term (even if the coefficients on them are insignificant)
- Testing whether or not a variable has an effect in this sort of context (with a term and a term^2 included), we will typically use a joint (F) test (test that both are equal to zero simultaneously)

- Marginal effects will usually be functions of the RHS variables (and not constants).

4.4 Logarithmic Transformations

Economists often utilize logs (in particular natural logs: ln) because they are a nice way of linearizing variables with exponential growth (eg. wages). There are three types of log-transformed regression models: log-log, log-linear, linear-log

Examples:

$$ln(salary_i) = \alpha + \beta hours_i + \varepsilon_i \quad \text{(log-linear)}$$
$$ln(wage_i) = \alpha + \beta ln(parentinc_i) + \varepsilon_i \quad \text{(log-log)}$$
$$deposits_t = \alpha + \beta ln(GDP_t) + \varepsilon_t \quad \text{(linear-log)}$$

Another (very) useful thing about logs is that they are a great way of approximating percentage changes. Say, for example, that we wanted to approximate the percentage change of going from 10 to 11 (10%). We can do the following:

$$ln(11) - ln(10) = 0.0953 \approx 0.10$$

Recall, also, some useful log rules:

- $ln(AB) = ln(A) + ln(B)$
- $ln(A/B) = ln(A) - ln(B)$

That is, we could have written our approximation as $ln(11/10)$. Further, the approximation gets better the closer are A and B. Now, let's return to our examples and interpret the β's.

$$\frac{dln(salary_i)}{dhours_i} = \beta \quad \longrightarrow \quad \text{a 1 unit inc. in hours increases wage by } 100 \times \beta\%$$

$$\frac{dln(wage_i)}{dln(parentinc_i)} = \beta \quad \longrightarrow \quad \text{a 1\% increase in parent income increases wages by } \beta\%$$

$$\frac{ddeposits_t}{dln(GDP_t)} = \beta \quad \longrightarrow \quad \text{a 1\% increase in GDP increases deposits by } \beta/100 \text{ dollars}$$

Remember to include *ceteris paribus* if we were doing multiple regressions. Also notice that in the second example (the coefficients in a log-log situation) are elasticities! Remember that d or ∂ denote differentials (or infinitesimally small increments / changes).

4.5 Coefficient Interpretation

A big part of this class is going to center around *interpreting* a regression model. In particular, we'll want to interpret the coefficients. Recall our general regression model:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

In words, α is the value of y_i when all of the right hand variables are zero and β_j is the effect on y_i obtained from increasing x_{ki} by one unit while holding all other variables constant (the *ceteris paribus* interpretation). If our RHS variables are continuous, then the coefficients have a partial derivative interpretation.

$$\beta_j = \frac{\partial y_i}{\partial x_{ji}}$$

Consider the following example. Suppose we have the following model of wages,

$$wage_i = \alpha + \beta_1 Age_i + \beta_2 Exper_i + \beta_3 Educ_i + \beta_4 SAT_i + \varepsilon_i.$$

Let's see how well you can interpret the coefficients. α and ε_i are measured in dollars. β_1, β_2, and β_3 are measured in dollars per year. β_4 is measured in dollars per point. The interpretations should always include: "holding all other variables constant."

4.6 Binary Variables

A very important type of variable we commonly use are binary, or "dummy," variables. These are discrete variables that take on one of two possible values: zero or one.

$$X_i = \begin{cases} 1 \\ 0 \end{cases}$$

Interpreting regressions with dummy variables is fairly straightforward, but can be confusing at first. Consider the following 3 regression:

$$ln(wage_i) = \alpha_1 + \alpha_2 FE_i + \varepsilon_i \tag{1}$$

$$ln(wage_i) = \beta_1 + \beta_2 M_i + \varepsilon_i \tag{2}$$

$$ln(wage_i) = \gamma_1 FE_i + \gamma_2 M_i + \varepsilon_i \tag{3}$$

Indeed, each of the three regressions above tell us the same thing, it's getting at that information which is slightly different. Going through, we can determine what the expected log wages for men and women and the difference in means.

Specification	Female	Male	Difference
(1)	$\alpha_1 + \alpha_2$	α_1	α_2
(2)	β_1	$\beta_1 + \beta_2$	$-\beta_2$
(3)	γ_1	γ_2	$\gamma_1 - \gamma_2$

4.7 Perfect Multicollinearity

When using specifications with sets of binary models, we have to be very careful with what we include in the regression. Notice, importantly, that an intercept term is not included in the third specification. Including it would result in *perfect multicollinearity*, a situation where one of the right hand side variables (covariates) can be written as a perfect linear combination of others. In our example, if we were to add in an intercept (which you can think of adding in the variable 1 with a coefficient α) then we can write:

$$1 = FE_i + M_i \qquad \Longleftrightarrow \qquad 1 = FE_i + (1 - FE_i) \qquad \Longleftrightarrow \qquad 1 = 1$$

This is a problem because the addition of that last variable (the one that would induce perfect multicollinearity) doesn't add any new information to the regression, and so we won't be able to determine what effect should be attributed to which variable. Indeed, if we tried to run the third regression with an intercept, we'd get an error (Eviews won't run). When there is perfect multicollinearity as a result of binary variables, it's called the "dummy variable trap." Here are some tips when dealing with dummy variables in regressions:

- if you include an intercept, exclude exactly 1 of all exhaustive sets of binary variables
- if you include an exhaustive set of binary variables, you must exclude the intercept

- if you have multiple exhaustive sets of binary variables and exclude an intercept, you can only have 1 full set of binary variables (you must omit 1 from the others)
- when interpreting coefficients, the omitted variable is the "base case" and is what the other variables of that set are relative to

4.8 F-tests

Now that we're in multiple regression, we might want to run *joint* hypotheses. That is, we might concern ourselves with a hypothesis with multiple *restrictions*. Take, for example, the following regression and an associated hypothesis test:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \qquad i = 1, \ldots, n$$

$$H_0 : \qquad \beta_1 = 0 \ \text{ and } \ \beta_2 = 0$$

$$H_1 : \qquad \beta_1 \neq 0 \ \text{ or } \ \beta_2 \neq 0 \ \text{ (or both).}$$

Notice that we are estimating 4 parameters and we have $q = 2$ restrictions in the hypothesis. To do an F-test, we calculate the F-statistic and compare it to the appropriate critical value (usually obtained from a table). Rule: reject if $\widehat{F} > F_{c.v.}$.

$$F = \frac{(SSR^* - SSR)/q}{SSR/(n-k)} \sim F_{q,n-k}$$

where SSR^* is the sum of squared residuals under the restricted regression, q is the numerator degrees of freedom (number of restrictions), and $n-k$ is the denominator degrees of freedom.

4.9 Interaction Terms

The next important type of variable that we often concern ourselves with are *interaction terms*. Simply put, they are just new variables that are created (generated) by multiplying two other variables together. For example, consider the variables FE_i and $Black_i$, the latter being a dummy variable indicating if someone is black or not black. Our model might be

$$ln(wage_i) = \alpha + \beta FE_i + \gamma Black_i + \delta FE_i \times Black_i + \varepsilon_i.$$

How are these coefficients interpreted?

	Men	Women	Difference
non-Black	α	$\alpha + \beta$	β
Black	$\alpha + \gamma$	$\alpha + \beta + \gamma + \delta$	$\beta + \delta$
Difference	γ	$\gamma + \delta$	δ

- α: the expected log wage of a non-black male
- β: the difference in wages between men and women who are non-black
- γ: the difference in wages between black and non-black men
- δ: the difference in wages of men and women who are black OR the difference in wages of black and non-black people who are women (difference in differences coefficient)

We will see another example of interaction terms when we discuss the Difference-in-Differences Model.

5 Violations of Classic Assumptions

Now we are going to start "breaking" the classic OLS assumptions. Not only can breaking some of these assumptions lead to biased estimates, but we would also ideally like to make use of Gauss-Markov Theorem (which, recall, states that OLS is BLUE if certain of those assumptions hold). This part of the class can get confusing as it is very easy to get bogged down in the math of it all, so here is a little table to fix "big picture" ideas.

Issue	Broken Assumption	Solution(s)
Misspecification	A2	Fixed Effects
Heteroskedasticity	A4 part 1	GLS or Huber-White SEs
Autocorrelation	A4 part 2	GLS or HAC SEs

5.1 Misspecification

Now we are going to start "breaking" the classic OLS assumptions. The first thing we are going to look at is really an implicit assumption, but we can think of it as breaking A2 (x_i is nonrandom). Assume that the DGP is $y = \beta x + \varepsilon$. Further, remember from before that we can rewrite $\widehat{\beta}$ as follows (using some of the other assumptions)

$$\widehat{\beta} = \frac{\sum yx}{\sum x^2} = \beta + \frac{\sum \varepsilon x}{\sum x^2} = \beta + \frac{\frac{1}{n}\sum \varepsilon x}{\frac{1}{n}\sum x^2} \approx \beta + \frac{cov(x, \varepsilon)}{var(x)}.$$

If $cov(x, \varepsilon) \neq 0$, then $\widehat{\beta}$ will be biased in the direction of the sign of $cov(\cdot)$. Recall, under A2, that this is automatically 0 (x doesn't covary with anything!). Indeed, this assumption is just to make some arithmetic easier, the real A2 should be that the covariance is zero. So, when is the covariance not equal to zero? Intuitively, if there is something left out of the regression (i.e. in the error term) that is related to x and determines y (*omitted variable bias*).

Let's say we are concerned with the effect of class size on student test scores. In particular let's assume that the real DGP (that is all of the classic assumptions hold for it) is given by

$$score_{it} = \beta size_{it} + \delta teacherquality_i + \varepsilon_{it} \qquad i = 1, \ldots, n \qquad t = 1, \ldots, T.$$

Notice the indexes (there are two). We have observations of individual classrooms (i) over time (t). We call this sort of data structure *panel data*. When we just have observations over i we call that *cross-sectional data*, and if we have observations over t we call that *time-series data*. In any regard, with our panel structure, notice that teacher quality doesn't vary over time (we might think that a particular teacher in classroom i doesn't improve over time).

Unfortunately, we don't observe $teacherquality_i$, so we run the regression:

$$score_{it} = \gamma size_{it} + u_{it},$$

where the coefficient of interest is called something else because we don't (at least right now) know if the above regression will give us the right value for the effect of class size (we want $\gamma = \beta$). Further, the error term u_{it} is named differently because $teacherquality_i$ is now in the error term (this is all just notation). Now we want to see if our estimator for the effect of class size is biased. Assume that better teachers are assigned to smaller classes (this will be important). Running the pooled regression (we won't have to worry about the double indexes) ...

$$\begin{aligned}
\mathbb{E}[\widehat{\gamma}] &= \mathbb{E}\left[\frac{\sum size * score}{\sum size^2}\right] \\
&= \mathbb{E}\left[\frac{\sum size * (\beta size + \delta teacherquality + \varepsilon)}{\sum size^2}\right] && \text{(plugging in the true DGP: A1)} \\
&= \mathbb{E}\left[\beta + \delta\frac{\sum size * teacherquality}{\sum size^2} + \frac{\sum size * \varepsilon}{\sum size^2}\right] \\
&= \beta + \delta\mathbb{E}\left[\frac{\sum size * teacherquality}{\sum size^2}\right] + 0. && \text{(A2 and A3)} \\
&= \beta + \delta\widehat{\lambda} && \text{(the way Thiel writes it)}
\end{aligned}$$

The two conditions that are necessary for omitted variable bias are:

$$1.\ \delta \neq 0 \iff cov(z_i, y_i) \neq 0$$

$$2.\ \widehat{\lambda} \neq 0 \iff cov(z_i, x_i) \neq 0$$

If these two conditions are met this implies $cov(x_i, \varepsilon_i) \neq 0$ which leads to bias.

Thus our estimator for the effect of size, $\widehat{\gamma}$, will be unbiased if that second term is equal to zero (the truth is β). Note that the messy part of the second term looks like the OLS estimator for a regression of teacher quality on class size, $teacherquality_{it} = \lambda size_{it} + v_{it}$, which is where λ comes from. Since we know that $cov(size, teacherquality) < 0$ (i.e. $\widehat{\lambda}$ will be negative) and intuit that δ is positive (better teachers yield higher test scores), we can be pretty sure that our estimator will be biased downward.

5.2 Fixed Effects

One way of fixing the problem from before is to use *fixed effects*. These work if the omitted (unobserved) effects only vary across one dimension (e.g. observation or time). To continue on with our earlier example, we would want to add "entity" or observation fixed effects to help control for the unobserved teacher quality. Our regression will then look like

$$score_{it} = \beta size_{it} + \delta_i + \varepsilon_{it},$$

where δ_i are the observation fixed effects. In order to implement this in practice, we would simply include dummy variables for all of the observations. That is, we would create a dummy variable for each observation (e.g. D_1 equals 1 if it's observation 1 and 0 if it's observation $2 - n$). We could rewrite this as

$$score_{it} = \beta size_{it} + \delta_1 D_1 + \cdots + \delta_n D_n + \varepsilon_{it}.$$

Intuitively, these dummies will capture all of the effects of each individual that are not explained by the included variables (class size for us). We won't interpret the δs because they will capture more than just teacher quality. They are simply used to "clean up" the error term.

Suppose we also thought there were unobserved things that varied only over time (but not across observation), like general economic conditions (recession, boom, etc.), that we thought were important that we didn't observe. We could also add in time fixed effects (which we could implement by adding in dummies for each year).

$$score_{it} = \beta size_{it} + \delta_i + \eta_t + \varepsilon_{it}$$

5.3 Difference-in-Differences

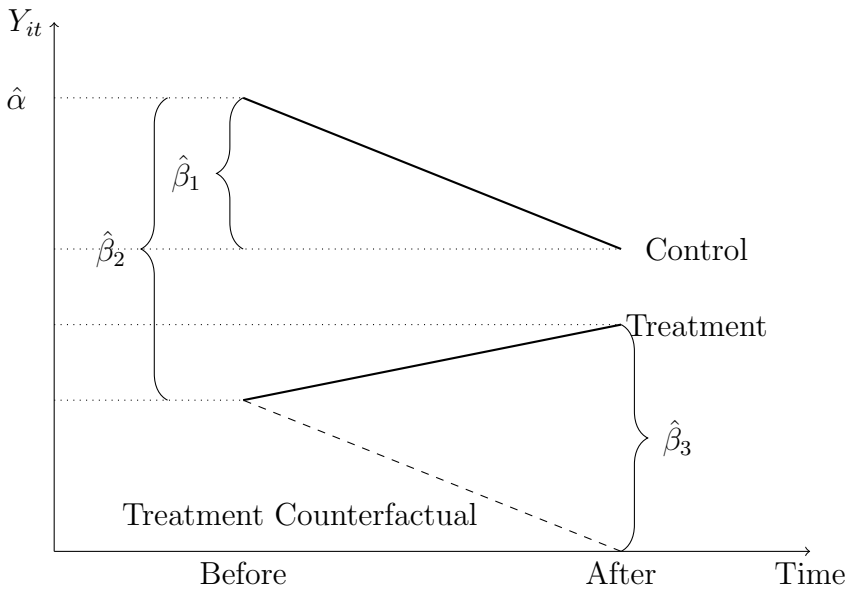
You need data with variation over time.

- Repeated cross-section – observe the same group but different individuals over time
- Panel Data – observe the same individual over time

A diff-in-diff framework is a useful tool for policy analysis. We use a control group as a counterfactual to the treatment group.

$$Y_{it} = \alpha + \beta_1 * After_{it} + \beta_2 * Treated_{it} + \beta_3 * After_{it} * Treated_{it} + \varepsilon_{it}$$

$\mathbb{E}[Y_{it}\|]$	**Treated**	**Control**	**Difference**
After	$\alpha + \beta_1 + \beta_2 + \beta_3$	$\alpha + \beta_1$	$\beta_2 + \beta_3$
Before	$\alpha + \beta_2$	α	β_2
Difference	$\beta_1 + \beta_3$	β_1	β_3



Underlying Assumptions:

- Parallel (Common) trends assumption – treatment group would be on the same trend as control if policy/intervention did not happen
- No other contemporaneous shocks to treatment group aside from treatment
- Policy change as good as random

5.4 Heteroskedasticity

The next assumption we will break is part 1 of A4, which states that the variance of the error term for each observation is the same (homoskedasticity). Indeed, there is no intuitively good reason to believe this, so we'll relax that now. That is, we will amend that part of assumption 4 to read: $var(\varepsilon_i) = \sigma_i^2$ (notice the index). Again, this is problematic because Gauss-Markov will no longer hold. The first solution I will mention is what is normally done in actual research. The second solution to this problem is fairly straightforward, and what you'd most likely be tested on. What you should know is that if there is heteroskedasticity, and you treat the problem like you have homoskedasticity, the SEs that you obtain will be artificially low and you will reject hypotheses too frequently (which is not good, it's better to be conservative in this regard).

Because the first solution is fairly anticlimactic, it doesn't get it's own section. Knowing that the SEs are wrong if we assume homoskedasticity when there is heteroskedasticity, we can use a different standard error formula to correct for that. This particular formula (which you don't need to know, just be aware of it) are called Huber-White "heteroskedasticity robust" SEs. Henceforth you should always correct for heteroskedasticity on homework assignments / projects (there is a way to change the SEs in Eviews).

5.5 Generalized Least Squares (GLS)

When we do GLS, what we are doing is transforming our original, problematic, regression with heteroskedasticity into one with homoskedasticity (which we already know how to work with). To do the transformation, we will weight each observation by something (call it w_i) that will make the variance of that observation's error the same for everyone. Intuitively, we will put less weight on observations with high variance, and more weight on observations with low variance (we prefer to be more accurate). Once we figure out the weights, we plug that into the transformed regression and then run OLS on that transformation:

$$w_i y_i = \beta w_i x_i + w_i \varepsilon_i \qquad \Longrightarrow \qquad \widehat{\beta}_{GLS} = \frac{\sum w_i^2 x_i y_i}{\sum w_i^2 x_i^2}$$
$$\widetilde{y}_i = \beta \widetilde{x}_i + \widetilde{\varepsilon}_i.$$

We are not done, though, we need to actually figure out what w_i will be. For each question you might get, there will be a different weight that you use that depends on the structure of the variance for your problem. I will go through an example here. Suppose we thought that the variance of the error term was $var(\varepsilon_i) = x_i \sigma^2$. We want to find a weight w_i that we multiply our observations by such that we remove anything that has an i index on it. That is

$$
\begin{aligned}
var(w_i\varepsilon_i) &= \sigma^2 \\
w_i^2 var(\varepsilon_i) &= \sigma^2 \\
w_i^2 &= \frac{\sigma^2}{x_i\sigma^2} \\
w_i &= \frac{1}{\sqrt{x_i}}
\end{aligned}
$$

And so by plugging in we can get our transformed regression and the GLS estimator:

$$
\frac{y_i}{\sqrt{x_i}} = \beta\sqrt{x_i} + \frac{\varepsilon_i}{\sqrt{x_i}} \qquad \Longrightarrow \qquad \widehat{\beta}^{GLS} = \frac{\sum y_i}{\sum x_i}. \quad \text{(they don't always come out this clean)}
$$

What happens when part 2 of A4 breaks (that is, when the covariance of different error terms is not zero). This topic is called serial correlation (sometimes autocorrelation) and is of particular (though not exclusive) interest when looking at time series data (i.e. we only have 1 "observation" but witness it over time). As an extension, we will get a glimpse of some very important time series topics, namely forecasting.

5.6 Serial Correlation

First, in time series, a typical regression might look like

$$
y_t = \alpha + \beta_1 y_{t-1} + \cdots + \beta_m y_{t-m} + \gamma_1 x_{t-1} + \cdots + \gamma_n x_{t-n} + \varepsilon_t,
$$

where (interestingly) we can have *lags* of the left-hand variable on the right side! Consider, for simplicity, the simple case with just one lag of the dependent variable for just a moment: $y_t = \alpha + \beta y_{t-1} + \varepsilon_t$. Because y_t is a function of y_{t-1}, y_t will also be a function of ε_{t-1}. That is, we should have a strong concern about A4:

$$
\begin{aligned}
\underline{A4}: \qquad & Var(\varepsilon_i) = \mathbb{E}[\varepsilon_i^2] = \sigma^2 \;\; i = 1, \ldots, n && \text{("homoskedastic")} \\
& Cov(\varepsilon_i, \varepsilon_j) = \mathbb{E}[\varepsilon_i \varepsilon_j] = 0 \;\; \forall i \neq j. && \text{("no correlation")}
\end{aligned}
$$

Having a lag of the dependent variable on the RHS isn't necessary for this, though. In general, when thinking about variables (like GDP) over time, we generally think that there are important things that we can't observe (in the error term) that are related over time. We

call this *serial correlation*. This is a problem because Gauss-Markov (which requires there to be no covariance) will no longer hold. If we think back from before (in the derivation of the variance of the OLS estimator), one of the steps involved invoking A4 to eliminate all of the covariance terms. When we break this assumption, however, we can't do that.

Moving forward, how will we fix this? Indeed, we'll use another GLS procedure (before when doing "GLS" we were performing the "WLS" procedure to fix heteroskedasticity) called Cochrane-Orcutt. Here, we'll transform the DGP such that the no serial correlation assumption is once again satisfied. Just like with WLS and the variance of the error terms, in order to do this we'll need to make an assumption about the structure of the serial dependence. A common assumption (one that we'll work with) is that the error term follows a "first-order autoregressive process" (AR(1)). Consider the following.

$$y_t = \beta x_{t-1} + \varepsilon_t \qquad \text{where} \qquad \varepsilon_t = \rho\varepsilon_{t-1} + v_t \qquad \text{where} \qquad |\rho| < 1$$

We will assume that $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ follows all of the usual (classic) assumptions: $\mathbb{E}[v_t] = 0$, $Var(v_t) = \sigma_v^2 \ \forall t$, and $cov(v_t, v_s) = 0 \ \forall t \neq s$. You might see this assumption written as v_t being independently and identically distributed (iid), which is a stronger assumption than what we usually need, but it implies the above. Intuitively, what we'll want to do is get the regression written in such a way so that v_t is the only error term. We can do this by *quasi-differencing* (to take a difference means to subtract one thing from another; to quasi-difference essentially means to subtract a scaled thing from another). To continue with the earlier example (of only one lagged dependent variable on the RHS), notice that we can write:

$$\begin{aligned} y_t &= \beta x_t + \varepsilon_t \\ y_{t-1} &= \beta x_{t-1} + \varepsilon_{t-1} \\ \rho y_{t-1} &= \beta\rho x_{t-1} + \rho\varepsilon_{t-1} \end{aligned}$$

Everything above is just a manipulation. Recalling the structure of the error term from before, what we want to do is get the $\rho\varepsilon_{t-1}$ out of ε_t so that we are left with v_t. To do this, simply subtract the third equation from the first equation above to get the transformed DGP. To get the Cochrane-Orcutt estimator (again, which is a type of GLS estimator), we will simply run OLS on the transformed regression.

$$\underbrace{y_t - \rho y_{t-1}}_{\text{new } y \text{ variable}} = \beta \underbrace{(x_t - \rho x_{t-1})}_{\text{new } x \text{ variable}} + \underbrace{\varepsilon_t - \rho\varepsilon_{t-1}}_{v_t} \qquad \text{("quasi-difference")}$$

$$\implies \qquad \widehat{\beta}^{CO} = \frac{\sum(y_t - \rho y_{t-1})(x_t - \rho x_{t-1})}{\sum(x_t - \rho x_{t-1})^2}$$

In practice, however, we'll need to get an estimate for ρ, $\widehat{\rho}$. To do this, what we could do is to follow these steps:

1. Run the original regression ($y_t = \beta x_t + \varepsilon_t$) and save the residuals
2. Run the regression $e_t = \rho e_{t-1} + \eta_t$ to obtain $\widehat{\rho}$
3. Then use that to get (and run) the transformed regression: $y_t - \widehat{\rho} y_{t-1} = \beta(x_t - \widehat{\rho} x_{t-1}) + v_t$

Eviews also provides a shortcut. Remember that we assumed that the error followed an AR(1) process? Well, we can simply include that in a least squares regression and it will do all of the work for us!

$$ls \;\; y \;\; c \;\; x \;\; AR(1)$$

Note that we could also include AR(2), AR(3), ... and so on if we thought that the error term had some correlation with more than one period prior.

5.7 Examples

Fall 2015, Final, Question 4. You are interested in estimating $Y_t = c + \beta X_t + \varepsilon_t$. You find evidence of serial correlation of the AR(1) type such that

$$\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t$$

(a) Suppose that $\rho = 0.9$ and that $\mathbb{E}[\eta_t \eta_s] = \begin{cases} 1 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases}$

i. What is the correlation between ε_t and ε_{t-1}? $\boxed{0.9}$

ii. What is the correlation between ε_t and ε_{t-4}? $\boxed{0.6561}$

(b) Suppose that $\rho = 0.4$ and that $\mathbb{E}[\eta_t \eta_s] = \begin{cases} 1 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases}$

i. What is the correlation between ε_t and ε_{t-1}? $\boxed{0.4}$

ii. What is the correlation between ε_t and ε_{t-4}? $\boxed{0.0256}$

Solution Method:

$$\begin{aligned} corr[\varepsilon_t, \varepsilon_{t-1}] &= corr[\rho\varepsilon_{t-1} + \eta_t, \varepsilon_{t-1}] \\ &= corr[\rho\varepsilon_{t-1}, \varepsilon_{t-1}] + corr[\eta_t, \varepsilon_{t-1}] \\ &= \rho corr[\varepsilon_{t-1}, \varepsilon_{t-1}] + corr[\eta_t, \varepsilon_{t-1}] \\ &= \rho(1) + 0 = \rho \end{aligned}$$

$$\begin{aligned} corr[\varepsilon_t, \varepsilon_{t-4}] &= corr[\rho\varepsilon_{t-1} + \eta_t, \varepsilon_{t-4}] \\ &= corr[\rho(\rho\varepsilon_{t-2} + \eta_{t-1}) + \eta_t, \varepsilon_{t-4}] \\ &= corr[\rho(\rho(\rho\varepsilon_{t-3} + \eta_{t-2}) + \eta_{t-1}) + \eta_t, \varepsilon_{t-4}] \\ &= corr[\rho(\rho(\rho(\rho\varepsilon_{t-4} + \eta_{t-3}) + \eta_{t-2}) + \eta_{t-1}) + \eta_t, \varepsilon_{t-4}] \\ &= \rho^4 corr[\varepsilon_{t-4}, \varepsilon_{t-4}] + \rho^3 corr[\eta_{t-3}, \varepsilon_{t-4}] + \cdots + corr[\eta_t, \varepsilon_{t-4}] \\ &= \rho^4(1) + 0 + \cdots + 0 = \rho^4 \end{aligned}$$

(c) Comment on and explain the reason for the difference between the answers for (a) and (b).

Answer: In part (a), the errors are highly correlated. As a result, the relationship between the errors over time is higher whereas in part (b), the correlation between errors is practically 0 after 4 periods.

(d) You have typed in "ls y c x" in Eviews. Which of the following is true about your regression output?

1. Your estimate of β is biased.
2. Your confidence intervals are valid.
3. Your standard errors are incorrect. ($\star$)
4. Your p-values are valid.

(e) When you include the AR(1) term, will the t-statistics change more (compared to their values from part (d)) in the case of $\rho = 0.9$ or $\rho = 0.4$? Why?

Answer: The t-statistics will change more when $\rho = 0.9$ because they are more highly correlated over time (and so will need to be corrected "more").

6 Hypothesis Testing

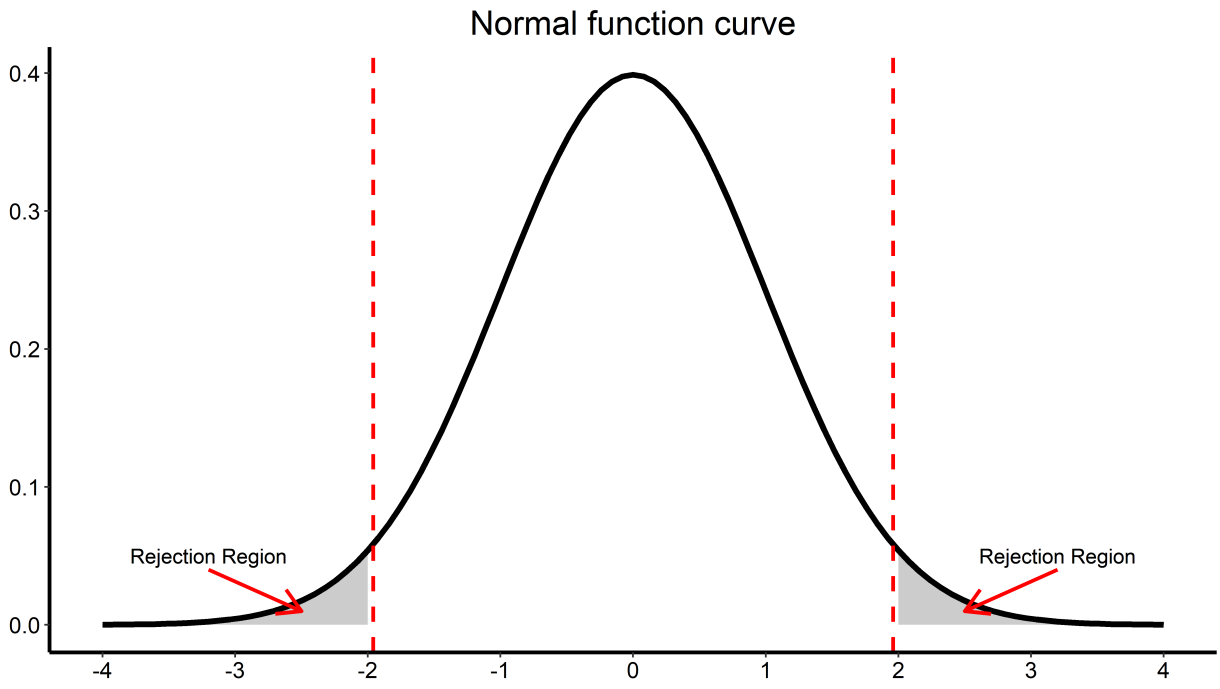
After estimating a parameter, we wish to know how close the population value is likely to be to an estimate. A common test we wish to perform is whether a regression coefficient is equal to zero or not equal to zero. First we set up the null and alternative hypothesis.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Next, construct the test statistic

$$\widehat{t} = \frac{\widehat{\beta}_1 - \beta_{H_0}}{SE(\widehat{\beta}_1)}$$

Now we compare $\widehat{t}$ to a critical value obtained from a t or z distribution. Finally, we either reject or fail to reject the null hypothesis. If $|\widehat{t}| > c.v.$ we reject the null hypothesis and conclude that there is evidence to suggest β_1 is significantly different than zero. If $|\widehat{t}| < c.v.$ we fail to reject the null hypothesis and conclude that there is not evidence to suggest β_1 is significantly different than zero.



We can also test a hypothesis by constructing a confidence interval.

$$\mathbb{P}\left(-c.v. \leq \frac{\widehat{\beta}-\beta}{SE(\widehat{\beta})} \leq c.v.\right) = 1-\alpha$$
$$\mathbb{P}\left(-c.v. * SE(\widehat{\beta}) \leq \widehat{\beta}-\beta \leq c.v. * SE(\widehat{\beta})\right) = 1-\alpha$$
$$\mathbb{P}\left(\widehat{\beta}-c.v. * SE(\widehat{\beta}) \leq \beta \leq \widehat{\beta}+c.v. * SE(\widehat{\beta})\right) = 1-\alpha$$

$$\implies \beta \in [\widehat{\beta}-SE(\widehat{\beta})*c.v., \widehat{\beta}+SE(\widehat{\beta})*c.v.]$$

6.1 Example

Hypothesis test of a difference of means. Let's say we were concerned about testing whether men and women have the same mean yearly incomes at the 5% level. In particular, we found that

$\overline{x}_M = 4.5$	$n_M = 12$	$\sigma_M^2 = 1$
$\overline{x}_F = 3.4$	$n_F = 15$	$\sigma_F^2 = 1.5$

where incomes are measured in thousands of dollars per pay period (2 weeks). Assume that the mean income for men and women are independent.

Answer: To answer this, we can use my 4 step procedure to answer hypothesis tests.

1. The null and alternative hypotheses will be given by the following. To test the hypothesis, we will use a difference of sample averages as our estimator.

$$H_0: \ \mu_M = \mu_F \qquad \text{vs.} \qquad H_1: \ \mu_M \neq \mu_F$$
$$\iff \qquad H_0: \ \mu_M - \mu_F = 0 \qquad \text{vs.} \qquad H_1: \ \mu_M - \mu_F \neq 0$$

2. Our test statistic will be slightly different than what we're used to. Recall that, generally, we have (note that, for this problem, $\widehat{\beta} = \overline{x}_M - \overline{x}_F$)

$$\widehat{t} = \frac{\widehat{\beta}-\beta_{H_0}}{SE(\widehat{\beta})} = \frac{(\overline{x}_M - \overline{x}_F) - (\mu_M - \mu_F)}{\sqrt{(\sigma_M^2/n_M) + (\sigma_F^2/n_F)}} = \frac{(4.5-3.4)-0}{\sqrt{(1/12)+(1.5/15)}} = \frac{1.1}{.4282} = 2.57$$

3. We are told to test it at the 5% level. The critical value we will want to use is ± 2.06.

4. Because the calculated test statistic falls in the rejection region, we reject the null hypothesis at the 5% level of significance. There is evidence to suggest that the mean incomes of men and women differ.

6.2 R-squared

R-squared is a statistical measure of how close the data are to the fitted regression line. R-squared $= \frac{\text{Explained variation}}{\text{Total variation}}$

R-squared is always between 0 and 100%:

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

7 Instrument Variables

One method to deal with the problem of endogenity ($Cov(x_i, \varepsilon_i) \neq 0$) is an Instrument Variable approach.

What we need for an instrument variable:

- An "exogenous" factor (something outside the model) that shifts x_i in such a way that ε_i is not affected.
- Alternatively, something randomly determined that affects x_i

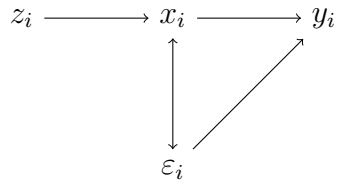
Suppose we have the following model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i,$$

and we are concerned that x_i and ε_i are correlated (i.e. x_i is endogenous). We can use an instrument variable z_i to "instrument" for x_i. There are two conditions that must be met for a variable, z_i, to be a valid instrument.

1. $Cov(x_i, z_i) \neq 0$ (The instrument is relevant or the first stage exists)
2. $Cov(z_i, \varepsilon_i) = 0$ (exclusion restriction)

The exclusion restriction can be thought of another way. The instrument, z_i, does not directly influence the dependent variable, y_i, its only affect is indirectly through x_i.



7.1 Two Staged Least Squares

Casual Relationship of interest:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i,$$

First Stage:

$$x_i = \alpha_0 + \alpha_1 z_i + \alpha_2 w_i + u_i,$$

Predicted First Stage:

$$\widehat{x}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 z_i + \widehat{\alpha}_2 w_i$$

Second Stage:

$$y_i = \beta_0 + \beta_1 \widehat{x}_i + \beta_2 w_i + \beta_1 u_i + \varepsilon_i,$$

Note: $x_i = \widehat{x}_i + u_i$.

You need at least as many instruments as endogenous right hand side variables in equation being estimated. In practice it is often difficult to find convincing instruments (in particular because many potential IVs do not satisfy the exclusion restriction). An example of a paper utilizing IV is "Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size" by Joshua Angrist and William Evans, American Economic Review, 1996. It turns out, parents typically have strong preferences for mixed-gender children. What this means is that parents of two same-sex children are more likely to have a third child than parents of mixed-sex children (about 6 percentage points more likely). Instrument, z_i, is a dummy indicating first two children are the same sex. We can only look at parents with at least 2 children, instrument shifts the probability of having a third child. y_i is labor supply and x_i is the number of children. Two stage least squares results show that a third child reduces hours per week by 4.5 hours.

8 Linear Probability Model

You need data with a binary dependent variable

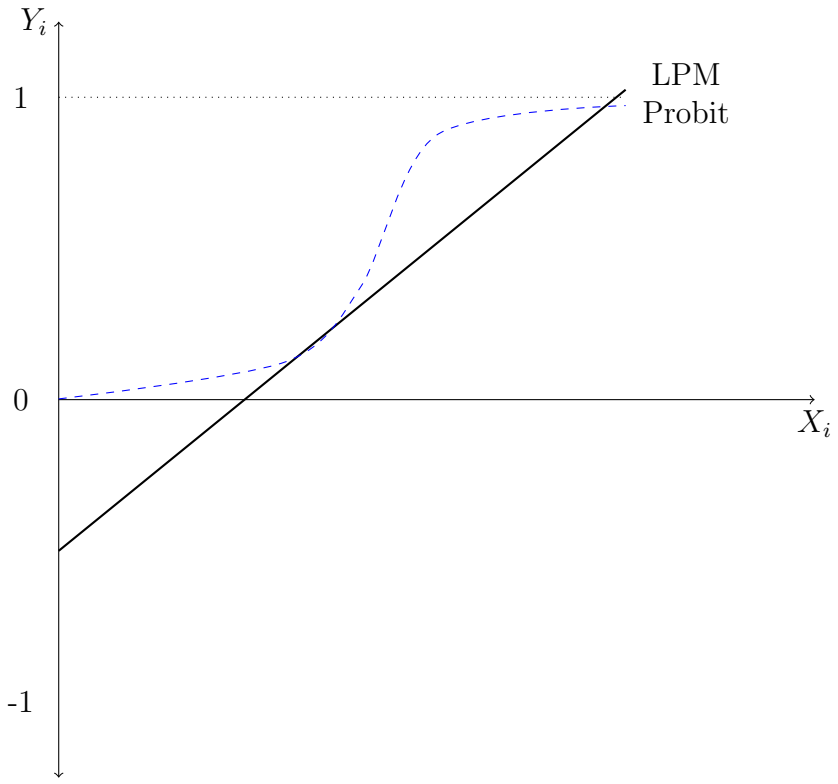
$$Y_i = \{0, 1\}$$

Binary outcome examples:

- Accepted/Rejected
- Win/Lose
- Mortality

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Interpreting the β_1 coefficient. Increasing X_i by one changes the probability of the outcome equaling 1 by β_1 percentage points, holding everything else constant. A predicted value $\hat{Y}$ is the predicted probability that the dependent variable equals one, given X.



Using the Probit or Logit model forces an s-curve on the prediction.

$$S_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\mathbb{P}(y_i = 1|X_i) = F(S_i)$$

$F(\cdot)$ is the cumulative distribution function for some probability distributions. CDFs have an S-shape and can only take values between zero and one.

- $F(\cdot)$ normal gives probit model
- $F(\cdot)$ logistic gives logit model

No matter how high the "score" is, the predicted probability can never be > 1 and no matter how low the "score" is, the predicted probability can never be < 0.

Probit and logit have nonlinear marginal effects. A coefficient is the change in probability due to a one-unit increase in a given X variable, BUT depends on both the values of other X's and the starting value of the given X. Yields unintuitive coefficients in regression output; we will focus on SIGN and relative size of coefficients with a probit/logit regression.