Econometrics

A Summary

The Joint Distribution

- The **joint distribution** of discrete RVs X and Y is the probability that the two RVs simultaneously take on certain values, say x and y: That is, $\Pr(X = x, Y = y)$, like a cross-tab.
- Example: weather and commuting time.
  - Let C denote commuting time. Suppose commuting time can be long ($C = 1$) or short ($C = 0$).
  - Let W denote weather. Suppose weather can be fair ($W = 1$) or foul ($W = 0$).
  - There are four possible outcomes: ($C = 0$, $W = 0$), ($C = 0$, $W = 1$), ($C = 1$, $W = 0$), ($C = 1$, $W = 1$).
  - The probabilities of each outcome define the joint distribution of C and W:

	Foul Weather (W=0)	**Fair Weather (W=1)**	**Total**
Short Commute (C=0)	0.15	0.25	**0.4**
Long Commute (C=1)	0.55	0.05	**0.6**
Total	**0.7**	**0.3**	**1**

Marginal Distributions

- When X,Y have a joint distribution, we use the term **marginal distribution** to describe the probability distribution of X or Y alone.
- We can compute the marginal distribution of X from the joint distribution of X,Y by adding up the probabilities of all possible outcomes where X takes a particular value. That is, if Y takes one of k possible values:

$$\Pr(X = x) = \sum_{i=1}^{k} \Pr\left(X = x, Y = y_i\right)$$

	Foul Weather (W=0)	Fair Weather (W=1)	Total
Short Commute (C=0)	0.15	0.25	**0.4**
Long Commute (C=1)	0.55	0.05	**0.6**
Total	**0.7**	**0.3**	**1**

The marginal distribution of weather is in blue. The marginal distribution of commuting time is in yellow.

Conditional Expectation

- The mean of the conditional distribution of *Y* given *X* is called the **conditional expectation** (or **conditional mean**) **of *Y* given *X***.
- It's the expected value of *Y*, given that *X* takes a particular value.
- It's computed just like a regular (unconditional) expectation, but uses the conditional distribution instead of the marginal.
  - If *Y* takes one of *k* possible values $y_1, y_2, \ldots, y_k$ then:

$$E(Y \mid X = x) = \sum_{i=1}^{k} y_i \Pr(Y = y_i \mid X = x)$$

- Example: in our commuting example, **suppose a long commute takes 45 minutes and a short commute takes 30 minutes**. What's the expected length of the commute, conditional on foul weather? What if weather is fair?
  - Foul weather: 30*0.15/0.7 + 45*0.55/0.7 = 41.79 minutes
  - Fair weather: 30*0.25/0.3 + 45*0.05/0.3 = 32.5 minutes

Independence

- Often, we're interested in quantifying the relationship between two RVs.
  - linear regression methods (the focus of this course) do exactly this.
- When two RVs are **completely** unrelated, we say they are **independently distributed** (or simply **independent**).
  - If knowing the value of one RV (say X) provides **absolutely no information** about the value of another RV (say Y), we say that X and Y are independent.
- Formally, X and Y are independent if the conditional distribution of Y given X equals the marginal distribution of Y:

$$\Pr(Y = y \mid X = x) = \Pr(Y = y) \qquad (*)$$

- Equivalently, X and Y are independent if the joint distribution of X and Y equals the product of their marginal distributions:

$$\Pr(Y = y, X = x) = \Pr(Y = y)\Pr(X = x)$$

  - This follows immediately from (*) and the definition of the conditional distribution:

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}$$

Covariance

- A very common measure of association between two RVs is their **covariance**. It is a measure of the extent to which to RVs "move together."
- $Cov(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$
- In the discrete case, if X takes one of m values and Y takes one of k values, we have

$$Cov(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{m}\left(x_j - \mu_X\right)\left(y_i - \mu_Y\right)\Pr\left(X = x_j, Y = y_i\right)$$

- Interpretation:
  - if X and Y are positively correlated ($\sigma_{XY} > 0$) then when $X > \mu_X$ we also have $Y > \mu_Y$, and when $X < \mu_X$ we also have $Y < \mu_Y$ (in expectation). This means X and Y tend to move "in the same direction."
  - Conversely, if $\sigma < 0$ then when $X > \mu_X$ we have $Y < \mu_Y$, and when $X < \mu_X$ we have $Y > \mu_Y$ (in expectation). This means X and Y tend to move "in opposite directions.
  - It is analogous to variance: the covariance of X and X is $Var(X)$.

Covariance and Correlation

- An unfortunate property of the covariance measure of association is that it is difficult to interpret: it is measured in units of X times units of Y.
- A "unit free" measure of association between to RVs is the **correlation** between X and Y:

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  - Notice that the numerator & denominator units cancel.
- $Corr(X,Y)$ lies between -1 and 1.
- If $Corr(X,Y) = 0$ then we say X and Y are **uncorrelated**.
- Note that if $Cov(X,Y) = 0$ then $Corr(X,Y) = 0$ (and vice versa).

Populations and Samples

- ECONOMETRIC INFERENCE ABOUT A POPULATION IS **ALMOST ALWAYS** BASED ON A SAMPLE!
- How do we choose which population members to sample?
- In a nutshell: choose them **randomly**.
- Example: Suppose I'm interested in the probability distribution of my commuting time to campus. Rather than recording my commuting time *every day*, I could randomly select five days each month to record my commuting time.
  - Population: every day
  - Sample: the days I record my commuting time
  - Use the sample data to estimate the population mean, variance, etc.
- Example: Political pollsters try to predict election outcomes. They ask questions like "If there was an election today, which of these candidates would you vote for?" Rather than asking *everyone in the country*, they randomly select a group of individuals to answer the question.
  - Population: everyone in the country
  - Sample: the group selected to answer the question
  - Use the sample to estimate the population mean, variance, etc.

Sampled Objects are Random Variables

- Suppose we're interested in a variable X.
- We're going to select a sample of individuals/businesses or whatever and measure their value of X.
- The observed measurements of X that comprise our sample are called **observations**. All the observations together are our **data**.
- Usually, we denote the n observations in the sample $X_1, X_2, \ldots, X_n$
  - If X was annual earnings, X_1 is the first person's response, X_2 is the second, etc
- Because we randomly select objects into the sample, the **values** of the observations $X_1, X_2, \ldots, X_n$ **are random.**
  - We don't know what values of X we'll get in advance
  - If we had chosen different members of the population, their values of X would be different.
- Thus, given random sampling, we treat $X_1, X_2, \ldots, X_n$ as random variables.

Statistics and Sampling Distributions

- A **statistic** is any function of the sample data.
  - A (scalar-valued) **function** *f(x1,…xN)* is a single number associated with each set of values that x1,…,xN can take on.
- Because the sample data are random variables, so are statistics.
- We know that all random variables have probability distributions.

  ➔**All statistics have probability distributions (pdfs&cdfs)**.
- In fact we have a special name for the probability distribution of a statistic: we call it a **SAMPLING DISTRIBUTION.**
- **THIS IS THE MOST IMPORTANT CONCEPT IN THIS COURSE!!!**
- Every statistic has a sampling distribution because **if we drew a different sample, the data would take different values, and hence so would the statistic.**
- The sampling distribution represents **uncertainty** about the **population value** of the statistic because it is based on a sample, and not based on the whole population.

What the Sampling Distribution Tells Us

- Like any probability distribution, the sampling distribution tells us what values of the statistic are possible, and how likely the different values are.
- For instance, the **mean of the sampling distribution** tells us the expected value of the statistic.
  - It is a good measure of what value we expect the statistic to take.
  - It also tells us where the statistic's probability distribution is centered.
- The **variance of the sampling distribution** tells us how "spread out" the distribution of the statistic is.
  - It is usually a function of the sample size.
  - It has a special name: the **sampling variance** of the statistic (note: this is NOT THE SAME AS THE **SAMPLE VARIANCE**!)
  - If the sampling variance is large, then it is **likely** that the statistic takes a value "far" from the mean of the sampling distribution.
  - If the sampling variance is small, then it is **unlikely** that the statistic takes a value "far" from the mean of the sampling distribution.
  - Usually, the sampling variance gets smaller as the sample size gets bigger.
- A picture shows this.

Estimation

- An **estimator** is a statistic that is used to infer the value of an unknown quantity in a statistical model
- The sample mean, sample variance, and sample covariance are all statistics. But, they are also all called **estimators**, because they can be used to **estimate** population quantities.
- That is, the thing we care about is a population quantity like the population mean μ.
- We don't get to observe μ directly, and we can't measure its value in the population.
- So we draw a sample from the population, and **estimate** μ using the sample. $\bar{X}$
- One way to do this is to compute the **sample mean** in our sample.
- It is a "good" estimate of the population mean, in a sense we'll now make precise.

Estimators and Their Properties: Bias

- There are lots and lots of estimators, but not all are equally "good."
  - The sample mean is an estimator of the population mean.
  - So is the median.
  - So is the value of one randomly selected observation.
- This is where the estimator's sampling distribution comes in – it tells us the estimator's properties.
  - Whether it gives "good" or "bad" estimates of a population quantity.
- Suppose we're interested in a population quantity Q and R is a sample statistic that we use to estimate Q.
  - e.g., Q might be the population mean, and R the sample mean
- We say R is an **unbiased estimator of Q** if $E(R) = Q$.
  - ➔**if R is an unbiased estimator of Q, then Q is the mean of the sampling distribution of R**
- The **bias** of R is $E(R) - Q$. An **unbiased** estimator has bias = 0.
- DRAW A PICTURE!!

Estimators and Their Properties: Efficiency

- Unbiasedness is a nice property, but it is "weak."
  - There can be many unbiased estimators of a given population quantity.
  - Example: suppose we want to estimate the population mean μ. In an iid sample, the sample mean is an unbiased estimator of μ:

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

  - because $E(X_i) = \mu$ for every observation.
  - Another unbiased estimator is the value of X_1, because $E(X_1) = \mu$.
- How do we choose between unbiased estimators?
  - We prefer the unbiased estimator with the smaller sampling variance. A picture shows the how the sampling distributions of the sample mean and a single observation's value differ.
  - Suppose we have two unbiased estimators of Q, call them R_1 and R_2. We say that R_1 is **more efficient** than R_2 if $Var(R_1) < Var(R_2)$.

Null and alternative hypotheses

- Suppose we're interested in evaluating a specific claim about the population mean. For instance:
  - "the population mean is 5"
  - "the population mean is positive"
- We call the claim that we want to evaluate the **null hypothesis**, and denote it H_0.
  - $H_0 : \mu = 5$
  - $H_0 : \mu > 0$
- We compare the null hypothesis to the **alternative hypothesis**, which holds **when the null is false**. We will denote it H_1.
  - $H_1 : \mu \neq 5$ (a "two-sided" alternative hypothesis)
  - $H_1 : \mu \leq 0$ (a "one-sided" alternative hypothesis)

How tests about the population mean work

- Step 1: Specify the null and alternative hypotheses.
- Step 2a: Compute the sample mean and variance
- Step 2b: Use the estimates to construct a new statistic, called a **test statistic**, that has a **known sampling distribution *when the null hypothesis is true*** ("under the null")
  - the sampling distribution of the test statistic depends on the sampling distribution of the sample mean and variance
- Step 3: Evaluate whether the calculated value of the test statistic is "likely" when the null hypothesis is true.
  - We **reject** the null hypothesis if the value of the test statistic is "unlikely"
  - We **do not reject** the null hypothesis if the value of the test statistic is "likely"
  - (Note: thanks to Popper, we never "accept" the null hypothesis)

Example: the t-test

- Suppose we have a random sample of *n* observations from a $N(\mu,\sigma^2)$ distribution.
- Suppose we're interested in testing the null hypothesis:

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis:

$$H_1 : \mu \neq \mu_0$$

- A natural place to start is by estimating the sample mean, $\overline{X}$
- We know that **if the null hypothesis is true**, then the sampling distribution of $\overline{X}$ is normal with mean μ_0 and variance σ^2/n.
  - We say: $\overline{X} \sim N(\mu_0,\sigma^2/n)$ **under the null**
  - (draw a picture)

Example: the t-test (continued)

- Because $\bar{X} \sim N(\mu_0, \sigma^2/n)$ under the null, we know that

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1) \text{ under the null}$$

(recall we can transform any normally distributed RV to have a standard normal distribution by subtracting off its mean and dividing by its standard deviation)

- If we knew σ^2, we could compute *Z*, and this would be our test statistic:
  - If *Z* is "far" from zero, it is unlikely that the null hypothesis is true, and we would reject it.
  - If *Z* is "close" to zero, it is likely that the null hypothesis true, and we would not reject it.
  - Why *Z*? Because we can look up its critical values in a table.
- Problems with this approach:
  - we don't know σ^2
  - how do we quantify "close" and "far"?

Interval Estimation

- We're done talking about hypothesis testing for now – but it will come up again soon in the context of linear regression.
- We talked earlier about estimators – statistics that we use to estimate a population quantity.
- The examples we saw (the sample mean, sample variance, sample covariance, etc.) are all called **point estimators** because they give us a single value for the population quantity.
- An alternative to a point estimator is an **interval estimator.**
- This is an interval that contains a population quantity with a known probability.
- An interval estimator of a population quantity Q takes the form $[L,U]$, where L and U are functions of the data (they're statistics).
- We use the interval estimator $[L,U]$ to make statements like:

$$\Pr[L \leq Q \leq U] = 1 - \alpha$$ (look familiar yet?)

Example: Confidence Interval for the Population Mean

- A 95% confidence interval for the population mean μ is an interval $[L,U]$ such that:

$$\Pr[L \leq \mu \leq U] = 0.95$$

- How do we find the interval $[L,U]$ such that this is true?
- An illustrative (but impossible) way:
  1. Pick a random value μ_1 and construct the T statistic to test $H_0 : \mu = \mu_1$ vs. $H_1 : \mu \neq \mu_1$.
  2. If we reject H_0, then μ_1 **is not** in the interval. If we do not reject H_0, then μ_1 **is** in the interval.
  3. Pick another value μ_2 and repeat.
  4. Do this for all possible values of μ (this is why it's impossible).
- Thankfully, there's an easier way.

Notation

- If we have more (say *k*) independent variables, then we need to extend our notation further.
- We could use a different letter for each variable (i.e., *X*, *Z*, *W*, etc.) but instead we usually just introduce another subscript on the *X*.
- So now we have two subscripts: one for the variable number (first subscript) and one for the observation number (second subscript).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

- What do the regression coefficients measure now? They are **partial derivatives, or marginal effects**. That is,

$$\beta_1 = \frac{\partial Y_i}{\partial X_{1i}} \quad \beta_2 = \frac{\partial Y_i}{\partial X_{2i}} \quad \cdots \quad \beta_k = \frac{\partial Y_i}{\partial X_{ki}}$$

So, β_1 measures the effect on Y_i of a one unit increase in X_{1i} **holding all the other variables X_{2i}, X_{3i}, ... , X_{ki} and ε_i constant.**

Simple Linear Regression

- Suppose now that we have a linear regression model with one independent variable and an intercept: $\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$
- Suppose also that

$$\boxed{E[\varepsilon_i] = 0 \; and \; E[(\varepsilon_i)^2] = \sigma^2 \; and \; E[(\varepsilon_i \varepsilon_j)] = 0 \; for \; all \; i, j}$$

- Now, define an estimator as the number $\hat{\beta}$ that minimises the sum of the squared prediction error

$$\boxed{e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i}$$

- Min $\hat{\beta}$ $\boxed{\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2}$

OLS Coefficients are Sample Means

- The estimated coefficients are weighted averages of the *Y*'s:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sum_{i=1}^{n}\left(\frac{\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} - \frac{1}{n}\right)Y_i$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}\left(\frac{\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} - \frac{1}{n}\right)\right)Y_i$$

- It is a function of the data (a special kind of sample mean), and so it is a *statistic*.
- It can be used to estimate something we are interested in: the population value of β
- Since it is a statistic, it has a sampling distribution that we can evaluate for bias and variance.

OLS estimator is unbiased

$$E\left[\hat{\beta}_1\right] = E\left[\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right] = E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\beta_0 + \beta_1 X_i + \varepsilon_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]$$

$$= E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 - \beta_1 \overline{X} - \overline{\varepsilon}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]$$

$$= \beta_1 E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right] + E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(X_i^{\,2}\right)}\right] - E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\overline{\varepsilon}}{\sum_{i=1}^{n}\left(X_i^{\,2}\right)}\right]$$

$$= \beta_1 + 0 + 0 = \beta_1$$

Variance of OLS estimator

- Variance is more cumbersome to work out by hand, so I won't do it:
- Top looks like the
- "even simpler" model.

$$Var(\beta_1) = \frac{1}{\left(\sum_{i=1}^{n} X_i^{\,2}\right) - n\overline{X}^2}\sigma^2$$

$$= \frac{1}{n\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2}\right) - \overline{X}^2\right)}\sigma^2$$

$$= \frac{1}{n\widehat{Var}(X)}\sigma^2$$

- Where V-hat is the
- sample variance of *X*
- $V(X)=E[X^2]-(E[X])^2$

How Do You Get Low Variance?

- The OLS estimator is unbiased, so it centers on the right thing.
- Its variance $Var(\beta_1) = \frac{1}{n\widehat{Var}(X)}\sigma^2$ has 3 pieces:
- N
- V(X)
- sigma-squared
- (draw them all)

The Classical Assumptions

1. The regression model is **linear in the coefficients**, **correctly specified**, and has an **additive error term**.
2. The error term has **zero population mean**: $E(\varepsilon) = 0$.
3. All independent variables are **uncorrelated with the error term**: $Cov(X_i,\varepsilon_i) = 0$ for each independent variable X_i.
4. Errors are uncorrelated across observations: $Cov(\varepsilon_i,\varepsilon_j) = 0$ for two observations *i* and *j* (no **serial correlation**).
5. The error term has constant variance: $Var(\varepsilon_i) = \sigma^2$ for every *i* (no **heteroskedasticity**).
6. No independent variable is a **perfect linear function** of any other independent variable (no **perfect multi-collinearity**).
7. The error terms are normally distributed. *We'll consider all the others, and see what we get. Then, we'll add this one.*

Specification

- Every time we write down a regression model (and estimate it!) we make some important choices:
  - what independent variables belong in the model?
  - what functional form should the regression function take (i.e., logarithms, quadratic, cubic, etc.)?
    - Dummy Variables
  - what kind of distribution should the errors have?
- Usually, we look to economic theory (and some common sense!) to guide us in making these decisions.
- The particular model that we decide to estimate is the culmination of these choices: we call it a **specification**
  - a regression specification consists of the model's independent variables, the functional form, and an assumed error distribution

Omitted Variables

- Suppose the **true DGP** is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

but we incorrectly estimate the regression model:

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^*$$

  - *example: Y is earnings, X_1 is education, and X_2 is "work ethic" – we don't observe a person's work ethic in the data, so we can't include it in the regression model*
- That is, we **omit** the variable X_2 from our model
- What is the consequence of this?
- Does it mess up our estimates of β_0 and β_1?
  - it definitely messes up our **interpretation** of β_1. With X_2 in the model, β_1 measures the marginal effect of X_1 on Y **holding X_2 constant**. We can't hold X_2 constant if it's not in the model.
  - Our estimated regression coefficients may be **biased**
  - The estimated β_1 thus measures the marginal effect of X_1 on Y **without holding X_2 constant**. Since X_2 is in the error term, the error term will covary with X_1 if X_2 covaries with X_1 .

Omitted Variables May Cause Bias

$$E\left[\hat{\beta}_1\right] = E\left[\frac{\sum_i \left(X_{1i} - \bar{X}_1\right)\left(Y_i - \bar{Y}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right] = E\left[\frac{\sum_i \left(X_{1i} - \bar{X}_1\right)\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i - \beta_0 - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right]$$

$$= E\left[\frac{\sum_i \left(X_{1i} - \bar{X}_1\right)\left(\beta_1\left(X_{1i} - \bar{X}_1\right) + \beta_2\left(X_{2i} - \bar{X}_2\right) + \varepsilon_i - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right]$$

$$= E\left[\frac{\beta_1 \sum_i \left(X_{1i} - \bar{X}_1\right)^2}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2} + \frac{\sum_i \left(X_{1i} - \bar{X}\right)\left(\beta_2\left(X_{2i} - \bar{X}_2\right) + \varepsilon_i - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}\right)^2}\right]$$

$$E[\hat{\beta}_1] = \beta_1 \text{ or } E[\hat{\beta}_1] < \beta_1 \; ?$$

$$= \beta_1 + \beta_2 E\left[\frac{\sum_i \left(X_{1i} - \bar{X}\right)\left(\left(X_{2i} - \bar{X}_2\right) + \varepsilon_i - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right] = \beta_1 + \frac{\beta_2}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2} E\left[\sum_i \left(X_{1i} - \bar{X}\right)\left(X_{2i} - \bar{X}_2\right)\right]$$

$$= \beta_1 + \frac{\beta_2}{\sum_i \left(X_{1i} - \bar{X}\right)^2} nCov\left[X_1, X_2\right] = \beta_1 + \beta_2 \frac{Cov\left[X_1, X_2\right]}{Var\left[X_1\right]}$$

The estimated parameter is **biased**, with bias linear in the true parameter on the left-out variable, and the covariance of the left-out variable with the included variable.

Violating Assumption 3: $Cov(X_i, \varepsilon_i) = 0$

- We saw that correlated missing regressors induce bias.
- So does biased sample selection and reverse causality.
- Consider correlated missing regressors.

Endogeneity in a Scatter-Plot

- Endogeneity is easy to draw.
- Consider a 1 variable (with intercept) model
- Let the conditional mean of the error term also rise linearly with the included variable
- Draw the true regression line and the data
- The OLS regression line will pick up both the slope of *Y* in *X* and the slope of the conditional mean of the error with respect to *X*.

General Endogeneity Bias

- Endogeneity bias shows up in the Ballentine diagrams.
- Correlated missing regressor: x2 is invisible.
- What does the OLS estimator do to the (x1,x2) overlap?
- more generally, some of what looks like (x1,y) is really the model error term, and not (x1,y).

Correcting for Endogeneity

- Endogeneity is like pollution in the *X*.
- You need information that allows you to pull the pollution out of the *X*.
- Including missing regressors is like identifying the pollution exactly, so that you can just use the *X* that is uncorrelated with that pollution.
- Alternatively, you could find a part of the variation in *X* that is unpolluted by construction.

Instrumental Variables

- *Instruments* are variables, denoted *Z*, that are correlated with *X*, but uncorrelated with the model error term by assumption or by construction.
- Cov(Z,e)=0, so in the Ballentine, Z and the error term have no overlap.
- But, (Z,X) do overlap

2-Stage Least Squares

- Regress *X* on *Z*
  - generate $\hat{X}$ =*E[X|Z]*, the predicted value of *X* given *Z*.
  - This is “clean”. Since *Z* is uncorrelated with the model error term, so is any linear function of *Z*.
- Regress *Y* on $\hat{X}$
- This regression does not suffer from endogeneity
- But it does suffer from having less variance in its regressor.

Violating Assumption 4

- Recall Assumption 4 of the CLRM: that all errors have the same variance. That is,

$$\text{Var}(\varepsilon_i) = \sigma^2 \text{ for all } i = 1,2,\ldots,n$$

- **Heteroskedasticity** is a violation of this assumption. It occurs if different observations' errors have different variances. For example,

$$\text{Var}(\varepsilon_i) = \sigma_i^2$$

  - In this case, we say the errors are **heteroskedastic.**
- Because heteroskedasticity violates an assumption of the CLRM, we know that least squares is not BLUE when the errors are heteroskedastic.
- Heteroskedasticity occurs most often in **cross-sectional** data. These are data where observations are all for the same time period (e.g., a particular month, day, or year) but are from different entities (e.g., people, firms, provinces, countries, etc.)

Inefficiency

- Why is OLS inefficient when we have pure heteroskedasticity?
- It is because there is another linear estimator that uses the data better, and can deliver a lower-variance estimated coefficient
- Eg, what if some observations had zero-variance on their errors, but others had positive variance
  - A linear estimator that delivers a lower-variance coefficient is to run OLS on *only* those observations with zero-variance. Trash all the rest of the data

What to do if errors are heteroskedastic …

- If you find evidence of heteroskedasticity – whether through a formal test by looking at residual plots – you have several options

1. Use OLS to estimate the regression and "fix" the standard errors
   A. We know OLS is unbiased, it's just that the usual formula for the standard errors is wrong (and hence tests can be misleading)
   B. We can get **consistent** estimates of the standard errors (as the sample size goes to infinity, a consistent estimator gets arbitrarily close to the true value in a probabilistic sense) called **White's Heteroskedasticity-Consistent** standard errors
   C. When specifying the regression in EViews, click the OPTIONS tab, check the "Coefficient Covariance Matrix" box, and the "White" button
   D. Most of the time, this approach is sufficient
2. Try Weighted Least Squares (WLS) – if you know the source of the heteroskedasticity and want a more efficient estimator
3. Try re-defining the variables – again, if you think you understand the source of the problem (taking log of dependent variable often helps)

Violating Assumption 5

- **Serial correlation** occurs when one observation's error term (ε_i) is correlated with another observation's error term (ε_j): $Corr(\varepsilon_i, \varepsilon_j) \neq 0$
- We say the errors are **serially correlated**
- This usually happens because there is an important relationship (economic or otherwise) between the observations. Examples:
  - **Time series data** (when observations are measurements of the same variables at different points in time)
  - **Cluster sampling** (when observations are measurements of the same variables on related *subjects*, e.g., more than one member of the same family, more than one firm operating in the same market, etc.)
    - Example: Suppose you are modeling calorie consumption with data on a random sample of families, one observation for each family member. Because families eat together, random shocks to calorie consumption (i.e., errors) are likely to be correlated within families.
- Serial correlation violates Assumption 4 of the CLRM. So we know that least squares is not BLUE when errors are serially correlated.

Consequences of Serial Correlation

- We know that serial correlation violates Assumption 4 of the CLRM, and hence OLS is not BLUE. What more can we say?

1. OLS estimates remain unbiased

   We only need Assumptions 1-3 to show that the OLS estimator is unbiased, hence a violation of Assumption 4 has no effect on this property

2. The OLS estimator is no longer the best (minimum variance) linear unbiased estimator

   Serial correlation implies that errors are partly predictable. For example, with positive serial correlation, then a positive error today implies tomorrow's error is likely to be positive also. The OLS estimator ignores this information; more efficient estimators are available that do not.

3. Standard formulae for standard errors of OLS estimates are wrong.

   Standard formulae for OLS standard errors assume that errors are not serially correlated – have a look at how we derived these in lecture 12 (we needed to use assumption 4 of the CLRM). Since our t-test statistic depends on these standard errors, we should be careful about doing t-tests in the presence of serial correlation.

What to do if errors are serially correlated …

- If you find evidence of serial correlation – whether through a formal test or just by looking at residual plots – you have several options available to you

1. Use OLS to estimate the regression and "fix" the standard errors
   A. We know OLS is unbiased, it's just that the usual formula for the standard errors is wrong (and hence tests can be misleading)
   B. We can get **consistent** estimates of the standard errors (as the sample size goes to infinity, a consistent estimator gets arbitrarily close to the true value in a probabilistic sense) called **Newey-West** standard errors
   C. When specifying the regression in EViews, click the OPTIONS tab, check the "coefficient covariance matrix" box, and the "HAC Newey-West" button
   D. Most of the time, this approach is sufficient
2. Try Generalized Least Squares (GLS) – if you want a more efficient estimator

Violating Assumption 6: $Cov(X_1, X_2) \neq 0$

- Recall we assume that no independent variable is a **perfect linear function** of any other independent variable.
  - If a variable X_1 can be written as a perfect linear function of X_2 , X_3 , etc., then we say these variables are **perfectly collinear**.
  - When this is true of more than one independent variable, they are **perfectly multicollinear.**
- Perfect multicollinearity presents technical problems for computing the least squares estimates.
  - Example: suppose we want to estimate the regression:
    $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ where $X_1 = 2X_2 + 5$.
    That is, X_1 and X_2 are perfectly collinear. Whenever X_2 increases by one unit, we see X_1 increase by 2 units, and Y increase by $2\beta_1 + \beta_2$ units. It is completely arbitrary whether we attribute this increase in Y to X_1, to X_2, or to some combination of them. If X_1 is in the model, then X_2 is completely redundant: it contains **exactly** the same information as X_1 (if we know the value of X_1, we know the value of X_2 **exactly**, and vice versa). Because of this, there is no unique solution to the least squares minimization problem. Rather, there are an infinite number of solutions.
  - Another way to think about this example: β_1 measures the effect of X_1 on Y, holding X_2 constant. Because X_1 and X_2 always vary (exactly) together, there's no way to estimate this.

Imperfect Multicollinearity

- It is quite rare that two independent variables have an **exact** linear relationship
  - it's usually obvious when it does happen: e.g., the "dummy variable trap"
- However it is very common in economic data that two (or more) independent variables are strongly, but not exactly, related
  - in economic data, everything affects everything else
- Example:
  - perfect collinearity: $X_{1i} = \alpha_0 + \alpha_1 X_{2i}$
  - imperfect collinearity: $X_{1i} = \alpha_0 + \alpha_1 X_{2i} + \zeta_i$ where ζ_i is a stochastic error term
- Examples of economic variables that are strongly (but not exactly) related:
  - income, savings, and wealth
  - firm size (employment), capital stock, and revenues
  - unemployment rate, exchange rate, interest rate, bank deposits
- Thankfully, economic theory (and common sense!) tell us these variables will be strongly related, so we shouldn't be surprised to find that they are ...
- But when in doubt, we can look at the **sample correlation** between independent variables to detect imperfect multicollinearity
- When the sample correlation is big enough, Assumption 6 is "almost" violated

Consequences of Multicollinearity

- Least squares estimates are still unbiased
- recall that only Assumptions 1-3 of the CLRM (correct specification, zero expected error, exogenous independent variables) are required for the least squares estimator to be unbiased
- since none of those assumptions are violated, the least squares estimator remains unbiased