CPP 523: Foundations of Eval I
Regression Model Mechanics

Prof. Jesse Lecy

Review Questions for the Final Exam

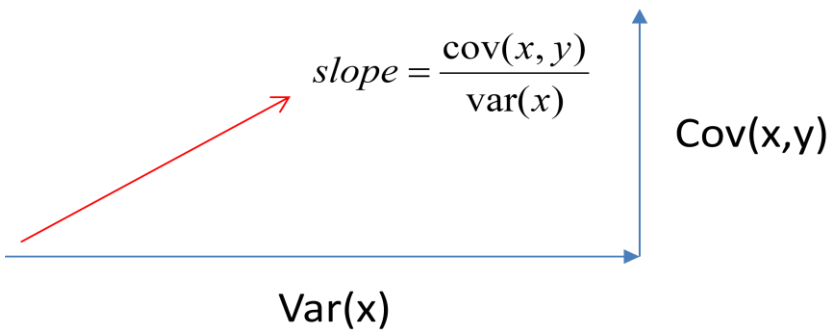
---

(1) Please give non-mathematical definitions of the following statistical concepts:

- Covariance - **covariance is a measure of how much X and Y move together** (if X moves to the right of its mean, does Y also move in the same direction from its mean?). The covariance has the same properties as a correlation except it is a raw measure, so the range can be from a very large negative number to a very large positive number versus negative one to positive one.

- Correlation - **the correlation is a standardized covariance so that it is easy to interpret** (think of it as the covariance in percentage terms). It ranges from negative one to positive one.

- Regression coefficient - **in a causal framework (which is what we care about in program evaluation), the coefficient represents the change in the dependent variable that we expect to see as a result of a change in the policy or treatment.** This information is captured by the formula for a slope, which is literally the change in Y associated with the change in X. The way that we measure change is through covariance and variance.



- Standard deviation - **on "average", how far is the data from the mean**. The term "average" is in parenthesis because it is not a true average, but rather squaring the distances and then later taking the square root. This is done for mathematical reasons and gives an approximation of the average distance, but it is not exact. (Recall that the exact average is calculated using absolute values, but it is not possible to take a derivative of an absolute value so they are rarely used.)

- Standard error - **on average, how far is the "best guess" from the "truth". The best guess is the statistic that we calculate from a sample - the sample mean or the slope of a regression. The truth is the population mean or slope.** We want to know how confident we can be about the guess so we calculate the average error that we would expect given the variance of the data. For the mean it is the variance of X that matters (the larger the variance the less confident we are about the sample mean matching the population mean). For the slope it is the variance of the error (residual) term. The error is the actual value of Y minus the prediction made by the regression model.

- Heterogeneity - **there are natural groups in the data and the group variable is correlated with the dependent variable and the policy variable.** Thus, omitting the group variable leads to bias. If there was no correlation than the population would be homogeneous - group traits would not predict the outcome. Recall that the fixed effects models were designed to deal with heterogeneity bias. So, for example, the state is a group-level variable that is correlated with economic output and spending on infrastructure so it should be included in the model.

- Multicollinearity - **high correlation of two independent variables.**

- The level of significance - **the percentage of times that we are willing to allow the true statistic (the mean or the slope) to fall outside of the confidence interval.**

- The confidence interval of a slope - **after choosing the level of significance, the confidence interval tells us the range over which we expect the true value to fall.**

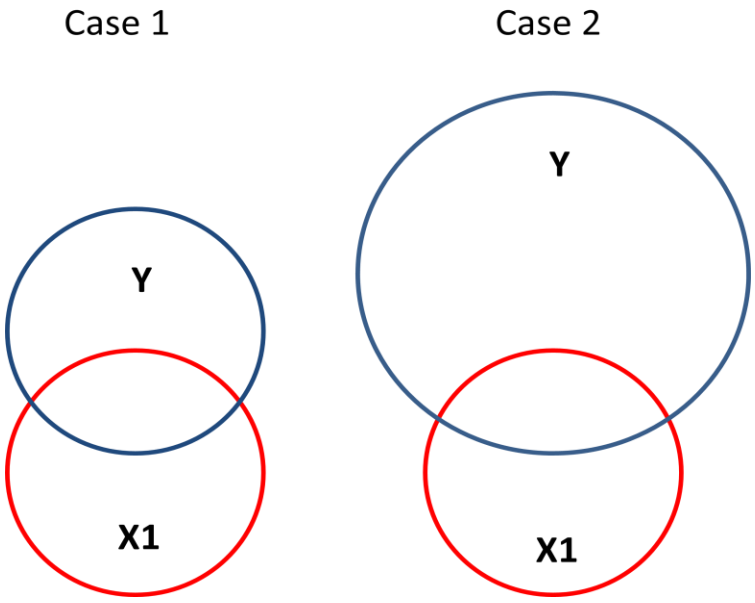
- R-square - **the percentage of the variance of the dependent variable we were able to explain with the regression model.**

- To "control" for a variable - **surgically removing the "contaminated" covariance of the control variable from the dependent variable and the policy variable so that only the independent effect of the policy variable will appear in the slope.**

(2) Standard Error of the Slope: Name three things that will reduce the standard error of a regression slope.

Larger sample size
More variation in the X variable
Additional control variables

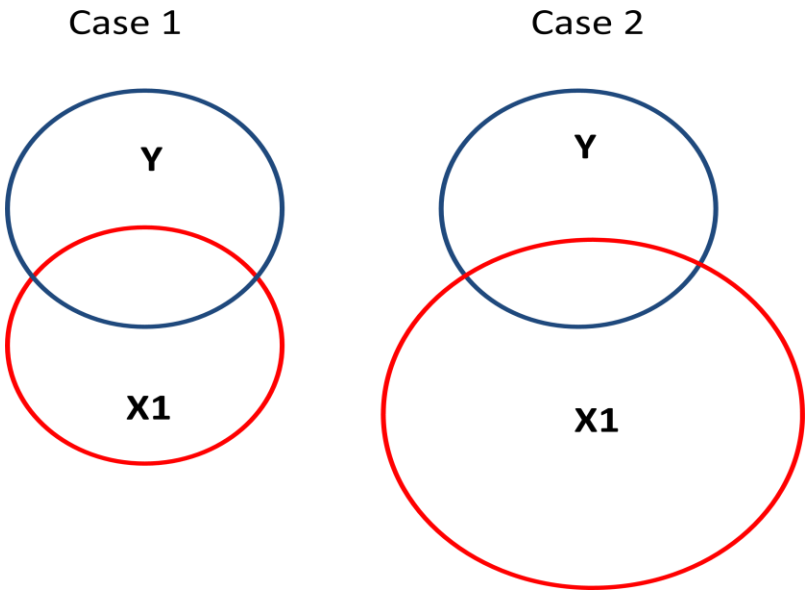
(3) Which will have the larger standard error, holding *cov(x,y)* and *var(x)* constant? Which will have the larger slope b_1?



The slope will stay the same because the slope comes from *cov(x,y)* and *var(x)*, both of which are constant in this example.

The standard error will be much larger in case #2 because there is more unexplained variance of Y. This leads to a larger error term, which is the numerator of the standard error. The denominator would remain constant since *var(x)* is not changing.

(4) Which will have the larger standard error, holding *cov(x,y)* and *var(y)* constant? Which will have the larger slope b_1?

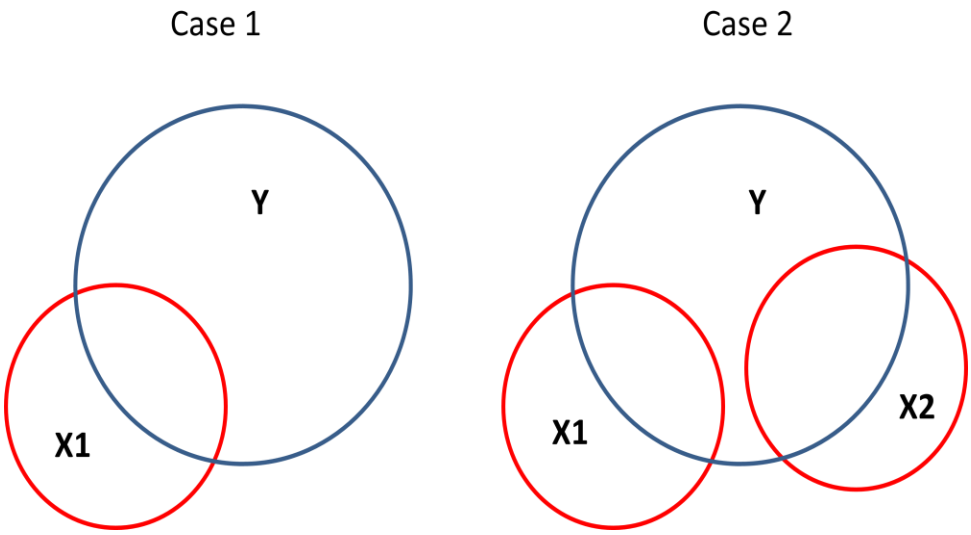


The slope of case #2 will be much smaller because *var(x)* is the denominator of the slope equation and it is larger. Specifically, smaller means closer to zero. It does not matter if the slope is positive or negative, the increased variance of x will make the slope move closer to zero.

The standard error of case #2 will, however, be much smaller. The amount of explained variance (the numerator of the standard error) remains constant. The variation in X (the denominator) increases though. As a result the standard error of case #2 is smaller than case #1.

This is an example of measurement error - poor measurement of X will increase the variance of the variable. It will attenuate the slopes but also decrease the standard errors. This happens when measurement error occurs in the independent variables. The previous question, (5), is an example of measurement error in the dependent variable.

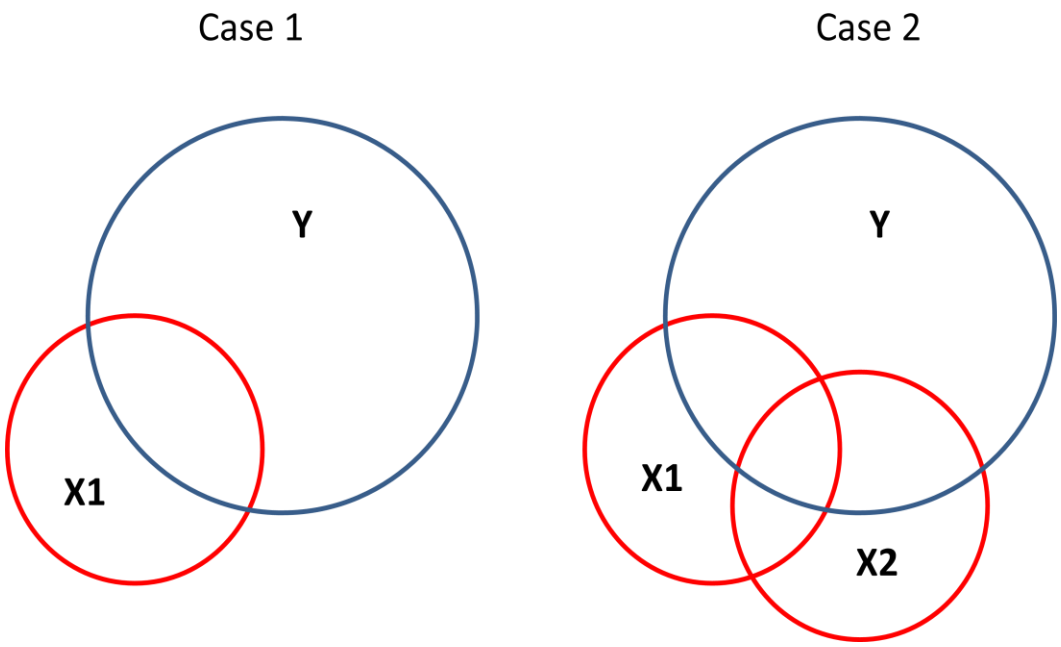
(5) Which will have the larger standard error, holding *cov*(x_1,*y*), *var*(x_1) and *var*(*y*) constant? Which will have the larger slope b_1?



The slope b_1 will be the same in both cases because the *cov(x,y)* and *var(x)* do not change.

Adding the variable in case #2 explains additional variance of Y, thus leading to a smaller standard error.

(6) Which will have the larger standard error, holding $cov(x_1,y)$, $var(x_1)$ and $var(y)$ constant? Which will have the larger slope b_1?



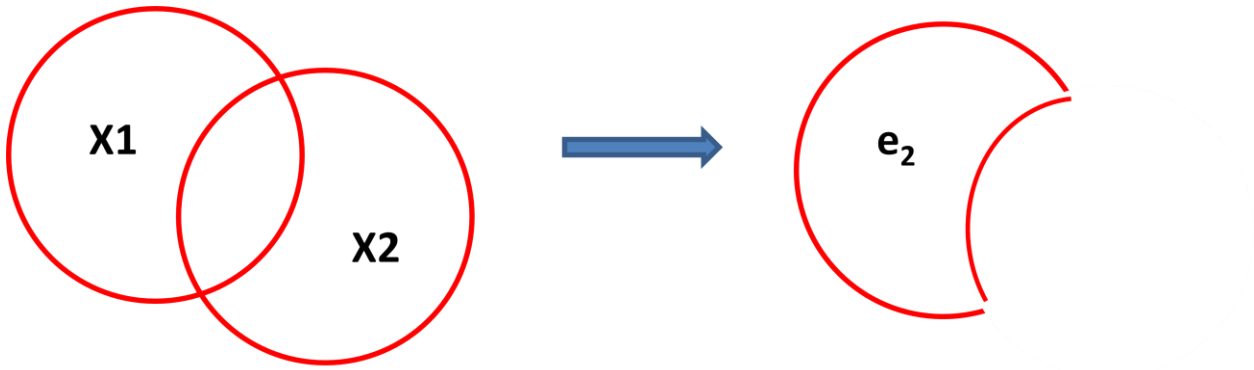
We cannot tell whether the slope will increase or decrease. All that we know is the slope in case #1 will be biased if the omitted variable X_2 is not included. The direction of the bias will determine whether the slope increases or decreases.

Similarly, it is impossible to tell how the standard error will change without doing the calculation. This is because the numerator and denominator of the standard error - the unexplained variance and the variation in X_1, will both be decreasing. They will not decrease the same amount, though, so you cannot tell the exact change in the standard error without doing the calculation.

It would be easy to draw two new scenarios, though, both that have three variables (Y, X_1 and X_2) and both in which X_1 and X_2 are correlated, but that it would be possible to tell from the Venn diagram which had the smaller standard error. Can you think of what this might look like?

(7) **Partitioned Regression:** In a lab we performed a partitioned regression to calculate the slope between mileage and price - the average amount that price will fall as a result of increased mileage in a used car. In the partitioned regression framework, explain why regressing RES1 (the residual of Price taking into account all controls) onto Mileage (versus RES2, the residual of Mileage taking into account the controls) will result in the wrong slope.

If you regress Mileage onto the RES1 then the denominator of the slope will be the whole variance of X_1. This is wrong because you only want to use the part that is uncontaminated by the shared covariance with X_2. You must first remove that portion in order to get the correct slope. RES2 (e_2 in the diagram) is the variable that has that portion removed. It comes from the regression of X_1 onto X_2.

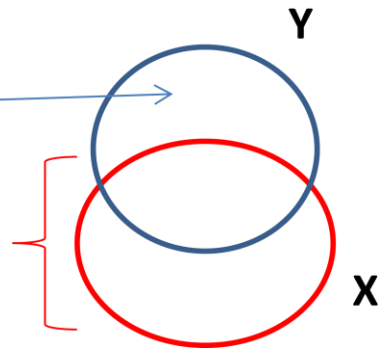


(8) **Measurement Bias:** Explain measurement bias using two different kinds of measurement error and four Venn diagrams.

See problems (5) and (6) above.

(9) **Standard Error:** Talk about the three parts of the standard error that drive the size of the statistic.

$$SE_{b_1} = \frac{\sqrt{\frac{SSE}{n-2}}}{\sqrt{(x_i - \bar{x})^2}}$$



The standard error of the slope comes from the variance of the error term (the numerator) and the variation of the independent variable (the denominator). Conceptually, it is:

$$SE_{b_1} \approx \frac{residual_y}{sample\,size \cdot \mathrm{var}(x)}$$

The phrases "variance of the error term" and "unexplained variance of Y" refer to the same thing - the top portion of Y that has no relationship with X. Because this portion of the standard error is in the numerator, explaining more of the variation in Y (usually by adding control variables) will lead to smaller standard errors.

The denominator is a raw measure of the variance of X and the sample size N. In general:

- **the larger the variance of X the smaller the standard errors.**
- **the larger the sample size, the smaller the standard errors.**

Note, though, that variance of X can be wiped out by including a highly correlated control variable in the model (the problem of multicollinearity).

(10) **Multicollinearity:** When do we care about multicollinearity? Specifically when it occurs in which variable (dependent, policy or control)? What is the tangible effect of multicollinearity?

There is no such thing as multicollinearity in the dependent variable. There is only the strength of correlation between the outcome of interest and the explanatory variables in the model. Multicollinearity specifically refers to correlation between independent variables.

Multicollinearity leads to inflated standard errors. This is because it removes a lot of the variation of X, thus making the denominator of the standard error smaller and the whole term bigger.

We don't care about multicollinearity if it occurs between two control variables. It might change their standard errors but we don't necessarily pay attention to the significance level of the controls, just of the policy variable. Highly correlated control variables will still do their job - accounting for extra variance in the dependent variable (and thus making the standard error of the policy variable smaller) and also making sure that there is not omitted variable bias in the policy slope.

The problem occurs when our policy variable is highly correlated with a control. In this case including the control could inflate the standard error of the policy variable and we could see a statistically non-significant slope. For example, in the first homework we had two measures of self-esteem that were highly correlated. If one of the measures was the policy variable then it would not have been wise to include the other in the model because they both could become statistically non-significant.

(11) **Slope:** If the covariance between x and y is positive, is it ever possible for the slope in a bivariate regression (only x and y) to be negative? Explain why. Use the formula for slope.

$$slope = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

No, it is not possible because the variance of X can never be negative. Similarly, if the covariance between X and Y is negative, the slope in a bivariate regression will never be positive.

In the case of multivariate regression the slope can change sign, but that is due to omitted variable bias (see the next question).

(12) **Bias:** If the covariance between x and y is positive, is it ever possible for the slope in a multivariate regression of y on x to be negative? Explain why. Use the formula for bias.

$$bias = \alpha_1 \cdot \beta_2$$
$$\approx cor(x_1, x_2) \cdot cor(x_2, y)$$

$$b_1 = \beta_1 + bias$$

The formula for bias takes into account the correlation between X_1 and X_2, and X_2 and Y (X_2 being the omitted variable in this case). Correlations can be positive or negative, so there are four possible scenarios which determine the direction of the bias:

Cor(X_1,X_2)	**Cor(X_2,Y)**	**Bias**
(+)	**(+)**	**(+)**
(-)	**(-)**	**(+)**
(+)	**(-)**	**(-)**
(-)	**(+)**	**(-)**

The question asks if it is possible for the covariance between X_1 and Y to be positive but the slope negative. This occurs when the omitted variable bias is both negative and it is larger than the true slope for X_1 and Y.

(13) **Omitted Variable Bias:** Consider a naive regression:

$$GPA = b_0 + b_1 GRE + e$$

And the true regression:

$$GPA = \beta_0 + \beta_1 GRE + \beta_2 MAT + \varepsilon$$

Calculate the bias that results from leaving out the MAT variable.

α_1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	38.299	19.654		1.949	.061
	GRE_Q	.051	.035	.267	1.466	.154

a. Dependent Variable: MAT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-2.129	.927		-2.297	.030
	GRE_Q	.006	.002	.484	3.756	.001
	MAT	.031	.008	.475	3.683	.001

a. Dependent Variable: GPA

β_2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.949	1.047		-.907	.372
	GRE_Q	.008	.002	.611	4.086	.000

a. Dependent Variable: GPA

b_1

$$bias = \alpha_1 \cdot \beta_2$$
$$0.051 \cdot 0.031$$
$$0.002$$

$$b_1 = \beta_1 + bias$$
$$\beta_1 + \alpha_1 \cdot \beta_2$$
$$0.006 + 0.051 \cdot 0.031$$
$$0.006 + 0.002$$
$$0.008$$

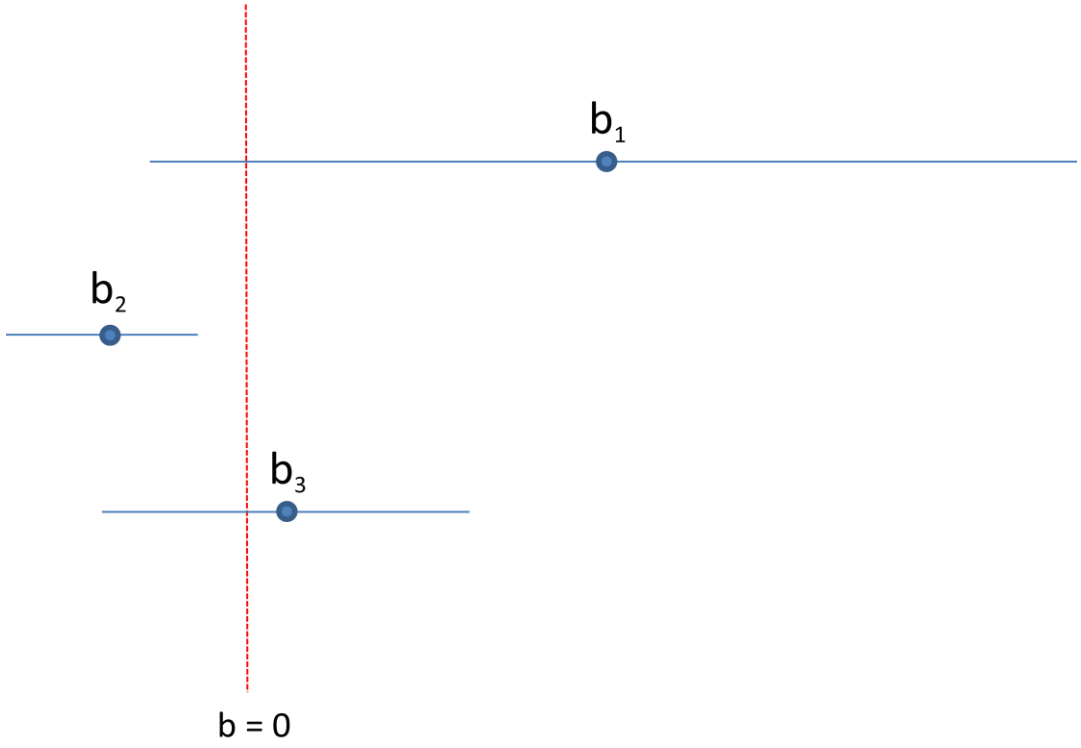
What does the bias tell us about the naive regression - is the slope too big or too small?

The slope b_1 is too big.

Consider a situation where there is another omitted variable, X3, that is correlated with both GPA and GRE scores. Under what conditions will the bias that results from omitting X3 from the regression be negative?

X3 must be positively correlated with the dependent variable and negatively correlated with the independent variable, or vice-versa.

(14) **Confidence Intervals:** Which variable, X_1, X_2 or X_3 has the largest variance and how do we know? (b_1 corresponds with X_1, etc)

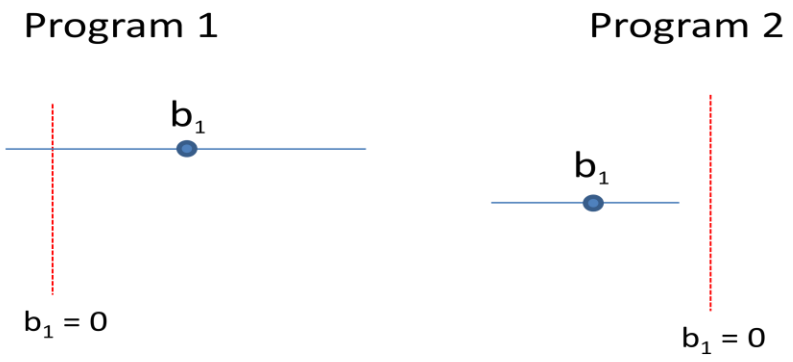


Think about the formula for the standard error. The numerator will be identical for the three slopes because it is always the variance of the error term. The denominator is the only thing that will be different, and this is the variation in each X. Since a larger denominator leads to a smaller standard error, X2 must have the largest variance in this example.

(15) What information do we get from the graphical representation of confidence intervals that we cannot get from the p-value? Stated differently, draw two confidence intervals that have the same p-values but very different policy impacts.

We not only get a level of significance, but we can see the range of policy outcomes that might be expected.

(16) Consider two programs that are meant to improve reading comprehension. The dependent variable is a score on a reading comprehension exam (higher being better). Which program do you prefer and why?



If you had to pick a program, Program 1 is a better bet because the range of program slopes includes many positive values. It also includes zero and negative values, so there is a chance the program might have no effect or a negative effect.

Program 2, however, will always have a negative impact so it is a bad option. Neither choice is great, but I would take a probability of a positive impact over a certainty of a negative impact.