Gov2k2 Final Review Jeopardy!

Put a (Beta) Hat On it	I Do it All for The (L)OLS	Assuming the Worst	Positive Interactions	Normally, I'm Pretty Confident
$200	$200	$200	$200	$200
$400	$400	$400	$400	$400
$600	$600	$600	$600	$600
$800	$800	$800	$800	$800

☠️ Final Jeopardy ☠️

Clue

Put a (Beta) Hat on It - $200

True or false:

You can eliminate **omitted variable bias** (OVB) by adding more observations to a regression model.

False!

OVB makes the OLS estimator both biased and inconsistent. Therefore reducing the variance of the estimator by increasing **n** does not eliminate the bias asymptotically (the OLS estimator does not converge on the true value).

Put a (Beta) Hat on It - $200

Back

True or false:

> You can eliminate **omitted variable bias** (OVB) by adding more observations to a regression model.

The answer is implicit in the name!

Put a (Beta) Hat on It - $400

Clue

Suppose we omit a variable from a regression model. That variable is negatively correlated with our independent variable **X** and negatively correlated with the outcome **Y**.

Will our OLS estimator tend to over-estimate, under-estimate, or correctly estimate the true coefficient on **X**?

Over-estimate (or be upwardly biased). By the omitted variable bias formula the bias is $\beta_2 \cdot$ **Cov(Z, X) / Var(X)**. The former is negative when **Y** and **Z** are negatively correlated and the latter is also negative since **Var(X)** is strictly positive. Therefore the bias is positive.

Put a (Beta) Hat on It - $400

Back

Suppose we omit a variable from a regression model. That variable is negatively correlated with our independent variable **X** and negatively correlated with the outcome **Y**.

Will our OLS estimator tend to over-estimate, under-estimate, or correctly estimate the true coefficient on **X**?

The OVB formula is $\beta_2 \cdot \mathbf{Cov(Z, X) / Var(X)}$

Clue

Put a (Beta) Hat on It - $600

Assume the following population model:

$$Y = \beta_0 + \beta_1 X + u$$

Your friend proposes the following estimator for β_1:

$$\beta_{1^*} = \beta_{1OLS} / (1 + \lambda)$$

Where $\lambda > 0$ is a positive constant and β_{1OLS} is the usual OLS estimator $(X'X)^{-1}Xy$.

What's the direction of the bias of this estimator β_{1^*}?

Does β_{1^*} have higher or lower variance than β_{1OLS}?

The bias tends toward zero since $1/(1 + \lambda) < 1$, which attenuates the OLS estimator toward zero. Since we know the OLS estimator is unbiased, this alternative estimator is biased.

The variance is lower than β_{1OLS}:

- $\text{Var}(\beta_{1OLS}/(1 + \lambda)) = 1/(1 + \lambda)^2 \cdot \text{Var}(\beta_{1OLS})$

- $\text{Var}(\beta_{1OLS})$ is strictly positive and $1/(1 + \lambda)^2$ is strictly less than 1.

Put a (Beta) Hat on It - $600

Assume the following population model:

$$Y = \beta_0 + \beta_1 X + u$$

Your friend proposes the following estimator for β_1:

$$\beta_{1^*} = \beta_{1OLS} / (1 + \lambda)$$

Where $\lambda > 0$ is a positive constant and β_{1OLS} is the usual OLS estimator $(X'X)^{-1}Xy$.

What's the direction of the bias of this estimator β_{1^*}? $E[\beta_{1OLS} / (1 + \lambda)]$?

Does β_{1^*} have higher or lower variance than β_{1OLS}? $V[\beta_{1OLS} / (1 + \lambda)]$

Put a (Beta) Hat on It - $800

If your OLS estimator is the Best Linear Unbiased Estimator, which of the following are true:

1. OLS has the lowest possible variance of any estimator

2. In small samples, test statistics for coefficient estimates will have *t*-distributions

3. Standard error estimates are unbiased and consistent for the true standard error

4. There exists no other unbiased estimator of the coefficients that has lower variance than OLS

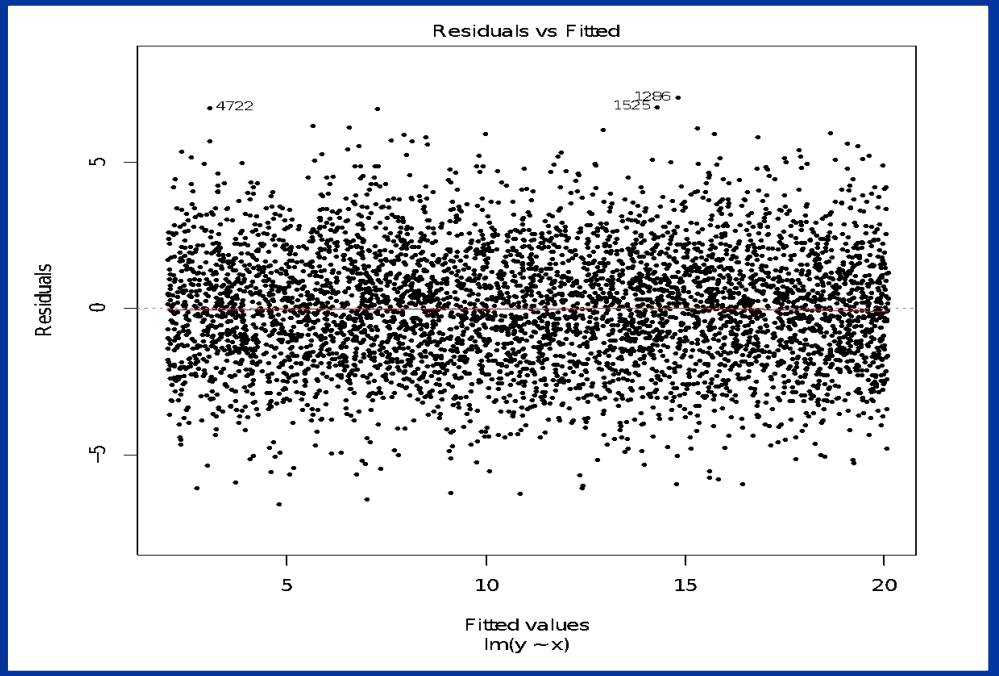
5. All of the above!

Just 4!

Recall from the hierarchy of assumptions that BLUE refers to the lowest variance among *unbiased* estimators. We can always trivially get lower variance by picking a biased estimator like $\beta_1^* = 5$, which has a variance of 0.

I Do it all for The (L)OLS- $200

Look at the following fitted values v. residuals plot. Does this plot suggest any evidence of **heteroskedasticity** or **non-linearity**?

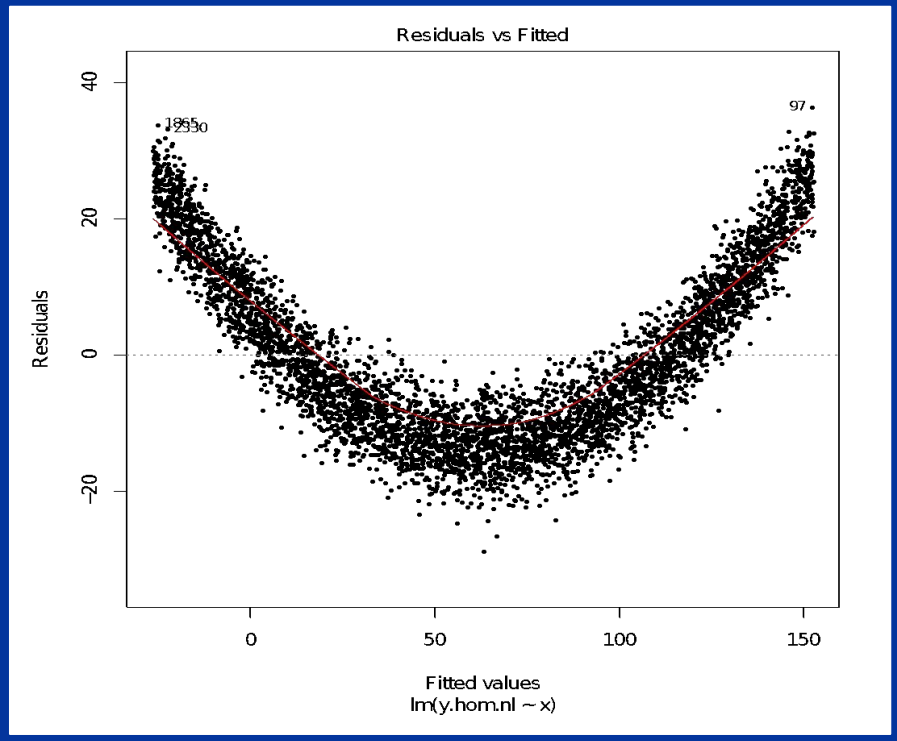


Neither!

This plot presents evidence of *neither* **non-linearity** nor **heteroskedasticity**. The **LOESS** curve through these points is essentially a straight line and there are roughly even numbers of residuals above and below the horizontal line at any given fitted value, so we don't see any visual evidence of nonlinearity. On average, the residual distances from the horizontal line stay constant as we travel up or down along our axis of fitted values, suggesting that there is no meaningful evidence of **heteroskedasticity** here.

I Do it all for The (L)OLS - $400

Take a look at the fitted values vs. residuals plot below. Does this plot suggest that there may be evidence of **heteroskedasticity** in your regression model?

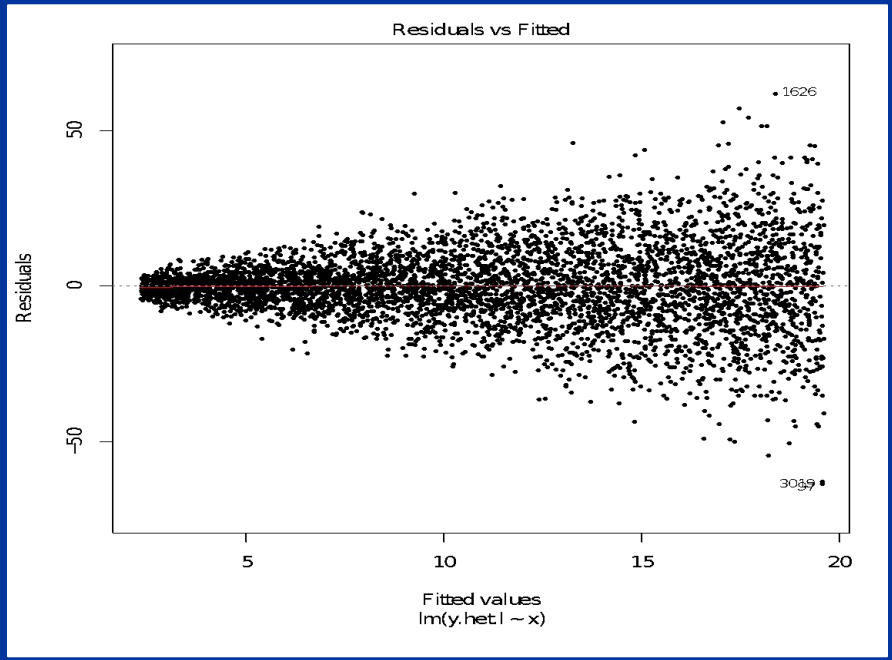


No!

While this plot clearly suggests **non-linearity** (the LOESS curve through these points are in the shape of a parabola), the the average spread of the residuals around the LOESS curve through these points doesn't actually change much as we go up or down along our axis of fitted values.

I Do it all for The (L)OLS - $600

Take a look at the fitted values vs. residuals plot below. Does this plot suggest that there may be evidence of **heteroskedasticity** in your regression model?



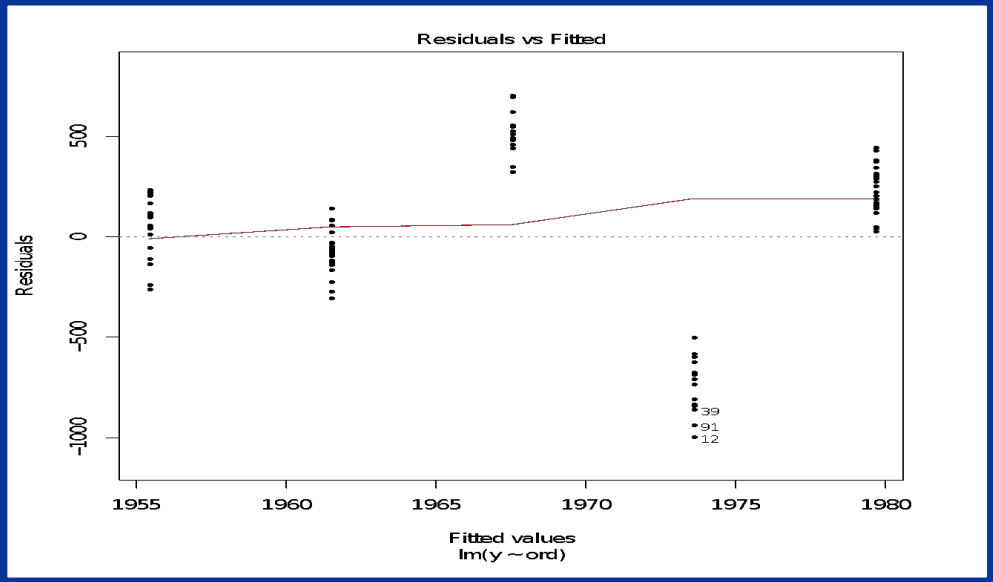
Yes!

Here, the average spread of the residuals from the horizontal line through 0 *is* getting larger with larger fitted values. That suggests non-constant error variance.

I Do it all for The (L)OLS - $800

Suppose you have an ordinal covariate **X** that takes on 5 different values for 5 different years, and a continuous outcome variable, **Y**.

In an attempt to estimate **E[Y|X]**, you run the regression `lm(Y ~ X)` and decide to do some diagnostics. `R` returns the fitted values vs. residual plot below for your model. List three issues here and how might you fix each one?



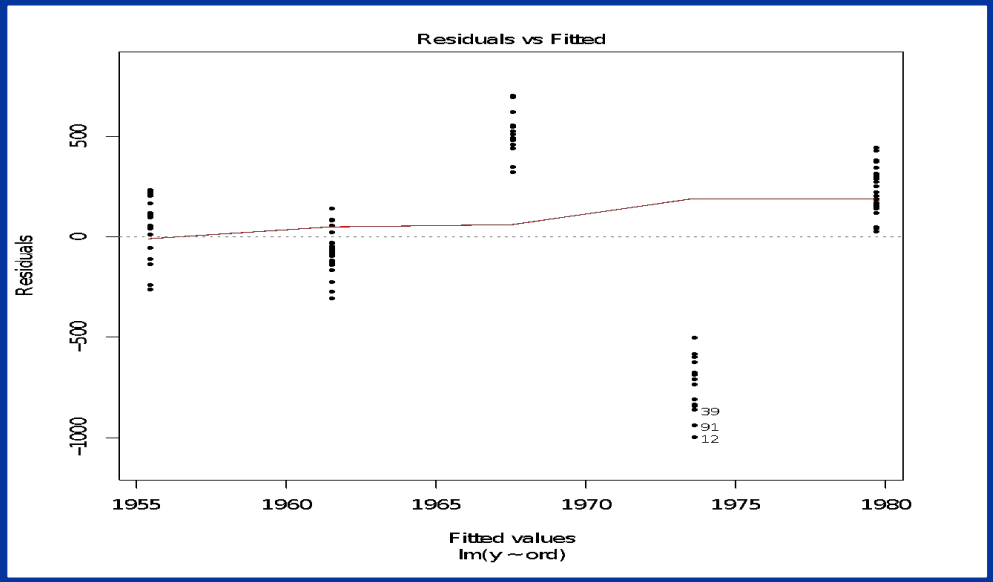
(1) **Non-linearity**, (2) **heteroskedasticity**, (3) treating **ordinal variable as continuous**. Fix (1) and (3) by turning **X** into **n – 1** binary dummy variables, e.g. `lm(Y ~ as.factor(X))`. Fix (2) with HC SEs with lmtest and sandwich.

Clue

I Do it all for The (L)OLS - $800

Suppose you have an ordinal covariate **X** that takes on 5 different values for 5 different years, and a continuous outcome variable, **Y**.

In an attempt to estimate **E[Y|X]**, you run the regression `lm(Y ~ X)` and decide to do some diagnostics. `R` returns the fitted values vs. residual plot below for your model. List three issues here and how might you fix each one?



With ordinal variables, you usually have to do an extra step before running `lm()`!

Back

Assuming the Worst - $200

Suppose you collected an i.i.d. sample for two variables: a continuous outcome variable, **Y**, and a binary covariate, **X.**

1. Is regressing **Y** on **X** a valid way to estimates **E[Y | X = 1] – E[Y | X = 0]**?

2. What two additional assumptions do you need for OLS to be unbiased?

3. What are the interpretations of β_0 and β_1 in a regression model for this data?

1. Yes, you can use OLS to calculate **E[Y | X = 1] – E[Y | X = 0]**

2. Recalling hierarchy of assumptions, we (1) **regularity**, meaning variation in **X** and (2) need **E[u|X]=0**, meaning **zero conditional mean** (i.e. no confounders of **X** and **Y**).

3. β_0 **= E[Y | X = 0]**, so our intercept is the conditional mean of **Y** when **X = 0**. β_1 **= E[Y | X = 1] -E[Y | X = 0],** so our slope is the average difference in **Y** between the **X = 1** and **X = 0** groups.

Assuming the Worst - $400

Suppose you are trying to estimate a model of the following form using OLS:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u$$

where **X**, **Z**, and **Y** are all continuous variables. You have **1,000** observations in your sample data. Which of the following assumptions have to hold in order for OLS to be **B.L.U.E.**?

1. Linearity

2. Random (i.i.d.) sample

3. Variation in **X**

4. $\mathbf{E[u] = 0}$

5. Homoskedasticity

6. $\mathbf{u \mid X \sim N(0, \sigma_u^2)}$

We need assumptions 1-5

We need to make the conditionally normal errors assumption (**6**) when we are using OLS to do inference with small samples. We rely on the CLT to get normally distributed errors in large samples.

Assuming the Worst - $600

Which of the following population regression models represent a violation of the **linearity** assumption:

1. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$
2. $Y = \beta_0 + \beta_1 X + \beta_2 \log(X) + \beta_3 X^3 + \beta_4 Z^4 + \beta_5 (1/Z) + u$
3. $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z + u$ (**X, Y, Z** are binary)
4. $Y = \beta_0 + \beta_1 X + \exp(\beta_2 Z) + u$ (**X** is continuous, **Y** is binary)
5. $Y = 1/(\beta_0 + \beta_1 X) + \beta_2 Z + u$

Just models 4 and 5.

Models can be non-linear in *variables*, but not in *parameters*.

Assuming the Worst - $800

Suppose you collected a data set that contains two variables: a continuous outcome variable, **Y,** and a binary covariate, **X**. You're interested in **E[Y | X = 1] - E[Y | X = 0]**, and you decide to calculate this quanitity of interest in two ways.

First, you calculate **E[Y | X = 1] - E[Y | X = 0]** nonparametrically.

Then, you run `lm()` and interpret β_1 as your difference in means.

Then you calculate confidence intervals for each of your difference in means estimators. But you get different values for the confidence intervals! Why might this happen?

`lm()` **SEs assumes homoskedasticity, the diff-in-means SE we used probably *doesn't*!**

$$\text{SE}_{\text{diff}} = \sqrt{\frac{s^2_{X=1}}{n_{X=1}} + \frac{s^2_{X=0}}{n_{X=0}}} \qquad \text{SE}_{\text{OLS}} = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

when **X** is discrete, we say that: "the above SE_{diff} uses **unpooled** variance, this SE_{OLS} uses **pooled** variance" (note, as you did in a pset, you can assume a pooled variance for SE_{diff} also)

Clue

Assuming the Worst - $800

Suppose you collected a data set that contains two variables: a continuous outcome variable, **Y,** and a binary covariate, **X**. You're interested in **E[Y | X = 1] - E[Y | X = 0]**, and you decide to calculate this quanitity of interest in two ways.

First, you calculate **E[Y | X = 1] - E[Y | X = 0]** nonparametrically.

Then, you run `lm()` and interpret β_1 as your difference in means.

Then you calculate confidence intervals for each of your difference in means estimators. But you get different values for the confidence intervals! Why might this happen?

Pset 6 Q4: you derived a standard error for a difference-in-means with the Gerber & Green data,

Pset 10 Q1: you derived a standard error for an OLS coefficient

... in what way were they different?

Back

Positive Interactions - $200

Suppose you estimate the following regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z + u$$

What's the expected marginal effect of a unit change in **X** conditional on **Z**?

$$\beta_1 + \beta_3 Z$$

Positive Interactions - $400

Suppose you estimate the following regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X \cdot Z + u$$

Where **Y** is a continuous outcome variable, **Z** is a binary variable, and **X** is a continuous variable.

What assumption are you implicitly making about the relationship between **X, Z, Y**?

What are the implications of that assumption?

The implicit assumption here is that the coefficient on your lower-order term, which might be $\beta_3 Z$, is zero.

This suggests that you think there is no difference between the **Z = 0** and **Z = 1** group when X is zero.

While that might be reasonable under some circumstances, you typically don't know enough to justify this assumption in practice. The implication is that your slope estimates for **X** at each level of **Z** will be distorted. Can conduct an **F Test** (see slide 21).

Clue

Positive Interactions - $400

Suppose you estimate the following regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X \cdot Z + u$$

Where **Y** is a continuous outcome variable, **Z** is a binary variable, and **X** is a continuous variable.

What assumption are you implicitly making about the relationship between **X, Z, Y**?

What are the implications of that assumption?

Hint: you're assuming *something* is zero in the population regression.

Back

Positive Interactions - $600

Suppose you estimate the following regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z + u$$

Where **X**, **Y**, and **Z** are all continuous variables. Suppose

- **Y** is wage income in USD,
- **X** is years of work experience and
- **Z** is # of merit-based promotions received while in workforce

What is the substantive interpretation of β_3?

β_3 tells us the *change in the effect that years of work experience has on wage income given a one unit change in the number of promotions an individual has received in the workforce*.

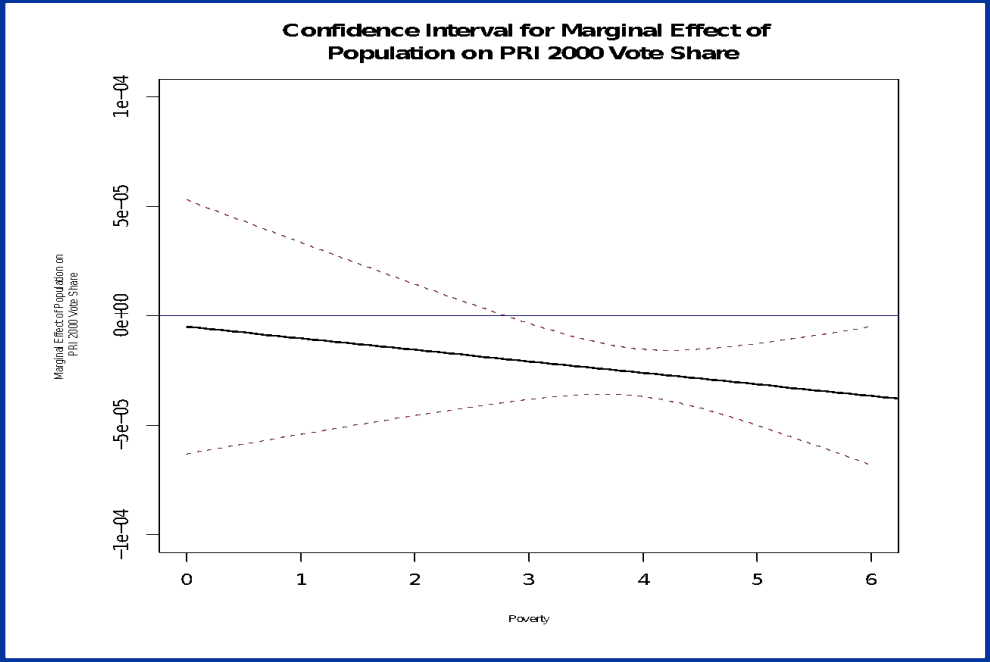
We might use a model like this if we think an individual's wage income depends on some interaction between their work experience (seniority) and some proxy for their productivity (promotions).

Positive Interactions - $800

Suppose you use the following interactive model to estimate the PRI (Mexico's Partido Revolucionario Institucional)'s vote share as a function of poverty and population:

$$\textbf{Vote share} = \beta_0 + \beta_1 \textbf{poverty} + \beta_2 \textbf{population} + \beta_3 \textbf{poverty} \cdot \textbf{population} + u$$

The following is a 95% CI plot of the marginal effect of population on the PRI's vote share. What function is being plotted here? What does this plot tell us about the statistical significance of the effect of population *conditional* on poverty?



The marginal effect of population, and the function plotted, is $\beta_2 + \beta_3 \textbf{poverty}$.
This plot tells us that the marginal effect of population on the PRI's vote share in 2000 is <u>not</u> significantly different from **0** when levels of poverty is **<3.**

Normally, I'm Pretty Confident - $200

Suppose I generate the following confidence intervals for some β_1:

- **85%** CI: [-0.278, 1.806]
- **80%** CI: [-0.163, 1.691]
- **75%** CI: [-0.068, 1.596]
- **70%** CI: [0.015, 1.514]
- **65%** CI: [0.089, 1.440]

Find a lower and upper bound on the *p*-value for a hypothesis test of the null: $H_0: \beta_1 = 0$, $H_A: \beta_1 \neq 0$.

DAILY DOUBLE!!!

The *p*-value is somewhere between .30 and .25 given that the 75% CI includes 0 and the 70% CI does not.

Clue

Normally, I'm Pretty Confident - $200

Suppose I generate the following confidence intervals for some β_1:

- **85%** CI: [-0.278, 1.806]
- **80%** CI: [-0.163, 1.691]
- **75%** CI: [-0.068, 1.596]
- **70%** CI: [0.015, 1.514]
- **65%** CI: [0.089, 1.440]

Find a lower and upper bound on the *p*-value for a hypothesis test of the null: $H_0: \beta_1 = 0$, $H_A: \beta_1 \neq 0$.

Hint: a 100(1-α) CI is all candidate β_1 values where we retain the null hypothesis (i.e. $p < \alpha$)

Hint: notice that the 75% CI (α=0.25) is the first α where we include **0** (i.e. retain the null)

Back

Normally, I'm Pretty Confident - $400

If the true β_1 equals **-3**, which of the following hypothesis tests has a higher probability of rejecting the null hypothesis?

1. $H_0: \beta_1 = 0, H_A: \beta_1 \neq 0$
2. $H_0: \beta_1 \leq 0, H_A: \beta_1 > 0$

The first one

In the first test, the true value of β_1 is *in* the null hypothesis.

So if the test statistic is negative (which it will be with high probability if β_1 is **-3**), Test 1 will reject while Test 2 will fail to reject.

Clue

Normally, I'm Pretty Confident - $400

If the true β_1 equals **-3**, which of the following hypothesis tests has a higher probability of rejecting the null hypothesis?

1. $H_0: \beta_1 = 0, H_A: \beta_1 \neq 0$
2. $H_0: \beta_1 \leq 0, H_A: \beta_1 > 0$

Hint: which of the null hypotheses actually contains the case $\beta_1 = -3$?

Back

Normally, I'm Pretty Confident - $600

Suppose in the population it is true that $\mathbf{Y = \beta_0 + \beta_1 X + u}$.

Assume that the errors u are normally distributed with a mean 0 and variance σ^2 and that the **Gauss-Markov** assumptions hold.

Suppose you estimate a regression model using OLS that includes **X** along with an irrelevant variable **Z** using a sample of **12** observations.

You find that the estimated standard error of β_1 is the *same* in this multivariate regression as it is in the simple regression of **Y** on **X**.

Is the *p*-value of your hypothesis test for $\mathbf{H_0: \beta_1 = 0, H_A: \beta_1 \neq 0}$ in the multivariate regression larger, smaller, or the same as the *p*-value for the same hypothesis test in the bivariate regression?

Larger

Even though the standard errors are the same, the distribution of the test statistic is slightly different – it's *t* with one less degree of freedom. Since the t distribution is fatter-tailed when the number of degrees of freedom is low, the probability of getting extreme test statistics is larger and therefore the p-value for the test will increase.

Clue

Normally, I'm Pretty Confident - $600

Suppose in the population it is true that $\mathbf{Y = \beta_0 + \beta_1 X + u}$.

Assume that the errors u are normally distributed with a mean 0 and variance σ^2 and that the **Gauss-Markov** assumptions hold.

Suppose you estimate a regression model using OLS that includes **X** along with an irrelevant variable **Z** using a sample of **12** observations.

You find that the estimated standard error of β_1 is the *same* in this multivariate regression as it is in the simple regression of **Y** on **X**.

Is the *p*-value of your hypothesis test for $\mathbf{H_0: \beta_1 = 0, H_A: \beta_1 \neq 0}$ in the multivariate regression larger, smaller, or the same as the *p*-value for the same hypothesis test in the bivariate regression?

Hint: because of **G-M** assumptions, your hypothesis test statistic is *t* distributed with **n-k** degrees of freedom … google what happens to the variance of that *t* distribution as degrees of freedom decreases.

Back

Normally, I'm Pretty Confident - $800

Suppose you are conducting hypothesis tests for **20** regression coefficients in a multivariate OLS regression model.

Assume that each *p*-value is independent of the others. You set the level for each individual test to **0.05**.

Suppose you choose to reject the null hypothesis that all of the coefficients equal **0** if you reject any one of the individual hypotheses.

What is the Type I error rate, Pr(Any rejected | All true). of your combined tests? What **α** level would you have to set to for each individual test in order to guarantee a Type I error rate of at most **0.05**?

Type I Error Rate: 0.64, α = 0.00256

P(Any rejected | All true) = 1 – P(None rejected | All True)

P(Any rejected | All true) = $1 - 0.95^{20} = 0.64$

Solving for α:

$0.05 = 1 - (1-\alpha)^{20} = 0.00256$

Clue

Normally, I'm Pretty Confident - $800

Suppose you are conducting hypothesis tests for **20** regression coefficients in a multivariate OLS regression model.

Assume that each *p*-value is independent of the others. You set the level for each individual test to **0.05**.

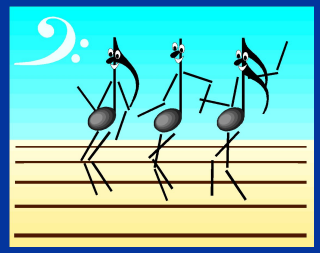
Suppose you choose to reject the null hypothesis that all of the coefficients equal **0** if you reject any one of the individual hypotheses.

What is the Type I error rate, Pr(Any rejected | All true). of your combined tests? What **α** level would you have to set to for each individual test in order to guarantee a Type I error rate of at most **0.05**?

Hint: P(Any rejected | All true) = 1 – P(None rejected | All True)
Hint: Use independence of the *p*-values to get the answer!

Back

☠ Final Jeopardy ☠



Suppose I start with a regression of **Y** on **X**.

I modify it to include a variable **Z** that is correlated with **Y** but uncorrelated with **X**.

Will the standard error of the coefficient for **X** increase, decrease, or stay the same?

Decrease!

Including **Z** decreases the variance of your residuals (for the same reason it increases your $\mathbf{R^2}$).

Therefore the standard error will decrease.