CPP 523: Foundations of Eval I   
Regression Specification

Prof. Jesse Lecy

PRACTICE EXAM

**NAME**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Instructions:** You have four hours to complete the exam once it is started. You can use notes, calculators, and statistical software. You are NOT allowed to work with anyone else, or share questions and solutions with others. Good luck!

Please give non-mathematical definitions to the following statistical concepts:

1. **The 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.**
2. **The 95% confidence interval of a slope:  
     
   The range over which we expect the true value of the statistic (the population statistic) to fall.**

1. **R-Squared:  
     
   The proportion or percentage of the variance of the dependent variable we were able to explain with the regression model.**
2. Name three things that will reduce the standard error of a regression slope.  
     
     
   **1. Increase the sample size.   
     
   2. Decrease the residual of Y (add control variables uncorrelated with the policy variable of interest).   
     
   3. Increase the variance of X (only possible if designing a sampling framework or an experiment where you have control over X – instead of caffeine dosages from 0 to 100mg, give dosages from 0 to 500mg – variance of X will increase as a result.**
3. Name two sins of the Seven Sins that will always increase the standard error of a regression slope.

**1. Measurement error in the dependent variable.**

**2. High multicollinearity (adding a control variable that is highly-correlated with X).**

1. Control variables that are uncorrelated with our policy variable will not cause omitted variable bias if we do not include them in a regression. Why do we include them in the model? For example, Teacher Quality in the Classroom Size example from class.  
     
   **They explain more of Y, or stated differently account for more of the residual of Y. As a result, including them will decrease the standard errors, which makes confidence intervals smaller, improving our likelihood of achieving statistically significant results and increasing the efficiency of our estimates.**
2. Calculate the slope and the intercept for a simple bivariate regression model (  ) from the following information:

: 4

: 11

*var(x):* 3

*var(y):* 7

*cov(x,y):* 6

b1 = cov(x,y)/var(x) = 6 / 3 = 2

b0 = 11 – 2(4) = 3

1. Now using the slope and intercept that you calculated above, calculate the residual (prediction error – column ***e***) for the following three cases.

|  |  |  |  |
| --- | --- | --- | --- |
| **X** | **Y** |  | e |
| 5 | 15 | 3 + (2)(5) = 13 | 2 |
| 6 | 11 | 3 + (2)(6) = 15 | -4 |
| 8 | 21 | 3 + (2)(8) = 19 | 2 |

1. Consider the following regression:



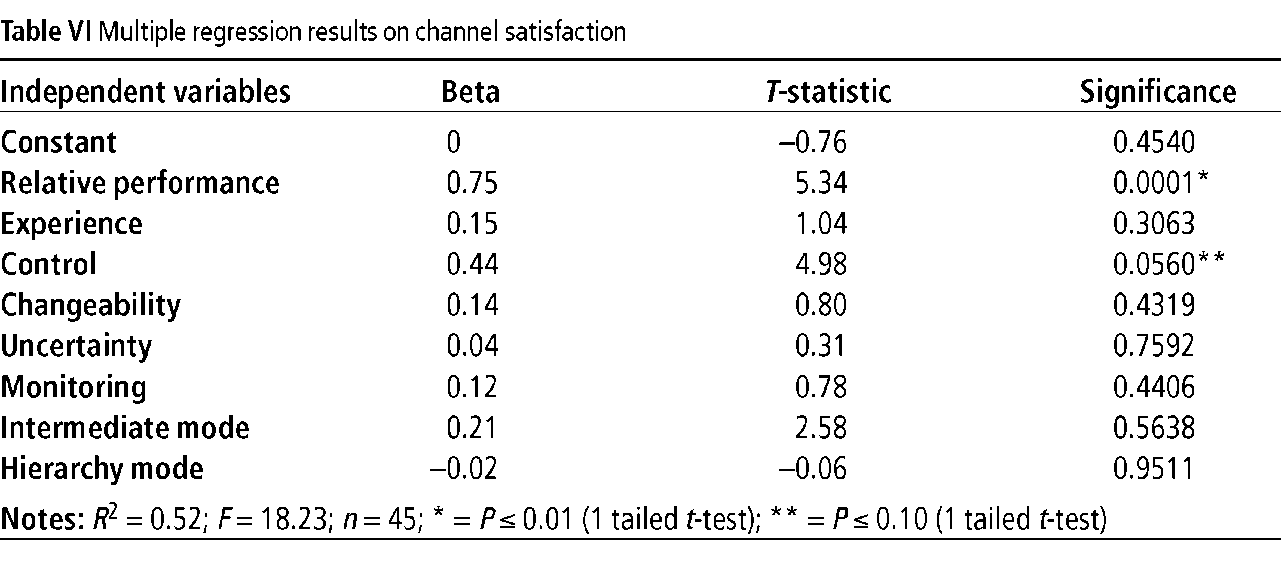
Consider the case where *B1* = 6, and *SEB1* = 2.61.  
  
Using t=2.58, calculate the 99% confidence interval for *B1.* Is the slope statistically significant at this level? How do we know?

**6 +/- (2.58)(2.61):**

**-0.7 < B1 < 12.7**

**It is NOT significant because it contains zero, meaning we can’t tell for certain whether the intervention has a positive or negative impact on the outcome (at a 99% level of confidence).**

1. Consider the following regression results.



What is the largest level of confidence you can chose for the confidence interval around the slope estimate for the **Experience** coefficient before it crosses zero?

**1 – 0.3063 = 0.6937 confidence interval**

1. Consider the following cases:



* 1. Holding *cov(x,y)* constant across all cases, which case(s) will have the largest standard error?

**SE ~ residual(y) / var(x)**

**Case 1: SE ~ a / b**

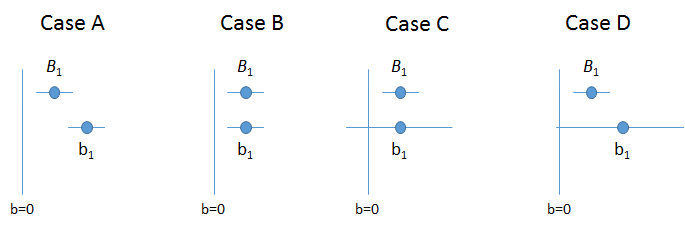
**Case 2: SE ~ A / b <<< CASE 2**

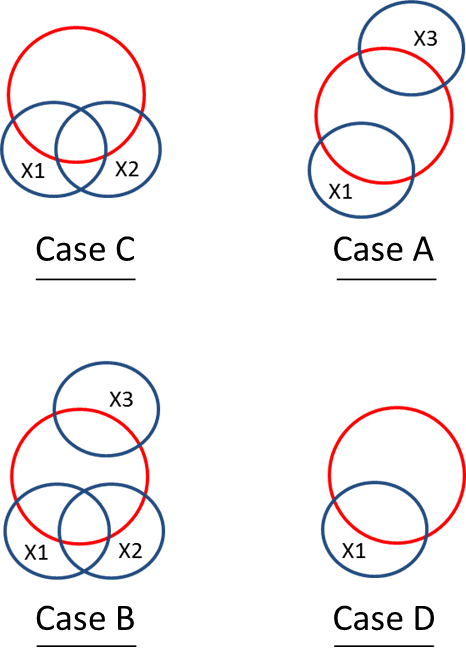
**Case 3: SE ~ a / B**

* 1. Holding *cov(x,y)* constant across all cases, which case(s) will have the smallest slope?  
       
     **slope ~ cov(x,y) / var(x)**

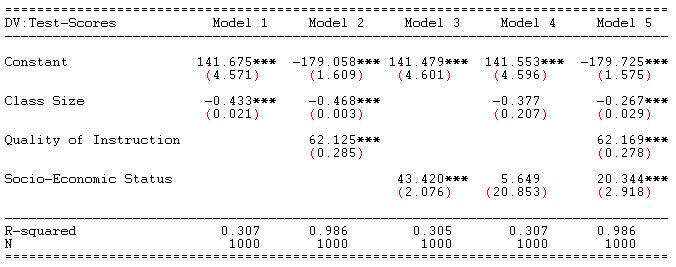
**Covariance is the same across all cases, so Case 3 has the biggest var(x) and thus the smallest slope.**

1. Consider four cases below. The full regression in this case is with X1 being the policy variable:  
     
       
     
   Write the correct case letter under each Venn diagram.





**BONUS (3pts):** Go back to the model that attempts to discern the effects of class size on test scores:



Now think about another model:



What is the exact slope for the regression of **SES** on **Class Size,** π1? Show your math.

**BONUS (4 pts):** Think back to the model that we have studied looking at the relationship between classroom size and test scores:

 (1)

Now think about a different way to run the regression model. What if we constructed it in the following way:

 (2)

 (3)

In this case the e1 in model (3) is the residual term from model (2). Using a Venn diagram to justify your response, answer the following questions:

Does ?

Does ?

Does ?

Does ?