14.32/14.320 Econometric Data Science
Midterm Exam
1pm, Thursday October 26, 2023

Part 1 (36 points) - USE BLUE BOOK #1

1. A student is considering increasing studying from $2\frac{1}{2}$ hours per day to 3 hours per day. Compute a 95% confidence interval for the predicted effect of this increase using:

   (a) (5 points) Table 2, regression (1). In real-world terms, is this effect large or small?
   Estimated increase $= 0.5 \times .038 = .019$ (increase of 0.5 hours)
   $SE = 0.5 \times .025 = .0125$
   95% confidence interval $= .019 \pm 1.96 \times .0125 = .019 \pm .0245 = (-.006, .044)$.
   Even the largest end of this range is a very small increase in GPA.

   (b) (5 points) Table 2, regression (2). In real-world terms, is this effect large or small?
   There are two methods to calculate it:
   Log method:
   Estimated increase $= \Delta \ln X \times \hat{\beta}_1 = (\ln(3) - \ln(2.5)) \times .114 = 0.182 \times .114 = 0.021$
   $SE = \Delta \ln X \times SE(\hat{\beta}_1) = 0.182 \times .086 = .0157$
   95% confidence interval$= .021 \pm 1.96 \times .0157 = .019 \pm .031 = (-.010, .052)$.
   Even the largest end of this range is a very small increase in GPA. (This confidence interval is similar to the confidence interval from part (a).)
   Percent method This method uses the approximation that the log change in X is approximately the decimal change in X, so that the estimated effect on Y of a one percent change in X is $.01\hat{\beta}_1$. An increase from 2.5 hours to 3 hours is a 20% increase. Thus,
   Estimated increase $= 20 \times .01\hat{\beta}_1 = 20 \times .00114 = 0.023$
   $SE = 20 \times .01 \times SE(\hat{\beta}_1) = .2 \times .086 = .0172$
   95% confidence interval $= .023 \pm 1.96 \times .0172 = .023 \pm .034 = (-.011, .057)$.
   Even the largest end of this range is a very small increase in GPA. (This confidence interval is similar to the confidence interval from part (a).)

2. (5 points) Suppose that the first hour of studying is more productive (in terms of grades) than the second, etc; that is, suppose that there are declining marginal returns to studying. Which specification is better suited to model these declining marginal returns, (1) or (2)? Briefly explain.

   The logarithmic specification is well suited to modeling declining marginal returns to studying, whereas the linear specification is not. In the linear specification (regression (1)), the effect of one hour studying does not depend on the number of hours studying

– that is, the slope is constant. Thus the linear model has constant marginal returns to studying. In the linear-log specification (regression (2)), the effect of a 1% increase in studying time does not depend on the number of hours studying. The larger the value of study time, the larger in hours is a 1% increase. Thus, to get the same payoff in terms of grades, you must study increasingly longer at higher values of Study; that is, the marginal return to studying is declining

3. (5 points) Test the null hypothesis that GPA1 is linear in *Study*, holding constant sex, race, ACT score, major, and health, against the alternative hypothesis that it is a polynomial of degree up to 3, at the 10% significance level. Explain your reasoning (be explicit).

   The hypothesis of linearity in *Study* implies that the coefficients on $Study^2$ and $Study^3$ in the population version of regression (3) are both zero; the alternative of a polynomial up to degree 3 implies that one or the other of those terms is nonzero. This can be tested using the F-statistic testing whether both those coefficients are zero. The value of the F-statistic is 2.03. The number of restrictions is $q = 2$, so from the table the 10% critical value is 2.30. Because $2.03 < 2.30$, we do not reject the null hypothesis at the 10% significance level.

4. (5 points) Test the hypothesis (at the 5% significance level) that Social Studies majors and Physical Education majors have the same GPA on average, holding constant hours studying, sex, race, ACT score, and health.

   Comparing the list of majors in Table 1 and Table 2 indicates that the omitted group in Table 2 is PE majors. Thus the coefficient on the major_ss binary variable is the mean difference in GPA1 between SS and PE majors, holding constant the other variables in the regression. Thus to test the hypothesis that these two majors have the same GPA on average (controlling for the other variables), we need to test whether the coefficient on major_ss is zero, against the hypothesis that it is nonzero. The double asterisk in the table indicates that the hypothesis is rejected at the 1%, and thus 5%, significance level. (The t- statistic is $2.96 > 1.96$, however because the asterisks are included in the table it was unnecessary to compute the t-statistic).

5. (5 points) How would you modify regression (1) to examine whether the effect on grades of studying is the same for men and women, holding constant race, ACT score, major, and health? Be specific.

   This can be done by adding an interaction term to the regression, specifically $Study \times male$. Then the coefficient on *Study* captures the effect for women, while the sum of coefficients on *Study* and $Study \times male$ capture the effect for men. Significance of coefficient on $Study \times male$ implies that two effects are statistically different.

6. (6 points) Provide an example of a variable omitted from regression (1) that plausibly would result in omitted variable bias in the estimated effect on GPA1 of Study, controlling for the other variables in regression (1).

   For a variable to cause OV bias, it must (i) be a determinant of Y (omitted from the regression) and (ii) be correlated with X. Unobserved student characteristics could have this effect. For example, ability is plausibly related to studying – perhaps more able students need to study less – and it is also plausibly an omitted variable from this regression. On the latter point, ACT score is included in the regression, but the ACT score reflects many things in addition to ability, such as the quality of the high school attended and ability to take standardized tests. Thus a very able student might not do well on the ACT – that is, the ACT is an imperfect measure of ability. Thus omission of ability would cause omitted variable bias.

   Some suggested that the number of hours of sleep would be an omitted variable. Understanding whether omission of Sleep would cause omitted variable bias is subtle. Sleep is surely a determinant of GPA1 and Sleep would be correlated with Study, so it would seem that omission of Sleep would cause omitted variable bias. However, including Sleep in the regression changes the question being asked, or more precisely the effect being estimated. Without Sleep, the question is, what is the effect of hours studying, holding constant the other variables in the regression? With Sleep, the question is, what is the effect of hours studying, holding hours sleeping (and the other variables) constant? In the first question, studying can come at the expense of sleep, but in the second question studying can come only at the expense of other non-sleep activities. It is plausible that these two effects would be different (if studying crowds out partying that is probably more beneficial to grades than if studying crowds out sleeping). A more extreme example is the omission of time spent on classwork including attending class. This is a determinant of GPA and also is correlated with Study. But including it changes the effect of Study to a question about whether it is more useful to go to class or to do homework, holding constant the total amount of time spent on classwork. This is clearly a different effect than simply varying Study without this constraint. So omission of Sleep does not cause omitted variable bias.

Part 2 (34 points) - USE BLUE BOOK #2

1. (6 points) As was shown in part 1 question 3, there is not much non-linearity and regression (3) in non-competitive. The choice between (1) and (2) should be made based on (i) interpretability and (2) R^2 (which is nearly identical here). Either choice between (1) and (2) will get a full credit if well argued.

2. Consider the following threats to internal validity:

(a) (7 points) sample selection bias

Sample selection bias arises when there is a selection process related to the dependent variable. In this case, the first-semester data consists of all 210 randomly selected students so there is no sample selection bias.

(b) (7 points) simultaneous causality bias

Simultaneous causality bias would arise if GPA1 affects study time. This seems quite likely – students who are doing poorly might be motivated to study more. This reverse causal effect would suggest a negative coefficient (worse grades induce more studying). The original causal effect plausibly has a positive coefficient (more studying leads to better grades). This suggests that the effect estimated by OLS combines these two and could well result in a coefficient near zero – with these two effects offsetting each other in the data. This is a major threat to internal validity and is a plausible explanation for the puzzlingly small effect of Study estimated in regressions (1) and (2).

3. (6 points) Consider the effect on GPA1 of a roommate bringing a video game box to school, holding constant sex, race, ACT, major, and health. In your judgment, is coefficient on video in Table 2, regression (4), a biased or unbiased estimate of this effect? Explain.

Unbiased. This is a good example of a variable, video, which arguably satisfies the conditional mean independence condition. The other regressors in the regression are presumably correlated with omitted variables – major is presumably correlated with creativity and native ability in a way that is not entirely captured by ACT, for example, so the causal effect of the variable major on grades is not captured by these regression coefficients. However, because roommates are randomly assigned, whether a roommate brings a video box can be viewed as randomly assigned. If it is randomly assigned, this means that it is distributed independently of the error tem, either with or without the W's. Thus, E(u|video, W) = E(u|W), which is the conditional mean independence condition. Because the conditional mean independence condition plausibly holds, the OLS estimator of the coefficient on video is plausibly unbiased (and consistent). Also, note in particular that the random assignment of roommates implies that whether a roommate brings a video game is distributed independently of any of the student's individual characteristics. Although these individual characteristics are a determinant of GPA, they are uncorrelated with video because video is (in effect) randomly assigned via the random assignment of roommates.

Actually, there is one subtlety here – roommates are not assigned entirely at random, just randomly within sex. If males have a higher proclivity to bring video games, then "video game" would be randomly assigned, conditional on sex, but would have a systematic relation to the error term if sex were not included in the regression (and

if sex were a determinant of GPA). Thus this is an example where at least one of the control variables is important for the conditional mean independence condition to hold.

4. Victor and Nathan are both new male Freshmen with the same race, ACT score, major, and health status; but Victor has a roommate with a video box whereas Nathan does not.

   (a) (4 points) What is the predicted difference between Nathan's study time and Victor's? Is this difference large or small in a real-world sense?

   The predicted difference is the coefficient on video in regression (5), which is a decrease in study time of 0.668 hours per day, or approximately 40 minutes per day. This is a large number: the average study time is 3 hours, or 180 minutes, so the effect of video is over 20% of the average study time.

   (b) (4 points) What is the predicted difference between Nathan's first-semester GPA and Victor's? Is this difference large or small in a real-world sense?

   The predicted difference is the coefficient on video in regression (4), which is a decrease in GPA1 of 0.241 points, or approximately a single gradation in a letter grade (A- to B+, for example). This is a moderately large number, for example it is approximately one-third of a standard deviation of GPA1 (see Table 1).

Part 3 (30 points) - USE BLUE BOOK #3 True, False, Uncertain. Explain your answer. Formal derivations are not necessary if you are able to fully explain your answer. Just guessing will earn you zero points, even if your guess is correct:

1. (7 points) Consider the regression:

$$\ln(Wage)_i = \beta_0 + \beta_1 Female_i + \beta_2 College_i + \beta_3 (Female_i \times College_i) + e_i \qquad (1)$$

   where $\ln(Wage)_i$ is log of wages, $Female_i$ is a dummy variable equal to one if female, and $College_i$ is a dummy variable equal to one if a person has a college diploma. You also decide to divide your observation in 4 groups: group 1- college educated women, group 2- college educated men, group 3- women without college diploma, and group 4- men without college diploma. You decide to run another regression:

$$\ln(Wage)_i = \gamma_0 + \gamma_1 Group1_i + \gamma_2 Group2_i + \gamma_3 Group3_i + e_i, \qquad (2)$$

   where $GroupK$ is a dummy for group $K = 1, 2, 3$. Then the R^2 for regressions (1) and (2) will be the same.

Answer: True. Optimization problem in (1) is equal to optimization problem in (2), coefficients are linear transformations of each other. In particular, regression (2) is

$$\begin{aligned}\ln(Wage)_i =& \gamma_0 + \gamma_1 Group1_i + \gamma_2 Group2_i + \gamma_3 Group3_i + e_i \\ =& \gamma_0 + \gamma_1 (Female_i \times College_i) + \gamma_2 ((1 - Female_i) \times College_i) \\ &+ \gamma_3 (Female_i \times (1 - College_i)) + e_i \\ =& \gamma_0 + \gamma_3 Female_i + \gamma_2 College_i + (\gamma_1 - \gamma_2 - \gamma_3)(Female_i \times College_i) + e_i\end{aligned}$$

so $\gamma_3 = \beta_1$, $\gamma_2 = \beta_2$ and $\beta_3 = \gamma_1 - \gamma_2 - \gamma_3$.

2. (7 points)Suppose you collect data from a survey on wages. In addition you ask for information about marijuana use. The original question is "On how many separate occasions last month did you smoke marijuana?" Write an equation that would allow you to estimate the effect of marijuana use on wages. You should be able to make statements such as "Smoking marijuana 5 more times per month is estimated to change wage by x%"

   Answer: An example of equation is

$$\ln(wage_i) = \alpha + \beta MarperMonth_i + x_i'\gamma + \epsilon_i$$

   where $MarperMonth$ is the number of occasions that an individual smoked marijuana last month, and $x_i's$ are other characteristics of the individual. We can answer to the question by looking at the quantity $500\hat{\beta}$.

3. (16 points) Assume that you have a regression $Y_i = \alpha + \beta X_i + e_i$ with heteroskedastic errors, which of the following statements are true?

   i. Running regression in STATA with robust option gives an unbiased estimator of β.

   ii. Running regression in STATA without robust option gives an unbiased estimator of β.

   iii. The OLS estimator of β is inconsistent.

   iv. The OLS estimator of β is not asymptotically normal.

   Answer: Heteroskedasticity-robust option only affects standard errors. It does not affects the estimate - it is the same OLS. Thus, we can easily see that (i), (ii) are true. Both (iii) and (iv) are not true because we know that the OLS estimator being consistent and asymptotically normal did not require the homoskedasticity assumption. Though it does affect how to calculate standard errors.