MOEC.0021
Empirical Methods
University of Zurich HS 2017
PROF: Greg Crawford / TA: E. Dicarlo, M.R. Greco, H.C. Boy Handout HS 2017

Problem Set 3

This problem set is due on the **12th of December** at **23:59**.
Solutions should be turned in via email to **emanuele.dicarlo@econ.uzh.ch** in PDF form. Please follow the following steps when submitting your solution:

1. Email Title: MOEC0021 Problem Set 3 Solutions
2. Attachment Title: GroupName_PS3.pdf
   For example, if my group was called 'DataMonkeys' I would name the attachment DataMonkeys_PS3.pdf

Remember, your goal is to communicate. Full credit will be given only to the correct solution which is described clearly. Convoluted and obtuse descriptions might receive low marks, even when they are correct. Also, aim for concise solutions, as it will save you time spent on write-ups, and also help you conceptualize the key idea of the problem.

---

1 Pencil and Paper

1. **Omitted Variable Bias**
   Suppose the true data generation process for a student's salary in their first job after they graduate with a Master's degree (their "starting salary") is:

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

   where Y_i = the starting salary of individual i, $X_{1i} = i$'s Grade Point Average (GPA) in their Master's coursework, and X_{2i} = whether their Master's degree was in economics or finance (versus, e.g., history or literature), and the standard CLRM assumptions hold, especially that $E(\epsilon_i | X_{1i}, X_{2i}) = 0$

   Suppose instead that you estimate the model

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \epsilon_i$$

   (a) Write down the formula for $\hat{\alpha}_1$ and calculate $E(\hat{\alpha}_1)$.

   (b) Is $\hat{\alpha}_1$ likely to be unbiased? Why or why not?

   (c) Given your answer to question (1b), do you think any bias will be positive or negative? Explain. If you need to make an assumption in order to answer the question, state your assumption clearly and give the reasons that you made it.

(d) Suppose the true data generation process included another variable, X_{3i}, which measured the time spent per week by student i on extracurricular activities (e.g. sports, travel, etc.), and that you were able to include this in your model, i.e.

$$\begin{aligned} \text{Truth:} \quad Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \\ \text{You estimate:} \quad Y_i &= \alpha_0 + \alpha_1 X_{1i} + \alpha_3 X_{3i} + \epsilon_i \end{aligned}$$

(Note that you are still omitting X_{2i}, the dummy variable indicating whether their Master's degree was in economics or finance)

- How does the addition of this new variable change your answer to question (1c), if at all? Explain.

2. **Measurement Error in y**

In class we showed that the OLS estimator is biased towards zero when there is measurement error in one of the x's. This question asks you to do a similar analysis for measurement error in y.

In particular, suppose the true model is

$$y_i^* = \alpha + \beta x_i + \epsilon_i^*$$

with $E(\epsilon_i^*|x_i) = 0$ and $V(\epsilon_i^*|x_i) = \sigma_*^2$ (as in the CLRM). Further suppose that there is measurement error in y_i. In particular, you don't observed y_i^*, but instead observe $y_i = y_i^* + \eta_i$, with $\eta_i \sim (0, \sigma_\eta^2)$. You then estimate the model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where I leave it to you to derive the relationship between ϵ_i, ϵ_i^*, and η_i as well as the mean and variance of ϵ_i.

Further assume (as we did in lecture) that this is "classical measurement error", i.e. it is uncorrelated with everything:

$$\begin{aligned} E(\eta_i|x_i) &= 0 \\ E(\eta_i|\epsilon_i^*) &= 0 \end{aligned}$$

(a) What is the mean and variance of ϵ_i?

Begin with the expression $\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$ and answer the following questions:

(b) Is $\hat{\beta}$ biased in this case?

(c) What is $V(\hat{\beta})$ in this case? Note that without measurement error, $V(\hat{\beta}) = \sigma_*^2(X'X)^{-1}$. How does $V(\hat{\beta})$ compare when there is measurement error in y?

(d) You decide to ask a friend who's taken the course before whether or not they think measurement error in y is something important to worry about and they say, "What? No way. It's no big deal." Given your answers to questions (2b) and (2c), do you agree? Explain.

2 Empirical Application

1. **Dealing with Measurement Error**
   Download the file *indicators.csv* from OLAT and import it into STATA or R. We have values for the following 8 variables in 2015.

   - *country*
   - *countrycode*
   - *mortalityun*, an index for child mortality rate under age 5 also reported by the UN
   - *hospital_deaths*, an index of mortality generated by number of deaths in hospitals
   - *govmort*, index of under 5 child mortality as reported by each government.
   - *corruptionun*, an index of corruption reported by UN observers
   - $ruleOfLaw$, another proxy for corruption based of the degree to which laws and regulations are actually enforceable in the country

   In this Exercise we will try to understand what are the different types of measurement error and what their consequences can be when estimating a model. To do so, we will analyze the relationship between corruption and child mortality. The corruption indexes are constructed such that higher values of the index indicate that the country is *more* corrupt. Furthermore, to ensure comparability, all indexes are standardized with a mean of zero and standard deviation of one.

   (a) Do you think these two variables are likely to be subject to measurement error? Explain.

   (b) Suppose you believe that the corruption and mortality scores reported by the UN are the most reliable. Regress mortality on corruption using these measures.

      i. What is your OLS estimate of the relationship between them? Call your estimate $\hat{\beta}$. What is the p-value from the one-sided hypothesis test that $\hat{\beta} = 0$?

      ii. Suppose the CLRM assumptions are satisfied: how do you interpret $\hat{\beta}$? Is this a large or small effect in your opinion?

      iii. Make a graph with the scatter plot of mortality and corruption together with the fitted regression line and confidence intervals. For the rest of the exercise, suppose this is the "true" relationship between the two variables.

   (c) Suppose now that official mortality data (i.e. *mortalityun*) are not available. However, you have access to hospitals records in each country from which you - with much time and effort - manually extracted the number of deaths of infants under the age of 5 to build your mortality index. Call this index "hospital_deaths". It's possible that you made mistakes doing this, but you are willing to assume that any such mistakes were probably random.

i. This setting is similar to that you studied in Question 2 in the Pencil-and-paper section! But is it just similar or is it really *the same*? In particular, do you think this variable is likely to satisfy the conditions of *classical measurement error* we invoked there? What findings do you expect from regressing *hospital_deaths* on *corruptionun*? Explain.

ii. Regress *hospital_deaths* on *corruptionun*. How does your coefficient estimate compare to that you estimated in question (1b)? Is this consistent with your expectations? Explain.

iii. Plot in a single figure the scatterplot of both of your mortality variables against *corruptionun* as well as each of your regression lines. How do they differ in terms of standard errors and *confidence intervals*? Is this consistent with your expectations? Explain.

(d) Suppose now that your UN mortality index is available but your UN corruption index (*corruptionun*) is not. Instead, you have the UN index for Rule of Law. This index is based on different measures of corruption and is highly correlated with the UN corruption index. You can safely assume that any error between the two is random.

i. Regress *mortalityun* on *ruleOfLaw*. How does the coefficient compare to that from question (1b)? Is this consistent with your expectations? Explain.

(e) As in question (1c), suppose that *mortalityun* is not available. Nor were you able to collect yourself the raw data. What is available is a mortality rate self-reported by each country in the data called *govmort*.

i. This setting is similar to that you studied in Question 2 in the Pencil-and-paper section! But is it just similar or is it really *the same*? In particular, do you think this variable is likely to satisfy the conditions of classical measurement error we invoked there? Explain.

ii. If yes, leave this question blank. If not, what is the likely sign of any bias in the coefficient on *corruptionun* from a regression of *govmort* on *corruptionun*? Show this using the same tools you used in Question 2 above.

iii. Regress *govmort* on *corruptionun*. How does the coefficient compare to 1b? Is this result consistent with your expectations? Why or why not?

(f) Which of the three cases covered in questions (1c), (1d), or (1e) do you believe to be most dangerous in terms of identification of the true causal effect of corruption on child mortality? Explain.

2. **IV Regression**

This question is based on the paper by Bonjour, Cherkas, Haskel, Hawkes and Spector ("Returns to Education: Evidence from UK Twins," The American Economic Review, 2003), which we will refer to in what follows as BCHHS. Start by reading the paper - it is available on OLAT and not very long. The dataset of BCHHS is available on to download here: `http://bit.ly/1YATkWe`.

The same data set is also available on the website of the American Economic Association (http://www.aeaweb.org/articles.php?doi=10.1257/000282803322655554), i.e. it is exactly the same dataset that BCHHS have submitted to the journal alongside their paper.

The data set contains the following variables: family (family number), twinno (twin number within the family: 1 or 2), earnings (hourly wage), highqua (each twin's reported years of schooling), age, and some other variables that we will not be using here. It also includes "twihigh", which is each twin's estimate of the years of schooling *of the other twin*. Open the data set in your preferred statistical software and familiarize yourself with the variables listed above.

Generate the variables 'lnearn '(log earnings) and 'agesq' (the square of age).

(a) Use the data set to reproduce the results in columns (2) and (3) of Table 2 in BCHHS. That is, perform the following regressions:

- Regress log earnings on years of schooling, age and age squared using OLS.
- Estimate the same model, but use twins estimated years of schooling as an instrument for years of schooling.

i. Do you find any discrepancy between your results and those reported in BCHHS? Is the discrepancy "serious"? Why or why not?

ii. The authors do not report the coefficient on the constant term. Is there any important information contained in that coefficient?

iii. The main coefficient of interest here is the coefficient on education (= years of schooling). Assume for the moment that your IV results are consistent and explain the interpretation of this coefficient.

(b) Let's think harder about the IV regression we just ran.

i. Try to provide at least *two* reasons why years of education ('highqua') might be endogenous.

ii. For *each* reason, conjecture about the likely sign of the bias the source of endogeneity would have on your estimate of the returns to schooling.

iii. For each reason, evaluate the relevance and exogeneity of using a twin's sibling's report of their years of schooling ('twihigh') as an instrument for their own report of their years of schooling ('highqua').

iv. Examine the difference between the OLS and IV results. Did the IV results move in the direction you expected? Why or why not?

v. Report the First Stage regression for your IV estimation and test whether the instrument is weak.

vi. Do you "believe" these results? Why or why not?