MOEC.0021
Empirical Methods
University of Zurich — HS 2018
PROF: Greg Crawford / TA: E. Dicarlo, M.R. Greco, S. Bagagli — Handout HS 2018

Problem Set 3

This problem set is due on the **19th of November** 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 'Winteriscoming' I would name the attachment Winteriscoming_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. **Coefficients Interpretation.**

   You have a sample from the US Consumer Expenditure Survey in the third quarter of 2016. It has the following variables:

   - *newid*, household id
   - *age*, age of head of household
   - fam_size, number of household components
   - *consumption*, consumption expenditure in last quarter USD
   - *income*, quarterly household income USD
   - *house*, dummy for owning a house

   You estimate several regression models to estimate how these variables can affect total *consumption* expenditures of a family. Your estimates are reported in Table 1

   (a) Write down the regression model estimated in column (1). What is the value of the coefficient on income? What is its interpretation?

(b) Now write down the regression model in column (2). What is the interpretation of the coefficient on income? What is the interpretation of the coefficient on family size?

(c) In the last column you add whether i owns a home. What is now the value of the coefficient on income? How do you interpret it?

(d) Why are your answers changing (if they are)? Which is "right", i.e. which interpretation should you be *trying* to get?

Table 1:

	Dependent variable:		
	consumption		
	(1)	(2)	(3)
income	0.267***	0.254***	0.254***
	(0.006)	(0.006)	(0.006)
fam_size		625.431***	625.445***
		(75.731)	(75.726)
house			1,395.781
			(1,021.504)
Constant	5,800.441***	4,429.220***	4,413.947***
	(132.779)	(212.165)	(212.445)
Observations	6,371	6,371	6,371
R^2	0.221	0.229	0.229
Note:			*p<0.1; **p<0.05; ***p<0.01

2. **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 $X_{2i} =$ a dummy for 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 $E(\hat{\alpha}_1 | X)$.

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

(c) Given your answer to question (2b), 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.

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

(Note that you are still omitting X_{2i}, the student's GPA)

- How does the addition of this new variable change your answer to question (2c), if at all? In particular, do you think any bias on $\hat{\alpha}_1$ is likely to be greater or less than your answer in part (2c)?

3. **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

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

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 $y_i : y_i = y_i^* + \eta_i$, with $\eta_i \sim (0, \sigma_\eta^2)$ and that you estimate the model

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

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?

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 on 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.