7110MK PhD ECONOMETRICS I (fall 2017)

Assignment 2 (October 6, 2017) Due on Oct. 15th and (revised) Oct. 20th

Instructions:

- You are encouraged to work in groups of up to 3 people for the assignment. Please note the members of your group on your write-up.
- Be sure to understand, that every member of the group needs to hand in an individual write up, exactly identical solutions will not be acceptable. (The idea is: Work on the solution together on paper, then go home and do the clean write up on your own...)
- You should send in the first version of the assignment *by* Oct. 15th, 8pm. The first version by no means needs to be perfect. You may revise this after class and resubmit the final version for Oct 20th, 8pm. Hence the deadline is set to Sunday, but resubmission is acceptable until the later date. (prefereably upload on T-square, or drop it at my mailbox, or drop it at the SoE office desk (this option is until 4pm).
- The total score are 100pts, which will translate to approx. 10 percent of your final grade.
- I strongly recommend to start working on the project part right away. Especially obtaining data (second part) can take weeks (years sometimes, but that's a different story).
- Do not hesitate to contact me via e-mail if in your group you encounter questions for clarification or suspect a typo. (typos are likely)
- BONUS: This time you can get up to 10 bonus points, if a colleague mentions your help in their problem set. (either in coding or computations.)

Reading (25pts.): .

Question 1: Reading Angrist and Pischke [10pts]

1.) Read:

- Reread Chapter 2 about the experimental ideal.
- Reread Chapter 3 "Making regression make sense."

2.) Answer:

- What is the Conditional Independence Assumption?
  - What do you think happens if it is not satisfied?
  - Provide an example (e.g. one coming up later) where it is not satisfied

- In which aspects does regression correspond to the experimental ideal in which does it not.
- What is matching? WhatŠs the relationship between matching and regression?

Question 2: Hansen [10pts]

Read chapters 8 and 9. For this class, you can read 9.4-9.6 superficially.
1.) Which chapters did we already cover in class?
2.) Answer Excercise 9.1

Question 3: Reading summary - Identification: [5pts]

- Formally, which of the OLS assumption that ensure unbiasedness is not satisfied, when there is an identification issue/an endogeneity problem?
- Name all threats to identification that we have mentioned in class.
- For each, briefly describe (use formulas or words) by which mechanism the assumptions are not met.

Theory (10pts.): .

Question 1: Sample Bias [4pts]

Feel free to refer to the last slide deck about Causality that we touched in class, and/or the Angrist Pischke book on samopling bias, when thinking about this excercise.
You are considering an advertisement campaign for beer at football events. To convince you of its effectiveness, the stadium owners provide you with the data on the amounts of beer and softdrinks sold at these events, and they give you the data of beer and soft drink sales at all publix stores in the same cities on the days following the matches. Now you are called upon to decide the effectiveness of the advertisements. These statistics show, that per person (shopper/visitor), the ratio of beer/softdrink consumption was 60% in the stadium and only 20% in the publix stores. Hence, they argue, the advertisement trippled the likelihood of a customer choosing beer.
*Note: These numbers are made up.

1. These estimates are corrupted by a specific bias, that stands in the way of identification. By which?
2. Would you expect the bias to work up- or downwards? i.e. do you think that the effectiveness of the advertisement is over- or underestimated?
3. Write the effect of advertisement in treatment notation and define treatment as seeing an advertisement for beer in the game, and no treatment as not seeing an advertisement in publix.

4. Next, decompose the observed difference into the unobserved treatment effect and the unobserved selection bias. – given your result, can you say anything about the treatment effect from the stadium/publix comparison? How? or Why not?

5. Assume you have to make a decide about the advertisements campaign's future today, and you are willing to make an assumption about how the beer/softdrinks consumption of a stadium visitor is going to differ from a publix consumer: Make such an assumption and discuss it. Can you now (given this assumption) derive the treatment effect?

6. If you can, derive it and discuss how it relates to the original observable statistics. How much does your estimate of the treatment effect depend on your assumption in the previous part? (i.e. what would you have gotten if you had assumed the difference to be half as large or zero?)

7. Can you think of a possible improvement of the initial estimation that compares Stadium and Publix beer purchases?

Question 2: Delta Method and Asymptotic Distribution: [6pts]

From the Hansen Book, answer Questions

- 5.4
- 5.6

Empirical Problems (25pts.) : .

Question E1: Omitted Variables [15pts]

1.) Generate a 5-dimensional multivariate normal (Y, X1, X2, Z1, Z2). 5347 observations, with bilateral correlations as follows:
Correlations:

	Y	x1	x2	Z1	Z2
Y	1.0	0.2	0.1	0.35	0.0
x1	0.2	1.0	0.0	0.4	0.0
x2	0.1	0.0	1.0	0.0	0.4
Z1	0.35	0.4	0.0	1.0	0.6
Z2	0.0	0.0	0.4	0.6	1.0

Add a vector of ones (or any other constant)

- a. Compute the sample covariances for all pairs and the sample variances.

2.) Univariate Regression:

- a. Use these to obtain the regression coefficient when regressing Y only on x1 (slide set from Sept. 14)

- b. Use standard regression commands to regress y only on x1. Compare the coefficients.
- c. The coefficients you obtained should be biased. Can you name the source of the bias?
- d. What is the theoretical bias of the regression coefficient? Provide the formula and compute the bias.

3.) Multivariate Regression on foot

- a. Next define your data (ones and all 4 X,Z columns as Matrix X)
- b. Compute X'X, and report it.
- c. Compute beta-hat, report the formula you used and the result.
- d. Now run the standard regression command
- e. Compare your own result to the result from the regression.
- f. Compare the results from the multivariate to the univariate regression.

4.) Finally: Compute the CEF for Y|x1 (discretize x1 as you see fit).

- a. Draw and report.
- b. When you compare points 2, 3 and 4, what do you observe?

P.S. Please try to set a "seed" at 420711 (cf. googling):

In what follows, the 'full' model considers x1 and x2, while the truncated model refers to a specification that only uses x1. Z1 and Z2 will be candidates for instrumental variables.

Question E2: IV for starters (E1 cont'd, 10pts)

- Now compare the estimates of the "truncated model" and the full model. - Why are the estimated coefficients different?
- Looking at the Var-Cov Matrix, how would you expect exclusion of Z_2 to affect the other coefficient estimates?, why?
- If one of the X Variables were endogenous with y (e.g. a simultaneity problem), could you use Z_2 as an instrument?
  - Consider X_1 and X_2 separately, if you cannot use Z_2 as instrument explain, why not, if you can, explain what the assumptions would be and explain the exclusion restriction.
  - Refer to the "IV-assumptions" in the Angrist book or on the slides for your argument.

Project 1 (20pts.): .

Take what we have seen to your project. Focus on one of your ideas, the one you will most likely take up.: Think about your estimation strategy, but also about the economic model.

- To fix ideas write down the estimation equation (a linear model) that you would estimate in a first look at the data. This will clarify your dependent variable, and the variables you plan to have available. (you can write it in formulas or in estimation command, whichever you prefer.)
- What is your unit of observation (Hint: each unit of observation will be one line of data in your dataset)? How many observations do you expect to get for your dataset? Do all x-variables vary at the same level of your units of observation (Hint: Teacher quality does not vary at the individual student level. Local GDP can be an important predictor of individual wages, but it varies at county or state-level)?
- Now, try to think of **all** the factors that matter for explaining dependent y-variable. Focus on the (currently) unobserved ones. Enlist them. Try to come up with at least one relevant but unobserved control variable. Discuss which ones you cannot observe, and whether that introduces a problem to your estimation strategy (under which conditions can the unobserved variables be omitted? What happens, if they cannot be omitted?)
- Now try to think of other potential sources of endogeneity. Are there any:
  - issues of simultaneity/equation-system?
  - selection biases? (where applicable, also think of selection into missingness (i.e. when sb would prefer not to give an answer or misreport, who would that be?)
  - likely measurement errors?
- Now think of an ideal experiment to show the causal relationship that you are interested in. Can you come up with one?
  - If yes, describe it focusing on how you would manipulate the X-variable of interest. Also verify whether the ideal experiment is robust to the 4 sources of endogeneity and clarify how.
  - If no, why not - in that case please revisit the sectionon FUQ and explain whether your question is fundamentally identified?
- What do you conclude about your project? In which ways can you already anticipate, that you will run into limitations? (Which is fine for the scope of this course, as long as you are aware of them!)
- Are there any steps you could take to improve your strategy?

If you do not like what you found when thinking about your project, feel free to think of your other ideas with the same framework and to (a.) switch project or (b.) modify your project, if you see solutions to your problem.

A final note: I greatly enjoyed reading about your ideas. While I am not giving you unsolicited feedback, I am happy to give you my two-cents on what I think will work better. Given the tight time, you will be better off by doing something easy where data are already available...

Project 2 Data Cleaning (15pts.): .

1. Attempt to obtain a (first) dataset for your project.
   - Try to get at least a proxy for the dependent variable and a proxy for the independent variable.
   - If you really cannot get any data yet, you may simulate data and use that to obtain 80% of the points.
2. Perform at least 5 of the steps in the slide set on data cleaning.
3. What does the distribution in you dependent variable like?
   - Does it look like standard normal?
   - Is it skewed?
   - What do you notice about outliers.
4. Repeat theses steps for at least 2 of your most relevant independent variables.
5. Report what you learned about your data, outliers etc.
6. Explain whether your findings require any corrections/adjustments to your data, or not (and why/why not)