ECON-7022MK PhD ECONOMETRICS I (Fall 2017)

Assignment 1 (Aug. 28th)
Due: September 09, 7.00pm.
Late Due: September 12, 7.00pm.

Instructions:

1. This Assignment IS MEANT TO BE A LOT OF WORK! I expect that most of you will have to work between 25-50 hours on it. Start ASAP (how about tonight?) and make sure you completed the reading assignments BEFORE Labour Day.
2. You may work in groups of up to 3 people for the assignment. Please **note the other members of your group on your write-up.**
3. 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.** (My suggestion is as follows: First attempt the problems alone. Then, meet in your group and work on the solution together. Then go home and do the clean write up on your own again...)
4. This Assignment is intended for making you go over the materials covered in class and make you think about them yourself. The questions should have you apply definitions yourself, to see how things connect. I will rigorously grade the reading assignments and the part about developing questions in view of the final paper, but will be less rigorous on your theoretical answers, accounting for the fact that class has just started and statistics is not our main focus.
5. I will reward the fact that you made an attempt, and the quality of your solution equally. Generally, please do not get stressed about this Assignment, but do make sure you read the readings once or even twice. The reading and research assignments, despite the weak verifiability, are more important for your success in class..
6. You should submit your answers in one zip file via T-Square by by Sept. 12th at 7pm. (You can get a 1 point positive credit for handing it in early on by Sunday, Sept. 9th by 7pm.) If you cannot upload to T-square, email, scan and email are acceptable, but T-Square is preferred. For this problem set there is only a 2 point discount if you hand in “late.” After that, both grading rules and discounts for belated submissions apply as stated on the syllabus. (a “day” refers to 24hrs, not dates)
7. The total score are 100pts, which I will translate to the corresponding percentage of your final grade (most likely 8%). Bonus Questions are intended to allow you to delve deeper into theory OR empirical. They are not required and might be quite time consuming. If you do both Bonus questions, I will count the one where you perform better.

I. Reading Assignments [40 pts.]: .

NOTE: The main part of this Assignment is reading. As said, I am consciously frontloading you with excess reading assignments. I will assign lower loads, when I expect the workload from the other courses to go up. I do not expect you to have full understanding for everything at this point, but I DO want you to see the materials for the first time NOW. Those who are not taking multiple classes, but only one or two courses may wish to delay some of the reading, just drop me a line.

First Readings of Hansen [20pts]

In Hansen, B. (2015). "Econometrics", Lecture notes University of Wisconsin. [HaL]

1. Chapters in Section 2
   - Focus on chapters 2.8 to 2.28
   - Skim over 2.1 to 2.7, they are a revision of the materials presented in class.
   - Feel free to skip or skim over the chapters with the stars and only read them, if you want to dig deeper. Especially so for the technical proofs.
   - TASK: write a short paragraph about which subchapter you thought was most difficult and explain what you struggled with (up to 3 subchapters)
2. Chapters in Section 3
   - Focus on Chapters 3.1-3.12
   - TASK as above: write a short paragraph about which subchapter you thought was especially difficult and what you struggled with (up to 2 subchapters)
3. It is useful to briefly read through the introduction before starting to read.
4. Outlook: Next week's assignment will be finishing chapter 3 and chapter 4
   - you might like to read ahead, before the workload in the other classes increases

First Readings of Angrist [20pts]

If you have not done so, make sure to get hold of the main book for part2, and familiarize with its structure and read the introduction: In Angrist, J. and S. Pischke (2009), Mostly Harmless Econometrics", Princeton University Press. [MHE]
Remember to give BRIEF ANSWERS of about 3 to 6 lines in length.

1. Read the Chapter about questions.
   - What is a "FUQ?" What is the problem with "FUQs" w.r.t econometric estimation?
   - Try to come up with an example "FUQ" (an example OF YOUR OWN) and explain why you think it is a "FUQ."
   - Try to think of an identified question explain how it differs from your previous example.

- NOTE: Keep these insights in mind for the Question assignment in part IV.

2. Browse through the chapter about regression analysis (ch3) superficially:

   - Report your impression: What are the main differences when comparing the style and the treatment in AP to Hansen?
   - Which of the books is more appealing to you and why?

II. Empirical (programming) Problems, [25pts]: .

Question 1: Coin Flips [10pts]

For this excercise, refer to the R-primer slides that I uploaded with this problem set. Reproduce the Coinflips analysis the R-primer.

a.) Write a source-files that performs all the steps for generating a random variable and generating samples. Carefully comment the steps you take in the file (i.e. write 1-3 short lines of explanation what each codeblock is doing)

b.) Comment on the following questions:

- Does the code or its output contain the analogon of a random variable? If yes which codebit or part of the output is it? Comment how you reach your conclusion.
- Does the code or its output contain the analogon of an outcome? If yes which codebit or part of the output is it? Comment how you reach your conclusion.
- Does the code or its output contain the analogon of a population/sample? If yes which codebit or part of the output is it? Comment how you reach your conclusion.

c.) Now simulate a random normal distribution with 5000 draws. Plot a histogram, calculate the sample mean and the sample variance.

Question 2: Do the 'Grades' analysis from the slides on a different dataset [15pts]

For this excercise, refer to the R-primer slides that I uploaded with this problem set. To open a dataset use the command "load", or the mouse-click menu of R-studio. Using the dataset dt_wages.RData:

a.) Write a single source-files that performs all the commands shown for analyzing the course grades, including the scatter plot with the linear fit and the conditional means for rounded values of the x-variable. Carefully comment on the steps you take in the file (i.e. write 1-3 short lines of explanation what each codeblock is doing). Instead of grades, we are now interested in the relationship of experience and wage. Do the analysis for the entire population.

b.) Try to run the analysis of (a) separately for men and women. Compare the slopes of the linear fit. (It's enough to visually interpret the slopes in the plot).

c.) Next, let's look whether there might be discrimination against unmarried people. First give a confidence interval for the earnings of unmarried people. Attempt to perform a hypothesis test that checks if unmarried have different earnings than married people.

NOTE: You cannot use all commands one for one, since some are outdated in newer versions. Most likely you will have to use online resources (like stackoverflow.com, or simply googling) to find the correct command, whenever the syntax on the slides is not exactly right.

Bonus Question: DIY: Simulate a multivariate distribution [8pts]

Finally attempt to simulate a bivariate random normal distribution with 12000 draws and Corr(X,Y) = 0.15.

a.) Calculate the sample means.

b.) Compute the sample Var-Cov Matrix (you can separately report variances and covariance).

c.) Compare the computed sample Var-Cov Matrix to the Var-Cov Matrix of (X,Y).

d.) Show a scatter-plot and fit a linear projection of Y on X.

Hint 1: Find a suitable Variance-Covariance Matrix for (X,Y) and use it in your code.
Hint 2: I strongly encourage using internet search to find out how to perform any specific trick, like generating a multivariate random variable.

III. Project: [35pts]

a.) To start thinking about your project, formulate 2-3 Questions that could potentially be answered by analyzing data.

b.) Once you have the questions, for each of them think about the following:

- Why do "we" care about the question? ("Why do you care?", is good for now)
- How would you answer the question using data?
- Specifically, which type of data do you think you would need to answer the question?
- Where/how would you try to get these data?
- Assume you get the data you want, think of your question in terms of a conditional expectation of Y given X. What would be your Y variables and what your X?
- Assume you could get the data and run a simple analysis of Covariance(Y,X) or the Conditional Expectation of Y given X. Assume further you found a relationship, suggest how to check whether the relationship is systematic ("statistically significant") or an accidental and non-generalizeable result of your sample. (base your thinking on the part of the slide set we have not yet covered.)
- Assume the relationship you uncovered is systematic: would you be willing to take "your" relationship at face value, or do you foresee any additional checks and balances that might become necessary?

Provide a short answer (ideally between 2-6 lines) to answer each point.

c.) Based on your considerations for both projects, which would you think is more promising, and give short reasons?

d.) All that said, briefly say which questions is more interesting/appealing to you?

e.) When you are done with all that, team up with one of your colleagues. Tell them your questions/research ideas, and ask them what they consider more promising. Report briefly who you talked to and what they considered more promising (and why)?.