Econometrics 2, Winter/Spring 2020
Final Exam

This is an individual exam, you are not supposed to use any outside help, but you are encouraged to consult the lecture slides and the recommended textbooks. You have 24 hours to answer all the questions. There are a total of 100 points. Try to keep verbal answers short. Submit a .pdf and a .do with solutions for simulations in Stata. The .do script should run out of the box without extra tweaking.

Single yes/no answers will not count; every answer should be accompanied with a short explanation.
Good Luck!

Question 1. (35 points)
In 2008, the Andalasian Ministry of Education (AME) subsidized the purchase of home computers. The program awarded approximately 10,000 vouchers worth of EUR 200 in 2008 toward the purchase of a personal computer for low-income students enrolled in Andalasian's public schools. The vouchers were allocated based on a simple ranking of family incomes. The income variable used is the monthly household income per family and families with income below 50 (Andalasian Currency) were eligible for computer vouchers. In 2018, AME wanted to understand short and long run effects of the policy on student outcomes.

1. How should they test whether computer access affect students test scores?

2. Write down the regression model. What type of regression design is this and why?

Unfortunately, some of the student records were lost. Hence, AME asked all public schools of Andalasia to send a list of all students that were enrolled in 2008. The schools were also asked to provide student test scores for years 2007, 2008, and 2009.

To get the family income of students, at the time of policy implementation, AME decided to conduct surveys. They contacted all individuals enrolled in 2008 about their family income in 2008.

3. Can we estimate longer run outcomes with the information they plan to collect? What data would you need for this?

4. Can the test scores of year 2007 be of any valid use?

5. What would be the biggest issue with survey data (especially recall data)?

6. Assume that people report their income with error and the measurement error in the running variable is additive and uniformly distributed with mean 1 and standard deviation of 2. Use the knowledge you acquired in Econometrics 2 to convince AME that there is an issue with this approach. For this you are going to use RD-data.dta. Create a do-file named RD.do simulate the measurement error, set the seed equal to 1234567, and add it to the running variable.

   (a) Run the regression you would run if the variable is not measured with error (use the variables already in the dataset). Plot the data to check for potential discontinuities. Estimate the causal effect of the policy on test-scores.

   (b) Now, run the regression in case the running variable is measured with error. What do you conclude?

   (c) Bonus 5 points: What happens when the measurement error is not classical?

Question 2. (25 points) True, false or no definite answer? Provide an explanation, a single yes/no/maybe answer won't count.

1. In order for propensity score matching to work well, the propensity score model should predict treatment with near certainty.

2. You try logit and the linear probability model (LPM) to estimate the relationship between Y and X. The coefficient on X obtained via logit is visibly different from that coming from the LPM. The logit estimate cannot be trusted as it relies on more assumptions than LPM does.

3. When sample selection is absent in the data (that is, $Y_i > 0$ for all i in the standard tobit), tobit and OLS produce identical estimates.

4. There are many ways to address the omitted variable bias. However, in practice, the most fruitful method is including more controls.

Question 3. (25 points)
You would like to estimate the effect of X_{it} on Y_i in a panel model with two alternating individual effects:

$$Y_{it} = \beta X_{it} + u_i + w_{it}, \text{ if } t \text{ is odd},$$
$$Y_{it} = \beta X_{it} + v_i + w_{it}, \text{ if } t \text{ is even}.$$

The individual effects (u_i, v_i) are likely to correlate with one another and with X_{it}. X_{it} may have strong serial correlation.

1. Would using pooled OLS be a good idea? Why or why not?

2. Propose a transformation that removes both u_i and v_i from the equations (Hint: use FE estimator as an inspiration). Under what assumption can you use pooled OLS to estimate β in the transformed equation?

3. What kind of standard errors would you use? Suppose nothing is significant if you use errors clustered at the level of i. At the same time, heteroskedasticity-robust errors (a.k.a. White errors) produce very tight confidence bounds. Is it safe to use the latter errors?

4. How about a middle ground — clustering at the level of (i, ODD_t), where ODD_t is the dummy for t being odd? This way, observations with odd and even values of t, but the same index i, are assigned to two separate clusters. Is this a good idea or not?

Question 4. (15 points)
A group of economists are estimating the following regression:

$$GAP_i = \beta_0 + \beta_1 VOTE_i + \beta_2 S_j + \beta_3 VOTE_i \times S_i + u_i$$

where GAP_i is a pay gap between Roma and non-Roma in municipality i, $VOTE_i$ is the Jobbik's (a far right Hungarian party) vote share in the same municipality, and S_i is the employment share of small firms in i. $VOTE_i$ measures the degree of anti-Roma sentiment.

The main hypothesis is that racial discrimination is harder to pull off in big firms due to reputational issues and higher visibility. The coefficient of interest is β_3.

1. Would you expect $VOTE_i$ to be endogenous to the outcome? What's the most plausible story why Jobbik vote and u_i, the unobservable factors affecting Roma pay gap, may correlate.

2. The authors propose the pre-2009 share of foreign currency loans in the municipality ($LOANS_i$) as an instrument for the vote share of Jobbik. Previous studies found a link between these variables: people whose loans got out of control due to forint depreciation were disproportionately attracted to Jobbik. The interaction $VOTE_i \times S_i$ is instrumented with $LOANS_i \times S_i$

   Suppose this instrument is exogenous and relevant. Given what you've learned about LATE and IV in the presence of heterogeneous treatment effects, what do you think about this strategy? Intuitively, would you expect the estimates of β_1 and β_3 to be too high or too low compared to the respective mean effects in the population?