NOVA SCHOOL OF BUSINESS & ECONOMICS

STUDENT NUMBER: _____

Microeconometrics – Final exam (May 29th, 2024)

Exam text:

1. **[25%]** A researcher is interested in the effect of attending a private high school on students' standardized test scores. Data include information on students' test scores, whether they attended a private high school, and several control variables including socioeconomic status, parents' education, and gender. The following model is proposed:

$$Y_i = \alpha + \delta P_i + \mathbf{X_i}\beta + \epsilon_i \tag{1}$$

where Y_i is the test score, P_i is a dummy variable equal to 1 if the student attended a private high school and 0 otherwise, X_i is a set of control variables, and ϵ_i is the error term.

   (a) [1] Discuss the potential endogeneity problem in estimating equation (1) with OLS.
   SOLUTION:
   - [1] P_i is correlated with the error term, therefore δ is biased.
   - [extra] Endogeneity could arise for different reasons. One reason could be omitted variable, with more connected families willing to have children in more selected (private) schools. Could also be reverse causality, with private schools accepting better students, and therefore improving the quality of classes, leading to improved learning.

   (b) [2] Suppose you have access to an instrumental variable (IV) for private high school attendance. Describe the characteristics of a good instrument in this context and provide an example of a potential IV.
   SOLUTION:
   - [1] Instrument needs to be relevant (correlated with attending private school) and respect the exclusion restriction (not being correlated with ϵ).
   - [1] Examples are open, as soon as the conditions are reasonably respected.

   (c) [2] Assume you have access to an experimental estimate of δ in equation (1), that is, someone ran an randomized experiment in a comparable sample randomly allocating students to private high schools and obtained δ by estimating equation (1) with OLS. The estimate is presented in Column (1) of Table (1). In column (2), you find an IV estimate of equation (1) from a random sample of students in the country (important: the two samples are different, but are draw n from the same population). Assume that the IV is a valid IV. Can you compare these two estimates?
   SOLUTION:
   - [2] It depends. Column (1) identifies an ATE (the average effect of attending private schools in the population of interest). Column (2) identifies ATE only if we assume homogeneity (i.e., each individual gains from attending private school in the exactly the same way). If there is heterogeneity, then IV identifies a LATE (the effect of attending private schools among compliers). The latter depends on how the IV is defined, meaning that every instruments has a group whose estimate refers to.
   - [extra] No need to discuss the consequences of working with an invalid IV, since we assume the IV is valid.

Table 1: Effect of attending a private high school

	Dep. var.: standardized test score	
	(1)	(2)
P_i	0.45***	0.83***
	(0.17)	(0.29)
Mean dependent variable	6.41	6.39
Observations	2,341	5,629

Note. Standard errors in parentheses. Standardized test score is a test score ranging from 0 to 10 in a nation-wide standard examination. Estimate in column (1) is based on the experimental sample. Estimate in column (2) is based on IV. * indicates p-value < 0.1, ** indicates p-value < 0.05, *** indicates p-value < 0.01.

2. **[25%]** Consider the problem of estimating the effect of a job training program on participants' earnings. You have access to panel data for a sample of individuals observed for three years before and three years after the training program, ($t = 1, .., 6$ with the program happening at time $t = 3.5$). Information consists of demographics and economics outcomes (including earnings) for each individual, and an indicator variable for whether an individual attended the training or not.

   (a) [1.5] Explain how you would use a Fixed Effects (FE) method to estimate the effect of attending the training, specifying the estimating equation, and the assumptions needed to identify the parameter of interest.
   SOLUTION:

   - [0.5] You can run the following regression:

   $$Y_{i,t} = \beta d_{i,t} + X_{i,t}\lambda + \gamma_i + \theta_t + \epsilon_{i,t} \quad (2)$$

   where γ_i is individual FE, and θ_t are dummies for time effects (one for each period). $d_{i,t}$ is an indicator equal to 1 if at time t the individual i has received the training, and 0 otherwise.
   - [1] Assumptions needed: a) FE version of orthogonality; b) FE version of rank conditions. These are obtained by subtracting the average over time of $Y_{i,t}$ from both the equation.
   - [extra] Discuss in detail the assumptions and the validity of these in this setting.

   (b) [1] You decide to also include some control variables in your FE model. In this approach, what constrains you in the choice of controls variables?
   SOLUTION:

   - [1] FE allows identifying only time-varying controls. Controls that are fixed over time are absorbed by the individual FE, and thus cannot be included in a FE model.

   (c) [1.5] Imagine you have access only to the following information: a table showing the average earnings for those who participated in training and those who did not. Can you can use a Difference-in-Differences (DiD) approach to estimate the effect of the job training program on earnings. Specify the assumptions needed, and their credibility in this setting.
   SOLUTION:

   - [0.5] Simply apply the DiD estimator: take first differences (post-pre for each group), then subtract the first difference in the bottom group to the first difference in the top group (the one that received the training).
   - [1] Discuss assumptions (similar to the FE assumptions), but specific to DiD.

   (d) [1] In this setting, what is the difference between using a FE method as in point a) versus the simple DiD estimate in point c)?
   SOLUTION:

Table 2: Average Earnings by group and period

	Period	
	$t = 1, 2, 3$	$t = 4, 5, 6$
Individual who participated in training	\$25,000	\$30,000
Individual who did not participate in training	\$24,000	\$25,000

- [1] FE method allows using controls, DiD in the second point does not.
- [extra] Adding controls might help with identification (assumptions are always conditional on the controls we add). Example, say you observe non-parallel trends when you do not add controls, but a control can capture differences in trends. Then controlling for this variable would solve the problem.
- [extra] Adding controls might help with precision, by absorbing residual variation.

3. **[25%]** A study investigates the impact of environmental regulations on manufacturing plant productivity. The key variable of interest is a standard on emission of pollutants. This is a (continuous) level of emissions that is considered acceptable, if a firm decides to emit more than the standard has to pay a fixed tax T. Each state has adopted a different standard, meaning that some states are more strict about allowing plants to emit, while others are less strict. The standard of emissions is given by the variable S_k, where k indicates the state where the plant is located. You have access to a random sample of manufacturing plants in the country, interviewed 3 years after the introduction of the standard. Information includes productivity ($Y_{i,k}$), emissions ($E_{i,k}$), and other plant characteristics.

   (a) [1.5] You are interested in comparing the average productivity of plants in high versus low standard states. You split states in two groups depending on whether S_k is above or below the median value in the country. Define this variable $HIGH_k$, equal to 1 if the standard is above median and 0 if the standard is below or equal to the median. You estimate the following using OLS:

   $$Y_{i,k} = \alpha + \beta \cdot HIGH_k + \epsilon_{i,k} \tag{3}$$

   where $\epsilon_{i,k}$ is an error term. The estimated β is provided in column (1) in Table 3. How do you interpret the parameter β?
   SOLUTION:
   - [0.5] β is the derivative of the expected value of Y (conditional on controls) with respect to $HIGH$. In OLS, this derivative is a constant. When HIGH goes from 0 to 1, then the productivity increases by 0.08 (or 8% because the dependent variable is in logs).
   - [0.5] Because $HIGH$ is a dummy variable, β can be interpreted as the difference in means of Y between $HIGH = 1$ and $HIGH = 0$.
   - [0.5] Because $HIGH$ is possibly not allocating at random, then we cannot interpret this as a causal effect. The difference can be interpreted as ATT + selection bias. ATT is the causal effect of high standards among those states that have high standards. Selection bias characterizes the difference in potential outcomes (how ex-ante plants would be productive in presence or absence of high standards) between plants in states with $HIGH = 1$ and plants in states with $HIGH = 0$.
   - [extra] Discuss what happens when $HIGH$ is allocated at random. Then selection bias is zero and ATT = ATE.

   (b) [1.5] You are interested in estimating the causal effect of introducing a tax for emissions on $Y_{i,k}$. In the previous 3 years, all plants that generated emissions higher than the standard did pay the tax, while the ones that generated emissions below did not pay any tax. Explain how you would use a Regression Discontinuity (RD) design to estimate this effect. Specify the RD estimating equation and describe how you would implement the analysis based on the information that is available.
   SOLUTION:

- [1] S_k can be used as running variable, meaning that there is a discontinuity in the allocation of the tax at S_k. Whoever produces emissions beyond that point pays a tax, whoever produces emissions below that point does not pay the tax. Exploiting this discontinuity, I can compare the productivity of those plants that produce emissions $E_{i,k}$ just below and just above S_k and estimate the effect.
- [0.5] There is no further information, so in order to apply the method, there are different alternatives. This is an open question. Discuss how in practice you would apply this approach choosing one or more methods (parametric or non-parametric) discussed in class.

(c) [1] In column (2) of Table 3, you have a RD estimate. How do you interpret it?
SOLUTION:

- [0.5] This is a local average treatment effect (LATE). It is the average effect of introducing a tax, but only for those plants that produced emissions at the level of the discontinuity, i.e. $E_{i,k} = S_k$.
- [0.5] Introducing the tax decreases the productivity by 2.5% among these plants. I cannot learn about the effect for the other plants (those emitting far away from S_k).
- [extra] I cannot compare estimate in column (1) with the estimate in column (2) because they measure different parameters. Column (1) is ATT + selection bias, column (2) is LATE.

(d) [1] In the RD design, do you foresee a problem if there would have been a mistake in the implementation of the program and you would observe 5% of plants having emitted less than the standard to have paid the tax by mistake?
SOLUTION:

- [1] In this case, the RD design is not sharp, in the sense that I do not observe a sharp discontinuity in the allocation of the tax (i.e., I have observations paying the tax in both sides of the discontinuity). Therefore I cannot just compare plants emitting just below the discontinuity with plants emitting just above the discontinuity.
- [extra] I would have to use a fuzzy RD design, which works pretty much like an IV.

Table 3: Fixed Effects Estimation Results

	Dependent variable: Plant-level productivity	
	(1)	(2)
$HIGH_k$	0.080**	
	(0.035)	
RD estimate		-0. 025*
		(0.013)
Estimation method	OLS	RD
Observations	1949	1949

Note. Productivity is plant-level productivity, defined as the logarithm of the total output produced divided by total inputs into production. Standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Column (1) is estimated using equation (3). Column (2) is an RD estimate of the effect of introducing the tax on emissions on productivity.

4. **[25%]** You are investigating the determinants of entrepreneurship among college graduates in business and economics degrees. You have access to a sample of college graduates that was interviewed 15 years after graduation. Information includes: an indicator variable E_i equal to 1 of the respondent started at least one own business since graduation, and 0 otherwise; $invest_i$ equal to the amount invested in own businesses since graduation; a set of individual characteristics, including performance indicators during college.

(a) [1.5] You are interested in whether the average grade in quantitative subjects during college ($quant_i$) influences entrepreneurship later on in life. Assume all respondents took at least one quantitative subject

($quant_i$ is never missing). You estimate the following specification with OLS:

$$E_i = \beta_0 + \beta_1 quant_i + X_i\gamma + \epsilon_i \tag{4}$$

Beyond the fact that $quant_i$ is not random, and thus orthogonality is not valid, what are the pros and cons of following this approach in this setting where the dependent variable is a dummy variable?

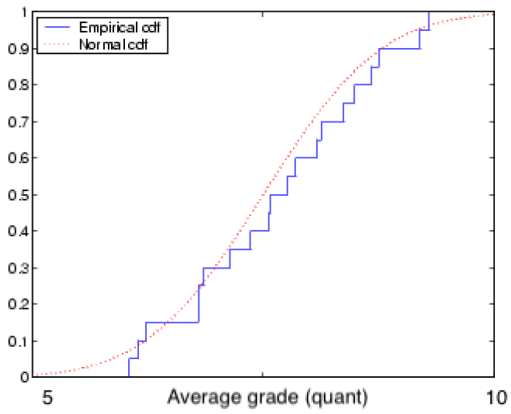
SOLUTION:

- [0.75] Pro: interpretability of the coefficients (i.e., β_1 is the average partial effect of the conditional mean of E with respect to $quant$).
- [0.75] Cons: not flexible, meaning I do not have the feature of a model predicting a dummy outcome variable (decreasing returns approaching 0 and 1 and not predicting values below 0 or above 1).

(b) [1.5] The empirical cumulative distribution (CDF) of $quant_i$ is presented in this graph. Together the figure also presents the CDF of a normal distribution using the sample average and the sample standard deviation. Based on this figure, explain how you would use a probit or a logit model to estimate the average partial effect of $quant_i$ on E_i?

SOLUTION:

- [1.5] The CDF has very tiny tails, it is therefore possible that probit would be preferable, and the logit would most possibly return the same estimates.
- [extra] Notice also that the CDF in the central part looks linear, so most probably the OLS regression will return very similar APE to probit/logit models.



(c) [2] You are now interested in estimating the relationship between $quant_i$ and $invest_i$. After carefully reviewing your data, you realize that $invest_i$ is missing if $E_i = 0$, meaning that information is filled only for those that have started at least one business. What can you learn from β_1 if you estimate the following regression with OLS in the sample with non-missing data on $invest_i$?

$$invest_i = \beta_0 + \beta_1 quant_i + X_i\gamma + \epsilon_i \tag{5}$$

SOLUTION:

- [2] β_1 describes the relationship between $quant$ and $invest$, but limitedly to the population without missing values (those with entrepreneural experience). These are most probably very selected as compared to a more general population.
- [extra] Discuss the fact that this estimate is non-causal, because of lack of orthogonality.

- [extra] Discuss what you can recover if you assume that missing data is actually equal to 0, and thus censored at 0, meaning you can apply Tobit to recover derivatives for a broader population of interest.
- [extra] Discuss the use of selection models to recover derivatives for a broader population of interest, without assuming that missing data can be set to 0.