NOVA SCHOOL OF BUSINESS & ECONOMICS

STUDENT NUMBER: ____________________

Microeconometrics – Final exam (May 30th, 2025)

Exam text:

1. **[25%]** You are evaluating the impact of broadband internet access on the annual revenues in a sample of 2,100 firms interviewed in year t in your country. The model is:

$$Y_i = \alpha + \delta D_i + X_i\beta + \varepsilon_i$$

where Y_i is log annual revenue, D_i is a dummy for broadband access, and X_i includes sector, firm size, and age.

(a) Explain why D_i may be endogenous in this setting and invalidate the OLS assumptions.

SOLUTION:

- Broadband access (D_i) may correlate with unobservables like management quality.
- If $\mathbb{E}[\varepsilon_i|D_i] \neq 0$ (i.e., the orthogonality assumption is not valid), then OLS estimates are biased.

(b) Assume distance to the nearest fiber optic cable is used as an instrument. Discuss the assumptions needed for identification and their validity.

SOLUTION:

- *Relevance*: distance to fiber affects likelihood of broadband access $\Rightarrow$ valid by definition.
- *Exclusion restriction*: distance affects revenues only through broadband access $\Rightarrow$ reasonable to assume that distance is not correlated with unobservable determinants of Y_i, but there can be reasons that invalidate this assumption (e.g., firms with better management quality position closer to infrastructure).

(c) Table 1 reports estimates using these two methods: OLS (column 1) and IV (column 2). In addition, in column 3, you have access to an experimental estimate from the following experiment: two years before the year t, a sub-set of 750 firms participated in a randomized experiment, in which the provider randomly connected half of these firms to broadband internet, while the other half was cut out. What do you learn about the validity of assumptions from the estimates in the table?

Table 1: Impact of Broadband Access on Revenues

	(1)	(2)	(3)
	OLS	IV	Experimental
Broadband Access	0.301***	0.250***	0.120***
	(0.055)	(0.070)	(0.035)
Mean Dependent Variable	10.25	10.25	10.18
Observations	2,100	2,100	750

Note. Standard errors in parentheses. * indicates p-value < 0.1, ** indicates p-value < 0.05, *** indicates p-value < 0.01.

SOLUTION:

- OLS (0.301) > IV (0.250) > RCT (0.120): suggests (selection) bias in OLS that is not addressed by IV (i.e., the instrument is possibly not a good instrument).
- Experimental estimate likely unbiased, and samples seems comparable (averages are similar) $\Rightarrow$ differences highlight limitations of OLS and IV.

(d) The coefficient δ captures the Average Treatment on the Treated (the effect of internet access on those that have access to the internet) plus a selection bias. How does the coefficients in column (2) and (3) differ with this interpretation?

SOLUTION:

- OLS captures ATT + selection bias.
- IV would capture ATE (in case of homegeneity) or LATE (in case of heterogeneity) if assumptions are valid, but in this case it is likely also biased.
- Experimental estimate captures the ATE, which in this case is also equal to ATT.

2. **[25%]** You want to study how receiving a subsidy for training impact the earnings of unemployed. You observe a random sample of individuals that were unemployed in 2018 and that have been interviewed yearly from 2014 to 2021. Your data includes individual characteristics, an indicator variable whether the person collected the training subsidy at time t ($T_{i,t}$), and yearly earnings ($Y_{i,t}$).

(a) Propose a Fixed Effects specification to estimate the causal impact of subsidy on earnings. Specify assumptions.

SOLUTION:

- FE specification:

$$Y_{it} = \alpha_i + \lambda_t + \delta T_{it} + \varepsilon_{it}$$

- Assumptions: strict exogeneity + rank condition in the transformed variable.

(b) You find out that only after 2020 training subsidies became available, while before they were not available. Table 2 shows the average earnings before and after the introduction of subsidies among those who received the subsidy and those who did not in different periods. You use the values in the periods 2 years before and 2 years after the introduction of the subsidy to estimate the effect of the subsidy using difference-in-differences and obtain an estimate of 200. Replicating this estimate using OLS confirms that the estimate is statistically significant at the 99% of confidence. Can you conclude that this estimate is the causal effect (an ATT) of the training subsidy?

Table 2: Average Monthly Earnings (in EUR)

	2014–2015	2016-2017	Pre (2018–19)	Post (2020–21)
Received Subsidy	950	1,000	1,050	1,350
No Subsidy	1,300	1,200	1,100	1,200

SOLUTION:

- Pre-trend differences suggest potential violation of the parallel trends assumption ⇒ the DiD estimate is possibly not capturing the ATT (caution in causal interpretation), but instead an ATT + Δ trends.

3. **[25%]** A government offers grants to be spent on innovation to firms with fewer than 25 employees and no grant for those with 25 employees or more. You have firm-level data on number of employees (E_i) and on investment levels (Y_i).

(a) The grant program is deployed with perfect compliance (no firm with 25 or more employees ended up obtaining a grant). Can you exploit a sharp of fuzzy RD design? Explain.

SOLUTION:

- Sharp RD (perfect compliance at 25-employee cutoff) ⇒ by definition, the probability of being offered a grant switches from 1 to 0 when the number of employees $z = 25$ (i.e., the discontinuity in the running variable).

(b) What are the identification assumptions in this design?

SOLUTION:

- Discontinuity: grant allocation is a function of the number of employees z discontinuous at $z* = 25$.
- Smoothness: potential outcomes are continuous at $z* = 25$ ($E[\alpha|z]$ and $E[\beta|z]$ are continuous at $z* = 25$).
- Local randomization: the effect of grants α_i is independent from the grant allocation in the neighbourhood of $z* = 25$.

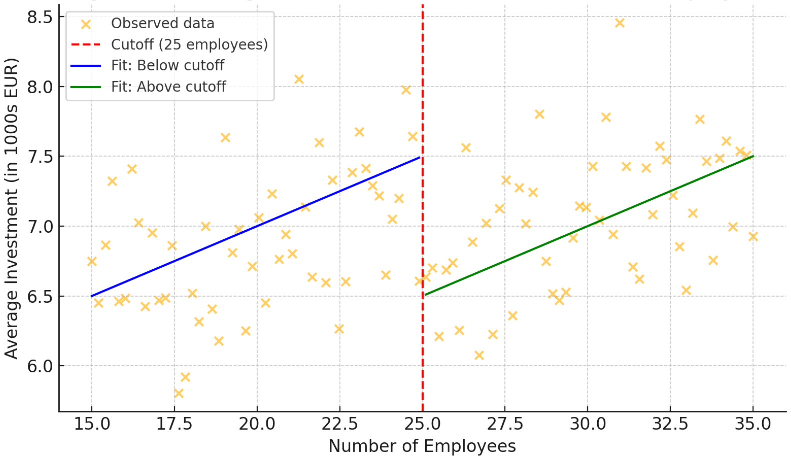
(c) The figure below shows average investment by firm size estimated using the following OLS model $Y_i = \alpha + \beta(E_i - 25) + \gamma D25_i + \delta(E_i - 25) \cdot D25_i + \epsilon_i$, where $D25_i$ is a dummy variable equal to 1 if the number of employees is smaller than 25. What parameter identifies the RD estimate and how you can interpret it?

SOLUTION:

- LATE: the effect of being offered a grant for firms with 25 employees.

(d) What can you conclude about the effect of the grant program if you have access only to the figure?

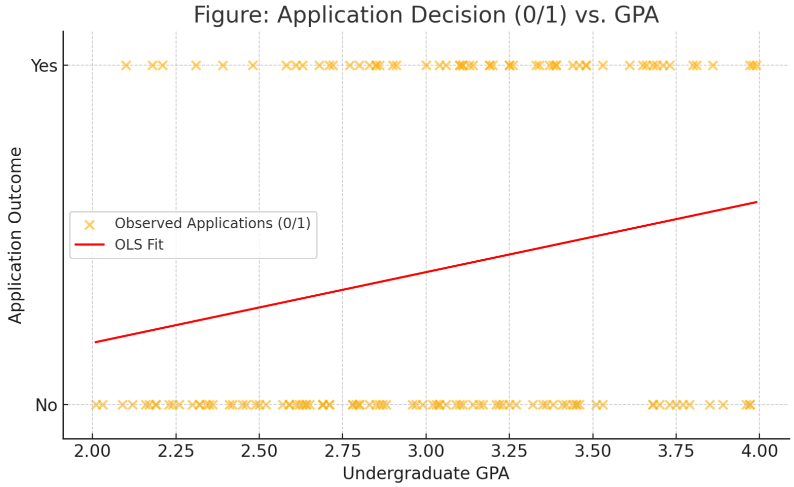
Figure: RD Design - Firm Investment vs. Number of Employees



SOLUTION:

- Visual jump at cutoff suggests positive effect of being offered a grant.
- Cannot conclude there is an effect because there is no information about standard errors. Extra: given the dispersion of points, it looks like the standard errors will be wide, and therefore the confidence intervals in both sides of the discontinuity likely overlap, meaning the impact estimate is not significant.

4. **[25%]** You study how undergraduate GPA affects the probability of applying to a Master's program (indicated by $Y_i = 1$ if applied to a program and $Y_i = 0$ if not). You use a sample of 10,000 undergrad students from all universities in the country. The following figure shows a scatter plot with GPA on the x-axis and Y_i on the y-axis, and a linear fit estimated with the following OLS specification: $Y_i = \alpha + \beta \cdot GPA_i + \epsilon_i$.



(a) What are the limitations of using OLS in this setting?

SOLUTION:

- Predictions of the model can be outside [0,1].
- Does not capture the decreasing returns when the predicted outcome approaches 0 or 1.
- Extra: requires assuming heteroskedasticity due to the nature of the dependent variable (Bernouilli).

(b) You decide to use a Probit or a Logit model to estimate the relationship between Y_i and GPA_i. Judging solely from the figure, would you use a Probit or Logit model. Explain.

SOLUTION:

- Choose based on tail behavior $\Rightarrow$ use Probit for thinner tails, Logit for fatter tails.
- Visual inspection indicates that points are distributed widely in terms of GPA for both $Y = 0$ and $Y = 1$ (i.e., for both values, the observations overlap in terms of distribution and are not concentrated in specific ranges of the distribution of GPA), indicating that there won't be a need to model fatter tails.

(c) You find out that data on GPA_i is never missing, but data on Y_i was obtained from an online survey sent to students, and only 25% replied to survey. Luckily, you also have access to a number of individual characteristics that were collected at enrolment, and therefore are never missing in the dataset. What problem do you face and how would you address it?

SOLUTION:

- A response rate of 25% does not seem like random, meaning that respondents decided to not to respond and are potentially different from the other 75% in both observable and unobservable characteristics $\Rightarrow$ potential incidental truncation (i.e., sample selection biases estimates by focusing only on selected observations).
- Because we observe all other variables for the full sample, we can apply the Heckman correction (if remaining assumptions are valid).