STUDENT NUMBER: _____

Microeconometrics – Final exam (May 23rd, 2022)

Exam text:

1. **[25%]** You are asked to estimate the causal effect of marriage on earnings using individual-level data for a representative sample of the population. We are interested in the following equation:

$$Y_i = \alpha + \delta M_i + \mathbf{X_i}\beta + A_i + \gamma_i + \epsilon_i \tag{1}$$

where Y_i is the wage, M_i is the marital status at the time of the interview (dummy variable equal to 1 if the individual i is married, and 0 otherwise), X_i is a set of control variables, which include ethnicity and gender of individual i, job tenure, and duration of marriage in years at the time of the interview. The variables A_i and γ_i are the (fixed) cognitive ability and non-cognitive ability of individual i, both are unobserved. ϵ_i captures remaining unobserved determinants of wages.

   (a) **[1.25 points]** Explain why you cannot estimate equation (1) using OLS if you are interested in the causal effect of marriage.
   SOLUTION:
   - Discussion about the endogeneity of M with respect to unobservable determinants of wage (0.75 points).
   - Detailed formulas for the orthogonality conditions (0.5 points).

   (b) **[1.5 points]** For the same sample of individuals, you obtain panel data, such that individual i is observed for four periods $t = 1, .., 4$. In Table 1 you estimate the panel version of equation (1) using in column (1) a random effect model, and in columns (2)–(4) a fixed effects model using alternative sets of control variables. Explains the main differences in terms of assumptions of these two methodologies.

Table 1: Estimated wage regressions

	Dependent variable: wage (in logs)			
	RE	FE	FE	FE
	(1)	(2)	(3)	(4)
Married	0.083***	0.056**	0.051**	0.033
	(0.022)	(0.026)	(0.026)	(0.028)
Job tenure	No	No	Yes	Yes
Quadratics in years married	No	No	No	Yes

Note. Standard errors in parenthesis. * significant at 10 % ** significant at 5 % *** significant at 1 %. *Quadratics in years married* introduces as controls the years of marriage and its squared term.

   SOLUTION:
   - Detailed assumptions of RE model (0.5 points).
   - Detailed assumptions of FE model (0.5 points).
   - Technical presentation of assumptions (0.5 points).

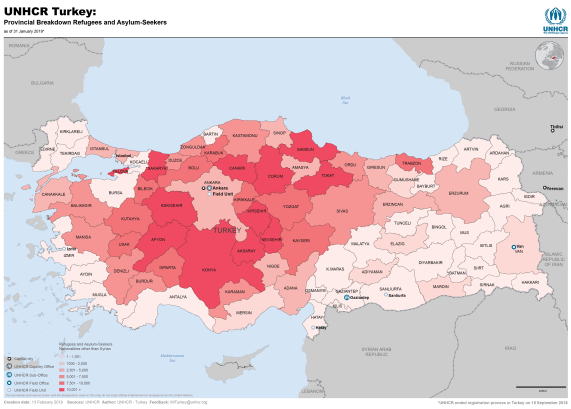
   (c) **[1.25 points]** Focus on columns (2) to (4). Why the coefficient decreases when you control for job tenure and for the years of marriage and its squared term?
   SOLUTION:
   - Conditional on individual (time-invariant) fixed effects, married is correlated with these variables (especially with years of marriage and its squared term). [0.75 points]

- Excluding them lead to the variable married to be correlated with the idiosynchratic error term ϵ, violating FE assumptions and leading to estimate in (2) to include a positive bias (0.5 points).
- EXTRA: comparison with (1) indicates the importance to control for individual (time-invariant) fixed effects (0.5 points).

(d) **[1 point]** From FE estimates, can you learn something about A_i and γ_i separately?
SOLUTION:

- The model we studied use only one individual FE, because they would all cancel out – we can therefore study only $\Omega_i = A_i + \gamma_i$ but we cannot distinguish them separately (1 point).
- EXTRA: Show how to estimate mean and variance of Ω_i as seen in class (0.5 points).

2. **[25%]** You are interested in estimating the effect of a large refugee influx that happened in 2018 in Turkey on the share of votes for party P in the last elections in Turkey. You are given the following map, which indicates the distribution of these refugees across provinces. You gain access to only a partial information from the map: a dummy variable R_i, which is equal to 1 if in January 2019 the province i received more than 5000 refugees, and 0 otherwise. In addition, you obtain electoral outcomes at provincial level for three rounds of elections (2011, 2016, and 2021), and some additional characteristics of the provinces in the same years (2011, 2016 and 2021), such as population size, public expenditure, and crime rate.



(a) **[2 points]** How can you use a Difference-in-Differences (DiD) approach to estimate the effect of refugee influx on the vote share for party P? Specify the estimating equation, the assumptions needed and their credibility in this setting.
SOLUTION:

- Set dummy variable as $post_t$ equal to 1 for observations in 2021, and 0 for observations in 2011 and 2016. Then estimate OLS with dependent variable P_i on R_i, $Post_t$ and the interaction $R_i \times Post_t$. The coefficient on the interaction captures the DiD estimate (1 point).
- Discuss assumptions about error decomposition and how to cancel them out: common trends, lack of idiosynchratic shocks (0.5 points).
- Discuss possibilities and motivate (0.5 points). Example: it looks like $R_i = 1$ is concentrated in the western part of the country, hard to assume common trends. However, in order to evaluate we need data to check common trends in voting pattern before the refugee influx.

(b) **[1.5 points]** Column (1) in the following table shows estimates of the following model:

$$P_{it} = \alpha + \beta R_i + \mathbf{X_{it}}\gamma + \mu_t + \epsilon_{it} \tag{2}$$

where μ_t are time fixed effects (two dummies for year 2016 and 2021). How can you interpret the coefficient in column (1).
SOLUTION:

- β in column (1) captures the difference in means (conditional on controls) in the vote share between provinces with $R_i = 1$ and $R_i = 0$, i.e. once we control for aggregate trends in vote shares, provinces with high level of refugee influx in 2019 have 7.5 percentage points higher vote share for party P (1.5 points).
- EXTRA: Clearly not a causal effect, but captures the effect of R plus bias driven by correlation between R and ϵ (0.5 points).

(c) **[1.5 points]** How would you interpret the DiD estimate in column (2) and why do you think it is larger than in (1)?

Table 2: Effect of refugee influx

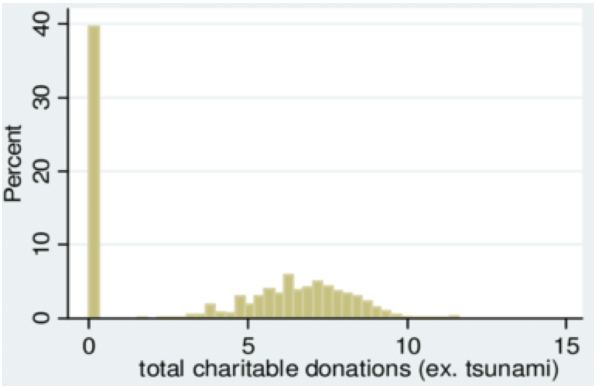
	Dependent variable: vote share for party ($P_i t$)	
	OLS	DiD
	(1)	(2)
R_i	0.075***	0.152***
	(0.022)	(0.046)
Controls	Yes	Yes

Note. Standard errors in parenthesis. * significant at 10 % ** significant at 5 % *** significant at 1 %. Observations are at province level and cover all periods (2011, 2016 and 2021).

SOLUTION:

- Assuming DiD assumptions are correct: DiD = ATT, i.e. effect of high refugee influx among the provinces with high refugee influx (0.5 points).
- Assuming DiD assumptions are NOT correct: DiD = ATT + difference in trends (+ difference in idiosynchratic shocks) (0.25 points).
- Technical presentation (0.25 points).
- Why is it larger (0.5 points).

3. **[25%]** Following the 2004 Indian Ocean Tsunami, an international NGO collected donations among its members by sending individual letters asking for a donation. Since the NGO was interested in measuring the effect on donations of giving a free poster of the Indian Ocean to the members together with the letter requesting the donation, it designed a randomized experiment. Among the 2500 members, 1250 randomly selected members received the letter and the poster and 1250 received only the letter. The two groups are indicated by the variable P_i, which is equal to 1 if the member i received the letter and the poster, and equal to 0 if the member i received only the letter. In addition, you have data on donations, y, and one individual characteristics, x, which is an important determinant of donations. The following figure shows the distribution of donations.



(a) **[1 point]** How can you estimate the average treatment effect (ATE) of providing the poster on donations?
SOLUTION:

- If randomization was carried out correctly, compute average donation for group with $P_i = 1$ and for the group with $P_i = 0$, and take the difference in means (0.5 points).
- If randomization was NOT carried out correctly, then you need to have a method to deal with bias, because the previous difference would capture only ATT + selection bias (0.5 points).

(b) **[1.5 points]** You decide to estimate the following Type I Tobit model:

$$\begin{aligned} y_i^* &= \alpha P_i + \beta x_i + u_i \\ y_i &= 1[y_i^* > 0] \end{aligned}$$

where u is distributed as a normal distribution $N(0, \sigma^2)$, and u is independent of x and P. Note that neither y_i^* nor u_i are observed. Why would you choose this model in the current setting?
SOLUTION:

- The observations in zero can be considered as corner solutions (i.e., they might be capturing negative willingness to contribute). Therefore I need to correct for censoring at zero (1.5 points).
- EXTRA: The treatment might influence two dimensions: probability to contribute, and the amount contributed. Previous point just focuses on the ATE on contributions, including both processes (0.5 points).

(c) **[1 point]** Derive the probability of zero donations as function of the parameters, $P[y = 0|P, x]$.
SOLUTION:

- Standard derivation used in a probit model (1 point).

(d) **[1.5 points]** Can you compare α and β estimated with the Tobit model with the following estimates computed with OLS?

$$y_i = \alpha P_i + \beta x_i + u_i$$

SOLUTION:

- No, because interpretation is different. OLS is a linear model, then coefficients have a direct interpretation in terms of marginal effects, but Tobit is a non-linear model (1.25 points).
- Explain why the magnitude of the coefficients in the Tobit model does not have a direct interpretation (0.25 points).
- EXTRA: Show derivatives in a Tobit model (marginal effects for $P[y = 0|P, x]$, $E[y|P, x, y > 0]$ and $E[y|P, x]$), their meanings and comparison with OLS (0.5 points).

4. **[25%]** A recent paper investigated the impact on schooling achievement of installing air filters in classrooms. The author used a unique setting arising from a gas leak in the United States, whereby the offending gas company installed air filters in every classroom, office and common area for all schools within 5 miles of the leak (but none beyond).

(a) **[1.5 points]** The author used a Regression Discontinuity (RD) design to identify the causal effect he wants to study. Is this a sharp or fuzzy RDD? Explain what variation does the author use and what are the keys assumptions of this identification strategy.
SOLUTION:

- Sharp RDD since the air filters are installed for all schools within the cutoff, but none beyond (should be verified with data though, to check is the rule was respected in practice) (0.5 points).
- The author uses variation in the installation of air filters to compare student achievement in schools receiving air filters relative to those that did not (0.5 points).

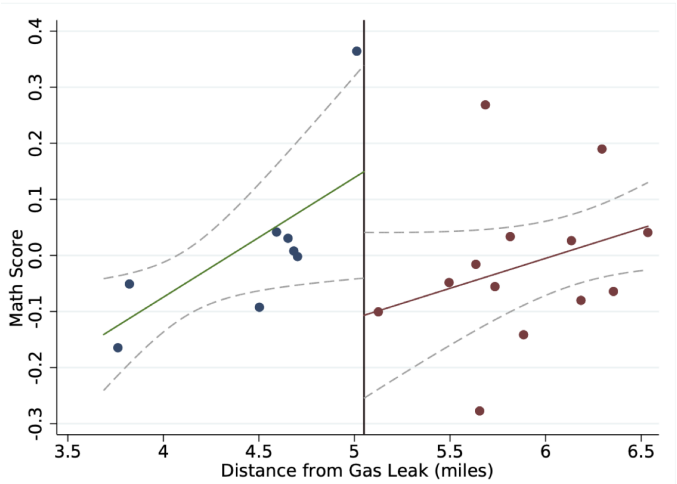
- The key assumption is that the schools/students on either side of the cutoff are similar (0.25 points).
- State assumptions about continuity of potential outcomes at the cutoff (0.25 points).

(b) **[1 point]** What causal effect is estimated using this strategy?
SOLUTION:

- Sharp RDD estimate the LATE at the cut off point (1 point).

(c) **[1.5 points]** The author's empirical analysis can be summarized using the following figure, which gives the scatter plot of Math score two years after the gas leak (aggregated on school level) against the distances of school from the gas leak. Based on the diagram, how do you think the author estimated the impact of the air filter on math achievement?

(a) Math Score



Note. Dashed lines represent 95% confidence intervals with standard errors clustered at the school level.

SOLUTION:

- Linear regression is used on both sides of the cutoff point at 5 (0.5 points).
- The causal effect is represented by the distance between two intersection points generated by the green line and the cut off line vs the red line and the cut off line (0.5 points).
- Although the overlapping of 95% CI does not necessarily imply that the difference is insignificant, the diagram does not provide very strong evidence that air filter has a positive effect (0.5 points).
- The linear regression seem to suggest that the score increases with the distance to gas leak (both within and outside the cutoff). EXTRA: Discussions on this should be given additional credit (0.5 points).

(d) **[1 point]** How would you estimate the effect of the air filter?
SOLUTION:

- Open question, the important is to motivate the answer. Some examples could be on linearity of functional form, or bandwidth restriction (1 point).