Graded Homework Exercise 1

Felix Weinhardt

May 2025

This problem set will be graded. Please work alone or together in groups of two. Send your solutions as Stata do-file and PDF using the appropriate github repository before May 16th 0:00am. The solutions will be discussed in class on May 16, 2025.

Problem 1

The Stata-code "ohmygod.do" on github generates a dataset with N=1,000 and heterogeneous treatments effects $\beta_1 = 1$ for the first 500 observations and $\beta_2 = 5$ for the remaining 500 observations. We are interested in estimating the average β for the true population specification: $Y = \alpha + \beta X + \epsilon$. X is a random variable with mean zero and standard deviation 1. There exists a further (control) variable W that correlates with X but that is not correlated with ϵ, i.e. does not belong to the model.

A friend from the Baccara School of Real Economics (BSoRE) shows you OLS estimates of β from six separate regressions: without and with W included as control using only observations 1-500, observations 501-1000 and the full sample.

Table 1: OLS and "irrelevant controls"

	(1)	(2)	(3)	(4)	(5)	(6)
Estimated effects of X	0.956***	0.956**	4.964***	5.017**	3.000**	2.466**
	(0.046)	(0.046)	(0.044)	(0.098)	(0.113)	(0.089)
Estimates effect of W	-	-0.008	-	-0.055	-	1.061**
		(0.043)		(0.091)		(0.073)
Controls included		✓		✓		✓
N	500	500	500	500	1000	1000

Notes: ** denotes significance at the 1%-level. For DGP and estimates, see Stata do-file "ohmygod.do".

1. Your friend gets pretty worried about the coefficient movement from column (5) to column (6) and asks you to implement the Oster-method to bound the coefficient movement. Under the standard assumptions (proportional selection on observables and unobservables) do the Oster bounds include zero?

2. Do you agree with your friend that bounding the coefficient movement using the Oster-methods is valuable for assessing how unobserved factors might bias the estimates in columns (5) and (6)?

3. Your friend from the BSoRE came up with a machine learning lasso procedure to generate a data-driven answer for the variables that should be included or not. Based on that result, he insists that W needs to be part of the model. Given the insights from columns (1) to (4), can you write down a specification for the full sample, which includes W, and which returns an estimate of the ATE that is close to three?

Problem 2

Please read the classic paper by David Card (1993) that examines the returns of education on earnings. This paper was key in the development and understanding of instrumental variables for estimation, which we will cover soon. In this exercise, we want to understand better what the OLS results presented in this paper are estimating.

1. Replicate the unconditional and conditional OLS result from the paper using the card.dta data provided on github.
2. We now want to examine empirically, if unobserved heterogeneity is at play in these OLS estimates. This might already be the case in the unconditional OLS. Focusing on the unconditional estimate and effect of college on wages, identify two strata using a latent group detection method. One way to do this is using the "gsm" command in Stata, i.e. "gsem (lwage <- educ)" but maybe you find something better.
3. Provide (unconditional) OLS estimates for these subgroups and calculate an estimate for the group-sized weighted OLS, β^{AE}. Does this differ from the unconditional OLS, β^{P}, obtained for the full sample?
4. Do you think this re-weighted estimate is a good measure for the returns of education?
5. A key insight from this paper was that the OLS and IV estimates differ. In this setting, does your analysis strengthen or weaken this result?

Problem 3

We have discussed in the lecture that the common interpretation of fixed effect models is that these capture any between-group (unobserved) factors (confusingly often also called "heterogeneity") and that only the within-group variation is used to identify effects. Read Murphy and Weinhardt (2020) and answer the following questions:

1. What is the key question of this paper, and why might this be of interest?
2. How many fixed effects do they have in their main specification? Are these included as regressors? If not, how are these included in the estimation?
3. The authors write that they include fixed effects at the school-subject-cohort-level, which can be interpreted as classroom-level fixed effects. Does this mean that these specifications rely only on within-classroom variation to estimate rank effects?
4. Do you still think that including group fixed effects means that only within-group variation is used? How could you modify the above statement to be more precise about what kind of between-group variation is being absorbed with fixed effects?

References

Card, D. (1993). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. NBER Working Papers 4483, National Bureau of Economic Research, Inc.

Murphy, R. and Weinhardt, F. (2020). Top of the class: The importance of ordinal rank. *The Review of Economic Studies*, 87(6):2777–2826.