# Econ 312: Problem Set 2

# Professor Magne Mogstad

# Due Thursday, April 16 before the lecture

## Problem 1

Consider the linear regression model:

$$Y = X_1'\beta_1 + X_2'\beta_2 + U$$

$Y$, $U$ - scalars; $X_1, X_2, \beta_1, \beta_2$ - vectors; $(X_1, X_2, U)$ i.i.d.; $(X_1, X_2) \perp\!\!\!\perp U$; $\mathbb{E}[X_1], \mathbb{E}[X_2], \mathbb{E}[U] < \infty$

**a)** What is the effect of deleting $X_2$ on the estimated coefficient of $\beta_1$

**b)** Suppose that you include $X_2$ in different ways:

1) Regress $X_1$ on $X_2$, work with the residual of $X_1$ purged of $X_2$ and regress $Y$ on this residual.

2) Regress $X_1$ on $X_2$ and $Y$ on $X_2$, and regress the residual from the second regression on the residual from the first regression.

How do these estimates compare with your result from **a)**? With the OLS estimates of the initial specification?

# Problem 2

Let $Y_i(1)$ and $Y_i(0)$ be potential outcomes of individual $i$ if treated or not treated, respectively. Let $Y_i$ be the actual outcome and let $D_i$ be the treatment indicator. We assume:

- $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | X_i$
- $0 < \mathbb{P}[D_i = 1 | X_i = x] < 1 \ \forall x \in \text{supp}(X_i)$

**a)** Propose an estimator of $\mathbb{E}[Y_i(1) - Y_i(0)]$

**b)** Show that the assumptions stated imply that $D_i$ is conditionally independent of $(Y_i(1), Y_i(0))$ given $\mathbb{P}[D_i = 1|X_i]$: $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbb{P}[D_i = 1|X_i]$. Why is this result important?

**c)** Define the propensity score $\mathbb{P}[D_i = 1|X_i]$ as the probability of receiving the treatment given the observables variables $X_i$. Propose an estimator of $\mathbb{E}[Y_i(1) - Y_i(0)]$ based on the propensity score.

# Problem 3

LaLonde (AER, 1986) investigated whether non-experimental methods could reproduce the experimental estimate based on the National Supported Work (NSW) Demonstration. The following dataset from Smith and Todd (J Ectrics, 2005):

`https://www.dropbox.com/s/dl/aw4yi13mz9z03yf/lalonde2.dta`

includes the NSW sample, as well as two non-experimental samples: one based on the Current Population Survey (CPS) and one on the Michigan Panel of Income Dynamics (PSID).

The variable -sample- identifies the relevant observations.

The variable -treated- identifies the observations that were treated (participate in a subsidized work experience program) in the NSW (from April 1975 to August 1977).

You are interested in the average effect on Real Earnings in 1978 of the treatment for the treated.

Start with the NSW sample:

**a)** Investigate whether the data is consistent with randomization of the treatment.

**b)** Estimate the effect using the experimental sample.

Now use the sample consisting in the treated from the NSW sample and the comparison individuals from the CPS sample.

**c)** Estimate the effect using OLS.

**d)** Investigate covariate balancing and support between the treated and the CPS sample.

**e)** Estimate the effect using 1 nearest neighbor propensity score matching. (Use -psmatch2- which can be installed using: ssc install psmatch2, if you use Stata).

**f)** Estimate the effect using the propensity score and local linear regression.

## Problem 4

You are interested in estimating the effect of doing mathematics homework (yes/no) on mathematics test scores. You have data on all 10th graders in Oslo for the school year 2010/2011. The dataset contains the following variables:

- students: end year test scores, homework, # of missed classes, gender, age, test scores in 2009/2010
- parents: education
- teacher: gender, age, education

About half of the students do their homework.

**a)** If you want to abolish homework, what effect would you want to estimate?

**b)** If you want to make homework mandatory, what effect would you want to estimate?

**c)** You want to compare the effect of doing homework as compared to an extra hour of math teaching. What effect of homework would you like to know?

You want to estimate how well students that are currently not doing their homework would do, if they did their homework. You decide to use matching, and will therefore rely on a conditional independence assumption (CIA).

**d)** Explain your CIA. Be explicit about the counterfactual outcomes and the variables that you want to control for. Why might your CIA not hold? Can you think of examples where you get upward biased estimates? And downward biased estimates?

**e)** Explain how you use the CIA to estimate the counterfactual outcome, how you take into account that students that do their homework have different characteristics, and what support condition you need.

**f)** How would you estimate your effect using OLS?

**g)** You see in your data that boys never do their homework. What implications does this have for your research?

You discover that not all teachers assign homework, and you get a new variable from Oslo municipality with information (0/1) on whether the teacher assigned homework or not. They tell you that teachers were assigned to give homework (or not) in a randomized experiment.

**h)**First you add this new information to your matching variables. What will happen to your estimates and standard errors?

**i)** How will you use this new data and what effects can you estimate?