Econ 312: Problem Set 2

Professor Magne Mogstad

Answer Key

Problem 1

a) Denote:

$$P_X = X(X^{'}X)^{-1}X^{'}$$

$$M_X = I - P_X$$

When both X_1 and X_2 are included we can use FW theorem to write the coefficient on X_1:

$$\hat{\beta} = (X_1^{'}M_{X_2}X_1)^{-1}X_1^{'}M_{X_2}Y$$

In the case without X_2:

$$\tilde{\beta} = (X_1^{'}X_1)^{-1}X_1Y$$

So, their difference is:

$$\tilde{\beta} - \hat{\beta} = \left((X_1^{'}X_1)^{-1}X_1 - (X_1^{'}M_{X_2}X_1)^{-1}X_1^{'}M_{X_2}\right)Y$$

The estimates are equal only in the case $X_1 \perp\!\!\!\perp X_2$

b) 1) Regressing Y on the residual of X_1 from its regression on X_2:

$$\hat{\beta} = \left(\underbrace{(M_{X_2} X_1)}_{\text{residual}}{}' M_{X_2} X_1 \right)^{-1} (M_{X_2} X_1)' Y \underbrace{=}_{M'_{X_2} M_{X_2} = M_{X_2}} \left(X_1' M_{X_2} X_1 \right)^{-1} X_1' M_{X_2} Y$$

2) Now we also use the residuals from the regression of Y on X_2:

$$\tilde{\beta} = \left((M_{X_2} X_1)' M_{X_2} X_1 \right)^{-1} (M_{X_2} X_1)' M_{X_2} Y \underbrace{=}_{M'_{X_2} M_{X_2} = M_{X_2}} \left(X_1' M_{X_2} X_1 \right)^{-1} X_1' M_{X_2} Y = \hat{\beta}$$

Problem 2

a) Using LIE:

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}(\mathbb{E}[Y_1 - Y_0 | X]) \underbrace{=}_{CIA} \mathbb{E}(\mathbb{E}[Y|D=1, X] - \mathbb{E}[Y|D=0, X])$$

In the case of discrete variable X (or very few observed realizations) we can compute averages in each cell and sum them with weights $\mathbb{P}[X = x]$. In the case of continuous X with substantial variation we can use a kernel estimator and integrate over the values of X.

b) *Proof:*

$$\begin{aligned}
\mathbb{P}[D=1|Y_1, Y_0, p(X)] &= \mathbb{E}[D|Y_1, Y_0, p(X)] \\
&\underbrace{=}_{LIE} \mathbb{E}[\mathbb{E}[D|Y_1, Y_0, p(X), X]|Y_1, Y_0, p(X)] \\
&= \mathbb{E}[\mathbb{E}[D|Y_1, Y_0, X]|Y_1, Y_0, p(X)] \\
&\underbrace{=}_{CIA} \mathbb{E}[\mathbb{E}[D|X]|Y_1, Y_0, p(X)] \\
&= p(X)
\end{aligned}$$

The result is important since $p(X)$ is 1 dimensional scalar, that could be parametrically estimated (reduction in dimensionality).

c) Given the result from the previous part we can use conditioning on $p(X)$ instead of X and apply the same logic as in **a)**. For example in the discrete case:

$$\hat{\mathbb{E}}[Y_1 - Y_0] = \sum_p \mathbb{P}[\hat{p}(X) = p] \left(\hat{\mathbb{E}}[Y|D = 1, p(X) = p] - \hat{\mathbb{E}}[Y|D = 0, p(X) = p] \right)$$

Where inside expectations are means across corresponding cells.

Problem 3

Below is the example of Stata code(since the majority used stata for this exercise).

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

```
clear all
use lalonde2

g d = 1 if treat==1
replace d = 0 if sample==2

// 1a.
global x age educ black married nodegree re74 re75 hisp w76 kids18 kidmiss metro
foreach v of var $x {
        qui ttest `v', by(treat)
        di"`v'" _col(15) %10.0g r(mu_1) %10.0g r(mu_2)  %6.3f r(p)
}

reg treat $x, robust
```

b) Estimate the effect using the experimental sample.

```
// (note all obs in nsw are in metro area)
// ==> support issue!!
drop if metro==0

// 1b.
reg re78 treat, robust
reg re78 treat $x, robust
```

Now use the sample consisting in the treated from the NSW sample and the comparison indi-

viduals from the CPS sample.

c) Estimate the effect using OLS.

```
// 1c.

// check some distributions and bivariate relationships
hist age, by(d, col(1)) discr
lpoly re78 age if d==0, deg(1) nosc
reg re78 educ black married nodegree re74 re75 hisp w76 kids18 kidmiss
predict res_re78
reg res_re78 c.age##c.age##c.age if d==0
predict p
twoway  (hist age if d==1, yaxis(2))(lpoly res_re78 age if d==0, deg(1)) (sc p age if d==0)


sum re74 if d==0
replace re74 = r(max) if d==1 & re74>r(max)
g max74 = re74 == r(max)
g zero74 = re74==0

sum re75 if d==0
replace re75 = r(max) if d==1 & re75>r(max)
g max75 = re75 == r(max)
g zero75 = re75==0

g age2 = age^2
g age3 = age^3 / 1000
g educ2 = educ^2

global x age age2 age3 educ educ2 black married nodegree ///
        re74 re75 hisp w76 kids18 kidmiss max74 max75 zero74 zero75

reg re78 black##c.($x) if d==0, robust
predict u, res // u = y - x*b0hat
sum u if d==1
```

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

```
// 1d
global x age age2 age3 educ educ2 black married nodegree re74 re75 hisp w76 kids18 kidmiss max74 max75 zero74 zero75
foreach v of var $x {
        qui ttest `v', by(d)
        di"`v'" _col(15) %10.0g r(mu_1) %10.0g r(mu_2)  %6.3f r(p)
}

reg d $x, robust
```

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).

```
// note that you want to iterate the following two lines until
// you achieve the best possible balancing
psmatch2 d $x
pstest $x


// note support issues
// also note that matching on X's which are not very predictive
// of the outcome in the control sample will reduce support
// at the cost of little bias reduction
psgraph

// once you're happy compute the effect
// for se's use stata's -teffect-
psmatch2 d $x, out(re78)
reg re78 d [aw=_weight]
reg re78 d $x [aw=_weight]

// impose common support by dropping treatment observations
// whose pscore is higher than the maximum or less than
// the minimum pscore of the controls.
psmatch2 d $x, out(re78) common
reg re78 d $x [aw=_weight]

// compare with nsw sample on support
reg re78 treated if _support
reg re78 treated $x if _support
```

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

```
// 1f.
probit d $x
predict pscore
g ps1 = pscore if d==1 // only predict for the treated sample
lpoly re78 pscore if _treated==0, deg(1) gen(y0llr) at(ps1)
g effect = re78 - y0llr
sum effect if _treated==1

// see Smith and Todd (J Ectrics, 2005) for a very detailed and
// thoughtful matching analysis of these data, as well as a good
// discussion of the benefits and limitations of the method
```

Problem 4

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

You then want to know the effect on people currently doing their homework - ATT.

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

You then want to know the effect on people currently not doing their homework - ATU.

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 then want to know the effect on all students (or a random student)-ATE

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?

You want to estimate $\mathbb{E}[Y_1 - Y_0|D = 0] = \mathbb{E}[Y_1|D = 0] - \mathbb{E}[Y_0|D = 0]$. The first term on the RHS is unobserved and you therefore need the following CIA: $Y_1 \perp\!\!\!\perp D|X$. Do not match on missed classes because this variable is potentially endogenous (it is not predetermined). Things like motivation and ability, which affect potential achievement with homework (Y_1), are not observed and may correlate with doing homework. If motivated students do their homework, you will get upward biased estimates. If weak students do their homework (because they think they need it, or their parents force them), then you get downward biased estimates. Stories about good/bad teachers are also possible.

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.

Let $\mathbb{E}[X|D = 0]$ be the expectations operator over the distribution of X on the support of $D = 0$.

You want to estimate:

$$\begin{aligned}\mathbb{E}[Y_1|D=0] &= \mathbb{E}_{X|D=0}[\mathbb{E}[Y_1|D=0,X]] \\ &\underbrace{=}_{CIA} \mathbb{E}_{X|D=0}[\mathbb{E}[Y|D=1,X]] \\ &= \mathbb{E}_{X|D=0}[\mathbb{E}[Y|D=1,X]] \\ &= \sum_{x\in S_0} \mathbb{E}[Y|D=1,X=x]\mathbb{P}(X=x|D=0)\end{aligned}$$

you therefore use the CIA to estimate $\mathbb{E}[Y_1|D=0,X] = \mathbb{E}[Y_1|D=1,X]$, you weight $\mathbb{E}[Y_1|D=0,X=x]$ using $\mathbb{P}(X=x|D=0)$, which will usually be different from $\mathbb{P}(X=x|D=1)$. If $\mathbb{P}(X=x|D=1)$ then you cannot estimate $\mathbb{E}[Y_1|D=1,X]$ which gives the support condition: $\mathbb{P}(X=x|D=1)>0 \forall x \in S_0$. (Notice that $\mathbb{P}(X=x|D=0)=0$ is not a problem, since then these are not part of the population for which we want to estimate the effect.)

f) How would you estimate your effect using OLS?

By estimating $\mathbb{E}[Y_1|D=1,X=x]$ with a flexible regression of Y on X in the sample of students doing their homework. You can then use the estimated coefficient, say $\hat{\beta}$, to estimate $\hat{\mathbb{E}}[Y_1|D=0] = \bar{X}_{D=0}\hat{\beta}_1$, and then estimate the $\bar{X}_{D=0}\hat{\beta}_1 - \bar{Y}_{D=0}$. To be concrete, you could ideally estimate a saturated model, such as:

$$Y_i = \sum_{x\in S_0} \beta_x \mathbb{1}[X_i = x] + \epsilon_i$$

And then use:

$$\hat{\mathbb{E}}[Y_1|D=0] = \sum_{x\in S_0} \hat{\beta}_x \frac{N_{X=x,D=1}}{N_{D=1}} = \sum_{x\in S_0} \hat{\beta}_x \mathbb{P}[X=x|D=1]$$

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

You have a support problem: You can only estimate the effect for girls.

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?

Nothing (in expectation) to the effect, but your standard errors will increase.

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

You can use this information to estimate the effect of homework assignment (the ITT), or use IV to estimate the effect of homework (since there is possibly not perfect compliance).