ECON 21020: Econometrics
The University of Chicago, Spring 2022
Instructor: Thomas Wiemann
Problem Set # 4: Selection on Observables & Multiple Linear Regression
Due: 11:59am on May 23, 2022

Problem 1 15 Points

The following are "True or false?"-questions. If the statement is true, provide a brief proof ($\approx$ 3 lines). If the statement is false, provide a counter example. There are no points awarded for answers without a proof or counter example.

a)

True or false? Let X, W, and U be random variables. If $W \perp\!\!\!\perp U$, then $W \perp\!\!\!\perp U|X$.

b)

True or false? Let $X_n, n \geq 1$, be a sequence of random vectors, and X be another random variable. If $X_n \overset{d}{\to} X$, then $X_n \overset{p}{\to} X$.

c)

True or false? Let Y be a random variable and $X = (1, X_1, \dots, X_k)$ be a random vector. If $\varepsilon = Y - \text{BLP}(Y|X)$, then $E[\varepsilon|X] = 0$.

Problem 2   10 Points

Prove Corollary 5 of Lecture 7.

(Hint: This should be a reasonably straightforward application of the CMT. You do not need to re-prove results from previous problem sets or from class!)

Problem 3 20 Points

Consider the all causes model. In particular, let (Y, W, X, U) be a random vector with joint distribution characterized by

$$Y = g(W, U), \tag{1}$$

where W is binary. Suppose further that both selection on observables – i.e., $W \perp\!\!\!\perp U|X$ – and common support – i.e., $\operatorname{supp} X|W = \operatorname{supp} X$ – hold.

In the lecture, we proved identification of common causal parameters under selection on observables and common support when the econometrician observes (Y, W, X). In this exercise, you will show that the econometrician need not condition on the entirety of X. Instead, it suffices to condition on the *propensity score* defined by

$$p(X) \equiv P(W = 1|X). \tag{2}$$

A version of this result was first shown in Rosenbaum and Rubin (1983).[1]

a)

Give a brief interpretation of $P(W = 1|X)$.

b)

Show that

$$W \perp\!\!\!\perp U|X \quad \Rightarrow \quad W \perp\!\!\!\perp U|p(X). \tag{3}$$

(Hint: It suffices to show that

$$P\left(U \leq u, W = 1|p(X)\right) = P\left(U \leq u|p(X)\right) P\left(W = 1|p(X)\right),$$

$\forall u \in \operatorname{supp} U|p(X)$.*)*

[1] The result continues to be of highest practical relevance because it allows researchers to condition on a scalar random variable $p(X)$ rather than a random vector X, which may have computational benefits. To give a hint at how influential the result is: The paper has almost 32,000 citations on Google scholar to date!

c)

Consider defining the parameter

$$E[g(1,U) - g(0,U)|p(X) = p], \tag{4}$$

$\forall p \in \operatorname{supp} p(X)$. Give a brief interpretation.

d)

Use part a) to show that

$$E[g(1,U) - g(0,U)|p(X) = p] \tag{5}$$

is point-identified $\forall p \in \operatorname{supp} p(X)$.

Problem 4   20 Points

Suppose an econometrician is interested in the effect of military service on lifetime earnings in the US. To structure the analysis, she considers the random vector (Y, W, X, U) with joint distribution characterized by

$$Y = g(W, U),$$

where

- $Y \equiv$ lifetime earnings;
- $W = 1$ if the individual served in the military and 0 otherwise;
- $X = 1$ if the individual is male and 0 otherwise;
- $U \equiv$ all determinants of Y other than W;
- $g : \operatorname{supp} W \times \operatorname{supp} U \to \operatorname{supp} Y$.

Suppose the econometrician observes a sample $(Y_1, W_1, X_1), \ldots, (Y_n, W_n, X_n) \overset{iid}{\sim} (Y, W, X)$.

a)

Give an example for an unobserved determinant U of Y.

b)

Give an example for a potential confounder in this setting.

c)

Define and interpret the potential outcomes for $w \in \{0, 1\}$.

d)

From 1940 to 1973, the US conscripted men to fill vacancies in the military that could not be filled with voluntary means. During the Vietnam war era, the conscription process – commonly known as "the draft" – relied on a birthday lottery that determined which men had to join the armed forces.[2]

[2] This exercise is crudely based on Angrist (1990), who exploits random variation from the draft lottery during the Vietnam war era.

Suppose that the econometrician focuses on the Vietnam war era in order to exploit the random variation from the draft lottery. In particular, she considers assuming selection on observables conditional on being male – that is,

$$W \perp\!\!\!\perp U | X.$$

Give a brief economic interpretation of the assumption. Does it appear plausible here? Explain briefly.

e)

Suppose for the remainder of this exercise that everyone serving in the US military during the Vietnam war era was randomly drafted, that only men were drafted, and that there are no draft-dodgers (or conscientious objectors).

Does this make the selection on observables assumption more/less plausible? Explain briefly.

f)

Define and interpret the CATE(x) for $x \in \{0, 1\}$.

g)

Is the CATE(1) point-identified under the assumptions of this exercise? Explain briefly.

h)

Is the CATE(0) point-identified under the assumptions of this exercise? Explain briefly.

i)

Define and interpret the ATE.

j)

Is the ATE point-identified under the assumptions of this exercise? Explain briefly.

Problem 5 10 Points

Let Y be a random variable and $X = (1, X_1, \dots, X_k)$ be a random vector. Consider

$$\min_{\beta \in \mathbb{R}^{k+1}} E\left[\left(E\left[Y|X\right] - X^\top \beta\right)^2\right], \tag{6}$$

and

$$\min_{\beta \in \mathbb{R}^{k+1}} E\left[\left(Y - X^\top \beta\right)^2\right]. \tag{7}$$

Show that the solutions to the minimization problems (6) and (7) are identical.[3]

(Hint: You can – but don't need to – take first order conditions to solve this problem.)

[3] This motivates why the best linear approximation to the CEF $E[Y|X]$ – as defined in (6) – is commonly referred to as the best linear predictor: Just as you showed in Problem Set 3 for *simple* linear regression.

Problem 6 25 Points

This exercise revisits the data of Abrevaya (2006), who analyzes the effect of smoking on birth outcomes. A cleaned version of the data is posted to Canvas (see the file `bw06.csv`).[4] The variables we focus on in this problem set are:

- `birthweight` $\equiv$ birth weight in grams;
- `cigsdaily` $\equiv$ cigarettes smoked per day by the mother;
- `boy` $\equiv$ indicator for a male infant;
- `age` $\equiv$ mother's age at birth;
- `highschool` $\equiv$ indicator for being a high school grad;
- `somecollege` $\equiv$ indicator for having completed some college;
- `college` $\equiv$ indicator for being a college grad;
- `married` $\equiv$ indicator for being married.

Once downloaded, you can load the data into R using the following code:

```r
# Load the bw06.csv data
dat <- read.csv("data/bw06.csv")
dat <- as.matrix(dat)
```

Suppose we are interested in the association between `cigsdaily` and `birthweight`, possibly controlling for some other determinants of infant birth weight. To think clearly about the relationship of interest, consider the random vector $(Y, W, X, \tilde{X}, U)$ with joint distribution characterized by

$$Y = g(W, U),$$

where

- $Y \equiv$ `birthweight`;
- $W \equiv$ `cigsdaily`;
- $X \equiv$ (`boy`, `age`, `highschool`, `somecollege`, `college`);

[4]The posted data is a subset of the full data used in Abrevaya (2006). In particular, the data on Canvas contains 9800 observations from the 1996 sample.

- $\tilde{X} \equiv$ `married`;
- $U \equiv$ all determinants of Y other than W;
- $g : \operatorname{supp} W \times \operatorname{supp} U \to \operatorname{supp} Y$.

Consider a sample $(Y_1, W_1, X_1, \tilde{X}_1), \ldots, (Y_n, W_n, X_n, \tilde{X}_n) \overset{iid}{\sim} (Y, W, X, \tilde{X})$, and suppose that the data is a realization of this sample for $n = 9800$.

It will be convenient to store the variables in dedicated R vectors/matrices.

```
# Select variables
y <- dat[, "birthweight"]
w <- dat[, "cigsdaily"]
x <- cbind(1, dat[, c("boy", "age", "highschool",
                      "somecollege", "college")])
x_tld <- dat[, "married"]
```

Note: This exercise must be completed in base R. That is, don't load any dependencies.

If you upload your solutions to a GitHub repository and share the link in your homework solutions, you earn an extra credit of 5 percentage points on this problem set.

a)

Compute an estimate of the BLP$(Y|W)$-coefficients. Give a brief economic interpretation the coefficient corresponding to W.

b)

Let β_W denote the BLP$(Y|W, X)$-coefficient corresponding to W. Compute an estimate of β_W. Give a brief economic interpretation.

c)

Does your estimate in Part a) differ from Part b)? Why or why not?

d)

Against your better judgment, you decide to apply for a Summer internship at the tobacco company *Dromedary*. They are not amused when you share your results during your interview. Your interviewer – who apparently did not take Econ 21020 – responds:

- “Don’t share these with anyone! If the public knew that smoking causes low birth weights, we’re done for.”

Explain briefly why the interviewer misinterpreted the results you shared.

e)

Somewhat appeased by your explanation, the interviewer wonders whether a causal interpretation of β_W may be warranted under reasonable assumptions.

State and interpret the common support and the selection on observables assumption where you condition on X.

f)

Can you verify common support using the observed data?[5]

g)

Use the variable `married` to conduct a balance test for assessing the plausibility of the selection on observables assumption. Does the test reject on a 1% significance level? Give a brief economic interpretation of the result.

[5] This is *not* a theoretical question: Check whether you can verify common support with the data.

Problem 7   20 Points (Extra Credit)

This is an optional extra credit exercise.

This exercise must be completed in base R *without* using the `lm`-command.

a)

Write a function `my_coef` that takes 1) a vector $\mathtt{y} \in \mathbb{R}^n$, and 2) a matrix $\mathtt{X} \in \mathbb{R}^{n,k+1}$, and that returns ols estimates $\hat{\beta}_n \in \mathbb{R}^{k+1}$ for the BLP($\mathtt{y}|\mathtt{X}$)-coefficients β.

```
# Define a custom function to compute the ols estimates
my_coef <- function(y, X) {
  # Compute and return estimates for beta
  # [INSERT YOUR CODE HERE]
}#MY_COEF

# Test the function using your solution to Problem 6
coef <- my_coef(y, X)
coef
```

b)

Write a function `my_blp` that takes 1) a vector $\mathtt{coef} \in \mathbb{R}^{k+1}$ containing estimates $\hat{\beta}_n$, and 2) a matrix $\mathtt{X} \in \mathbb{R}^{n,k+1}$, and that returns estimates of BLP($\mathtt{y}|\mathtt{X}$).

```
# Define a custom function to compute the blp estimates
my_blp <- function(coef, x) {
  # Compute and return BLP estimates
  # [INSERT YOUR CODE HERE]
}#MY_BLP

# Test the function
mean(y - my_blp(coef, X)) # 0
```

c)

Write a function `my_se` that takes 1) a vector $\mathtt{coef} \in \mathbb{R}^{k+1}$ containing estimates $\hat{\beta}_n$, 2) a vector $\mathtt{y} \in \mathbb{R}^n$, and 3) a matrix $\mathtt{X} \in \mathbb{R}^{n,k+1}$, and that returns a vector of standard errors $se\left(\hat{\beta}_n\right)$.

Your solution *must* make use of your function `my_blp`.

```
# Define a custom function to compute the standard error
```

```r
my_se <- function(coef, y, X) {
  # Compute and return the standard error
  # [INSERT YOUR CODE HERE]
}#MY_SE

# Test the function using your solution to Problem 6
se <- my_se(coef, y, X)
se
```

d)

Write a function `my_teststat` that takes 1) a vector `coef` $\in \mathbb{R}^{k+1}$ containing estimates $\hat{\beta}_n$, and 2) a vector `se` $\in \mathbb{R}^{k+1}$ containing the standard errors $se\left(\hat{\beta}_n\right)$, and that returns a vector of test statistics $T_{j,n} = |\hat{\beta}_{j,n}/se\left(\hat{\beta}_{j,n}\right)|$ for $j = 1, \ldots, k+1$, as well as the corresponding p-values.

```r
# Define a custom function to compute the test stat and p-value
my_teststat <- function(beta, se) {
  # Compute and return the test stats and p-values
  # [INSERT YOUR CODE HERE]
}#MY_TESTSTAT

# Test the function
my_teststat(coef, se)
```

e)

Write a function `my_ols` that takes 1) a vector `y` $\in \mathbb{R}^n$, and 2) a matrix `X` $\in \mathbb{R}^{n,k+1}$, and that returns a matrix containing the ols-estimates $\hat{\beta}_n$, the standard errors $se\left(\hat{\beta}_n\right)$, the test statistics T_n, and the corresponding p-values.

Your solution *must* make use of your functions defined in earlier parts: `my_coef`, `my_se`, and `my_teststat`.

```r
# Define a custom function to compute and characterize ols estimates
my_ols <- function(y, X) {
  # Compute and return the the ols estimate, se, Tn, and p-val
  # [INSERT YOUR CODE HERE]
}#MY_OLS

# Test the function using your solution to Problem 6
my_ols(y, X)
```

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics*, 21(4):489–519.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *American Economic Review*, pages 313–336.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.