ECON 21020: Econometrics
The University of Chicago, Spring 2022
Instructor: Thomas Wiemann
Problem Set # 3: Random Assignment & Simple Linear Regression
Due: 11:59am on May 4, 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? If $X_1, \ldots, X_n \sim X$, then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right) \xrightarrow{d} N\left(0, Var(X)\right). \tag{1}$$

b)

True or false? Let W, X, and U be random variables. If $E[U|W] = 0$, then $E[U|W, X] = 0$.

c)

True or false? Let Y and X be random variables. If $E[Y|X] = X^2$ and $X \sim U(0, 1)$, then $\beta = 1$, where β is the BLP$(Y|X)$-coefficient.

Problem 2 20 Points

Consider the all causes model discussed in class:

$$Y = g(W, U), \tag{2}$$

where (Y, W, U) is a random vector.

Suppose we are interested in studying the effect of sentencing on recidivism of juvenile offenders. Let $Y = 1$ if the offender committed another crime and $Y = 0$ otherwise, and let $W = 1$ if the offender's sentence included confinement (e.g., time in prison) and $W = 0$ otherwise (e.g., only a monetary fine).

a)

Give two examples for unobserved determinants U of Y.

b)

State and interpret the random assignment assumption in context of this exercise. Does it seem plausible here? Explain briefly.

c)

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

d)

Define and interpret the ATE between $w = 1$ and $w = 0$.

e)

Suppose a judge is re-evaluating cases of juvenile offenders who are about to be sentenced to confinement. Would knowing the ATE as defined in Part c) help her make an effective decision? Explain briefly.

f)

Suppose that an eager clerk of the judge informs her that the ATE is -0.2 and that the share of juvenile offenders sentenced to confinement is 60%.[1] What are the possible values that

[1] These numbers are made-up! For an actual analysis of the topic, see, for example, Manski and Nagin (1998), who find that confinement may result in higher recidivism rates for juvenile offenders.

the ATT may take?

g)

Suppose that in addition to the information provided by the clerk, the judge knows that her colleagues did not make poor sentencing decisions on average. In particular, that the effect of confinement for those sentenced to confinement is not greater than of those not sentenced to confinement: $\text{ATT} \leq \text{ATU}$. What are the possible values the ATT may take?

h)

Using your solution Part f), can the judge be certain that changing sentences of juvenile offenders who are about to be sentenced to confinement will increase expected recidivism? Explain briefly.

Problem 3   15 Points

The University of Chicago recently set the maximum class size for advanced undergraduate courses to 32 students. Suppose we are interested in studying how class sizes for Econ 21020 are associated with learning outcomes for undergraduate students.

To carefully think about the causal effect of class size on student's learning outcomes, we use the all causes model discussed in the lectures:

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

where Y is the student's score in the Econ 21020 final, W is the size of the class the student participated in, and U are all determinants of Y other than W.

a)

Give an example for an unobserved determinants U of Y.

b)

Define and interpret the potential outcomes for $w \in \mathbb{N}$. Can you define the potential outcomes for non-integer valued w? Explain briefly.

c)

Let (α, β) denote the coefficients of the best linear predictor $\text{BLP}(Y|W)$. Give a brief economic interpretation of β.

d)

Suppose that the registrar was willing to share data on class sizes and students' final scores with us. We regress students' final scores on a constant and the corresponding class size. Our OLS estimate for β is

$$\hat{\beta}_n = -0.7 \qquad \text{and} \qquad se\left(\hat{\beta}_n\right) = 0.2. \tag{3}$$

Construct a 95% confidence interval for β.

e)

Conduct a test of $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ on a 5% significance level. Give a brief economic interpretation of your result.

f)

You share your results with the student-run newspaper *The Chicago Clinker*. The following day, the front page reads:

- "Do Smaller Class Sizes Make Better Econometricians? New Data Says: Clearly!"

Explain briefly why the headline does not accurately describe the results you shared.

g)

The editor of The Chicago Clinker is unwilling to change the headline. Instead they are looking for motivating their interpretation of the results you shared.

State an assumption on the joint distribution of (Y, W, U) that would warrant a causal interpretation of β. Does it seem plausible here? Explain briefly.

Problem 4 10 Points

Let Y and X be random variables. Consider

$$\min_{\alpha,\beta\in\mathbb{R}} E\left[(E[Y|X]-(\alpha+X\beta))^2\right], \tag{4}$$

and

$$\min_{\alpha,\beta\in\mathbb{R}} E\left[(Y-(\alpha+X\beta))^2\right]. \tag{5}$$

Show that the solutions to the minimization problems (4) and (5) are identical.[2]

[2]This problem motivates why the best linear approximation to the CEF $E[Y|X]$ – as defined in (4) – is commonly referred to as the best linear predictor.

Problem 5 15 Points

Let Y and X be random variables. Consider the Yitzhaki weights $\omega(t)$ as discussed in the lectures.

a)

Suppose that X is continuous and $\operatorname{supp} X = \mathbb{R}$. Show that

$$E[X] = \arg\max_{t \in \mathbb{R}} \omega(t). \tag{6}$$

(Hint: Use that the expression we derived in class before simplifying the weights further. In particular, $\omega(t) = \frac{1}{Var(X)} \int_t^\infty (x - E[X]) f_x(x) dx$*.)*

b)

Suppose for the rest of the exercise that $X \sim U(-1, 1)$. Calculate $E[X|X \geq t]$, $E[X|X < t]$, $P(X \geq t)$, and $Var(X)$.

c)

Use Part b) to give an expression for the Yitzhaki weights $\omega(t)$ for $t \in [-1, 1]$ and graph the function. (You may graph the function by hand, no need to use R.)

d)

Suppose that $E[Y|X] = X^3$. Use Yitzhaki's Theorem to calculate the BLP-coefficient β.

e)

Does β equal the average derivative of the conditional expectation function $E\left[\frac{\partial}{\partial X} E[Y|X]\right]$?

Problem 6 15 Points

This exercise revisits the data of Angrist and Krueger (1991) considered in Problem 5 of Problem Set 2. A cleaned version of the data is posted to Canvas (see the file `ak91.csv`). It contains 329,509 observations of American men born between 1930 and 1939. The variables we focus on in this problem set are:

- `YRS_EDUC` $\equiv$ years of education;
- `WKLY_WAGE` $\equiv$ the weekly wage.

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

```
# Load the ak91.csv data
df <- read.csv("data/ak91.csv")
n <- nrow(df)

# Store years of education and the weekly wage in separate variables
yrs_educ <- df$YRS_EDUC
wkly_wage <- df$WKLY_WAGE
```

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)

Let Y and X be two random variables, denoting the weekly wage and the years of education, respectively. Consider a sample $(Y_1, X_1), \ldots, (Y_n, X_n) \overset{iid}{\sim} (Y, X)$, and suppose that the data of Angrist and Krueger (1991) is a realization of this sample for $n = 329,509$.

Give OLS estimators $(\hat{\alpha}_n, \hat{\beta}_n)$ for the BLP$(Y|X)$-coefficients (α, β).

b)

State the estimates $(\hat{\alpha}_n, \hat{\beta}_n)$ computed using the Angrist and Krueger (1991) data. (Use your solution to Part a): Do *not* use the `lm`-command in R!)

Give a brief economic interpretation of your estimate for β.

c)

Use your result to Part b) to compute and state an estimate for $\text{BLP}(Y|X = 16)$. Does your result differ from your solution to Problem 5b) of Problem Set 2? Why or why not? Explain briefly.

d)

Compute and state the standard error $se\left(\hat{\beta}_n\right)$.

e)

Consider testing $H_0 : \beta = 31$ versus $H_1 : \beta \neq 31$. Compute and state an appropriate test statistic.

f)

Compute and state the p-value associated with the test considered in Part e).

g)

Does the test reject on a 10% significance level? Give a brief economic interpretation of the result.

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_simplecoef` that takes two vectors, 1) a vector $\mathtt{y} \in \mathbb{R}^n$, and 2) a vector $\mathtt{x} \in \mathbb{R}^n$, and that returns OLS estimates $(\hat{\alpha}_n, \hat{\beta}_n)$ for the BLP($\mathtt{y}|\mathtt{x}$)-coefficients (α, β).

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

# Test the function using your solution to Problem 6
coef <- my_simplecoef(wkly_wage, yrs_educ)
coef
```

b)

Write a function `my_simpleblp` that takes two vectors, 1) a vector $\mathtt{coef} \in \mathbb{R}^2$ containing estimates $(\hat{\alpha}_n, \hat{\beta}_n)$, and 2) a vector $\mathtt{x} \in \mathbb{R}^n$, and that returns estimates of BLP($\mathtt{y}|\mathtt{x}$).

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

# Test the function
mean(wkly_wage) - mean(my_simpleblp(coef, yrs_educ)) # 0
```

c)

Write a function `my_simplese` that takes three vectors, 1) a vector $\mathtt{coef} \in \mathbb{R}^2$ containing estimates $(\hat{\alpha}_n, \hat{\beta}_n)$, 2) a vector $\mathtt{y} \in \mathbb{R}^n$, and 3) a vector $\mathtt{x} \in \mathbb{R}^n$, and that returns $se\left(\hat{\beta}_n\right)$.

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

```
# Define a custom function to compute the standard error
my_simplese <- function(coef, y, x) {
```

```
  # Compute and return the standard error
  # [INSERT YOUR CODE HERE]
}#MY_SIMPLESE

# Test the function using your solution to Problem 6
se <- my_simplese(coef, wkly_wage, yrs_educ)
se
```

d)

Write a function `my_simpleteststat` that takes two scalars, 1) an estimate $\hat{\beta}_n$, and 2) a standard error $se\left(\hat{\beta}_n\right)$, and that returns a test statistic $T_n = |\hat{\beta}_n/se\left(\hat{\beta}_n\right)|$ and the corresponding p-value.

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

# Test the function using your solution to Problem 6
my_simpleteststat(coef[2] - 31, se)
```

e)

Write a function `my_simpleols` that takes two vectors, 1) a vector $\mathbf{y} \in \mathbb{R}^n$, and 2) a vector $\mathbf{x} \in \mathbb{R}^n$, and that returns a vector containing the ols-estimate $\hat{\beta}_n$, the standard error $se\left(\hat{\beta}_n\right)$, the test statistic T_n, and the corresponding p-value.

Your solution *must* make use of your functions defined in earlier parts: `my_simplecoef`, `my_simplese`, and `my_simpleteststat`.

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

# Test the function using your solution to Problem 6
my_simpleols(wkly_wage, yrs_educ)
```

References

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Manski, C. F. and Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociological methodology*, 28(1):99–137.