EC3303 Tutorial 1 Suggested Answers

1. Describe how one can design a hypothetical randomized control experiment to study the effect of having immigrant classmates on the test scores of native students. Suggest some impediments to implementing this experiment in practice.

Answer: There are many ways to design such an experiment. Here's one: At the beginning of the year, randomly assign immigrant and native students in a school to many classes. Then let students take a standardized exam at the end of the year. Then see whether there is a relationship between the test scores received by native students and the proportion of immigrant students in the class.

In practice, such an experiment may not provide unbiased estimates of the causal effect of having immigrant classmates. The reason is that parents of native children may move them from classes with many immigrants to classes with few immigrants if they believe that having immigrant classmates hurts the learning progress of their children. Such movements will invalidate the causal estimates.

2. Suppose that a researcher, using data on class size(CS) and average test scores from 100 primary 3 classes, estimates the OLS regression

$$(\widehat{TestScore} = 731.4 - 3.42 \times CS, \qquad R^2 = 0.10, \qquad SER = 11.0)$$

a) A classroom has 19 students. What is the regression's prediction for that classroom's average test score?
b) Last year, a classroom had 17 students and this year it has 22 students. What is the regression's prediction for the change in the classroom average test score?
c) The sample average class size across the 100 classroom is 21.4. What is the sample average of the test scores across the 100 classrooms?

Answer:

a) The predicted average test score is

$$\widehat{Testscore} = 731.4 - 3.42 \times 19 = 666.4$$

b) The predicted change in the classroom test score is

$$\Delta\widehat{Testscore} = -3.42 \times 22 - (-3.42 \times 17) = -75.24 + 58.14 = -17.1$$

c) We know that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$. So The sample average of the test scores across the 100 classroom is

$$\overline{Testscore} = \hat{\beta}_0 + \hat{\beta}_1 \times \overline{CS} = 731.4 - 3.42 \times 21.4 = 658.2$$

3. You are interested in examining the relationship between earnings and height. Accordingly you run a regression of *Earn* on *Height* using a sample of American workers (where the variable *Earn* represents annual labour earnings in US dollars in 2015; and where *Height* represents the height of the worker in inches in 2015). The height of individuals in your sample ranges from 48 inches to 84 inches.

   You obtained the following regression output:

```
. regress  Earn Height, robust

Linear regression                               Number of obs =    17870
                                                F(  1, 17868) =   197.19
                                                Prob > F      =   0.0000
                                                R-squared     =   0.0109
                                                Root MSE      =    26777
```

Earn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Height	707.6716	50.39502	14.04	0.000	608.8924	806.4507
_cons	-512.7336	3379.864	-0.15	0.879	-7137.594	6112.126

For all questions below, provide your **final answers** to 2 decimal places.

a) How much more or less do we expect a person whose height is 72 inches to earn in annual earnings compared to a person whose height is only 62 inches?

b) Suppose a person, Jane, actually earns $33,712.97 per year. Jane is 59 inches tall. How large is the residual specific for Jane?

Answer:

a) A person whose height is 72 inches is predicted to earn $707.6716(72 - 62) = \$7{,}076.72$ more in annual earnings than a person whose height is 62 inches.

b) Jane's predicted annual earnings is $-512.7336 + 707.6716 \times 59 = \$41{,}239.89$. The residual specific for Jane is $33,712.97 - $41,239.89 = - $7,526.92.

Stata Exercise (to be done in tutorial with the tutor)

4. The data file CPS08.dta contains data for full-time workers, aged 25-34, with a high school diploma (equivalent to Secondary, JC, and Poly qualifications) or Bachelor's as their highest degree. In this exercise, you will investigate the relationship between a worker's age and earnings (generally, older workers have more job experience, leading to higher productivity and earnings)

a) Run a regression of average hourly earnings (AHE) on age (Age). What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How much do earnings increase as workers age by 1 year?

b) Ah Teck is a 26 year-old worker. Predict Ah Teck's earnings using the estimated regression. Ravin is a 30 year-old worker. Predict Ravin's earnings using the estimated regression.

c) Does age account for a large fraction of the variance in earnings across individuals? Explain.

Answer:

a) $\widehat{AHE} = 1.08 + 0.60 \times Age$
Earnings increase, on average, by 0.6 dollars per hour when workers age by 1 year.

b) Ah Teck's predicted earnings = $1.08 + (0.60 \times 26) = \$16.68$
Ravin's predicted earnings = $1.08 + (0.60 \times 30) = \$19.08$

c) The regression R-squared is 0.03. This means that age explains a small fraction of the variability in earnings across individuals.

Supplementary Questions (serves as practice questions, but will not be discussed in the tutorial)

5. Labour economists are typically interested in knowing whether racial discrimination exists in the labour market. Discrimination is said to exist if someone is treated differently on the basis of his/her perceived race. Suppose you want to know whether being Asian as opposed to being Caucasian in Singapore affects your employment opportunities:

a) How would you design a potentially workable randomized controlled experiment to study this?
b) Can you compare the employment rates of Asian workers to employment rates of Caucasian workers in Singapore to reliably evaluate whether there is racial bias in hiring? Explain.

Answer:

a) There are many ways to do this. Here is one: Draw a random sample of say 1,000 companies from the population of companies in Singapore (or any other country). Next, create 2 fake resumes which differ only in the applicant's names (Terence Tan for one and William Smith for the other) and stated race (Chinese vs. Caucasian) but which are otherwise identical. Randomly send half of the 1,000 companies the resume with the Asian applicant and the other half the resume with the Caucasian applicant. Then, observe and record the number of times the Asian applicant receives an interview offer. Do the same for the Caucasian applicant. If the number of phone calls received by the Asian applicant is statistically significantly lower (higher) than the Caucasian applicant, then we conclude that being Asian lowers (enhances) the probability of getting a job interview (and hence a job, since getting called for an interview is a major first step in getting a job). Otherwise, we conclude that race has no effect on hiring decisions.

No. Employment rates between Asians and Caucasians may differ because of other differences in attributes. For instance, employment rates might be greater for Caucasians than Asians because Caucasians may have higher average levels of education or ability. Hence, any difference between employment rates, is in itself, at best weak evidence for racial discrimination in hiring.

6. In the review of statistics lecture, we showed that we could use the first observation (from a sample of *n*) as an estimator of the population mean. This first observation estimator is unbiased but has a variance of σ_Y^2, which makes it less efficient than the sample mean $\bar{Y}$. Suppose you now develop another estimator, which is the simple average of the first and last observation in your sample.

a) Show that this new estimator is also unbiased and show that it is more efficient than the estimator which only uses the first observation.

Let $\tilde{Y} = \frac{1}{2}(Y_1 + Y_n)$

$$E(\tilde{Y}) = E\left[\frac{1}{2}(Y_1 + Y_n)\right]$$

$$= \frac{1}{2} E(Y_1 + Y_n)$$

$$= \frac{1}{2}\left[E(Y_1) + E(Y_n)\right]$$

$$= \frac{1}{2}(\mu_Y + \mu_Y) \quad \text{Since } Y_1 \text{ and } Y_n \text{ are i.i.d.}$$

$$= \frac{2}{2}\mu_Y$$

$$= \mu_Y$$

$\therefore$ $\tilde{Y}$ is unbiased

b) Is this new estimator consistent?

Let $\tilde{Y} = \frac{1}{2}(Y_1 + Y_n)$

$$\begin{aligned}
\text{Var}(\tilde{Y}) &= \text{Var}\left[\tfrac{1}{2}(Y_1 + Y_n)\right] \\
&= \left(\tfrac{1}{2}\right)^2 \text{Var}(Y_1 + Y_n) \\
&= \tfrac{1}{4}\left[\text{Var}(Y_1) + \text{Var}(Y_n)\right] \quad \text{Since } Y_1 \text{ \& } Y_n \text{ are independent} \\
&= \tfrac{1}{4}\left(\sigma_Y^2 + \sigma_Y^2\right) \quad \text{Since } Y_1 \text{ \& } Y_n \text{ are identically distributed} \\
&= \tfrac{1}{4}\left(2\sigma_Y^2\right) \\
&= \frac{\sigma_Y^2}{2}
\end{aligned}$$

$\therefore$ $\tilde{Y}$ is not consistent since as $n \to \infty$, $\text{Var}(\tilde{Y})$ does not tend to 0. //

$$\text{Var}(Y_1) = \sigma_Y^2$$

$\therefore$ Since $\text{Var}(\tilde{Y}) = \frac{\sigma_Y^2}{2} < \text{Var}(Y_1) = \sigma_Y^2$, $\tilde{Y}$ is a more efficient estimator of the population mean than Y_1.

7. Your friend has developed a new estimator $\tilde{Y}$ to estimate the population mean.

$$\tilde{Y} = \frac{1}{n}\left(\frac{421}{2}Y_1 + \frac{123}{50}Y_2 + \frac{421}{2}Y_3 + \frac{123}{50}Y_4 + \cdots + \frac{421}{2}Y_{n-1} + \frac{123}{50}Y_n\right)$$

where $Y_1, Y_2, \ldots\ldots, Y_n$ are i.i.d observations, drawn from a population with mean μ_Y and variance σ_Y^2. Assume *n* is even.

a) Is the estimator, $\tilde{Y}$, unbiased? Show **clearly** how you reached your conclusion.

Answer: $E(\tilde{Y}) = E(\frac{1}{n}\left(\frac{421}{2}Y_1 + \frac{123}{50}Y_2 + \frac{421}{2}Y_3 + \frac{123}{50}Y_4 + \cdots + \frac{421}{2}Y_{n-1} + \frac{123}{50}Y_n\right)$

$$= \frac{1}{n}E\left(\frac{421}{2}Y_1 + \frac{421}{2}Y_3 + \cdots + \frac{421}{2}Y_{n-1} + \frac{123}{50}Y_2 + \frac{123}{50}Y_4 + \cdots + \frac{123}{50}Y_n\right)$$

$$= \frac{1}{n}E\left(\frac{421}{2}Y_1 + \frac{421}{2}Y_3 + \cdots + \frac{421}{2}Y_{n-1}\right) + \frac{1}{n}E\left(\frac{123}{50}Y_2 + \frac{123}{50}Y_4 + \cdots + \frac{123}{50}Y_n\right)$$

$$= \frac{1}{n}.\frac{421}{2}E(Y_1 + Y_3 + \cdots + Y_{n-1}) + \frac{1}{n}.\frac{123}{50}E(Y_2 + Y_4 + \cdots + Y_n)$$

$$= \frac{1}{n}.\frac{421}{2}[E(Y_1) + E(Y_3) + \cdots + E(Y_{n-1})] + \frac{1}{n}.\frac{123}{50}[E(Y_2) + E(Y_4) + \cdots + E(Y_n)]$$

$$= \frac{1}{n}.\frac{421}{2}\left[\frac{n}{2}\mu_Y\right] + \frac{1}{n}.\frac{123}{50}\left[\frac{n}{2}\mu_Y\right]$$

$$= 106\frac{12}{25}\mu_Y \text{ or } \frac{2662}{25}\mu_Y \neq \mu_Y$$

$\tilde{Y}$ is biased.

b) Derive an expression for the variance of the sampling distribution of $\tilde{Y}$, in terms of σ_Y^2.

Answer: $Var(\tilde{Y}) = Var(\frac{1}{n}\left(\frac{421}{2}Y_1 + \frac{123}{50}Y_2 + \frac{421}{2}Y_3 + \frac{123}{50}Y_4 + \cdots + \frac{421}{2}Y_{n-1} + \frac{123}{50}Y_n\right)$

$$= \left(\frac{1}{n}\right)^2 Var(\left(\frac{421}{2}Y_1 + \frac{123}{50}Y_2 + \frac{421}{2}Y_3 + \frac{123}{50}Y_4 + \cdots + \frac{421}{2}Y_{n-1} + \frac{123}{50}Y_n\right)$$

$$= \left(\frac{1}{n}\right)^2 Var(\left(\frac{421}{2}Y_1 + \frac{421}{2}Y_3 + \cdots + \frac{421}{2}Y_{n-1} + \frac{123}{50}Y_2 + \frac{123}{50}Y_4 + \cdots + \frac{123}{50}Y_n\right)$$

$$= \left(\frac{1}{n}\right)^2 Var\left(\frac{421}{2}Y_1 + \frac{421}{2}Y_3 + \cdots + \frac{421}{2}Y_{n-1}\right) + \left(\frac{1}{n}\right)^2 Var\left(\frac{123}{50}Y_2 + \frac{123}{50}Y_4 + \cdots + \frac{123}{50}Y_n\right)$$

Since the observations are independent, covariance between the Y_is are 0.

$$= \left(\frac{1}{n}\right)^2 \left(\frac{421}{2}\right)^2 Var(Y_1 + Y_3 + \cdots + Y_{n-1}) + \left(\frac{1}{n}\right)^2 \left(\frac{123}{50}\right)^2 Var(Y_2 + Y_4 + \cdots + Y_n)$$

$$= \left(\frac{1}{n}\right)^2 \left(\frac{421}{2}\right)^2 ((Var(Y_1) + Var(Y_3) + \cdots + Var(Y_{n-1})) + \left(\frac{1}{n}\right)^2 \left(\frac{123}{50}\right)^2 ((Var(Y_2) + Var(Y_4) + \cdots + Var(Y_n))$$

$$= \left(\frac{1}{n}\right)^2 \left(\frac{421}{2}\right)^2 (\frac{n}{2}\sigma_Y^2) + \left(\frac{1}{n}\right)^2 \left(\frac{123}{50}\right)^2 (\frac{n}{2}\sigma_Y^2)$$

$$= 22{,}158.15.\frac{\sigma_Y^2}{n}$$

c) Suppose your friend now develops another estimator of the population mean:

$$\ddot{Y} = \frac{1}{n}\left(\frac{57}{20}Y_1 + \frac{25}{30}Y_2 + \frac{57}{20}Y_3 + \frac{25}{30}Y_4 + \cdots + \frac{57}{20}Y_{n-1} + \frac{25}{30}Y_n\right)$$

Should your friend use $\ddot{Y}$ or $\tilde{Y}$ to estimate the population mean if she had to pick one? Explain your answer clearly.

Answer: $E(\ddot{Y}) = E(\frac{1}{n}\left(\frac{57}{20}Y_1 + \frac{25}{30}Y_2 + \frac{57}{20}Y_3 + \frac{25}{30}Y_4 + \cdots + \frac{57}{20}Y_{n-1} + \frac{25}{30}Y_n\right))$

$$= \frac{1}{n}E\left(\frac{57}{20}Y_1 + \frac{57}{20}Y_3 + \cdots + \frac{57}{20}Y_{n-1} + \frac{25}{30}Y_2 + \frac{25}{30}Y_4 + \cdots + \frac{25}{30}Y_n\right)$$

$$= \frac{1}{n}E\left(\frac{57}{20}Y_1 + \frac{57}{20}Y_3 + \cdots + \frac{57}{20}Y_{n-1}\right) + \frac{1}{n}E\left(\frac{25}{30}Y_2 + \frac{25}{30}Y_4 + \cdots + \frac{25}{30}Y_n\right)$$

$$= \frac{1}{n}.\frac{57}{20}E(Y_1 + Y_3 + \cdots + Y_{n-1}) + \frac{1}{n}.\frac{25}{30}E(Y_2 + Y_4 + \cdots + Y_n)$$

$$= \frac{1}{n}.\frac{57}{20}[E(Y_1) + E(Y_3) + \cdots + E(Y_{n-1})] + \frac{1}{n}.\frac{25}{30}[E(Y_2) + E(Y_4) + \cdots + E(Y_n)]$$

$$= \frac{1}{n}.\frac{57}{20}\left[\frac{n}{2}\mu_Y\right] + \frac{1}{n}.\frac{25}{30}\left[\frac{n}{2}\mu_Y\right]$$

$$= \frac{221}{120}\mu_Y$$

Answer: $Var(\ddot{Y}) = Var(\frac{1}{n}\left(\frac{57}{20}Y_1 + \frac{25}{30}Y_2 + \frac{57}{20}Y_3 + \frac{25}{30}Y_4 + \cdots + \frac{57}{20}Y_{n-1} + \frac{25}{30}Y_n\right)$

$$= \left(\frac{1}{n}\right)^2 Var(\left(\frac{57}{20}Y_1 + \frac{25}{30}Y_2 + \frac{57}{20}Y_3 + \frac{25}{30}Y_4 + \cdots + \frac{57}{20}Y_{n-1} + \frac{25}{30}Y_n\right)$$

$$= \left(\frac{1}{n}\right)^2 Var(\left(\frac{57}{20}Y_1 + \frac{57}{20}Y_3 + \cdots + \frac{57}{20}Y_{n-1} + \frac{25}{30}Y_2 + \frac{25}{30}Y_4 + \cdots + \frac{25}{30}Y_n\right)$$

$$= \left(\frac{1}{n}\right)^2 Var\left(\frac{57}{20}Y_1 + \frac{57}{20}Y_3 + \cdots + \frac{57}{20}Y_{n-1}\right) + \left(\frac{1}{n}\right)^2 Var\left(\frac{25}{30}Y_2 + \frac{25}{30}Y_4 + \cdots + \frac{25}{30}Y_n\right)$$

Since the observations are independent, covariance between the Y_is are 0.

$$= \left(\frac{1}{n}\right)^2 \left(\frac{57}{20}\right)^2 Var(Y_1 + Y_3 + \cdots + Y_{n-1}) + \left(\frac{1}{n}\right)^2 \left(\frac{25}{30}\right)^2 Var(Y_2 + Y_4 + \cdots + Y_n)$$

$$= \left(\frac{1}{n}\right)^2 \left(\frac{57}{20}\right)^2 ((Var(Y_1) + Var(Y_3) + \cdots + Var(Y_{n-1})) + \left(\frac{1}{n}\right)^2 \left(\frac{25}{30}\right)^2 ((Var(Y_2) + Var(Y_4) + \cdots + Var(Y_n))$$

$$= \left(\frac{1}{n}\right)^2 \left(\frac{57}{20}\right)^2 (\frac{n}{2}\sigma_Y^2) + \left(\frac{1}{n}\right)^2 \left(\frac{25}{30}\right)^2 (\frac{n}{2}\sigma_Y^2)$$

$$= 4.408\frac{\sigma_Y^2}{n}$$

Since the bias in $\ddot{Y}$ is smaller than the bias in $\tilde{Y}$. Also $Var(\ddot{Y}) < Var(\tilde{Y})$, $\ddot{Y}$ is a better estimator than $\tilde{Y}$. Therefore, use $\ddot{Y}$.