Problem Set 1

Sample Solutions

10/23/2018

Question 2: Child Health and School Attendance

2. Child Health and School Attendance

Part a)

> Why was the health intervention in the Primary School Deworming Project (Miguel and Kremer 2004) randomized across schools? How does the project's randomized design affect the estimation of treatment effects and help address omitted variable bias? (Please illustrate your points with the discussion presented in class, including the equations that illustrate potential bias.) [2 points]

Answer

The health intervention in the Primary School Deworming Project was randomized across schools (as opposed to within schools) because within school randomization may understate the effectiveness of deworming drugs if untreated students experience health gains due to local treatment externalities.

The randomized design implies that (if externality effects are localized within schools), treatment effects can be estimated by computing the difference in outcomes between pupils in treatment schools and pupils in comparison schools. The randomization of the program helped to address omitted variable bias (OVB) by equalizing the expected values of unobserved variables for treatment and comparison schools. In this case, estimating the treatment effect by computing the difference in outcomes between the treatment and comparison schools will not be biased.

We illustrate how randomization affects OVB with the following example from lecture. Consider the following estimation equation:

$$y_i = a + bT_i + cX_i + e_i$$

where y_i is the average test score in school i, T_i is a dummy variable indicating whether school i received treatment, X_i is a measure of average student ability in school i that is unobserved by the researcher, and e_i is white noise error (with mean zero, $E(e_i) = 0$).

We can compute a naive estimator by taking the difference in expected outcomes, conditional on treatment assignment:

$$\begin{aligned}
E[y_i|T=1] - E[y_i|T=0] =& E[a + bT_i + cX_i + e_i|T=1] - E[a + bT_i + cX_i + e_i|T=0] \\
=& E[a|T=1] + E[bT_i|T=1] + E[cX_i|T=1] + E[e_i|T=1] \\
& - (E[a|T=0] + E[bT_i|T=0] + E[cX_i|T=0] + E[e_i|T=0]) \\
=& a + b + cE[X_i|T=1] + 0 - a + 0 - cE[X_i|T=0] - 0 \\
=& b + c\{E[X_i|T=1] - E[X_i|T=0]\}
\end{aligned}$$

With randomization of treatment, we expect students in treated and non-treated schools not to differ in their average ability, so $E[X_i|T=1] = E[X_i|T=0] = E[X_i]$. Thus, the bias term in the equation above drops out and our estimator becomes:

$$E[y_i|T=1] - E[y_i|T=0] = b$$

Part b)

Download `Econ-172_PSet1-data.csv` from the bCourses page, and use the `read.csv` command to open it in R (or RStudio). Using the `lm` command, determine the average difference between 1998 treatment schools (`group1=1`) and comparison schools (`group1=0`) in the following three characteristics: Proportion of female children (`female`), Average child year of birth (`yob`), and Involvement in another school assistance program (`sap`).

Then separately determine the average difference between Group 2 (`group2=1`) versus Group 3 (`group3=1`) schools in the same three characteristics. (In this analysis, restrict attention to just these two groups of schools using a logic statement.)

Report the regression output for the six regressions, and interpret the coefficients. Please also discuss the standard errors and t-statistics. You should electronically turn in the actual R "history" file print-out (including the file time stamp and computer directory path, etc.) along with formatted tables using the output from "stargazer" or a similar command. Taken together, did the randomization appear to succeed in creating comparable groups at baseline? [1 point]

Answer

Loading the data and presenting descriptive statistics

```
worms <- read.csv("data/Econ-172_PSet1-S17-data.csv")
worms2 <- read.csv("data/Econ-172_PSet1-Spring2018-data.csv")

tab1 <- worms %>%
  select(-"schid") %>%
  descr(stats = c("mean", "sd", "min", "med", "max"),
        transpose = TRUE, omit.headings = TRUE)
```

This data set has 74 observations and 8 variables. Table 1 presents descriptive statistics for all the variables excluding the identifier (`schid`).

Table 1: Table 1: Descriptive Statistics

	Mean	Std.Dev	Min	Median	Max
group1	0.34	0.48	0.00	0.00	1.00
group2	0.34	0.48	0.00	0.00	1.00
group3	0.32	0.47	0.00	0.00	1.00
part98	0.80	0.16	0.39	0.81	1.00
female	0.38	0.08	0.28	0.38	0.98
yob	1986.75	0.70	1985.35	1986.67	1988.82
sap	0.36	0.48	0.00	0.00	1.00

Assesing balance of covariates

To start, we want to estimate the following equation:

$$Y = a + bX + e$$

where Y is the characteristic of interest (`female, yob, sap`), and X is group1 or group2. In this equation,

b describes the average difference between treatment and comparison schools, and a gives us the average value of the variable in comparison schools (the group left out, which is group1==0 in the first regression, or group2==0 in the second).

```
m1.1=lm(female ~ group1, data=worms)
m1.2=lm(yob ~ group1, data=worms)
m1.3=lm(sap ~ group1, data=worms)

m1.4=lm(female ~ group2, data=worms[worms$group1==0,])
m1.5=lm(yob ~ group2, data=worms[worms$group1==0,])
m1.6=lm(sap ~ group2, data=worms[worms$group1==0,])

m1.7=lm(female ~ group2, data=subset(worms,group1==0))
```

```
stargazer::stargazer(m1.1,m1.2,m1.3,m1.4,m1.5,m1.6, type = "text",
          title="Table 2: Covariate Analysis", align = TRUE,
          omit.stat=c("LL","ser","f","adj.rsq", "rsq"), no.space=TRUE,
          header=FALSE, column.sep.width = "0pt")
```

Table 2: Covariate Analysis

	Dependent variable:					
	female	yob	sap	female	yob	sap
	(1)	(2)	(3)	(4)	(5)	(6)
group1	-0.007	0.005	-0.128			
	(0.020)	(0.173)	(0.119)			
group2	0.036	0.779***	0.147			
	(0.027)	(0.190)	(0.142)			
Constant	0.386***	1,986.746***	0.408***	0.368***	1,986.348***	0.333***
	(0.011)	(0.100)	(0.069)	(0.019)	(0.136)	(0.101)

Observations 74 74 74 49 49 49

Note: *p<0.1;* ***p<0.05;*** p<0.01

Notice that this is equivalent to running a set of t-test comparisons of means of the three different outcomes between the two corresponding groups (group 1 v. groups 2+3 and group 2 v. group 3). For example the code below computes the difference between group 1 and groups 2+3 for the outcome variable `female`

```
t1 <- t.test(female ~ group1, data=worms, var.equal = TRUE)
est_t1 <- diff(t1$estimate)
st_t1 <- t1$statistic
```

This gives us a estimated difference of -0.0066 and a t-statistic of 0.3391, wich correspond to colum 1 and row 1 of table 2 (the t-statistic is the estimated difference divided by the standard error: -0.0066 / 0.0195 = -0.3391)

Group 1 vs. Groups 2 and 3 [replace hard coded numbers]

The proportion of female children in comparison schools is 38.6%, and the average difference in the proportion of female students between treatment and comparison schools is –0.7 percentage points. Thus, treatment schools have on average 37.9% female students.

The average child year of birth in comparison schools is approximately 1986.7, and the average difference in the child year of birth between treatment and comparison schools is 0.005 years, so the average year of birth of children in treatment schools is 1986.7.

The average participation in other school assistance programs (sap) is 40.8% for the control group and the average difference between treatment and comparison is -12.8 percentage points. Treatment schools had on average 28% participation in other programs

We can calculate t-statistics for the coefficients using the estimated coefficients (b) and the estimated standard errors (se(b)) provided in the regression output. The formula is: t = b/se(b). The t-statistics on the three estimated average differences are substantially less than 1.96 in absolute value, so the differences are not statistically significant at the 95% confidence level. We can thus say that the randomization succeeded in creating comparable groups.

Group 2 vs. Group 3

In the comparison of average proportion of females, we obtain that schools in Group 2 have on average 3.75 percentage points higher proportion, but the t-statistic is less than 1.96, so the difference is not statistically significant.

However, students in Group 2 schools were born on average 0.779 years later than those in Group 3, so they are about 9 months younger. This difference is statistically significant, with a standard error of 0.190 and a t-statistic greater than the critical value 1.96.

Finally, schools in group2 had on average a 14.7 percentage point higher participation in the other school program but the difference is not significant.

We conclude from this second part of the analysis that Groups 2 and 3, which are both the control group in this setting as of 1998, may not be perfectly comparable: even though the division into these groups was randomized and they have a similar proportion of females and sap, schools in Group 2 have significantly younger students on average than schools in Group 3. Thus it may be important to control for these characteristics in the analysis.

This does not mean however that randomization has failed: when comparing characteristics between treatment and control groups and using the 95% significance cutoff, on average 5% of the comparisons will still produce significant differences.

Part c)

> Determine the average difference between treatment and comparison schools in: Average school participation in 1998 after the program had started (`part98`).
>
> Report the regression results – you should again electronically turn in the history file and formatted tables – and interpret the coefficients. What is the impact of attending a treatment school on average school participation? Is this significantly different than zero at 95% confidence? Then determine the average difference between Group 2 and Group 3 schools in `part98` (again restricting analysis to just these two groups). Would you expect to find differences between Groups 2 and 3, and do you find any? [1 point]

Answer