Problem Set 1 - Solutions

Statistics and Econometrics

Question 1

Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved performance of students in Year 6. (fyi, "in schools in England Year 6 is the sixth year after Reception. It is the sixth full year of compulsory education, with children being admitted who are aged 10 before 1 September in any given academic year. It is also the final year of Key Stage 2 in which the National Curriculum is taught." source: Wikipedia)

1. Suppose you can collect observational data on several thousands of Year-6 students in England. You can obtain the size of their class and a standardized test score taken at the end of Year 6. Would you expect any correlation between class size and test score?
2. Suppose you find a negative correlation between class size and test score. Does it necessarily indicate that smaller class sizes cause better performance? Explain.

Solutions

1. One might expect a negative correlation between class size and test score, meaning that a larger class size is associated with lower performance. We might find a negative correlation because a larger class size actually hurts performance.
2. However, with observational data, there are other reasons we might find a negative relationship. For example, children from more affluent families might be more likely to attend schools with smaller class sizes, and affluent children generally might score better on standardized tests. Another possibility is that, within a school, a principal might assign the better students to smaller classes. Or, some parents might insist their children to be placed in smaller classes, and these same parents tend to be more involved in their children's education. Given the potential for confounding factors, finding a negative correlation would not be strong evidence that smaller class sizes actually lead to better performance. Some way of controlling for the confounding factors is needed, and this is the subject of multiple regression analysis.

Question 2

The data in fertil2.RData were collected on women living in the Republic of Botswana in 1988. The variable *children* refers to the number of living children. The variable *electric* is a binary indicator equal to one if the woman's home has electricity, and zero if not. The variable *heduc* refers to the husband's years of education.

1. Find the smallest and largest values of children in the sample. What is the average of children?
2. Find the average of husband's years of education in the sample. How many observations are used to compute this average? Explain.
3. Create a graph to examine the relationship between *children* and *heduc*. Comment.
4. What percentage of women have electricity in the home?
5. Compute the average of children for those without electricity and do the same for those with electricity. Comment on what you find.
6. Estimate the simple regression model

$$children = \beta_0 + \beta_1 electric + u,$$

and report your results.

7. Does this simple regression necessarily capture a causal relationship between the number of children and the presence of electricity in the home? Explain.

Solutions

1.

```
load("fertil2.RData")
max(data$children)
```

```
## [1] 13
```

```
min(data$children)
```

```
## [1] 0
```

```
mean(data$children)
```

```
## [1] 2.267828
```

The smallest and largest values of *children* are 0 and 13, respectively. The average is about 2.27.

2.

```
sum(!complete.cases(data$heduc))
```

```
## [1] 2405
```

```
sum(complete.cases(data$heduc))
```

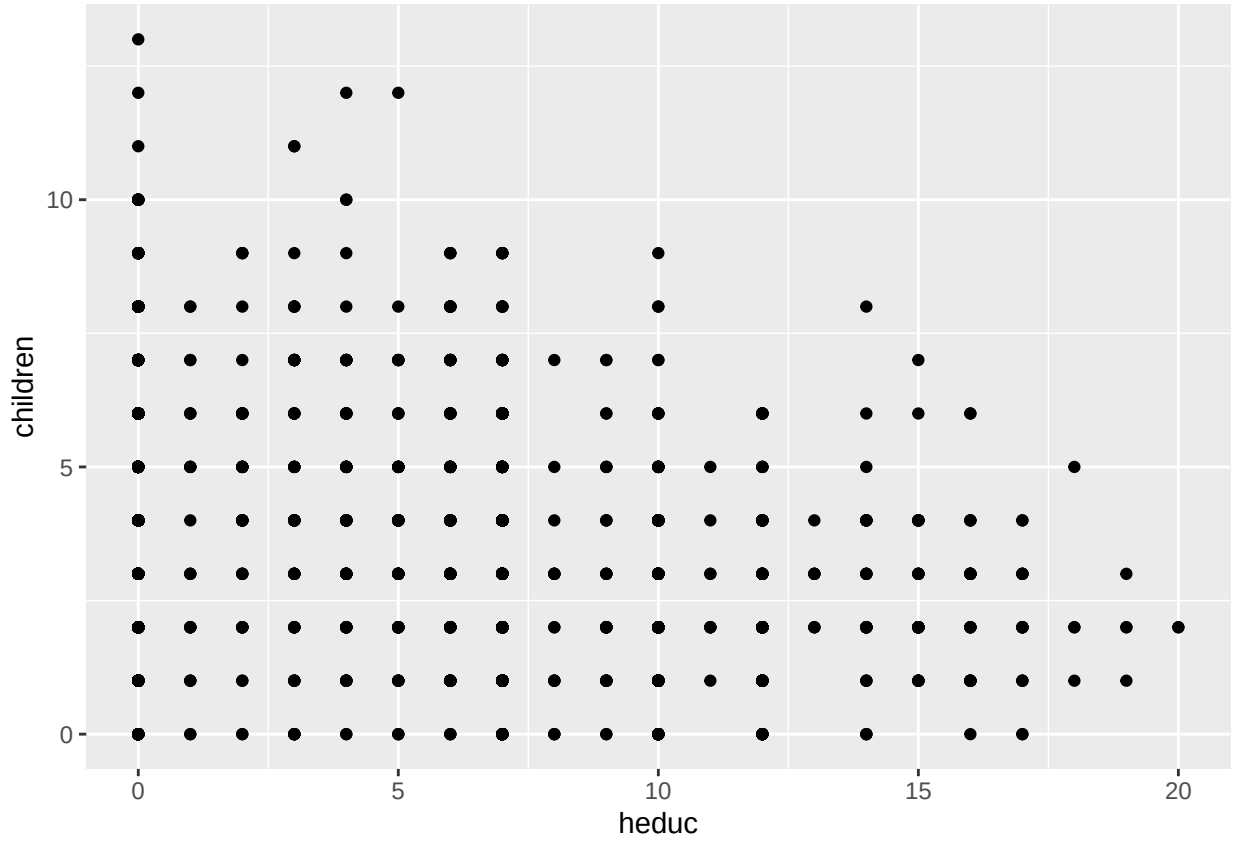
```
## [1] 1956
```

Only 1,956 observations are used to compute this average. This is because there are 2,405 observations with missing values in husband's education level.

3.

```
ggplot(data = data, aes(x = heduc, y = children)) + geom_point()
```

```
## Warning: Removed 2405 rows containing missing values (geom_point).
```



A couple of observations: (1) overall, there is a downward trend in the number of children as husband's education level increases; (2) variations in the number of children also decreases as husband's education level increases, partly driven by fewer observations at high education level.

4.

```
sum(data$electric == 1, na.rm = TRUE)
```

```
## [1] 611
```

```
sum(data$electric == 0, na.rm = TRUE)
```

```
## [1] 3747
```

Out of 4,358 women, only 611 have electricity in the home, or about 14.02 percent.

5.

```
with.electric <- data %>% filter(electric == 1)
mean(with.electric$children)
```

```
## [1] 1.898527
```

```
without.electric <- data %>% filter(electric == 0)
mean(without.electric$children)
```

```
## [1] 2.327729
```

The average of *children* for women without electricity is about 2.33, and for those with electricity it is about 1.90. So, on average, women with electricity have .43 fewer children than those who do not.

6.

```r
fitted.model <- lm(children ~ electric, data = data)
stargazer(fitted.model, header = FALSE, type = 'latex', title = "Question 2.6")
```

Table 1: Question 2.6

	Dependent variable:
	children
electric	-0.429^{***}
	(0.097)
Constant	2.328^{***}
	(0.036)
Observations	4,358
R^2	0.004
Adjusted R^2	0.004
Residual Std. Error	2.217 (df = 4356)
F Statistic	19.686^{***} (df = 1; 4356)
Note:	*p<0.1; **p<0.05; ***p<0.01

7. We cannot infer causality here. There are many confounding factors that may be related to the number of children and the presence of electricity in the home; household income and level of education are two possibilities. For example, it could be that women with more education have fewer children and are more likely to have electricity in the home (the latter due to an income effect).