Problem Set 4 - Solutions

Statistics and Econometrics

Question 1

You need to use two data sets for this exercise, jtrain2.RData and jtrain3.RData. jtrain2 was obtained from the National Supported Work Demonstration job-training program conducted by the Manpower Demonstration Research Corporation in the mid 1970s in US. Training status was randomly assigned, so this is essentially experimental data. On the other hand, jtrain3 contains observational data, where individuals themselves largely determine whether they participate in job training. The data sets cover the same time period.

Source:

jtrain2: R.J. Lalonde (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, 604-620.

jtrain3: R.H. Dehejia and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94, 1053-1062

1. In the data set jtrain2, what fraction of the men received job training? What is the fraction in jtrain3? Why do you think there is such a big difference?
2. Using jtrain2, run a simple regression of $re78$ on $train$. What is the estimated effect of participating in job training on real earnings?
3. Now add as controls to the regression in part 2 the variables $re74$, $re75$, $educ$, age, $black$, and $hisp$. Does the estimated effect of job training on re78 change much? How come?
4. Do the regressions in parts 2 and 3 using the data in jtrain3, and report the results. What is the effect now of controlling for the extra factors, and why?
5. Define $avgre = (re74 + re75)/2$. Find the sample averages, standard deviations, and minimum and maximum values in the two data sets. Are these data sets representative of the same populations in 1978?
6. Almost 96% of men in the data set jtrain2 have $avgre$ less than $10,000. Using only these men, regress $re78$ on $train$, $re74$, $re75$, $educ$, age, $black$, and $hisp$ and report the result. Run the same regression for jtrain3, using only men with $avgre \leq 10$. For the subsample of low-income men, how do the estimated training effects compare across the experimental and nonexperimental data sets?
7. Now use each data set to run the simple regression $re78$ on $train$, but only for men who were unemployed in 1974 and 1975. How do the training estimates compare now?
8. Using your findings from the previous regressions, discuss the potential importance of having comparable populations underlying comparisons of experimental and nonexperimental estimates.

Solutions

1.

```r
load("jtrain2.RData")
jtrain2.data <- data
mean(jtrain2.data$train)
```

```
## [1] 0.4157303
```

```r
load("jtrain3.RData")
jtrain3.data <- data
mean(jtrain3.data$train)
```

```
## [1] 0.06915888
```

About .416 of the men receive training in JTRAIN2, whereas only .069 receive training in JTRAIN3. The men in JTRAIN2, who were low earners, were targeted to receive training in a special job training experiment. This is not a representative group from the entire population. The sample from JTRAIN3 is, for practical purposes, a random sample from the population of men working in 1978; we would expect a much smaller fraction to have participated in job training in the prior year.

2.

```r
jtrain2.m1 <- lm(re78 ~ train, data = jtrain2.data)
jtrain2.m2 <- lm(re78 ~ train + re74 + re75 + educ
                 + age + black + hisp, data = jtrain2.data)
jtrain3.m1 <- lm(re78 ~ train, data = jtrain3.data)
jtrain3.m2 <- lm(re78 ~ train + re74 + re75 + educ
                 + age + black + hisp, data = jtrain3.data)
stargazer(jtrain2.m1, jtrain2.m2, jtrain3.m1, jtrain3.m2, font.size = "small",
          header = FALSE, type = 'latex', title = 'Questions 1.2 and 1.3')
```

Because *re*78 is measured in thousands, job training participation is estimated to increase real earnings in 1978 by $1,794 - a nontrivial amount.

3. Adding all of the control listed changes the coefficient on *train* to 1.680 (se = .631). This is not much of a change from part 2, and we would not expect it to be. Because train was supposed to be assigned randomly, it should be roughly uncorrelated with all other explanatory variables. Therefore, the simple and multiple regression estimates are similar. (Interestingly, the standard errors are the same to two decimal places.)

4. The simple regression coefficient on *train* is -15.205 (se = 1.155). This implies a huge negative effect of job training, which is hard to believe. Because training was not randomly assigned for this group, we can assume self-selection into job training is at work. That is, it is the low earning group that tends to select itself (perhaps with the help of administrators) into job training. When we add the controls, the coefficient becomes .213 (se = .853). In other words, when we account for factors such as previous earnings and education, we obtain a small but insignificant positive effect of training. This is certainly more believable than the large negative effect obtained from simple regression.

5.

```r
jtrain2.data <- jtrain2.data %>% mutate(avgre = (re74 + re75)/2)
summary(jtrain2.data$avgre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.740   1.492  24.376
```

```r
sd(jtrain2.data$avgre)
```

```
## [1] 3.900095
```

```r
summary(jtrain3.data$avgre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.829  16.873  18.040  25.257 146.901
```

```r
sd(jtrain3.data$avgre)
```

```
## [1] 13.29345
```

Table 1: Questions 1.2 and 1.3

	Dependent variable:			
	re78			
	(1)	(2)	(3)	(4)
train	1.794***	1.680***	−15.205***	0.213
	(0.633)	(0.631)	(1.155)	(0.853)
re74		0.083		0.281***
		(0.077)		(0.028)
re75		0.047		0.569***
		(0.131)		(0.028)
educ		0.404**		0.520***
		(0.175)		(0.075)
age		0.054		−0.075***
		(0.044)		(0.020)
black		−2.180*		−0.648
		(1.156)		(0.492)
hisp		0.144		2.203**
		(1.541)		(1.093)
Constant	4.555***	0.674	21.554***	1.648
	(0.408)	(2.423)	(0.304)	(1.301)
Observations	445	445	2,675	2,675
R^2	0.018	0.055	0.061	0.586
Adjusted R^2	0.016	0.040	0.061	0.584
Residual Std. Error	6.580 (df = 443)	6.499 (df = 437)	15.152 (df = 2673)	10.077 (df = 2667)
F Statistic	8.039*** (df = 1; 443)	3.617*** (df = 7; 437)	173.415*** (df = 1; 2673)	538.356*** (df = 7; 2667)
Note:				*p<0.1; **p<0.05; ***p<0.01

For JTRAIN2, the average is 1.740, the standard deviation is 3.900, the minimum is 0, and the maximum is 24.376. For JTRAIN3, the average is 18.040, the standard deviation is 13.293, the minimum is 0, and the maximum is 146.901. Clearly these samples are not representative of the same population. JTRAIN3, which represents a much broader population, has a much larger mean value and a much larger standard deviation.

6.

```
jtrain2.low <- jtrain2.data %>% filter(avgre < 10)
jtrain3.low <- jtrain3.data %>% filter(avgre <= 10)
jtrain2.m3 <- lm(re78 ~ train + re74 + re75 + educ
                 + age + black + hisp, data = jtrain2.low)
jtrain3.m3 <- lm(re78 ~ train + re74 + re75 + educ
                 + age + black + hisp, data = jtrain3.low)
jtrain2.unem <- jtrain2.data %>% filter(unem74 == 1 & unem75 == 1)
jtrain3.unem <- jtrain3.data %>% filter(unem74 == 1 & unem75 == 1)
jtrain2.m4 <- lm(re78 ~ train, data = jtrain2.unem)
jtrain3.m4 <- lm(re78 ~ train, data = jtrain3.unem)
stargazer(jtrain2.m3, jtrain3.m3, jtrain2.m4, jtrain3.m4, font.size = "small",
          header = FALSE, type = 'latex', title = 'Questions 1.6 and 1.7')
```

Table 2: Questions 1.6 and 1.7

	Dependent variable:			
	re78			
	(1)	(2)	(3)	(4)
train	1.583^{**}	1.844^{**}	1.842^{***}	3.803^{***}
	(0.632)	(0.893)	(0.689)	(0.884)
re74	-0.117	0.313^{***}		
	(0.124)	(0.069)		
re75	0.173	0.774^{***}		
	(0.189)	(0.076)		
educ	0.358^{**}	0.328^{***}		
	(0.176)	(0.110)		
age	0.044	-0.083^{***}		
	(0.044)	(0.031)		
black	-2.384^{**}	-1.973^{***}		
	(1.168)	(0.721)		
hisp	-0.369	-1.101		
	(1.551)	(1.432)		
Constant	1.737	3.448	4.112^{***}	2.151^{***}
	(2.446)	(2.141)	(0.430)	(0.560)
Observations	427	765	280	271
R^2	0.046	0.234	0.025	0.064
Adjusted R^2	0.030	0.227	0.022	0.061
Residual Std. Error	6.377 (df = 419)	7.962 (df = 757)	5.623 (df = 278)	7.134 (df = 269)
F Statistic	2.904^{***} (df = 7; 419)	33.107^{***} (df = 7; 757)	7.144^{***} (df = 1; 278)	18.520^{***} (df = 1; 269)
Note:				*p<0.1; **p<0.05; ***p<0.01

For JTRAIN2, which uses 427 observations, the estimate on *train* is similar to before, 1.583 (se = .632). For JTRAIN3, which uses 765 observations, the estimate is now much closer to the experimental estimate: 1.844 (se = .893).

7. The estimate for JTRAIN2, which uses 280 observations, is 1.842 (se = .689); it is a coincidence that this is the same, to two digits, as that obtained for JTRAIN3 in part 6. For JTRAIN3, which uses 271 observations, the estimate is 3.803 (se = .884).
8. When we base our analysis on comparable samples – those with average real earnings less than $10,000 in 1974 and 1975, roughly representative of the same population – we get positive, nontrivial training effects estimates using either sample. Using the full data set in JTRAIN3 can be misleading because it includes many men for whom training would never be beneficial. In effect, when we use the entire data set, we average in the zero effect for high earners with the positive effect for low-earning men. Of course, if we only have experimental data, it can be difficult to know how to find the part of the population where there is an effect. But for those who were unemployed in the two years prior to job training, the effect appears to be unambiguously positive.

Question 2

Consider a regression model $y = \beta_0 + u$. Suppose we have a variable x, and we want to decide whether or not to include it as an independent variable. Suppose that the model including x, i.e., $y = \beta_0 + \beta_1 x + u$, has a lower AIC than the model with only the intercept β_0. Is it possible that x is statistically insignificant even at 10% level in the model $y = \beta_0 + \beta_1 x + u$? Explain.

Solutions

The answer is yes. Consider the case when we have a sample of $n = 102$ observations.

Define the unrestricted model and the restricted model as $y = \beta_0 + \beta_1 x + u$ and $y = \beta_0 + u_r$, respectively. Including x leads to a lower AIC implies that

$$\begin{aligned} & AIC_{ur} - AIC_r < 0 \\ \Leftrightarrow\quad & n \ln\left(SSR_{ur}/n\right) + 4 - n \ln\left(SSR_r/n\right) - 2 < 0 \\ \Leftrightarrow\quad & n \ln\left(SSR_{ur}/SSR_r\right) + 2 < 0 \\ \Leftrightarrow\quad & \frac{SSR_r}{SSR_{ur}} > \exp(2/n) \approx 1.02. \end{aligned}$$

On the other hand, if we test the significance of x with an F test, the F statistic is given by

$$\text{F statistic} : \frac{(SSR_r - SSR_{ur})/1}{SSR_{ur}/(n-k-1)} = 100\left(\frac{SSR_r}{SSR_{ur}} - 1\right).$$

We fail to reject the null $H_0 : \beta_1 = 0$ at the 10% level if and only if

$$\begin{aligned} & 100\left(\frac{SSR_r}{SSR_{ur}} - 1\right) < c = 3.936 \text{ (The 10\% } F_{1,100} \text{ critical value is } c = 3.936.) \\ \Leftrightarrow\quad & \frac{SSR_r}{SSR_{ur}} < 1.04. \end{aligned}$$

Consequently, if the ratio between SSR_r and SSR_{ur} is between 1.02 and 1.04, we will choose to include x based on AIC, but x is statistically insignificant at 10% level.