

Final Question Bank

Prof. Beheshti

December 7, 2020

Define and Discuss

Problem 1. Define and explain the following terms. Use no more than three sentences. (2 points each)

1. Exclusion restriction (in context of IV)
2. 2SLS
3. Serial Correlation
4. Bias
5. Statistical significance

True/False/Uncertain For each of the following statements, decide whether it is true, false, or uncertain. Justify your answer. *Note: Your score will be determined entirely by your justification.* (2 points each)

Problem 2. For the next several problems, consider the following regression model

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

where we are concerned that $\mathbb{E}[\epsilon|X] \neq 0$, and Z is an instrument for X .

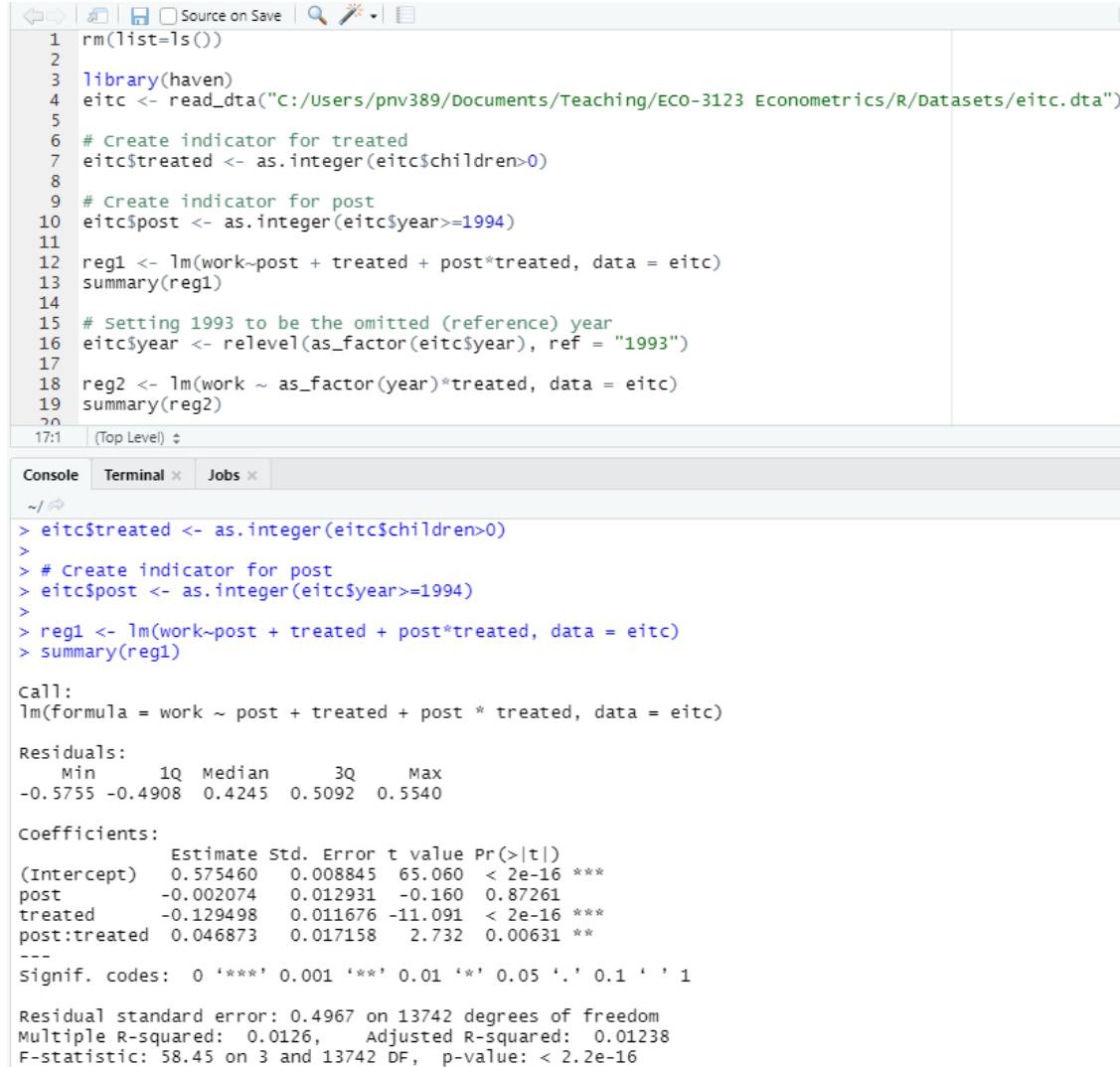
1. If $\text{Corr}(Z, \epsilon) < 0$, then Z is an invalid instrument.
2. In general, the 2SLS regression coefficient will be larger than the reduced form coefficient.
3. The IV estimator for β_1 is obtained by regressing Y on \hat{X} .

For the next several problems, consider a difference-in-differences model of the form

$$Y = \beta_0 + \beta_1 \cdot \text{Post} + \beta_2 \cdot \text{Treated} + \beta_3 \cdot \text{Post} \cdot \text{Treated} + \epsilon \quad (2)$$

1. Equation 2 does not allow us to evaluate the parallel trends assumption.
2. The parallel trends assumption is that, before the treatment, the treated and control groups were trending in parallel.
3. If Y is trending up (or down) prior to treatment, then $\hat{\beta}_3$ will be biased.

Application and Interpretation



```
1 rm(list=ls())
2
3 library(haven)
4 eitc <- read_dta("C:/Users/pnv389/Documents/Teaching/ECO-3123 Econometrics/R/Datasets/eitc.dta")
5
6 # Create indicator for treated
7 eitc$treated <- as.integer(eitc$children>0)
8
9 # Create indicator for post
10 eitc$post <- as.integer(eitc$year>=1994)
11
12 reg1 <- lm(work~post + treated + post*treated, data = eitc)
13 summary(reg1)
14
15 # Setting 1993 to be the omitted (reference) year
16 eitc$year <- relevel(as_factor(eitc$year), ref = "1993")
17
18 reg2 <- lm(work ~ as_factor(year)*treated, data = eitc)
19 summary(reg2)
20
```

Console

```
> eitc$treated <- as.integer(eitc$children>0)
>
> # Create indicator for post
> eitc$post <- as.integer(eitc$year>=1994)
>
> reg1 <- lm(work~post + treated + post*treated, data = eitc)
> summary(reg1)

Call:
lm(formula = work ~ post + treated + post * treated, data = eitc)

Residuals:
    Min       1Q   Median       3Q      Max
-0.5755 -0.4908  0.4245  0.5092  0.5540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.575460   0.008845   65.060 < 2e-16 ***
post        -0.002074   0.012931   -0.160  0.87261
treated     -0.129498   0.011676  -11.091 < 2e-16 ***
post:treated  0.046873   0.017158   2.732  0.00631 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4967 on 13742 degrees of freedom
Multiple R-squared:  0.0126, Adjusted R-squared:  0.01238
F-statistic: 58.45 on 3 and 13742 DF, p-value: < 2.2e-16
```

Figure 1: RStudio Screenshot

Figure 1 presents a screenshot from RStudio. The .dta file “eitc” contains data on 13,746 randomly sampled women’s labor supply and number of children between 1991 and 1996 (a new random sample of women each year, not a panel). The variable “children” reports the number of children the woman has, and “work” is an indicator variable that is equal to 1 if the woman worked that year and 0 if not.

Problem 3. What is the point of running the regression in line 18 instead of just the regression in line 12? (3 points)

Problem 4. Interpret the value of -0.129498. (2 points)

Problem 5. Can we reject the null hypothesis that there was no difference in labor supply between women with and without children prior to the EITC expansion? Explain. (2 points)

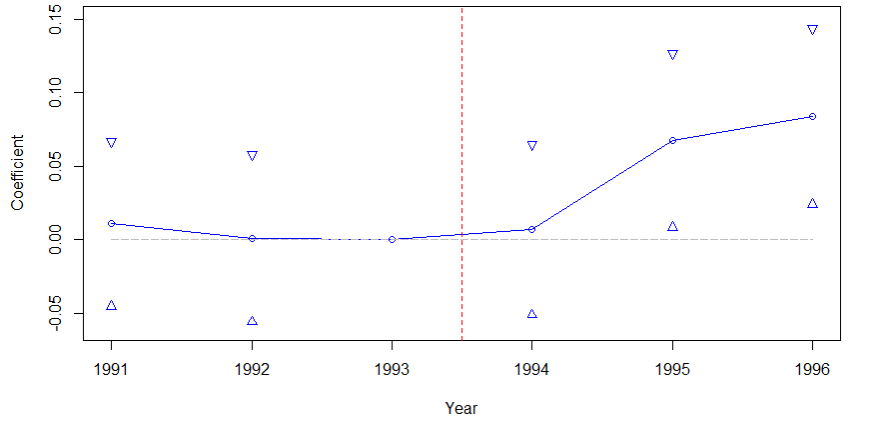


Figure 2: RStudio Screenshot

Problem 6. The above figure plots the coefficients (circles) and 95 percent confidence intervals (triangles) for the interactions between year and treated from line 18. Explain how to interpret this figure. (3 points)

Problem 7. Does this figure tell us anything about the plausibility of the parallel trends assumption? (3 points)

For the next several problems consider the following scenario. Suppose you wanted to test whether attending office hours improved students' test scores. Since office hours attendance is likely endogenous (i.e., correlated with other things that affect test scores), you instrument for office hours attendance with the distance between the student's dorm room and the professors office, with the idea that living closer to the professors office increases office hours attendance.

Problem 8. Explain what the relevance condition means in this context. (2 points)

Problem 9. Explain what the exclusion restriction means in this context. (3 points)

Problem 10. Can you think of any potential violations of the exclusion restriction? (2 points)

Problem 11. Do you think an OLS regression of test scores on office hours attendance is likely to be biased up or down? Explain. (2 points)

Derivations Consider the following difference-in-differences regression

$$Y = \beta_0 + \beta_1 \cdot Post + \beta_2 \cdot Treated + \beta_3 \cdot Post \cdot Treated + \epsilon \quad (3)$$

Problem 12. Interpret each regression coefficient. (4 points)

Problem 13. Suppose β_0 and β_1 are both positive and β_2 and β_3 are both negative. Sketch a plot of Y against time that would lead to these estimates. (4 points)