Applied Regression and Time Series Analysis (2016 Fall): HW5 - Week 7

Jeffrey Yau and Devesh Tiwari

October 5, 2016

Instructions:

The weekly assignment serves two purposes: (1) Review concepts, techniques, theories, statistical models covered during the week. (2) Extend the materials taught in the asynchronized lectures, assigned readings, and live sessions; some new concepts and/or techniques are introduced in the weekly assignment.

Below are specific instructions:

- **Due: 10/23/2016 (11:59pm PST)**
- You may complete this assignment on your own or in a group of no more than 3 students.
- When working in a group, you are strongly encouraged to complete the assignment on your own before discussing your group mates. Do not use the "division-of-labor" approach to complete the assignment.
- The homework is designed as a quantitative analysis. The instructions and questions are designed to guide you through the analysis of data using regression techniques. As such, you should think of it as a quantitative case study and the result of the study is a report with a set of well-written codes that can be used to reproduce the results in the report.
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. R-script or R markdown file
    2. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    * **SectionNumber_hw02_LastNameFirstInitial.fileExtension**
    * Examples:
      · Section1_hw02_YauJ.Rmd
      · Section1_hw02_YauJ.pdf
      · Section1_hw02_TiwariD_YauJ.Rmd
      · Section1_hw02_TiwariD_YauJ.pdf

Objective:

The key objective of this homework is to practice the use of the difference-in-difference technique to handle potential bias arising from omitted variables.

Description:

The data set in VOUCHER.DTA can be used to estimate the effect of school choice on academic achievement. Attendance at a choice school was paid for by a voucher, which was determined by a lottery among those who applied. The data subset was chosen so that any student in the sample has a valid *1994* math test score. Unfortunately, many students have missing test scores, possibly due to attrition (that is, leaving the Milwaukee public school district). These data include students who applied to the voucher program and were accepted, students who applied and were not accepted, and students who did not apply. Therefore, even though the vouchers were chosen by lottery among those who applied, we do not necessarily have a random sample from a population where being selected for a voucher has been randomly determined. (An important consideration is that students who never applied to the program may be systematically different from those who did, and in ways that we cannot know based on the data.)

This exericse asks you to do a cross-sectional analysis where winning the lottery for a voucher acts as an *instrumental variable* for attending a choice school. Actually, because we have multiple years of data on each student, we construct two variables. The first, *choiceyrs*, is the number of years from 1991 to 1994 that a student attended a choice school; this variable ranges from zero to four. The variable *selectyrs* indicates the number of years a student was selected for a voucher. If the student applied for the program in 1990 and received a voucher then selectyrs = 4; if he or she applied in 1991 and received a voucher then selectyrs = 3; and so on. The outcome of interest is mnce, the student's percentile score on a math test administered in 1994.

Question 1: Of the 990 students in the sample, how many were never awarded a voucher? How many had a voucher available for four years? How many students actually attended a choice school for four years?

```
print(paste("Number of Students never awarded a voucher:",990 -sum(data$select)))
```

```
## [1] "Number of Students never awarded a voucher: 468"
```

```
print(paste("Number of students who had voucher for four years:",sum(data$selectyrs4)))
```

```
## [1] "Number of students who had voucher for four years: 108"
```

```
print(paste("Number of students who attended choice school for four years:",sum(data$choiceyrs4)))
```

```
## [1] "Number of students who attended choice school for four years: 56"
```

Question 2: Run a simple regression of *choiceyrs* on *selectyrs*. Are these variables related in the direction you expected? How strong is the relationship? Is selectyrs a sensible IV candidate for *choiceyrs*?

The model is significant and the coefficient on selectyrs makes sense. It's not a perfect relationship, but it's significant. The coefficient on the intercept indicates that there were some students who attended choice schools without a voucher.

```
model = lm(choiceyrs~selectyrs,data=data)

summary(model)
```

```
##
## Call:
## lm(formula = choiceyrs ~ selectyrs, data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08725 -0.01992 -0.01992  0.21325  1.21325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01992    0.02461   0.809    0.419
## selectyrs    0.76683    0.01259  60.931   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.576 on 988 degrees of freedom
## Multiple R-squared:  0.7898, Adjusted R-squared:  0.7896
## F-statistic:  3713 on 1 and 988 DF,  p-value: < 2.2e-16
```

Question 3 Run a simple regression of mnce on *choiceyrs*. What do you find? Is this what you expected? What happens if you add the variables black, hispanic, and female?

When just regressing on choiceyrs, the coefficient is negative. This is unexpected, as one would think that going to a choice private school would increase test scores

```
model = lm(mnce~choiceyrs,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = mnce ~ choiceyrs, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.234 -13.234   0.603  12.766  60.114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.2344     0.8507  54.348  < 2e-16 ***
## choiceyrs    -1.8370     0.5255  -3.495 0.000494 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.75 on 988 degrees of freedom
## Multiple R-squared:  0.01222,    Adjusted R-squared:  0.01122
## F-statistic: 12.22 on 1 and 988 DF,  p-value: 0.0004943
```

Including the race and gender variables reduces the magnitude and statistical significance of the choiceyrs variable. This suggests that choiceyrs is correlated with one or more of the race and gender variables

```
model = lm(mnce~choiceyrs+black+hispanic+female,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = mnce ~ choiceyrs + black + hispanic + female, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.122 -12.507   0.108  12.156  60.156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.1219     1.6567  34.479  < 2e-16 ***
## choiceyrs    -0.5652     0.5307  -1.065    0.287
## black       -16.0174     1.7944  -8.926  < 2e-16 ***
## hispanic    -13.4029     2.3168  -5.785 9.73e-09 ***
## female        1.3527     1.2758   1.060    0.289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.99 on 985 degrees of freedom
## Multiple R-squared:  0.08677,    Adjusted R-squared:  0.08307
## F-statistic:  23.4 on 4 and 985 DF,  p-value: < 2.2e-16
```

Question 4 Why might choiceyrs be endogenous in an equation such as

$$mnce = \beta_0 + \beta_1 choiceyrs + \beta_2 black + \beta_3 hispanic + \beta_4 female + u_1?$$

There may be other factors that are not captured in this equation that are both correlated with choiceyrs and causal to mnce. If this is the case, it will show up as correlation between choiceyrs and the residual.

Question 5 Estimate the equation in question 4 by instrumental variables, using *selectyrs* as the IV for choiceyrs. Does using IV produce a positive effect of attending a choice school? What do you make of the coefficients on the other explanatory vairables?

Using selectyrs as an instrument for choiceyrs still produces a negative coefficient for attending a choice school. The coefficients on the other variables are approximately the same as in previous problem

```
predicted = lm(choiceyrs~selectyrs + black+hispanic+female,data=data)$fitted.values
summary(lm(mnce~predicted+black+hispanic+female,data=data))
```

```
##
## Call:
## lm(formula = mnce ~ predicted + black + hispanic + female, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.073 -12.550  -0.061  11.954  58.988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.0680     1.6582  34.415  < 2e-16 ***
## predicted    -0.2413     0.6055  -0.399    0.690
## black       -16.3169     1.8154  -8.988  < 2e-16 ***
## hispanic    -13.7754     2.3420  -5.882 5.55e-09 ***
## female        1.3197     1.2767   1.034    0.302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20 on 985 degrees of freedom
```

```
## Multiple R-squared:  0.08587,    Adjusted R-squared:  0.08216
## F-statistic: 23.13 on 4 and 985 DF,  p-value: < 2.2e-16
```

Question 6 To control for the possibility that prior achievement affects participating in the lottery (as well as predicting attrition), add *mnce90* — the math score in 1990 — to the equation in Question 4. Estimate the equation by OLS and IV, and compare the results for β_1. For the IV estimate, how much is each year in a choice school worth on the math percentile score? Is this a practically large effect?

In the OLS regression, each year is worth .41 percentile score points.

```
#First drop rows that are null for mnce90
data_mnce90 = data[is.finite(data$mnce90),]
summary(lm(mnce~choiceyrs + black+hispanic+female+mnce90,data=data_mnce90))
```

```
##
## Call:
## lm(formula = mnce ~ choiceyrs + black + hispanic + female + mnce90,
##     data = data_mnce90)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.921 -11.669   0.773  10.686  50.838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.1529     3.6204   6.119 2.73e-09 ***
## choiceyrs     0.4106     0.7359   0.558  0.57726
## black        -8.3052     2.5461  -3.262  0.00123 **
## hispanic     -4.1050     3.3624  -1.221  0.22303
## female       -0.8829     1.7760  -0.497  0.61945
## mnce90        0.6204     0.0484  12.817  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.03 on 322 degrees of freedom
## Multiple R-squared:  0.4237, Adjusted R-squared:  0.4147
## F-statistic: 47.34 on 5 and 322 DF,  p-value: < 2.2e-16
```

In the IV regression, each year is worth 1.79 percentile score points. This is a pretty large effect

```
predicted = lm(choiceyrs~selectyrs + black+hispanic+female+mnce90,data=data_mnce90)$fitted.values
print("IV regression")
```

```
## [1] "IV regression"
```

```
summary(lm(mnce~predicted+black+hispanic+female+mnce90,data=data_mnce90))
```

```
##
## Call:
## lm(formula = mnce ~ predicted + black + hispanic + female + mnce90,
##     data = data_mnce90)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -49.02 -11.20   0.92  10.83  47.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.53886    3.60226   5.979 5.95e-09 ***
## predicted    1.79938    0.84999   2.117 0.035032 *
## black       -9.06711    2.54093  -3.568 0.000414 ***
## hispanic    -5.00373    3.35256  -1.493 0.136545
## female      -1.02048    1.76512  -0.578 0.563574
## mnce90       0.62881    0.04816  13.056  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.93 on 322 degrees of freedom
## Multiple R-squared:  0.431,  Adjusted R-squared:  0.4222
## F-statistic: 48.79 on 5 and 322 DF,  p-value: < 2.2e-16
```

Question 7 The variables choiceyrs1, choiceyrs2, and so on are dummy variables indicating the different number of years a student could have been in a choice school (from 1991 to 1994). The dummy variables selectyrs1, selectyrs2, and so on have a similar definition, but for being selected from the lottery. Estimate the equation

$$mnce = \beta_0 + \beta_1 choiceyrs1 + \beta_2 choiceyrs2 + \beta_3 choiceyrs3 + \beta_4 choiceyrs4 + \beta_5 black + \beta_6 hispanic + \beta_7 female + \epsilon?$$

by IV, using as instruments the *four selectyrs dummy variables*. Describe your findings. Do they make sense?

The findings show the improvement in test score differs depending on how many years you go to the choice school. What's most unexpected is that predicted3 (corresponding to IV selected3) has a negative coefficient, which seems counter intuitive. Then again, maybe kids who only got 3 years of their choice school were unusual in some other way. Note that none of the IVr coefficients are significant.

```
predicted1 = lm(choiceyrs1~selectyrs1 + selectyrs2+selectyrs3+selectyrs4+ black+hispanic+female,data)$fi

predicted2 = lm(choiceyrs2~selectyrs1 + selectyrs2+selectyrs3+selectyrs4+ black+hispanic+female,data)$fi

predicted3 = lm(choiceyrs3~selectyrs1 + selectyrs2+selectyrs3+selectyrs4+ black+hispanic+female,data)$fi

predicted4 = lm(choiceyrs4~selectyrs1 + selectyrs2+selectyrs3+selectyrs4+ black+hispanic+female,data)$fi


summary(lm(mnce~predicted1 + predicted2+predicted3+predicted4+ black+hispanic+female,data))
```

```
##
## Call:
## lm(formula = mnce ~ predicted1 + predicted2 + predicted3 + predicted4 +
##     black + hispanic + female, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.025 -12.607   0.115  12.187  60.453
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.8858     1.6677  34.110  < 2e-16 ***
## predicted1    0.3900     2.3601   0.165    0.869
## predicted2    0.7737     4.1574   0.186    0.852
## predicted3   -4.2848     3.6949  -1.160    0.246
## predicted4    2.4071     4.1592   0.579    0.563
## black       -16.2972     1.8761  -8.687  < 2e-16 ***
## hispanic    -13.3660     2.4599  -5.433 6.97e-08 ***
## female        1.3664     1.2783   1.069    0.285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.01 on 982 degrees of freedom
## Multiple R-squared:  0.08733,    Adjusted R-squared:  0.08083
## F-statistic: 13.42 on 7 and 982 DF,  p-value: < 2.2e-16
```