Hayden C. Hancock

4/28/2022

Dr. Hwu, GBA 6230

Assignment 6

1. In 1985, neither Florida nor Georgia had laws banning open alcohol containers in vehicle passenger compartments. By 1990, Florida had passed such a law, but Georgia had not.

(a) Suppose you can collect random samples of the driving-age population in both states, for 1985 and 1990. Let arrest be a binary variable equal to unity if a person was arrested for drunk driving during the year. Without controlling for any other factors, write down a linear probability model that allows you to test whether the open container law reduced the probability of being arrested for drunk driving. Which coefficient in your model measures the effect of the law?

$Arrest_{ti}$=B0 + B1*$PolicyChange_{it}$+u

Effect of law would be B1, treatment or control

B1*Policy Change would measure the effect of the model

This represents the instrumental variable function. So the E(z|u)=0 & E(z|X1)=/0

(b) Why might you want to control for other factors in the model? What might some of these factors be?

You would want to control for other factors in the model because we don’t want omitted variable bias.

Some other variables we would like to control would be age distribution, gender distribution and law enforcement efforts in the area.

(c) Now, suppose that you can only collect data for 1985 and 1990 at the county level for the two states. The dependent variable would be the fraction of licensed drivers arrested for drunk driving during the year. How does this data structure differ from the individual-level data described in part (i)?

It differs from the previous example because it is county level data not a random sample from two different states. This would be the same set of county for both years.

What econometric method would you use?

We would use DID data, sub setting for the fraction of licensed drivers arrested for drunk driving during the year in those two counties. This is a much more controlled experiment because we have the same individuals over time.

2. Use the data in KIELMC.RAW for this question.

(a) The variable dist is the distance from each home to the incinerator site, in feet. Consider the model

$\log(price) = \beta 0 + \delta 0 y81 + \beta 1 \log(dist) + \delta 1 y81 * \log(dist) + u$

If building the incinerator reduces the value of homes closer to the site, what is the sign of $\delta 1$? What does it mean if $\beta 1 > 0$?

Ans: $\delta 1$ is positive with a value of 0.04819. In 1981, people may not have thought as much about the environmental concerns and more about the job opportunities that an incinerator would offer.

$\beta 1$ is positive with a value of 0.31669, this means each incremental step of distance away from the incinerator increases value of the price of the property, this is for the whole dataset, not just 1981.

(b) Estimate the model from part (a) and report the results in the usual form. Interpret the coefficient on y81 $*\log(dist)$. What do you conclude?

$\text{Log}(price)= 8.05848- 0.01133\ \delta 1+ 0.31669\ \beta 1 + 0.04819(\delta 1 y81 *\log(dist))+ u$

(0.50844) (0.80506) (0.05153) (0.08179)

This is the model with a created interaction term, below is with the interaction term provided

```
> summary(lm(kiel$lprice~kiel$y81+kiel$ldist+kiel$y81ldist))

Call:
lm(formula = kiel$lprice ~ kiel$y81 + kiel$ldist + kiel$y81ldist)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0196 -0.2059  0.0158  0.2032  1.5012 

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   7.878243   0.393109  20.041 < 0.0000000000000002 ***
kiel$y81      0.491484   0.041147  11.945 < 0.0000000000000002 ***
kiel$ldist    0.334979   0.039810   8.414  0.0000000000000137 ***
kiel$y81ldist -0.018118  0.009406  -1.926                0.055 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3404 on 317 degrees of freedom
Multiple R-squared:  0.4021,    Adjusted R-squared:  0.3965 
F-statistic: 71.07 on 3 and 317 DF,  p-value: < 0.00000000000000022
```

Ans: We can conclude that distance was not as impactful in 1981 due to the reduced estimate of the regression equation and significance.

(c) Add age, age2, rooms, baths, log(intst), log(land), and log(area) to the equation. Now, what do you conclude about the effect of the incinerator on housing values?

```
Call:
lm(formula = kiel$lprice ~ kiel$y81 + kiel$ldist + kiel$y81ldist + 
    kiel$age + kiel$agesq + kiel$rooms + kiel$baths + kiel$lintst + 
    kiel$lland + kiel$larea)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.19691 -0.10124  0.01316  0.11687  0.79409 

Coefficients:
                Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)    7.421284791 0.450642336  16.468 < 0.0000000000000002 ***
kiel$y81       0.402639831 0.025624266  15.713 < 0.0000000000000002 ***
kiel$ldist     0.028192945 0.038666105   0.729             0.466467    
kiel$y81ldist -0.008630171 0.005716873  -1.510             0.132165    
kiel$age      -0.007817529 0.001403121  -5.572       0.000000054860 ***
kiel$agesq     0.000034735 0.000008633   4.024       0.000072065907 ***
kiel$rooms     0.042986799 0.017398265   2.471             0.014021 *  
kiel$baths     0.104884698 0.027600607   3.800             0.000174 ***
kiel$lintst   -0.057764667 0.031640524  -1.826             0.068864 .  
kiel$lland     0.092314397 0.024491354   3.769             0.000196 ***
kiel$larea     0.351239700 0.051871591   6.771       0.000000000064 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2052 on 310 degrees of freedom
Multiple R-squared:  0.7875,    Adjusted R-squared:  0.7807 
F-statistic: 114.9 on 10 and 310 DF,  p-value: < 0.00000000000000022
```

Ans: The entire regression models p-value is significant, concluding that the price with all the variables included accounts for 78% of the variables that attribute to price. However, the interaction term of I(y81 * ldist) is now positive. This means that with the other variables included, the housing price of being near an incinerator is a good thing in the year of 1981.

(d) Why is the coefficient on log(dist) positive and statistically significant in part (b) but not in part (c)?

What does this say about the controls used in part (c)?

Ans: That would account for some omitted variable bias. In the model previously, we had a very small r squared of .2228, in the second example we accounted for a lot more of the attributes to price up to 78%. The controls used in part c were much better.

3. On July 15, 1980, Kentucky raised the cap on weekly earnings that were covered by workers compensation. An increase in the cap has no effect on the benefit for low-income workers, but it makes it less costly for a high-income worker to stay on workers compensation. Using random samples both before and after the policy change, Meyer, Viscusi, and Durbin (1995) were able to test whether more generous workers compensation causes people to stay out of work longer (everything else fixed). They started with a difference-in-differences analysis, using log(durat) as the dependent variable, where durat denotes the duration they workers stay on workers' compensation. Let afchnge be the dummy variable for observations after the policy change and highearn the dummy variable for high earners. Use the data in INJURY for this question.

(a) Estimate the following equation

$\log(durat) = \beta 0 + \beta 1afchnge + \beta 2highearn + \beta 3afchnge * highearn + u$

report the results in usual form. Which coefficients represent the policy effect? Is it significant?

$Log(durat)= 1.19934+ 0.02364\ \beta 1+ 0.21520\ \beta 2 + 0.18835\ \beta 3$

(0.02711) (0.03970) (0.04336) (0.06279)

R2= 0.01543

N=7,150

Ans: The interaction term represents the combination of high earn and after change, it is significant at the 1% level. The second I think would represent the policy effect would be afchnge, this is not significant.

(b) Estimate the following equation

$\log(durat) = \beta 0 + \beta 2highearn + \beta 3afchnge * highearn + u$

report the results in usual form. Do you find the estimate of β3 very different from part (a)? Why do you think this is the case?

$Log(durat)= 1.21036+ 0.20418\ \beta 2+ 0.21198\ \beta 3$

(0.01980) (0.03921) (0.04865)

$R2 = 0.01552$

N=7,150
Ans: This is very different from part a because both factors increase the duration. Both are statistically significant aswell. The interaction term rose in this equation from the first, meaning there was a difference in duration. Meaning the high earners would typically spend a longer duration due to the pay and would up until their benefits ran out.

(c) Estimate the following equation

$\log(durat) = \beta 0 + \beta 1afchnge + \beta 2afchnge *highearn + u$
report the results in usual form. Do you find the estimate of β3 very different from part (a)? Why do you think this is the case?

$Log(durat)= 1.28345 -0.06048\beta 1+ 0.40355\beta 3$

(0.02119) (0.03596) (0.04549)

$R2 = 0.01217$

N=7,150

Ans: This is the most important case, from here you can see that after change is negative, meaning that it has worked, and people are not as likely to spend as much of a duration on workers comp. However, the interaction term is higher, that people who earn high are also willing to stay on workers comp for longer.

4. Use the data in RENTAL for this question. The data on rental prices and other variables for college towns are for the years 1980 and 1990. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is
$\log(rentit) = \beta 0 + \delta 0y90t + \beta 1 \log(popit) + \beta 2 \log(avgincit) + \beta 3pctstuit + ai + uit$
where pop is city population, avginc is average income, and pctstu is student population as a percentage of city population (during the school year).

(a) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable?

```
Pooling Model

Call:
plm(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental,
    model = c("pooling"), index = c("city", "year"))

Balanced Panel: n = 64, T = 2, N = 128

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.242331 -0.078237 -0.016417  0.043890  0.480819

Coefficients:
              Estimate Std. Error t-value              Pr(>|t|)
(Intercept) -0.5688071  0.5348815 -1.0634                0.2897
y90          0.2622267  0.0347633  7.5432      0.000000000008782 ***
lpop         0.0406864  0.0225154  1.8070                0.0732 .
lavginc      0.5714460  0.0530981 10.7621 < 0.00000000000000022 ***
pctstu       0.0050436  0.0010192  4.9486      0.000002401411329 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    14.058
Residual Sum of Squares: 1.9501
R-Squared:      0.86128
Adj. R-Squared: 0.85677
F-statistic: 190.922 on 4 and 123 DF, p-value: < 0.000000000000000222
```

What do you get for β_3?

The 1990 dummy variable is positive which means that rent rate increased. Beta 3 is very small positive number

(b) Now, difference the equation and estimate by OLS. Compare your estimate of β_3 with that from part (a). Does the relative size of the student population appear to affect rental prices?

```
> summary(rent2)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental,
    model = c("fd"), index = c("city", "year"))

Balanced Panel: n = 64, T = 2, N = 128
Observations used in estimation: 64

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.186972 -0.062161 -0.014384  0.055182  0.237830

Coefficients:
             Estimate Std. Error t-value            Pr(>|t|)
(Intercept) 0.3855215  0.0368245 10.4692 0.000000000000003661 ***
lpop        0.0722453  0.0883435  0.8178            0.416720
lavginc     0.3099604  0.0664771  4.6627 0.000017884232063119 ***
pctstu      0.0112033  0.0041319  2.7114            0.008727 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.7191
Residual Sum of Squares: 0.48736
R-Squared:      0.32226
Adj. R-Squared: 0.28837
F-statistic: 9.50991 on 3 and 60 DF, p-value: 0.000031362
```

Yes the relative size of the student population appears to affect the rental prices. It is larger than the pooled model going from 0.005 to 0.011

(c) Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part (c).

```
> summary(rent3)
Oneway (individual) effect Within Model

Call:
plm(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental,
    model = c("within"), index = c("city", "year"))

Balanced Panel: n = 64, T = 2, N = 128

Residuals:
                   Min.                 1st Qu.                  Median                 3rd Qu.
-0.11891516636976454113 -0.02955918974968736040 -0.00000000000000027582  0.02955918974968768306
                   Max.
 0.11891516636976444399

Coefficients:
         Estimate Std. Error t-value             Pr(>|t|)
y90     0.3855215  0.0368245 10.4692 0.000000000000003661 ***
lpop    0.0722453  0.0883435  0.8178             0.416720
lavginc 0.3099604  0.0664771  4.6627 0.000017884232063119 ***
pctstu  0.0112033  0.0041319  2.7114             0.008727 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    10.383
Residual Sum of Squares: 0.24368
R-Squared:      0.97653
Adj. R-Squared: 0.95032
F-statistic: 624.146 on 4 and 60 DF, p-value: < 0.0000000000000000222
> 
```

The estimates are identical .