[redacted]DID

Jolly Ogbole

2023-04-23

R Markdown

#load library

```
#install.packages("plm")
library(plm)
library(xlsx)
```

Question 1

part a

A linear probability model for testing the effect of the open container law on the probability of being arrested for drunk driving can be written as:

arrest = β0 + β1(open container law) + u

where “arrest” is a binary variable equal to unity if a person was arrested for drunk driving during the year, and “open container law” is a binary variable equal to unity if the state has a law banning open alcohol containers in vehicle passenger compartments. The coefficient β1 measures the effect of the open container law on the probability of being arrested for drunk driving.

part b

Other factors that might affect the probability of being arrested for drunk driving include demographic characteristics (age, gender, race), socioeconomic status, driving behavior, and law enforcement efforts. It is important to control for these factors in the model to isolate the effect of the open container law on the probability of being arrested for drunk driving.

part c

The county-level data structure differs from individual-level data in that the dependent variable is the fraction of licensed drivers arrested for drunk driving during the year, rather than the binary variable “arrest” at the individual level. To analyze this data, we would

need to use a regression method called the ordinary least squares (OLS) regression. The dependent variable would be the fraction of licensed drivers arrested for drunk driving during the year, and the independent variable would be the presence of the open container law. We would also control for other factors that might affect the fraction of licensed drivers arrested for drunk driving, such as demographic characteristics, socioeconomic status, and law enforcement efforts. The coefficient on the open container law variable in the OLS regression would measure the effect of the law on the fraction of licensed drivers arrested for drunk driving.

Question 2

part a

If building the incinerator reduces the value of homes closer to the site, the sign of $\delta 1$ will be negative.

If $\beta 1 > 0$, it means that there is a positive relationship between the distance from the incinerator site and the log(price) of homes. This indicates that, on average, homes farther away from the incinerator site have higher prices than those closer to the site, all else being equal.

part b

```
kielmc.df <-read.xlsx("kielmc.xls", sheetIndex = 1)
summary.df <- summary(lm(lprice ~ y81+ldist+y81ldist, data=kielmc.df))
summary.df

##
## Call:
## lm(formula = lprice ~ y81 + ldist + y81ldist, data = kielmc.df)
##
## 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  < 2e-16 ***
## y81          0.491484   0.041147  11.945  < 2e-16 ***
## ldist        0.334979   0.039810   8.414 1.37e-15 ***
## 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: < 2.2e-16
```

The coefficient of y81∗log(dist) = -0.018118 This is the DID estimate of the treatment effect of the distance of a house to the incinerator location. What this suggest is that proximity to the incinerator does have a negative effect on a house price by the stated magnitude of the coefficient, although the significance of that effect is only at the 10 percent level according to this model.

part c

```
kielmc.df <-read.xlsx("kielmc.xls", sheetIndex = 1)
summary.df <- summary(lm(lprice ~
y81+ldist+y81ldist+age+agesq+baths+rooms+lintst+lland+larea, data=kielmc.df))
summary.df

##
## Call:
## lm(formula = lprice ~ y81 + ldist + y81ldist + age + agesq +
##     baths + rooms + lintst + lland + larea, data = kielmc.df)
##
## 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)  7.421e+00  4.506e-01  16.468  < 2e-16 ***
## y81          4.026e-01  2.562e-02  15.713  < 2e-16 ***
## ldist        2.819e-02  3.867e-02   0.729 0.466467
## y81ldist    -8.630e-03  5.717e-03  -1.510 0.132165
## age         -7.818e-03  1.403e-03  -5.572 5.49e-08 ***
## agesq        3.473e-05  8.633e-06   4.024 7.21e-05 ***
## baths        1.049e-01  2.760e-02   3.800 0.000174 ***
## rooms        4.299e-02  1.740e-02   2.471 0.014021 *
## lintst      -5.776e-02  3.164e-02  -1.826 0.068864 .
## lland        9.231e-02  2.449e-02   3.769 0.000196 ***
## larea        3.512e-01  5.187e-02   6.771 6.40e-11 ***
## ---
## 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: < 2.2e-16
```

The conclusion we can reach here is that after we control for other variables in this model, the effect of the incinerator on a house value remains negative but that negative effect is not statistically significant. i.e not significantly different than zero.

part d

In part b, it was because of other variables that could affect price but had not been accounted for. When these variables were controlled for in part c, it reduced the magnitude of the effect of log(dist). In essence, it suggests that the controls used in part c have largely significantly effects on the price of houses and their omission will lead to important omitted variables bias.

Question 3

part a

```
injury.df <-read.xlsx("injury.xls", sheetIndex = 1)
summary.df <- summary(lm(ldurat ~ afchnge+highearn+afhigh, data=injury.df))
summary.df

##
## Call:
## lm(formula = ldurat ~ afchnge + highearn + afhigh, data = injury.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0128 -0.7214 -0.0171  0.7714  4.0047
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.19934    0.02711  44.241  < 2e-16 ***
## afchnge      0.02364    0.03970   0.595  0.55164
## highearn     0.21520    0.04336   4.963 7.11e-07 ***
## afhigh       0.18835    0.06279   2.999  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 7146 degrees of freedom
## Multiple R-squared:  0.01584,    Adjusted R-squared:  0.01543
## F-statistic: 38.34 on 3 and 7146 DF,  p-value: < 2.2e-16
```

The coefficient β3 represents the policy effect, as it captures the interaction between the two dummy variables: af chnge (whether the observation is after the policy change) and highearn (whether the worker is a high earner). This interaction term shows how the effect of the policy change varies for high earners compared to low earners.

Going by the test statistics above, β3(afhigh) is not significant at the 5% level.

part b

```
injury.df <-read.xlsx("injury.xls", sheetIndex = 1)
summary.df <- summary(lm(ldurat ~ highearn+afhigh, data=injury.df))
summary.df

##
## Call:
## lm(formula = ldurat ~ highearn + afhigh, data = injury.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0128 -0.7214 -0.0171  0.7714  3.9937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.21036    0.01980  61.116  < 2e-16 ***
## highearn     0.20418    0.03921   5.207 1.97e-07 ***
## afhigh       0.21198    0.04865   4.357 1.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 7147 degrees of freedom
## Multiple R-squared:  0.01579,    Adjusted R-squared:  0.01552
## F-statistic: 57.34 on 2 and 7147 DF,  p-value: < 2.2e-16
```

Yes, there is a larger estimated effect of β3. I think this is because by estimating the model without β1afchnge, I removed the main effect of the policy change from the model. As a result, the estimated effect of the policy change on the duration of workers' compensation is now captured entirely by the interaction effect (β3afchnge * highearn). This can lead to a larger estimated effect of the policy change on the duration of workers' compensation for high earners, as the effect is no longer partially offset by the main effect of the policy change.

part c

```
injury.df <-read.xlsx("injury.xls", sheetIndex = 1)
summary.df <- summary(lm(ldurat ~ afchnge+afhigh, data=injury.df))
summary.df

##
## Call:
## lm(formula = ldurat ~ afchnge + afhigh, data = injury.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0128 -0.5903  0.1028  0.7960  3.9810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.28345    0.02119  60.561   <2e-16 ***
## afchnge     -0.06048    0.03596  -1.682   0.0927 .
## afhigh       0.40355    0.04549   8.870   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.301 on 7147 degrees of freedom
## Multiple R-squared:  0.01245,    Adjusted R-squared:  0.01217
## F-statistic: 45.05 on 2 and 7147 DF,  p-value: < 2.2e-16
```

Yes the estimate of β3 is even more larger. I think this is because when I estimated the model without β2highearn, I removed the effect of being a high earner from the model. This means that the effect of being a high earner on the duration of workers' compensation is not directly controlled for in the model.Therefore, it is possible that the coefficient β3 increased this much when I removed β2highearn from the model because the estimated effect of the policy change on the duration of workers' compensation for high earners became larger in the absence of the control for being a high earner.

Question 4

part a

```
rental.df <-read.xlsx("rental.xls", sheetIndex = 1)
summary.df <- summary(plm(lrent ~ y90+lpop+lavginc+pctstu,
model=c("pooling"), data=rental.df, index=c("city", "year")))
summary.df

## Pooling Model
##
## Call:
## plm(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental.df,
##     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 8.782e-12 ***
## lpop         0.0406864  0.0225154  1.8070    0.0732 .
## lavginc      0.5714460  0.0530981 10.7621 < 2.2e-16 ***
## pctstu       0.0050436  0.0010192  4.9486 2.401e-06 ***
## ---
## 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: < 2.22e-16
```

From the test statistics above, the estimate of the 1990 dummy variable, $\delta 0y90 = 0.2622267$ in the model, it means that rental prices in the college towns increased by approximately 26.22% between 1980 and 1990, all else being equal.

The model returned the estimate of $\beta 3 = 0.0050436$. This suggests that that a unit increase in the percentage of students in the town population, for town t, at time, i is associated with a 0.504% increase in rental prices, all else held constant.

part b

```
rental.df <-read.xlsx("rental.xls", sheetIndex = 1)
summary.df <- summary(plm(lrent ~ y90+lpop+lavginc+pctstu, model=c("fd"),
data=rental.df, index=c("city", "year")))
summary.df

## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental.df,
##     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 3.661e-15 ***
## lpop        0.0722453  0.0883435  0.8178  0.416720
## lavginc     0.3099604  0.0664771  4.6627 1.788e-05 ***
## 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: 3.1362e-05
```

Going by the test statistics above, the estimate on β3 = 0.0112033 and statistically significant. This suggests that relative size of student population does appear to affect rental prices.

part c

```
rental.df <-read.xlsx("rental.xls", sheetIndex = 1)
summary.df <- summary(plm(lrent ~ y90+lpop+lavginc+pctstu, model=c("within"),
data=rental.df, index=c("city", "year")))
summary.df
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental.df,
##     model = c("within"), index = c("city", "year"))
##
## Balanced Panel: n = 64, T = 2, N = 128
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -1.1892e-01 -2.9559e-02 -2.7582e-16  2.9559e-02  1.1892e-01
##
## Coefficients:
##          Estimate Std. Error t-value  Pr(>|t|)
## y90     0.3855215  0.0368245 10.4692 3.661e-15 ***
## lpop    0.0722453  0.0883435  0.8178  0.416720
## lavginc 0.3099604  0.0664771  4.6627 1.788e-05 ***
## 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: < 2.22e-16
```

Yes, the fixed-effects model gives identical estimates and standard errors as those in the first-difference model.