Lecture7notebook

February 6, 2025

1 DA- Lecture 7- Spring 2025

1.1 Villas-Boas

1.2 Daily Assignment for after Lecture 7 EEO 118 Spring 2025

Can old election outcomes predict new ones across US counties?

Suppose we are tasked to estimate what is the marginal effect of past election ratio of republican to democrat candidate (x1) in 2012 across US counties on the predicted ratio of Trump to Clinton 2016 votes (y) in the US counties, holding everything else constant.

[ ]:
```
install.packages("pacman")
# Load the 'pacman' package
library(pacman)
#packages to use load them now using the pacman "manager"
p_load(dplyr, readr)
#Another great feature of p_load(): if you try to load a package that is not␣
 ↪installed on your machine, p_load() install the package for you, rather than␣
 ↪throwing an error. For instance, let's install and load one final package␣
 ↪named ggplot2.
p_load(ggplot2)



#set scientific display off, thank you Roy
options(scipen=999)


pacman::p_load(lfe, lmtest, haven, sandwich, tidyverse)
# lfe for running fixed effects regression
# lmtest for displaying robust SE in output table
# haven for loading in dta files
# sandwich for producing robust Var-Cov matrix
# tidyverse for manipulating data and producing plots
```

[ ]:
```
library(haven)
library(ggplot2)
```

```
[ ]: #----------------------------------------
     #1. Read in data
     #----------------------------------------
     my_data <- read_dta("dataLecture7.dta")
     head(my_data)
```

1.2.1 The Sample

Data for 3110 counties on many variables pertaining to election outcomes. The source is the MIT election lab.

1.3 Population Model

Suppose there is a true relationship, across US counties, between a variable y and a variable x1 where

y = the Number Votes for Trump relative to votes for Clinton in 2016

and

x1 = the number of votes for Rodney relative to number of votes for Obama in the 2012/previous election

$y= _0 + _1 x_ + $

Call the true marginal effect of x1 on y the parameter β_1

Does a larger number of Rodney votes relative to Obama votes (larger x_1) leads to the occurrence a larger y, holding all else constant, i.e., leads to votes for Trump relative to Clinton to increase?

Figuring out whether β_1 is significant and what is its size is interesting.

Finally, thinking about what other factors could affect y (the number of Trump votes over Clinton votes) that we should control for in this analysis is also interesting

that is, do we need to worry about Omitted variable Bias (OVB) ?

As you can see in the dataset, there are other variables in the dataframe beyond x1, namely,

Population by country = pop

Whether democratic or republican senate

Whether dem or rep house

Whether rep or dem governor

Ratio Rodney2012/Obama2012—> x1

Income

white percentage of the population in the county

Generate y and x1

[ ]:
```r
#generate variables Y and x1
my_data$Y<-my_data$trump16/my_data$clinton16
my_data$x1<-my_data$romney12/my_data$obama12
```

[ ]:
```r
scatterPlot<-plot(my_data$x1,my_data$Y)
```

Looks like there is a positive relationship in terms of a y and x1 scatter plot

Where Y= Number Trump Votes/ Number Clinton Votes

and

x1 = =Number Rodney votes/ Number Obama votes

Sometimes there are missing values, below is how we don't use them, if we so wish

[ ]:
```r
#use non-missing values only from now on
my_data2<-my_data[complete.cases(my_data),]
my_data3 <- my_data2[my_data2$Y !=Inf,]
#my_data3 <-my_data2[my_data2$white_pct!=NA]
```

1.4 Regression of y on x1 and white percentage of population in the county

[ ]:
```r
#regression model with x1 and white percentage in the county using my_data3

regfull<-lm(Y ~ x1 + white_pct, my_data3)
summary(regfull)
```

1.5 Regression of y on x1 only

[ ]:
```r
#regression no percent white as control
regsmall<-lm(Y ~ x1, my_data3)
summary(regsmall)
```

1.5.1 What do you see when white_percentage is omitted from the model? what happens to the estimate of the coefficient of x1?

The coefficient changes. Why does it change like that? Let's go through the OVB formula and then look at this case

2 Omitted variable Bias (OVB)

Beta1 tilde is not equal to the true β_1. there is a BIAS.

The sign of the Bias consists of the sign of a product of two things

The first is the sign of the correlation between x1 and x2, the variable we care about (x1) and the omitted one x2, which is the same sign as ρ

The second thing is the sign of the correlation between the outcome y and x2 that we omit which is the same sign as β_2.

2.0.1 write the above code below and get the needed correlations to figure out whether you have a positive or negative bias of beta1 tilde when you omit white percentage fromt he model

[ ]:
```
#type code here
```

3 Please estimate the model below and interpret

$y = \beta_o + \beta_1 x_1 + \beta_2 female\_perc + \beta_3 white\_perc + u$

[ ]:
```
#type your code here
regLast<-lm(Y ~ x1+female_pct+white_pct, my_data3)
summary(regLast)
```

4 Multicollinearity

[ ]:
```
#collinearity Slides

#Baseline Model
regBase <- lm(Y~x1+female_pct+white_pct, my_data3)
summary(regBase)
```

4.1 perfect collinearity

Note in the regression below, that male_pct coefficient cannot be estimated because male_pct is 1-female_pct, they are perfectly collinear.

[ ]:
```
#Alternative Model - Perfect collinearity
my_data3$male_pct=1-my_data3$female_pct
regPC <- lm(Y~x1+female_pct+male_pct+white_pct, my_data3)
summary(regPC)

#note in the regression output below,  that male_pct coefficient is not␣
 ↪estimated
# because male_pct is 1-female_pct, they are perfectly collinear
#with each other,
#the line male_pct has NA   NA
#Non available NA
```

4.2 Alternative Model – Multi collinearity

Note in the regression below, we have white percentage and black percentage in the regression

These two are not perfectly collinear but highly correlated

corr(white_pct,bh_pct)=-0.92

[ ]:
```
#Alternative Model - Multi collinearity
regMC <- lm(Y~x1+female_pct+white_pct+bh_pct, my_data3)
#where corr(white_pct,bh_pct)=-0.92
summary(regMC)
```

4.2.1 put Y and Y hat on same graph and x1 on horizontal axis

make combined scatter plot of Y data and fitted values of Y (Yhat) given regression estimates usinf X1, white_pct and female_pct

[ ]:
```
#for graph
#get the predicted Y hats
my_data3$Yhat<-regBase$fitted.values

#put Y and Y hat on same graph and x1 on horizontal axis
#make combined scatter plot of Y data and fitted values of Y (Yhat)
#given regression estimates usinf X1, white_pct and female_pct
scatter_data_fittedVals <- ggplot(data = my_data3) + geom_point(aes(x=x1, y=Y,␣
 ↪color = "data")) +
  geom_point(aes(x=x1, y=Yhat, color = "fitted")) +
  xlab("x1=Ratio Rodney to Obama Votes in 2012") + ylab("Y=Ratio Trump to␣
 ↪Clinton Votes in 2016 ") +
  ggtitle("Y (Red) and Predicted Y (Blue) and x1")
scatter_data_fittedVals
```

5 THE END