cdf5579_econometrics_Rproblemset_2

2022-11-11

This is the R-problem set 2 for Econometrics (Fall 2022), New York University Author: Carlos Figueroa (cdf5579)

Let's begin.

```
#Lets load some libraries first

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(readxl)
library(tinytex)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(tibble)
library(ggplot2)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(sandwich) #robust SEs
```

Part 1————————————————————————————-

1.1) questions

a. In any regression, what will be the interpretation of β1?

Aside from being the slope of the regression line, it is interpreted as the marginal change on our prediction based on the independent datapoint. Thus, if there is one unit increase in the x, our prediction will shift by that beta 1 value. It's interpretation also changes depending on the units of the regression, amount of regressors, and the nature of the data (if dummy, it could be the average treatment effect in some cases). But as a general, its just the slope of the linear fit.

b. In a regression with no controls, what will be the interpretation of β0?

It will be considered the baseline of our approximations, which is mathematically intrepreted as the intercept of the regression. In many cases, it will be our best linear approximation when the input is zero. But again, its interpretation changes dramatically. In a model of dummy variables, it could represent as the average of the control group. (non-smokers if beta 1 was 1 if smokers).

1.2) Load the data

```
data = data.frame(read.csv(file = "foxData.csv", header = TRUE , sep =","))

head(data)
```

```
##   foxnews2000 income2000 channels   rep1996   rep2000 region urban
## 1           0     2.5000        8 0.2376238 0.3804348   West     0
## 2           0     3.1250        6 0.5460123 0.6086956   West     0
## 3           0     4.7500       18 0.6956522 0.7500000   West     0
## 4           0     4.0577       33 0.5994118 0.6627713   West     1
## 5           0     5.7163       25 0.4790941 0.5672609   West     1
## 6           0     2.3977        8 0.6381579 0.7171053   West     0
```

1.3) Calculate the average vote share (unweighted) across towns, as well as the fraction of towns with access to Fox News in 2000.

```
avg_voteshare_1996 <- mean(data$rep1996)

avg_voteshare_2000 <- mean(data$rep2000)

print("Average voteshare (unweighted) in 1996 across towns")
```

```
## [1] "Average voteshare (unweighted) in 1996 across towns"
```

```
avg_voteshare_1996
```

```
## [1] 0.4696955
```

```
print("Average voteshare (unweighted) in 2000 across towns")
```

```
## [1] "Average voteshare (unweighted) in 2000 across towns"
```

```
avg_voteshare_2000
```

```
## [1] 0.5377708
```

```
fraction_of_towns_FN <- mean(data$foxnews2000)

print("Fraction of towns with access to Fox News in 2000")
```

```
## [1] "Fraction of towns with access to Fox News in 2000"
```

```
fraction_of_towns_FN
```

```
## [1] 0.1952247
```

1.4) We are interested in sources of omitted variable bias. Can you think of a reason that region will be an important source of omitted variable bias? As a hint: News Corporation is run out of New York City and firms tend to expand radially from their home base

There are many reasons why region is an important source of omitted variable bias, specially in the US. The most obvious to point out will be socioeconomical levels, education, political history, etc. But one of the most important is also the access to different news' channels. Some regions of the country are dominated by their hometown news companies (since these firms tend to expand radially form their home base). OBV appears when a model omits an independent variable that is a determinant of the dependent variable and correlated with one or more of the included independent variables.

In this case, aside from omitting the variables stated at first, we are omitting the ramifications News corporations have on their expansion. Thus, since Fox news is located in the northeast, there's gonna be a lot of people that watch fox news in that region. But that wouldnt be representing what we are searching for.

1.5) To see if region might be an omitted variable, calculate the share of towns with Fox News in the Northeast in 2000 and compare it other towns. Do a t-test to test if this is statistically significant. Hint: The region variable includes 4 regions–you will need to combine 3. You'll want a command like this:

```
print("Share of towns with Fox News by region")
```

```
## [1] "Share of towns with Fox News by region"
```

```
tapply(data$foxnews2000, data$region, mean)
```

```
##   Midwest Northeast     South      West
## 0.1770724 0.2173697 0.1420690 0.2220610
```

```
#Lets define it as a table
data_ = as.data.table(data)

data_[ ,neDummy:= region == "Northeast"]

t.test(foxnews2000 ~ neDummy, data_)
```

```
##
##  Welch Two Sample t-test
##
## data:  foxnews2000 by neDummy
## t = -4.6797, df = 8324.1, p-value = 2.918e-06
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not eq
ual to 0
## 95 percent confidence interval:
##  -0.05565132 -0.02279269
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           0.1781477           0.2173697
```

What do you think would be the sign of the bias coming from omitting Northeast?

It is statistically significant. There seems a higher consumption of Fox News in the Northeast of the US. And it might be connected to the ideas we explained. Hence, there seems to be a sign of bias coming from omitting Northeast.

1.6) Explain why the GOP vote share in 1996 is likely also an important omitted variable. To ground thinking here, notice that voter preferences tend to be correlated over time, and Fox News likely would want to enter markets with viewers who want their product. What do you think will be the sign of the bias from ignoring this variable??? (Optional: Can you think of a t-test to run in R to check this?)

```
tapply(data$rep1996, data$region, mean)
```

```
##   Midwest Northeast     South      West
## 0.4728409 0.4647344 0.4597701 0.4917551
```

```
tapply(data$rep2000, data$region, mean)
```

```
##   Midwest Northeast     South      West
## 0.5551923 0.5197919 0.5310795 0.5535838
```

```
cor(data$rep1996, data$rep2000,  method = "pearson", use = "complete.obs")
```

```
## [1] 0.9112099
```

```
print("We can see that ")
```

```
## [1] "We can see that "
```

```
#it grew everywhere, but by diff nums. West was already high
```

We see that these variables are indeed correlated.Then, Fox news is is putting a bigger effort in joining regions where GOP dominates, so that their material is accepted. And this as a result might fuel the support for GOP in these regions. So we would not know if the increase in GOP support comes from the results on the elections of 1996 (voter preferences being correlated over time), or Fox News influencing preferences in a state. So it will be an important source of omitted variable that we would want to control for.

Part 2: Regressions and Plots————————————————————

1.7) Run a regression of the GOP vote share in 2000 on whether Fox News is present in 2000. Test the coefficient on Fox News. Your code should look like this:

```
m1 <- lm(rep2000 ~ foxnews2000, data = data)

m1
```

```
##
## Call:
## lm(formula = rep2000 ~ foxnews2000, data = data)
##
## Coefficients:
## (Intercept)  foxnews2000
##   0.5378214   -0.0002593
```

```
coeftest(m1 , vcov = vcovHC(m1, type = c("HC3")))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value Pr(>|t|)
## (Intercept)  0.53782141  0.00150808 356.6268   <2e-16 ***
## foxnews2000 -0.00025935  0.00339684  -0.0763   0.9391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remind yourselves what the meaning of the vcovHC command is.

vcovHC stands for Heteroscedasticity-Consistent Covariance Matrix Estimation. That is the way we would like to test our model with the power of a computer, where Heteroscedasticity might be the most realistic assumption for the nature of our data (as we know heteroskedasticity can be used when discussing variables that have identifiable seasonal variability).

1.8) Let's begin adding control variables. First, add in the fraction of GOP voters in the town. What happens to the coefficient? Does the coefficient move in the direction you expected given the discussion in part 1?

```
m2 <- lm(rep2000 ~ foxnews2000 + rep1996, data = data)

m2
```

```
##
## Call:
## lm(formula = rep2000 ~ foxnews2000 + rep1996, data = data)
##
## Coefficients:
## (Intercept)  foxnews2000      rep1996
##     0.09456     -0.01180      0.94852
```

```
coeftest(m2 , vcov = vcovHC(m2, type = c("HC3")))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  0.0945560  0.0025146  37.6021 < 2.2e-16 ***
## foxnews2000 -0.0117950  0.0013572  -8.6907 < 2.2e-16 ***
## rep1996      0.9485240  0.0049981 189.7760 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient moved from -0.0002593 to -0.01180, which aligns with what we talked about in part 1. In that way, we can isolate the control variable's effects from the relationship between the variables of interest. And now we have a better view of the effect of the introduction of fox news in the GOP 2000 population.

1.9) Now add in the dummies for region. Interpret the coefficient on South. Why isn't Midwest in the regression?

```
m3 <- lm(rep2000 ~ foxnews2000 + rep1996 + factor(region), data = data)

m3
```

```
##
## Call:
## lm(formula = rep2000 ~ foxnews2000 + rep1996 + factor(region),
##     data = data)
##
## Coefficients:
##             (Intercept)              foxnews2000                  rep1996
##                 0.10971                  -0.01017                  0.94595
## factor(region)Northeast      factor(region)South       factor(region)West
##                -0.02732                  -0.01210                 -0.01904
```

```
coeftest(m3 , vcov = vcovHC(m3, type = c("HC3")))
```

```
##
## t test of coefficients:
##
##                            Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)              0.1097090  0.0025622  42.8182 < 2.2e-16 ***
## foxnews2000             -0.0101656  0.0013297  -7.6448 2.301e-14 ***
## rep1996                  0.9459490  0.0049264 192.0170 < 2.2e-16 ***
## factor(region)Northeast -0.0273225  0.0011071 -24.6796 < 2.2e-16 ***
## factor(region)South     -0.0121044  0.0022423  -5.3983 6.895e-08 ***
## factor(region)West      -0.0190431  0.0029633  -6.4262 1.373e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The reason why Midwest is not there is because when you include a categorical/factor variable in a model, one category of the variable is always excluded and serves as the reference category (due to multicollinearity, R does this automatically). The estimates for the categories that you do see in the output are in reference to the category that was excluded.

The coefficient on south is -0.01210. which means that if we hold foxnews2000 and the other covariates constant, a town in the south has a 0.12 lower percentage point decrease in the GOP vote share in 2000 compare toa town in the midwest.

1.10) Now do an F-test on whether the region dummies are significant. I will PARTIALLY fill in the code and you must do the rest:

```
m3 = lm(rep2000 ~ foxnews2000 + rep1996 + factor(region), data = data)

linearHypothesis(m3, c("factor(region)Northeast", "factor(region)South", "factor(region)West"),
 vcov = vcovHC(m3, type = c("HC3")))
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(region)Northeast = 0
## factor(region)South = 0
## factor(region)West = 0
##
## Model 1: restricted model
## Model 2: rep2000 ~ foxnews2000 + rep1996 + factor(region)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   9253
## 2   9250  3 204.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This means that we disprove the null hypothesis with the F obtained.

1.11) Now consider the variable channels. This is a variable that tells you how many channels are available in the region. Cable providers may strategically place popular channels up front. It is better to enter big markets early, even at the expense of competing with many channels. Hence, Fox News Entry is positively correlated with number of channels available (you can check this). For the number of channels to be an omitted variable, what else must it be correlated with? Can you think of a reason this may be the case?

```
cor(data$foxnews2000, data$channels,  method = "pearson", use = "complete.obs")
```

```
## [1] 0.5358817
```

They are indeed correlated. As we stated before, OBV appears when a model omits an independent variable that is a determinant of the dependent variable and correlated with one or more of the included independent variables. Then, channels must be a determinant of the dependent variable, which is rep2000 in our case, and has to be correlated with one of more of the independent variables, which is the case as we proved above (is it correlated with the presence of fox news for the reasons explained in the question)

1.12) Run the regression including the number of channels as a control. What happens to the coefficient on Fox? Is it statistically significant?

```
m4 <- lm(rep2000 ~ foxnews2000 + rep1996 + factor(region) + channels, data = data)

m4
```

```
##
## Call:
## lm(formula = rep2000 ~ foxnews2000 + rep1996 + factor(region) +
##     channels, data = data)
##
## Coefficients:
##             (Intercept)              foxnews2000                  rep1996
##               0.1259920                0.0037069                0.9452314
## factor(region)Northeast      factor(region)South       factor(region)West
##              -0.0233195               -0.0123765               -0.0185763
##                channels
##              -0.0007138
```

```
coeftest(m4 , vcov = vcovHC(m4, type = c("HC3")))
```

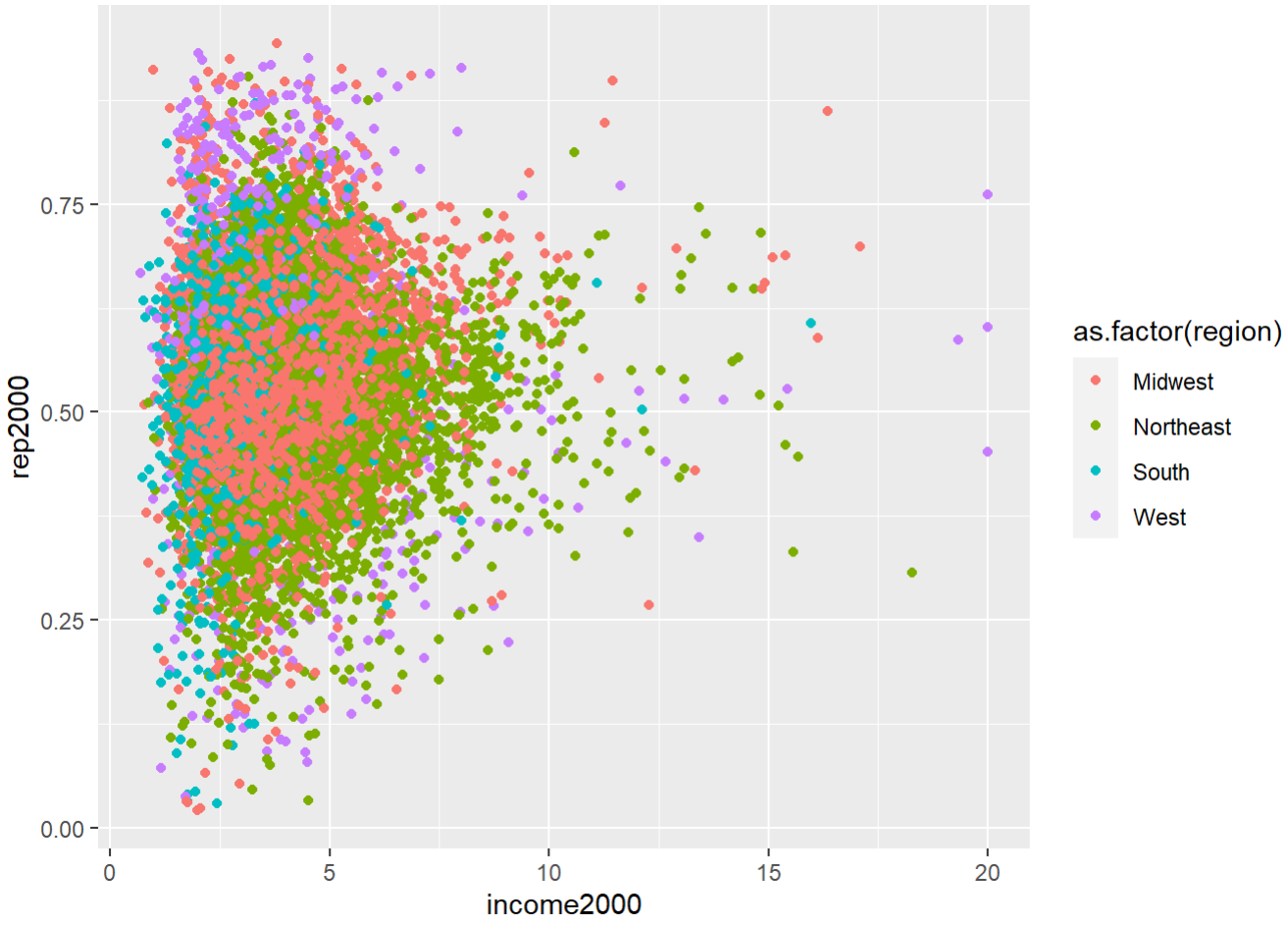
```
##
## t test of coefficients:
##
##                             Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)              1.2599e-01  2.7518e-03  45.7851 < 2.2e-16 ***
## foxnews2000              3.7069e-03  1.5543e-03   2.3849    0.0171 *
## rep1996                  9.4523e-01  4.8680e-03 194.1730 < 2.2e-16 ***
## factor(region)Northeast -2.3319e-02  1.1293e-03 -20.6500 < 2.2e-16 ***
## factor(region)South     -1.2376e-02  2.2293e-03  -5.5518 2.906e-08 ***
## factor(region)West      -1.8576e-02  2.9265e-03  -6.3476 2.290e-10 ***
## channels                -7.1375e-04  4.2882e-05 -16.6444 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is statistically significant at 5%, but not at 1%. Note that P value went higher than in our model without controlling for channels, connecting with what we discussed about. However, Fox News coefficient is now positive while controlling for channels, which adjust to our initial expectations, of republican votes going up by the introduction of Fox news.

1.13) Finally, consider the relationship between income, cable news, and voting. There are clear differences in voting patterns across the income distribution. Moreover, cable was quite expensive—costing about 360 dollars per year when the median income was 42000 per year.

At this time, there was no bundling with internet and a limited choice of channels. Hence, Fox News may choose to enter wealthier markets where people are more likely to have cable (richer customers also demand a higher premium from advertisers). To the first point, plot the relationship between GOP share in 2000 and income in 2000 —color code by region. To do this, you want a command like this (where you need to fill in the appropriate dataset and varaible names):

```
ggplot(data, aes(x = income2000 , y = rep2000, color = as.factor(region))) + geom_point()
```



Now include income as a regressor and test the fox news effect.

```
m5 <- lm(rep2000 ~ foxnews2000 + rep1996 + factor(region) + channels + income2000, data = data)

m5
```

```
##
## Call:
## lm(formula = rep2000 ~ foxnews2000 + rep1996 + factor(region) +
##     channels + income2000, data = data)
##
## Coefficients:
##             (Intercept)              foxnews2000                 rep1996
##                0.147938                 0.003846                0.975625
## factor(region)Northeast      factor(region)South       factor(region)West
##                -0.019408                -0.024029                -0.018518
##                 channels               income2000
##                -0.000535                -0.010333
```

```
coeftest(m5 , vcov = vcovHC(m5, type = c("HC3")))
```

```
##
## t test of coefficients:
##
##                             Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)              1.4794e-01  2.6973e-03  54.8468 < 2.2e-16 ***
## foxnews2000              3.8462e-03  1.4332e-03   2.6836  0.007297 **
## rep1996                  9.7562e-01  4.8138e-03 202.6717 < 2.2e-16 ***
## factor(region)Northeast -1.9408e-02  1.0725e-03 -18.0961 < 2.2e-16 ***
## factor(region)South     -2.4029e-02  2.2431e-03 -10.7124 < 2.2e-16 ***
## factor(region)West      -1.8518e-02  2.6728e-03  -6.9283 4.544e-12 ***
## channels                -5.3502e-04  4.0858e-05 -13.0949 < 2.2e-16 ***
## income2000              -1.0333e-02  2.8607e-04 -36.1204 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, the fox news variable is positive at 0.003846 which a higher statistical significance than before. This means that by accounting for all the control variables, we finally see a relation that states that Fox News presence does boost support for the GOP party.

1.14) You've seen the coefficient change dramatically as we entered new and different controls. Given our discussions in class on both external and internal validity, assess this study. This is a subjective question and requires you to think like a social scientist and not a statistician

What are the pros of this research design? What are the drawbacks? Are there omitted factors we've missed?

Internal validity is the one that we can discuss here. Properly, this validity refers to the belief that the OLS assumptions are met (E(U|X)). And in order to prove this, we will requiere better explanations than the ones we have to including control variables, that are substained by numbers and not hypothesis.

By the amount of variables that we are introducing as control, the clearest drawback will be the posibility of overfitting. Overfitting occurs when too many variables are included in the model and the model appears to fit well to the current data. Because some of variables retained in the model are actually noise variables, the model cannot be validated in future dataset.Then, also responding to the question if there are any omitted factors that we missed, even if we did, this is a good amount of variables to control for. On the top of my mind, education could be an interesting one, where people with less education will be more proned to be influenced by a media outcast.

Now, in the context of external validity, is is almost impossible, since fox news and these behavior is kind of unique in the US. This is stronger when we control for more variables (that are unique in context). Thus, assuming that internal validity is strong, it would be harsh to say that any news outlet would influence political preferences outside of the US, and more studies will be requiered in that manner, to see if it is a common denominator accross nations. It might be appropiate to say that our findings must be only extrapolated to countries that undergo a similar political division and media diversity and dispersion than the US.