Chapter 3 - The Multiple Regression Analysis - Estimation

Thalles Quinaglia Liduares

03/03/2022

Exercise 3.8

Upload packages

```
library(dplyr)
library(lmreg)
library(wooldridge)
```

Upload database

```
data<-wooldridge::discrim

attach(data)
```

Use the data in DISCRIM.RAW to answer this question. These are zip code–level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

(i) Find the average values of `prpblck` and `income` in the sample, along with their standard deviations. What are the units of measurement of `prpblck` and `income` ?

```
prpblck_avg<-round(mean(prpblck, na.rm=TRUE),2)

prpblck_sd<-round(sd(prpblck, na.rm = TRUE),2)

income_avg<-round(mean(income, na.rm=TRUE),2)

income_sd<-round(sd(income, na.rm = TRUE),2)
```

The mean value and sd for `prpblck` is 0.11 and 0.18, respectively. Measured in percentual.

The mean value and sd for `income` is $47,053.78 and $13,179.29, respectively. Measured in Dollars.

(ii) Consider a model to explain the price of soda, psoda, in terms of the proportion of the population that is black and median income:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u$$

Estimate this model by OLS and report the results in equation form, including the sample size and R-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on prpblck. Do you think it is economically large?

```
options(scipen=999) # To avoid scientific notation

lm1<-lm(psoda~prpblck+income)

summary(lm1)
```

```
##
## Call:
## lm(formula = psoda ~ prpblck + income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29401 -0.05242  0.00333  0.04231  0.44322
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) 0.9563196258 0.0189920097  50.354 < 0.0000000000000002 ***
## prpblck     0.1149881907 0.0260006361   4.423            0.0000126 ***
## income      0.0000016027 0.0000003618   4.430            0.0000122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08611 on 398 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.06422,    Adjusted R-squared:  0.05952
## F-statistic: 13.66 on 2 and 398 DF,  p-value: 0.000001835
```

The estimated equation is given by

$$\widehat{psoda} = 0.95 + 0.11prpblck + 0.000001income$$

The `R-Squared` and `Adjusted R-Squared` are equal to 6.4% and 5.9%, respectively.

The sample size is equal to 401 observations.

The coefficient of `prpblck` is equal to 0.11. Hence, the price medium of soda, increases $0.11 for each additional percent on proportion of black people.

(iii) Compare the estimate from part (ii) with the simple regression estimate from `psoda` on `prpblck`. Is the discrimination effect larger or smaller when you control for income?

```
lm2<-lm(psoda~prpblck)

summary(lm2)
```

```
##
## Call:
## lm(formula = psoda ~ prpblck)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30884 -0.05963  0.01135  0.03206  0.44840
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  1.03740    0.00519  199.87 < 0.0000000000000002 ***
## prpblck      0.06493    0.02396    2.71              0.00702 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0881 on 399 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.01808,    Adjusted R-squared:  0.01561
## F-statistic: 7.345 on 1 and 399 DF,  p-value: 0.007015
```

In this case, the estimated equation is given by

$$\widehat{psoda} = 1.03 + 0.06prpblck$$

The discrimination effect is smaller when controlling for income.

(iv) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + u$$

If `prpblck` increases by .20 (20 percentage points), what is the estimated percentage change in psoda? (Hint: The answer is 2.xx, where you fill in the "xx.")

```
lm3<-lm(data$lpsoda~prpblck+lincome)

summary(lm3)
```

```
##
## Call:
## lm(formula = data$lpsoda ~ prpblck + lincome)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33563 -0.04695  0.00658  0.04334  0.35413
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) -0.79377    0.17943  -4.424 0.00001254 ***
## prpblck      0.12158    0.02575   4.722 0.00000324 ***
## lincome      0.07651    0.01660   4.610 0.00000543 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0821 on 398 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.06809,    Adjusted R-squared:  0.06341
## F-statistic: 14.54 on 2 and 398 DF,  p-value: 0.0000008039
```

The estimated equation is given by

$$\widehat{log(psoda)} = -0.79 + 0.12prpblck + 0.07log(income)$$

If `prpblck` increases by 20%, then the medium price of soda increases by $2.40.

(v) Now add the variable `prppov` to the regression in part (iv). What happens to $\hat{\beta}_{prpblck}$?

```
lm4<-lm(lpsoda~prpblck+lincome+prppov)

summary(lm4)
```

```
##
## Call:
## lm(formula = lpsoda ~ prpblck + lincome + prppov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32218 -0.04648  0.00651  0.04272  0.35622
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) -1.46333    0.29371  -4.982 0.00000094 ***
## prpblck      0.07281    0.03068   2.373     0.0181 *
## lincome      0.13696    0.02676   5.119 0.00000048 ***
## prppov       0.38036    0.13279   2.864     0.0044 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08137 on 397 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.08696,    Adjusted R-squared:  0.08006
## F-statistic:  12.6 on 3 and 397 DF,  p-value: 0.00000006917
```

In this case, the coefficient of `prpblck` reduces from 0.12 to 0.07.

(vi) Find the correlation between log(income) and prppov. Is it roughly what you expected?

```
cor(lincome, prppov, use="complete.obs")
```

```
## [1] -0.838467
```

Hence, there's a strong negative correlation between this two variables.

(vii) Evaluate the following statement: "Because log(income) and prppov are so highly correlated, they have no business being in the same regression."

In econometric models, when two variables are high correlated, might cause bias in estimation process. However, in each specific case, the analyst must interpret if the inclusion of these variables are necessary for a efficient estimation process.