Chapter 4 - Inference

Thalles Quinaglia Liduares

23/03/2022

Upload packages

```
library(wooldridge)
library(car)
library(lmreg)
```

Upload database

```
data<-wooldridge::discrim

attach(data)
```

Use the data in DISCRIM.RAW to answer this question. (See also Computer Exercise C8 in Chapter 3.)

(i) Use OLS to estimate the model

$$log(psoda) = \beta_0 + \beta_1 prpblack + \beta_2 log(income) + \beta_3 prppov + u$$

and report the results in the usual form. Is $\hat{\beta}_1$ statistically different from zero at the 5% level against a two-sided alternative? What about at the 1% level?

```
summary(lm1<-lm(lpsoda~prpblck+lincome+prppov))
```

```
##
## 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  9.4e-07 ***
## prpblck      0.07281    0.03068   2.373   0.0181 *
## lincome      0.13696    0.02676   5.119  4.8e-07 ***
## 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: 6.917e-08
```

The estimated equation is expressed as follows

$$\widehat{log(psoda)} = -1.46 + 0.07prpblck + 0.13log(income) + 0.38prppov$$

```
linearHypothesis(lm1, c("prpblck=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## prpblck = 0
##
## Model 1: restricted model
## Model 2: lpsoda ~ prpblck + lincome + prppov
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    398 2.6657
## 2    397 2.6284  1  0.037296 5.6333 0.0181 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient $\hat{\beta}_1$ is statistically different from zero at the 5% significance level, but not at the 1% level.

(ii) What is the correlation between log(income) and prppov? Is each variable statistically significant in any case? Report the two-sided p-values.

```
round(cor(lincome, prppov, use="complete.obs"),3)
```

```
## [1] -0.838
```

Hence, the correlation between these two variables are equal to -83.8%.

The variable `lincome` is statistically significant at the 5% level. Otherwise, the variable `prppov` is statistically significant at the 1% level.

(iii) To the regression in part (i), add the variable `log(hseval)` . Interpret its coefficient and report the two-sided p-value for $H_0 : \beta_{log(hseval)} = 0$.

```
summary(lm2<-lm(lpsoda~prpblck+lincome+prppov+lhseval))
```

```
##
## Call:
## lm(formula = lpsoda ~ prpblck + lincome + prppov + lhseval)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30652 -0.04380  0.00701  0.04332  0.35272
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84151    0.29243  -2.878 0.004224 **
## prpblck      0.09755    0.02926   3.334 0.000937 ***
## lincome     -0.05299    0.03753  -1.412 0.158707
## prppov       0.05212    0.13450   0.388 0.698570
## lhseval      0.12131    0.01768   6.860 2.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07702 on 396 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.1839, Adjusted R-squared:  0.1757
## F-statistic: 22.31 on 4 and 396 DF,  p-value: < 2.2e-16
```

The coefficient of `lhseval` is statistically different from zero at the 1% signifficance level, because his p-value<2.67e-11.

```
linearHypothesis(lm2, c("lhseval=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## lhseval = 0
##
## Model 1: restricted model
## Model 2: lpsoda ~ prpblck + lincome + prppov + lhseval
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    397 2.6284
## 2    396 2.3493  1   0.27915 47.054 2.668e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(lm2)
```

```
##                   2.5 %      97.5 %
## (Intercept) -1.41642791 -0.26660192
## prpblck      0.04002437  0.15507597
## lincome     -0.12676568  0.02078482
## prppov      -0.21229889  0.31654466
## lhseval      0.08653924  0.15607205
```

(iv) In the regression in part (iii), what happens to the individual statistical significance of log(income) and prppov? Are these variables jointly significant? (Compute a p-value.) What do you make of your answers?

These both variables are not statistically significant individually at any level.

```
linearHypothesis(lm2, c("lincome=0","prppov=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## lincome = 0
## prppov = 0
##
## Model 1: restricted model
## Model 2: lpsoda ~ prpblck + lincome + prppov + lhseval
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    398 2.3911
## 2    396 2.3493  2  0.041797 3.5227 0.03045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, its possible to reject the null hypothesis only at the 5% level of significance.