Chapter 6 - Multiple Regression Analysis: Further Issues

Thalles Quinaglia Liduares

03/04/2022

Exercise 6.4

Upload packages

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

Upload database

```
data<-wooldridge::gpa2

attach(data)
```

Use the data in GPA2.RAW for this exercise. (i) Estimate the model

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u$$

where `hsize` is the size of the graduating class (in hundreds), and write the results in the usual form. Is the quadratic term statistically significant?

```
summary(lm1<-lm(log(sat)~hsize+hsizesq))
```

```
##
## Call:
## lm(formula = log(sat) ~ hsize + hsizesq)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77744 -0.08493  0.00557  0.09465  0.40946
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  6.8960291  0.0061515 1121.032  < 2e-16 ***
## hsize        0.0196029  0.0039572    4.954 7.57e-07 ***
## hsizesq     -0.0020872  0.0005444   -3.834 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1377 on 4134 degrees of freedom
## Multiple R-squared:  0.007773,   Adjusted R-squared:  0.007293
## F-statistic: 16.19 on 2 and 4134 DF,  p-value: 9.885e-08
```

The estimated equation is expressed as follows

$$\widehat{log(sat)} = 6.89 + 0.019hsize - 0.002hsize^2$$

The coefficient associated to `hsize^2` is statistically significant at the 1% level.

(ii) Using the estimated equation from part (i), what is the "optimal" high school size? Justify your answer.

Deriving the equation in relation to the `hsize` and setting equal to 0 to find the optimal value

$$\frac{\partial log(sat)}{\partial hsize} = 0.019 - 2 \cdot 0.002hsize = 0$$

$$hsize^* = \frac{0.019}{2 \cdot 0.002}$$

$$hsize^* = 4.75$$

(iii) Is this analysis representative of the academic performance of all high school seniors? Explain.

No. In terms of R^2 the model explain just 0.7% of the variability of `sat` , which is in fact, a very low power of explanation.

(iv) Find the estimated optimal high school size, using log(sat) as the dependent variable. Is it much different from what you obtained in part (ii)?

Already done, to smooth the estimated coefficients, the `log(sat)` was used.