Chapter 3 - Linear Regression

Thalles Quinaglia Liduares

04/03/2022

Applied Exercise 3.10

Upload packages

```
library(lmreg)
library(ISLR)
```

upload database

```
data<-ISLR::Carseats
```

10. This question should be answered using the Carseats data set

(a) Fit a multiple regression model to predict `Sales` using `Price` , `Urban` , and `US` .

```
lm1<-lm(Sales~Price+factor(Urban)+factor(US), data)

summary(lm1)
```

```
##
## Call:
## lm(formula = Sales ~ Price + factor(Urban) + factor(US), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13.043469   0.651012  20.036  < 2e-16 ***
## Price            -0.054459   0.005242 -10.389  < 2e-16 ***
## factor(Urban)Yes -0.021916   0.271650  -0.081    0.936
## factor(US)Yes     1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

The estimated equation is given by

$$\widehat{Sales} = 13.04 - 0.05Price - 0.02Urban + 1.20US \tag{1}$$

If the price of carseat, represented by the variable `Price` , increases \$1, the mountant of Sales decreases \$0.05, *ceteris paribus*.Controlling for store in US, the mountant of sales is \$1.20 higher in relation one store out of US.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly

Given by (1).

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

The following variables have statistical significance to the 1% level: `Intercept` , `Price` and `US` .

The variable `Urban` do not have statistical significance.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
lm2<-lm(Sales~Price+US, data)

summary(lm2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

In this case, there’s only one slightily difference between the estimated coefficients.

(f) How well do the models in (a) and (e) fit the data?

In both cases, with base on R-Squared, the models is aproximatelly 23% explaneid by the predictor variables.

(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

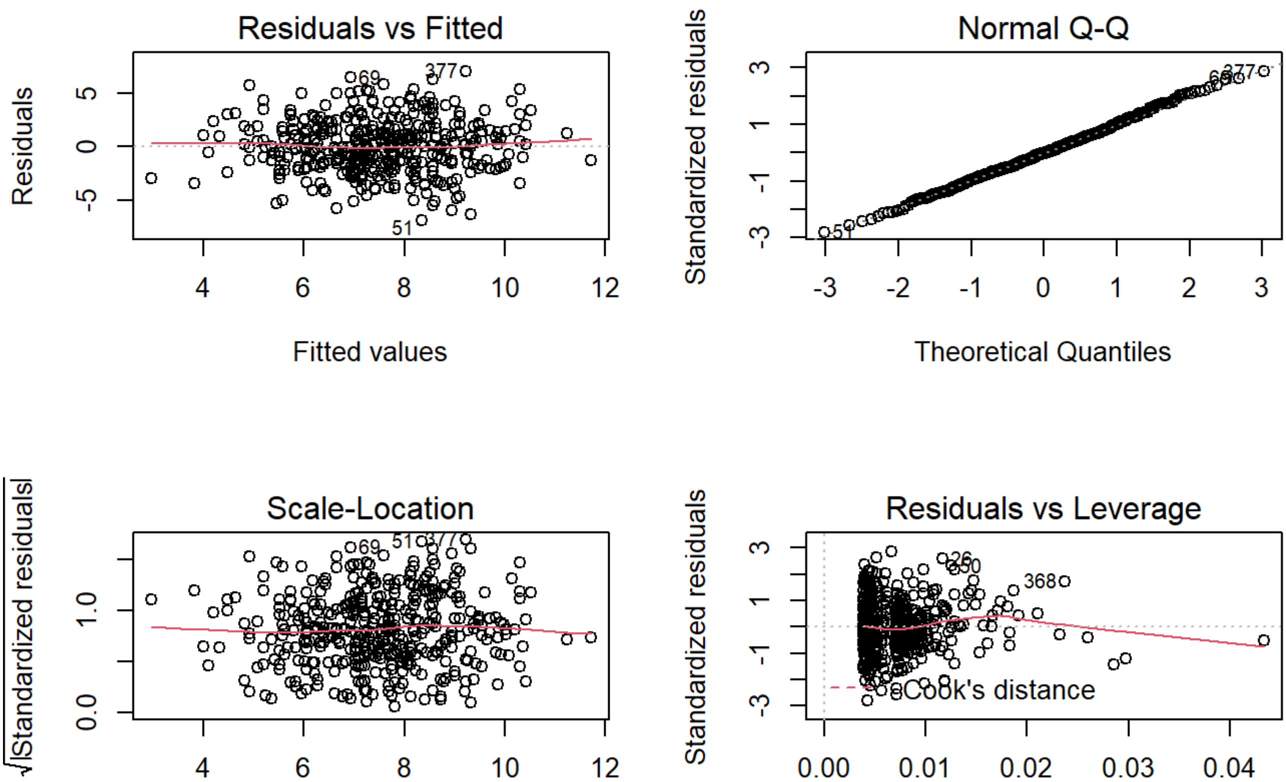
```
confint(lm2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

(h) Is there evidence of outliers or high leverage observations in the model from (e)?

```
par(mfrow=c(2,2))

plot(lm2)
```



As showed by the `Residuals vs Leverage` plot, the observations #26 and #368 might be outliers, as measured by Cook distance.