Individual_assignment2

Yuhan_Xu_474154

2019/9/7

Prefix

This question should be answered using the Carseats data set.

```
library("ISLR")
fix(Carseats)
attach(Carseats)
```

(a)

Q: Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
lm.fit1 = lm(Sales~Price+Urban+US)
```

(b)

Q: Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

```
summary(lm.fit1)

##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## 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 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        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
```

A:

-Price negatively correlates with Sales considering the small p-value. That is, when price goes up, the sales will decrease.

-Urban doesn't have a significant relationship with Sales considering the high p-value. So, whether the store is in an urban or rural location has no impact on sales.

-USYes positively correlates with Sales considering the small p-value. That is, whether the store is in the US or not will affect the sales.

(c)

Q: Write out the model in equation form, being careful to handle the qualitative variables properly.

A: To write out the model in equation form, first we need to know the coding that R uses for US, the qualitative variable.

```
contrasts(Urban)

##     Yes
## No    0
## Yes   1

contrasts(US)

##     Yes
## No    0
## Yes   1
```

Sales = 13.04 - 0.05 * Price - 0.02 * UrbanYes + 1.2 * USYes

Specifically, When a store is in US urban area, Sales = 14.22 - 0.05 * Price

When a store is in US rural area, Sales = 14.24 - 0.05 * Price

When a store is in urban area but not in US, Sales = 13.02 - 0.05 * Price

When a store is in rural area but not in US, Sales = 13.04 - 0.05 * Price

(d)

Q: For which of the predictors can you reject the null hypothesis H0 : βj = 0 ?

A: For Price and USYes, we can reject the null hypothesis, since two p-value are both extremely small.

(e)

Q: 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.

```
lm.fit2 = lm(Sales~Price+US)
```

(f)

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

```
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US)
##
## 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
```

A:	lm.fit1	lm.fit2
Residual standard error	2.472	2.469
Adjusted R-squared	0.2335	0.2354

Considering the Residual standard error and R-squared of these two models, the model in (e) fit the data better.

(g)

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

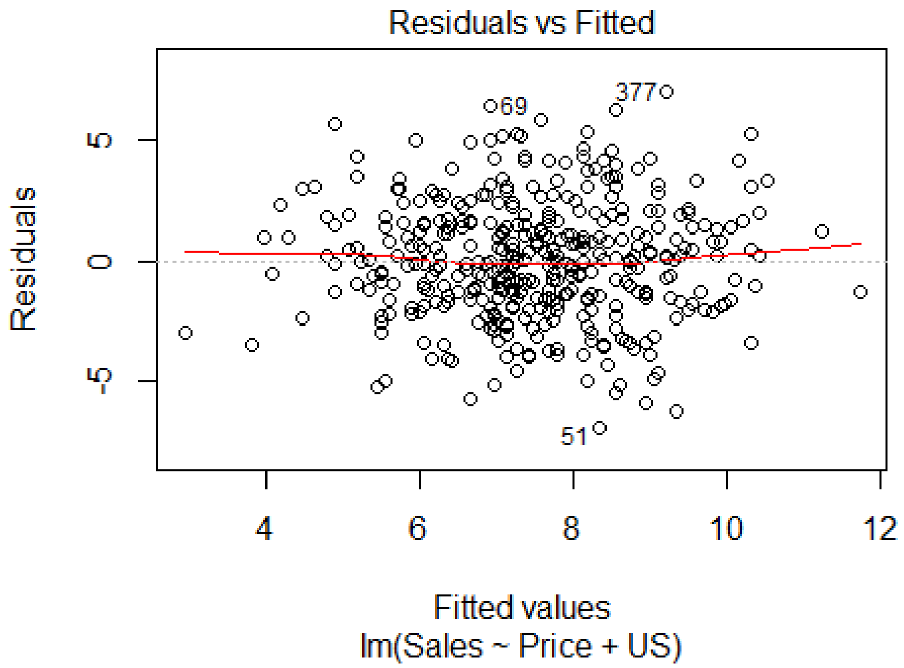
```
confint(lm.fit2)
```

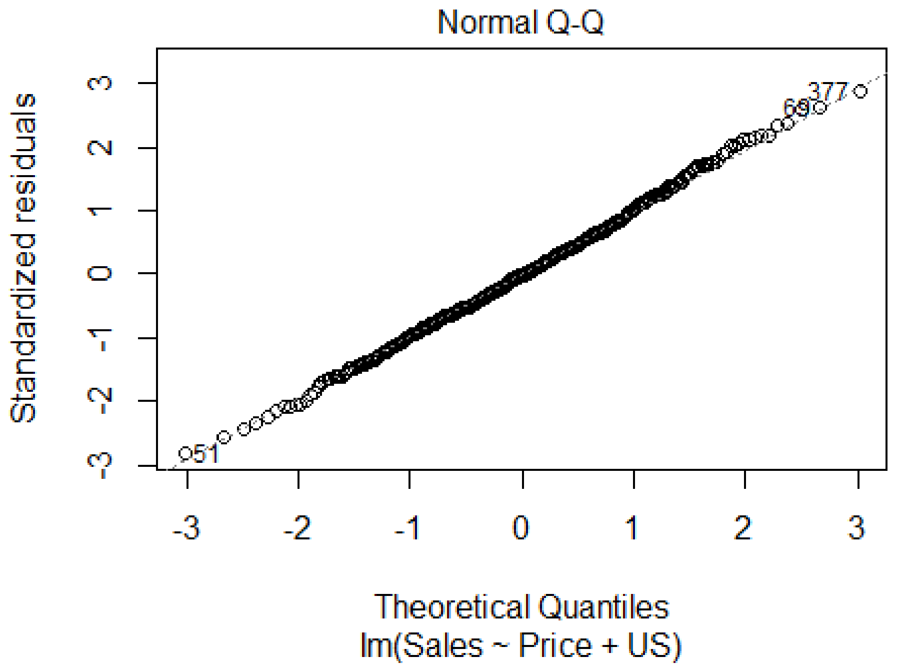
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

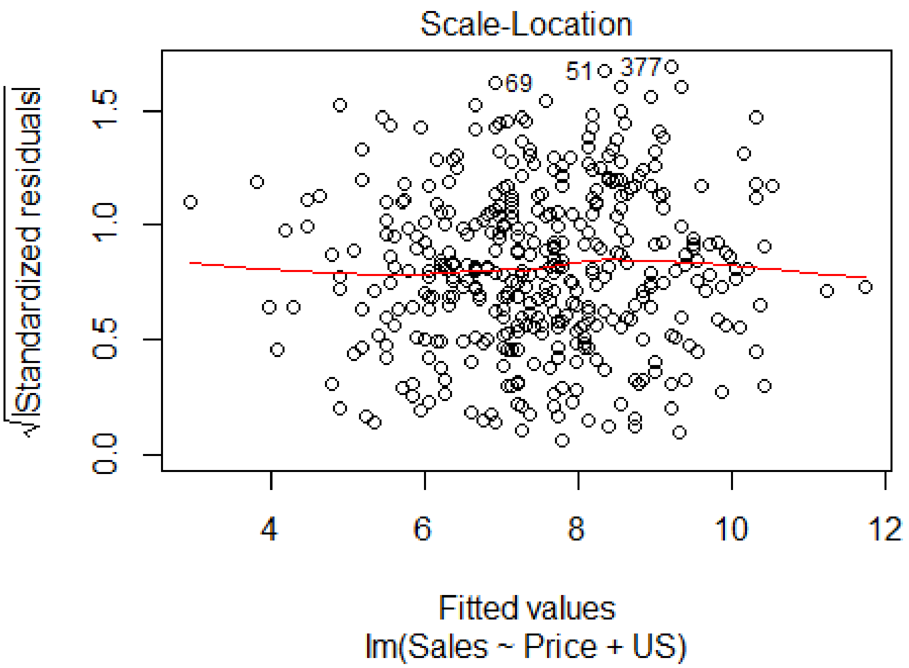
(h)

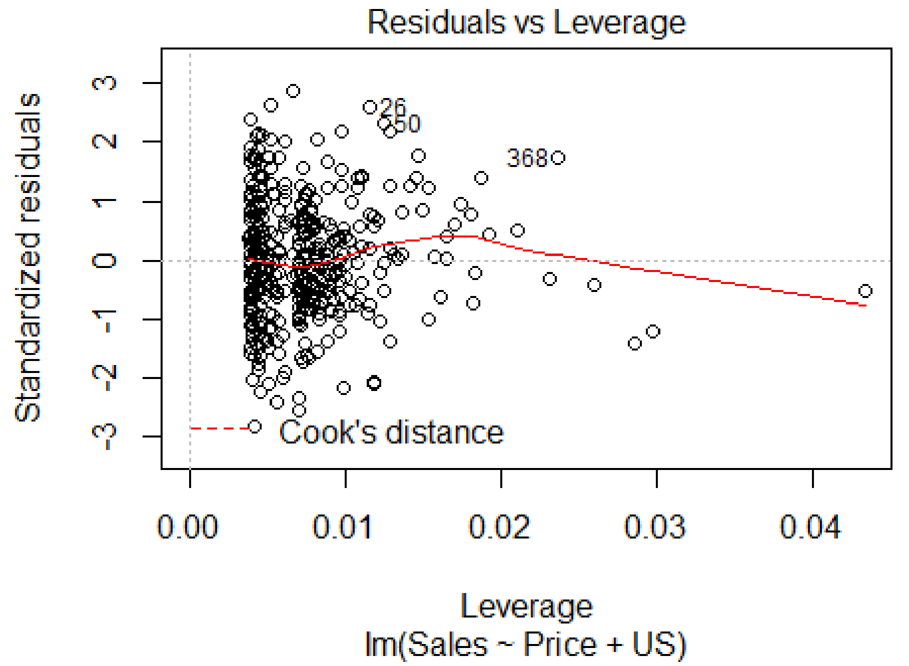
Q: Is there evidence of outliers or high leverage observations in the model from (e)?

```r
plot(lm.fit2)
```









The residual plot do not have any pattern, so there is not particular outliers. However, from the standardized residuals vs leverage plot, we can see that some observations are approaching the right side of the graph, thus have a high leverage.