ISL Exercise 3.7.10

Exercise 3.7.10

This question should be answered using the `Carseats` dataset.

a) Fit a multiple regression model to predit `Sales` using `Price`, `Urban` and `US`

Let's have a look at the dataset first:

```
library(ISLR2)
attach(Carseats)
?Carseats

fit1 <- lm(Sales ~ Price + Urban + US, data=Carseats)
```

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

```
summary(fit1)
```

```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

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
```

What we see first is that the variables `Urban` and `US` are both qualitative, the first representing whether the store is in an urban or rural location, and the second indicating whether the store was in the US or not.

The `Price` coefficient being slightly negative suggests that higher prices have a mildly negative effect on the unit sales.

The `UrbanYes` coefficient being close to 0 suggests that it has little to no effect, which is supported by the high p-value.

The `USYes` coefficient being close to 1 with a small p-value suggests that US locations have a higher chance of selling more car seats.

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

$Sales = \beta_0 + \beta_1 Price + \beta_2 UrbanYes + \beta_3 USYes$, where

$$UrbanYes = \begin{cases} 1 & \text{if urban location} \\ 0 & \text{if rural location} \end{cases}$$

and US yes is similarly encoded as a dummy variable.

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

As mentioned above, `UrbanYes`.

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.

```
fit2 <- lm(Sales ~ Price + US, data=Carseats)
```

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

```
# We should check the r2 coefficients of both models
summary(fit1)$r.squared
```

```
[1] 0.2392754
```

```
summary(fit2)$r.squared
```

```
[1] 0.2392629
```

Seems both are terrible.

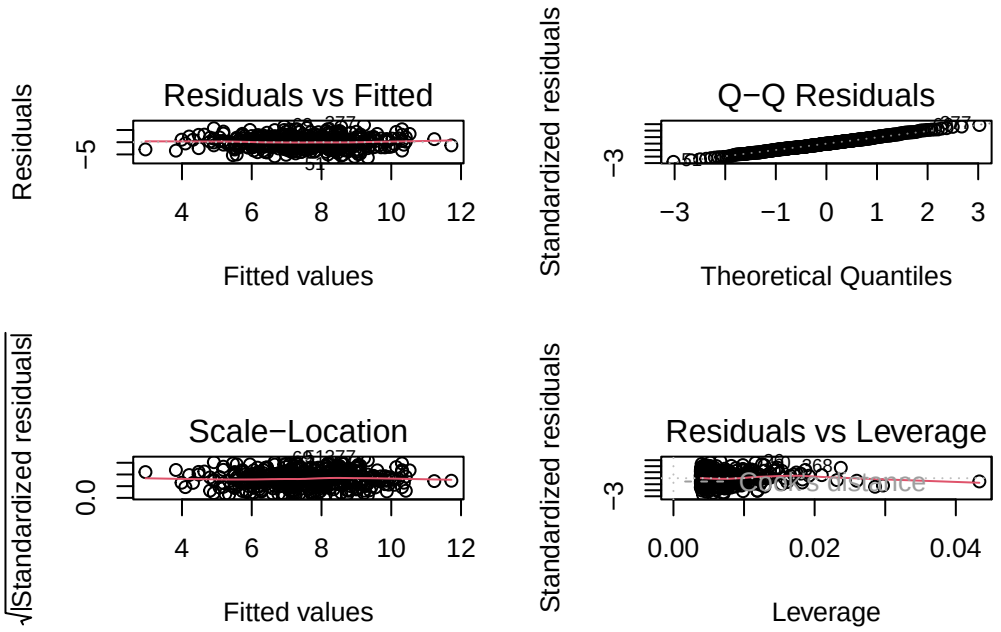
g) Using the model from e) give confidence intervals

```
confint(fit2)
```

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

f) Are there any high-leverage points?

```
par(mfrow=c(2,2))
plot(fit2)
```



Yes there's a really obvious one on the bottom right. Consider removing it.