DSO 530 - Homework 1

Xueyan Gu

2/2/2019

ISLR Chapter 3

3.7 Exercises

1. Answer:

(1) The null hypotheses related to Table 3.4 are that "TV", "radio" and "newspaper" advertising budgets do not have effects on product sales. In other words, the null hypotheses can be:

$$H_0^1 : \beta_1 = 0$$

$$H_0^2 : \beta_2 = 0$$

$$H_0^3 : \beta_3 = 0$$

(2) Based on Table 3.4, we can see that the p-values for the advertising budgets of "TV" and "radio" are highly significant since they are less than 0.05. On the other hand, the p-value for the advertising budgets of "newspaper" is not significant since it is greater than 0.05. So we reject H_0^1 and H_0^2 and we do not reject H_0^3, which mean that "TV" and "radio" advertising budgets have effects on product sales while "newspaper" advertising budgets do not affect product sales.

3. Answer:

(a) The answer is iii. Reasons are as follows: The model we get is as follows:

$$\hat{y} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA\ddot{O}IQ - 10GPA\ddot{O}Gender$$

For the females, since it is 1 for Female in the model, so we have:

$$\hat{y} = 85 + 10GPA + 0.07IQ + 0.01GPA\ddot{O}IQ$$

For the males, since it is 0 for Male in the model, so we have:

$$\hat{y} = 50 + 20GPA + 0.07IQ + 0.01GPA\ddot{O}IQ$$

Thus, we can see that if $\hat{y} = 50 + 20GPA > \hat{y} = 85 + 10GPA$, which is $GPA > 3.5$, then GPA is high enough and answer iii is correct.

(b) When IQ = 110 and GPA = 4.0, if we put the values into the model, we can get:

$$\hat{y} = 85 + 40 + 7.7 + 4.4 = 137.1$$

Thus we can predict that the salary is \$137,100.

(c) False. If we want to infer whether GPA/IQ has an effect on the entire model, we need to test the null hypothesis $H_0 : \beta\hat{4} = 0$ and to look at the p-value based on t test or F test. And we can not draw the conclusion only based on the coefficient of the model.

4. Answer:

(a) We expect that one may be lower than the other. Since we know that the relationship between X and Y is linear, we can infer that the linear regression line we get is relatively close to the true regression line. So we can also infer that the training RSS for the linear regression is relatively lower than the training RSS for the cubic regression.

(b) We do not have enough information to conclude but we can make some assumptions. Since the test RSS depends on the test data and we have no information about the test data, so there is not enough information for us to conclude. However, we can make assumptions that the test RSS for cubic regression may be higher than the test RSS for linear regression since there may be more deviations from the training cubic regression model.

(c) We expect that one may be lower than the other. Since we know that the relationship between X and Y is not linear and we have no idea how far it is from linear, so the linear regression we get is hard to be close to the true regression line. Besides, since the cubic regression is more flexible than linear regression, so we can assume that there are less deviations for cubic regression and the training RSS for cubic regression is lower than the training RSS for linear regression.

(d) We do not have enough information to conclude given the known information. Since it is hard to define "how far it is from linear", we do not know it is closer to linear or cubic. If it is closer to linear, the test RSS for linear regression may be lower than the test RSS for cubic regression. Vice Versa. So we have not enough information to conclude.

5. Answer:
If we substitute with the equation of $\hat{\beta}$ into the equation $\hat{y_i} = x_i\hat{\beta}$, we can get:

$$\hat{y_i} = x_i \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} {x_{i'}}^2} = \sum_{i'=1}^{n} \frac{x_i {x_i}'}{{x_{i'}}^2} {y_i}' = \sum_{i'=1}^{n} {a_i}' {y_i}'$$

6. Answer:
The least squares line is:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

If we substitute x with $\bar{x}$, then we can get:

$$y = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

Thus, we can conclude that the least squares line always passes through the point $(\bar{x}, \bar{y})$.

7. Answer:
The equation of R^2 is (given $\bar{y} = 0$):

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{\sum_{j=1}^{n} y_j^2}$$

If we substitute $\hat{y_i} = \hat{\beta} x_i$, we can get:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \sum_{j=1}^{n} x_j y_j / \sum_{j=1}^{n} x_j^2 x_i)^2}{\sum_{j=1}^{n} y_j^2}$$

$$= \frac{\sum_{j=1}^{n} y_j^2 - (\sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i (\sum_{j=1}^{n} x_j y_j / \sum_{j=1}^{n} x_j^2) x_i + \sum_{i=1}^{n} (\sum_{j=1}^{n} x_j y_j / \sum_{j=1}^{n} x_j^2)^2 x_i^2)}{\sum_{j=1}^{n} y_j^2}$$

So we can get:

$$R^2 = \frac{2(\sum_{i=1}^{n} x_i y_i)^2 / \sum_{j=1}^{n} x_j^2 - (\sum_{i=1}^{n} x_i y_i)^2 / \sum_{j=1}^{n} x_j^2}{\sum_{j=1}^{n} y_j^2} = \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2} = r(X, Y)^2$$

Thus, we can conclude that in the case of linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y.

8. Answer:

(a)

```
library(ISLR)
lm.fit = lm(mpg ~ horsepower, data = Auto)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. If we test the null hypothesis that is:

$$H_0 : \beta_1 = 0$$

we can see than the p-value from F test is less than 2.2e-16, which means that there is a relationship between "mpg" and "horsepower".

ii. We can see that R-squared is 0.6059, which means that 60.59% of the variability in the response "mpg" can be explained by using the predictor "horsepower".

iii. Since the coefficient of "horsepower" is -0.157845, which is negative, so the relationship between "mpg" and "horsepower" is negative.

iv.

```
predict(lm.fit, data.frame(horsepower = 98), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

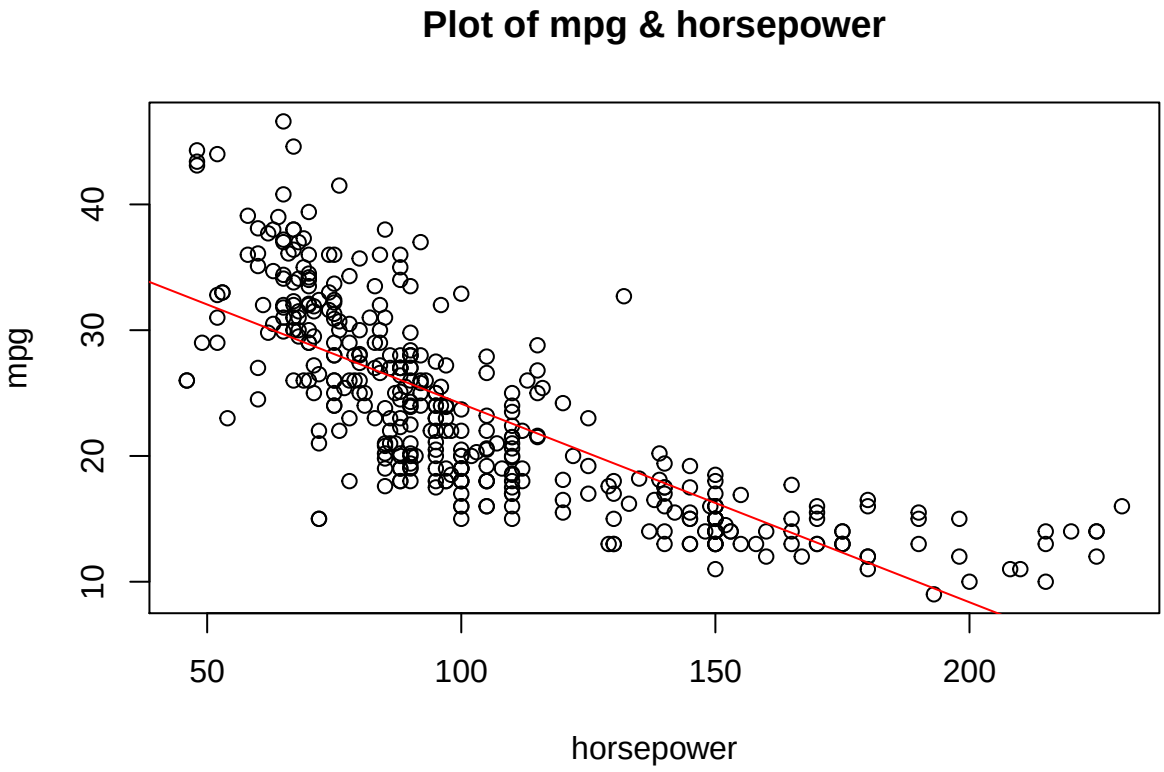
```
predict(lm.fit, data.frame(horsepower = 98), interval = "prediction")
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

1) As we can see, the predicted "mpg" associated with a "horsepower" of 98 is 24.46708.
2) The associated 95% confidence intervals are from 23.97308 to 24.96108.
3) The associated 95% prediction intervals are from 14.8094 to 34.12476.

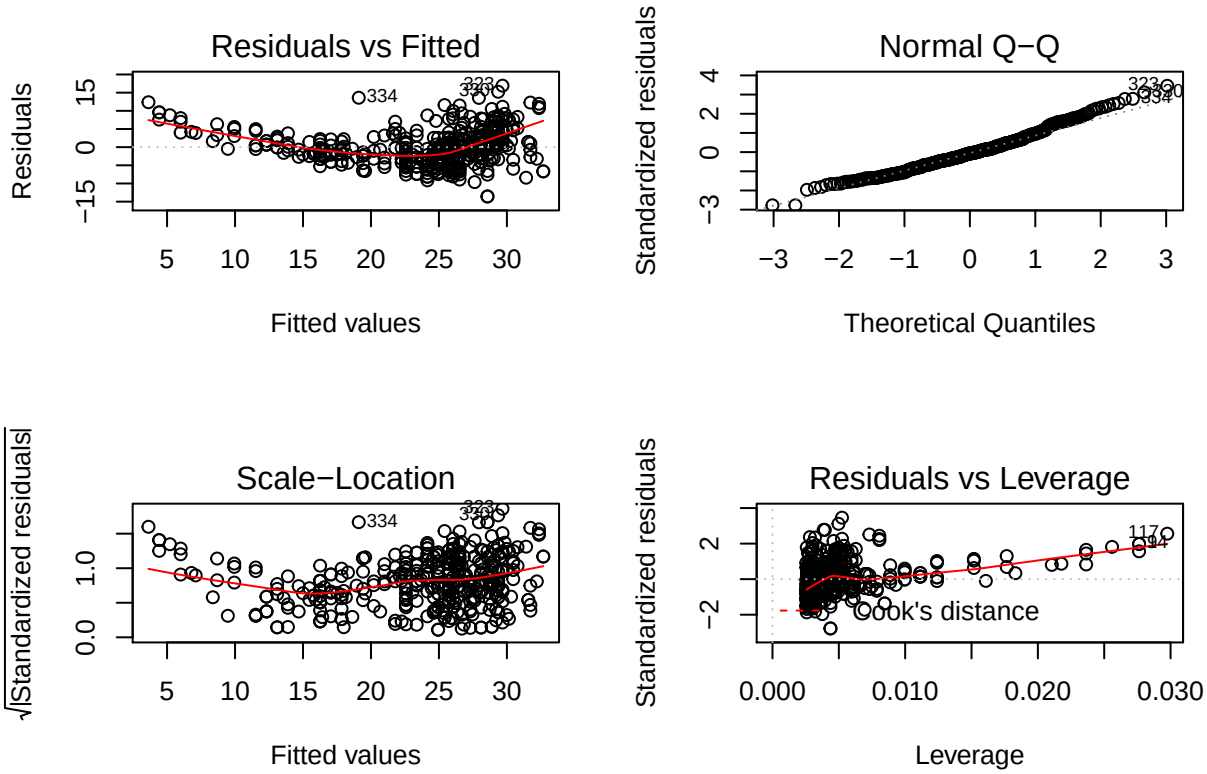
(b)

```
plot(Auto$horsepower, Auto$mpg, main = "Plot of mpg & horsepower", xlab = "horsepower", ylab = "mpg")
abline(lm.fit, col = "red")
```



(c)

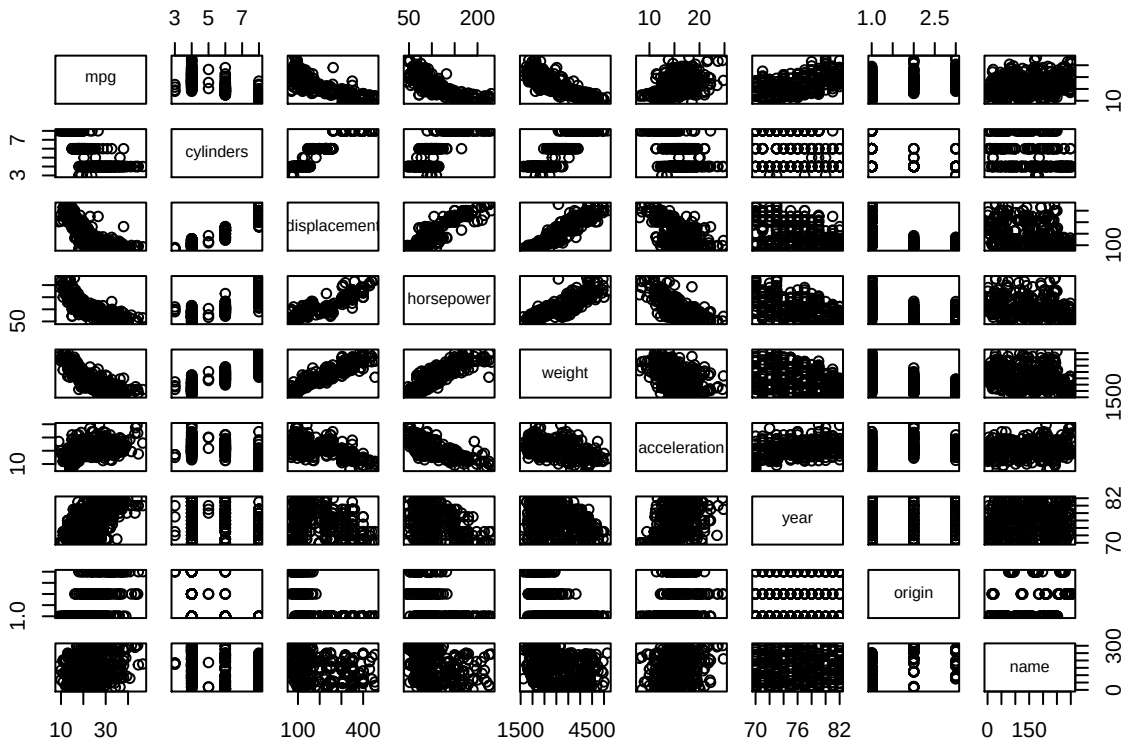
```
par(mfrow = c(2, 2))
plot(lm.fit)
```



As we can see, the first plot which displays the residuals vs. fitted values, shows that there is no linear tendency in the "mpg" and "hoursepower". Besides, the forth plot which displays the residuals vs. leverage values, shows that there are a few outliers and a few leverage points in the data.

9. Answer: (a)

```
pairs(Auto)
```



(b)

```
names(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

```
#We can see that "name" is the ninth variable.
```

```
cor(Auto[1:8])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##               acceleration       year     origin
## mpg              0.4233285  0.5805410  0.5652088
## cylinders       -0.5046834 -0.3456474 -0.5689316
## displacement    -0.5438005 -0.3698552 -0.6145351
## horsepower      -0.6891955 -0.4163615 -0.4551715
## weight          -0.4168392 -0.3091199 -0.5850054
## acceleration     1.0000000  0.2903161  0.2127458
## year             0.2903161  1.0000000  0.1815277
```

```
## origin          0.2127458  0.1815277  1.0000000
```

(c)

```
lm.fit2 = lm(mpg ~ . - name, data = Auto)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i. If we test the null hypothesis that is:

$$H_0 : \beta_1 = 0$$

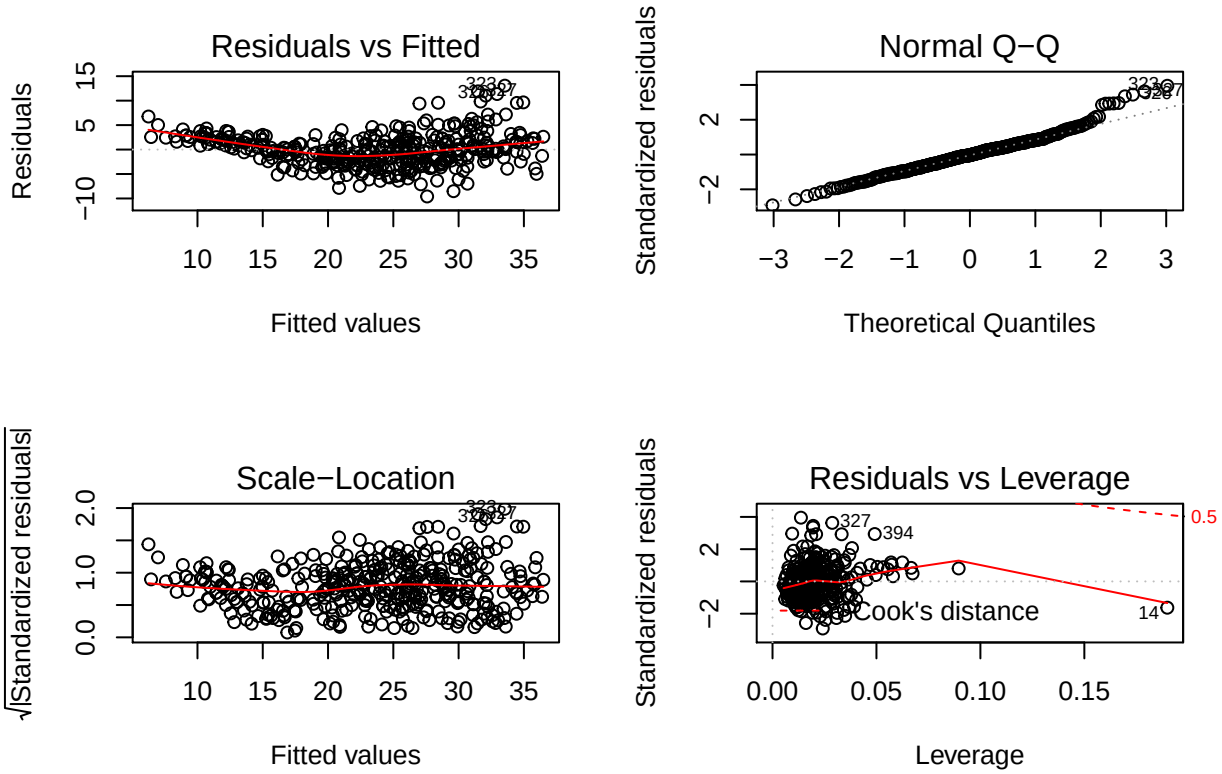
we can see than the p-value from F test is less than 2.2e-16, which means that there is a relationship between "mpg" and the other predictors.

ii. If we check the p-values of each predictor's t-statictic, we can conclude that "displacement", "weight", "year", "origin" respectively have a statistically significant relationship to the response "mpg".

iii. The coefficient of the "year" variable is 0.750773, which suggests that given all other predictors remaining constant, an increase of 1 year results in an increase of 0.7507727 in "mpg".

(d)

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



As we can see, the first plot which displays the residuals vs. fitted values, shows that there is no linear tendency in the data. Besides, the forth plot which displays the residuals vs. leverage values, shows that there are a few outliers and one leverage point in the data.

(e) Based on the correlation matrix from (b), we can see that the correlation between "cylinders" and "displacement" and the correlation between "displacement" and "weight" are the highest correlations in the matrix. So we use these two pairs to fit the model with interaction effects.

```
lm.fit3 = lm(mpg ~ cylinders * displacement + displacement * weight, data = Auto[, 1:8])
summary(lm.fit3)
```

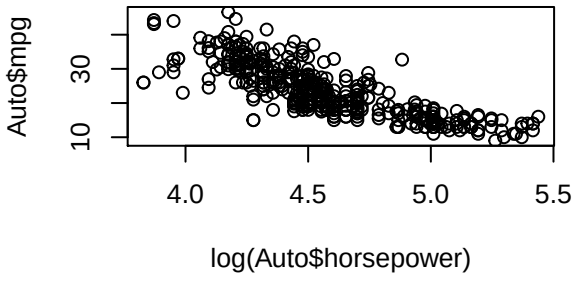
```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto[, 1:8])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders                    7.606e-01  7.669e-01   0.992    0.322
## displacement                -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                      -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement      -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight          2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

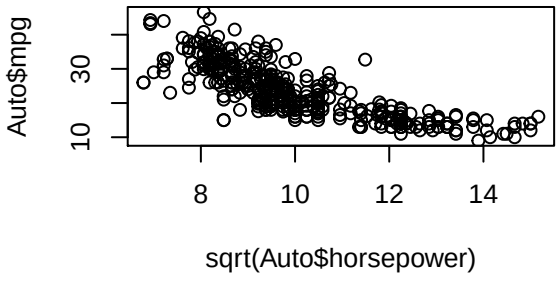
```
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

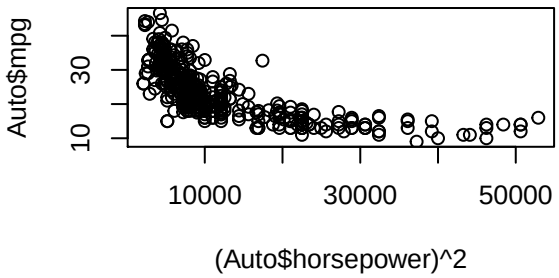
Based on the p-values, we can see that the interaction between "displacement" and "weight" is statistically significant, while the interaction between "cylinders" and "displacement" is not statistically significant.

(f)

```
par(mfrow = c(2, 2))
plot(log(Auto$horsepower), Auto$mpg)
plot(sqrt(Auto$horsepower), Auto$mpg)
plot((Auto$horsepower)^2, Auto$mpg)
```







If we select "horsepower" as our only predictor, we can see that the log transformation displays a more linear looking plot.

10. Answer:

(a)

```
lm.fit4 = lm(Sales ~ Price + Urban + US, data = Carseats)
summary(lm.fit4)
```

```
##
## 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
```

(b)

(1) The coefficient of the "Price" variable shows that given other predictors remaining fixed, the average effect of a price increase of \$1 is a decrease of 54.459 units in sales.
(2) Since "Urban" is a qualitative variable, the coefficient of the "Urban" shows that given other predictors remaining fixed, on average, the unit sales in urban are 21.916 units less than the unit sales in rural.
(3) Since "US" is a qualitative variable, the coefficient of the "US" shows that given other predictors remaining fixed, on average, the unit sales in the US are 1200.573 units more than the unit sales in the non US area.

(c) Based on the result, the model can be written as:

$$Sales = 13.043469 + (-0.054459)\text{Ö}Price + (-0.021916)\text{Ö}Urban + (1.200573)\text{Ö}US + \epsilon$$

For "Urban" variable, if Urban = 1, it means that the store is in an urban area, while if Urban = 0, it means that the store is in a rural area. For "US" variable, if US = 1, it means that the store is in the US, while if US = 0, it means that the store is in non US area.

(d) Based on the result, we can reject the null hypothesis for "Price" and "US" variable.

(e)

```
lm.fit5 = lm(Sales ~ Price + US, data = Carseats)
summary(lm.fit5)
```

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

(f) Since the adjusted R^2 for smaller model is higher than the adjust R^2 for bigger model, so we can say that the adjusted R^2 for smaller model is better than the one for bigger model. For both model, about 23.93% of the variability can be explained by the models.

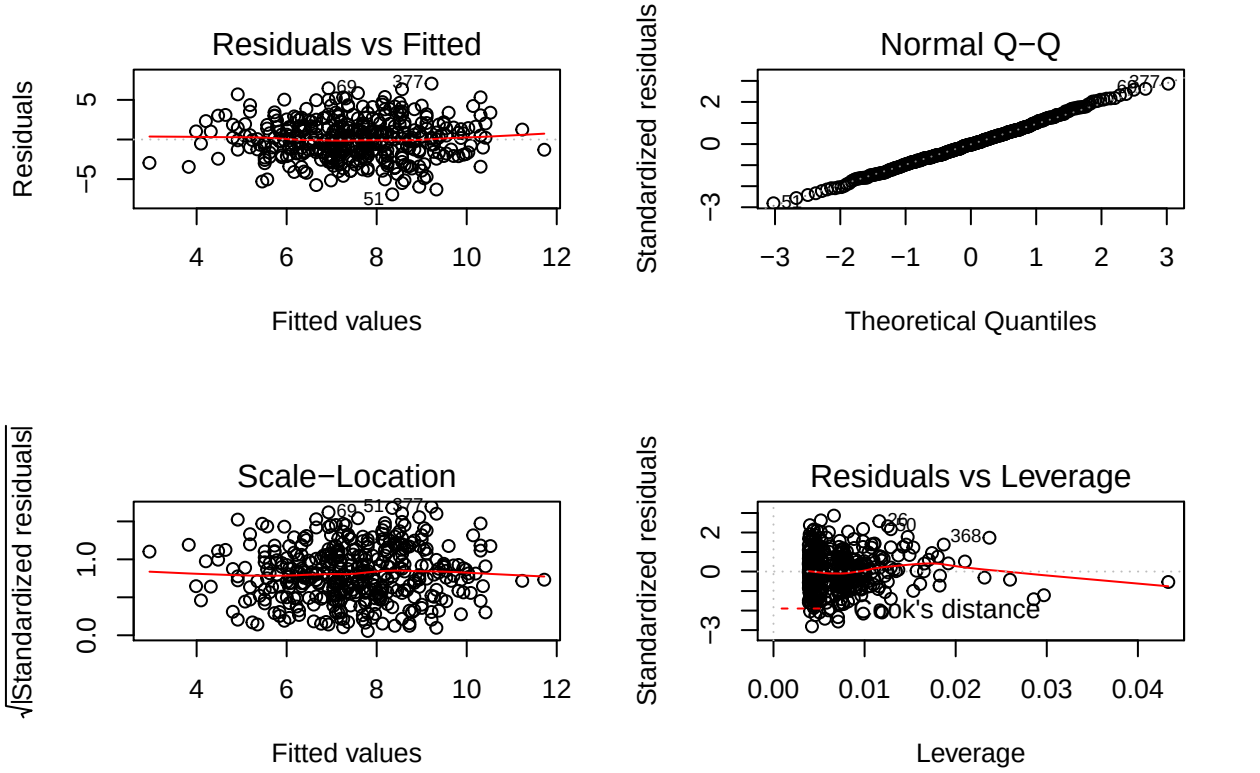
(g)

```
confint(lm.fit5)
```

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

(h)

```
par(mfrow = c(2, 2))
plot(lm.fit5)
```



There is evidence of outliers or high leverage observations in the smaller model. As we can see, the forth plot which displays the residuals vs. leverage values, shows that there are a few outliers (higher than 2 or lower than -2) and leverage points (points exceed (p+1)/n (0.01)) in the data.