90/90

Good job!

Homework 01

Hongfa Huang

8/8

1

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of *sales*, *TV*, *radio*, and *newspaper*, rather than in terms of the coefficients of the linear model.

Answers: Null hypotheses is that advertising on TV, radio or newspaper does not have effect on sales. Based on the p-value, we can conclude that it is statistically significant for TV and Sales to affect the sales but not for newspaper.

maybe radio here?

9/9

3

Suppose we have a dataset with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender(1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0$ =50, $\hat{\beta}_1$ = 20, $\hat{\beta}_2$ = 0.07, $\hat{\beta}_3$ = 35, $\hat{\beta}_4$ = 0.01, $\hat{\beta}_5$ = -10.

(a) Which answer is correct, and why?

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or False: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Answers:

(a) iii is correct. Given IQ and GPA, the difference between male and female can be $\Delta_{income} = 35 - 10X_1$. Therefore, males earn more on average than females provided that GPA is high enough.

(b) A female with IQ of 100 and GPA of 4.0 will have salary of 137.1 thousands of dollars.

(c) False. The number of the interaction between GPA/IQ will be large. Thus, the samll coefficient will also have large effect. Plus, we can see the p-value of the coefficient to see the effect.

12/12

4

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

Answers:

(a) Though the true relationship between X and Y is linear, the training RSS for cubic regression would be a bit smaller than the RSS for the linear regression. Because by adding more variables, we overfit the model and lead to decrease of RSS.

(b) Since the true relationship between X and Y is linear, the for testing data, the cubic regression will have the overfitting problem. Therefore, it would have more error than the simple linear regression.

(c) The training RSS for cubic regression would be a smaller than the RSS for the linear regression. Because by adding more variables, we can always get better match with the observations and thus reduce the training RSS.

(d) There is not enough information to tell which test RSS would be greater because we don't know the relation between X and Y. If it is closer to linear, the test RSS for cubic regression will be greater. On the other hand, if it is closer to polynomial regression, the test RSS for cubic regression will be smaller then the test RSS for the linear regression.

8/8 **5**

Consider the fitted values that result from performing linear regres- sion without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y}_i = x_i\hat{\beta}$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right)$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

What is $a_{i'}$?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations *of the response values.*

Answers:

To input $\hat{\beta}$ in the first equation, we can get:

$$\hat{y}_i = x_i \times \frac{\sum_{i'=1}^{n} (x_{i'} y_{i'})}{\sum_{j=1}^{n} x_j^2}$$

Then it can be transformed into:

$$\hat{y}_i = \sum_{i'=1}^{n} \left(\frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2} \times y_{i'}\right)$$

It can be shown that $a_{i'}$:

$$a_{i'} = \frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2}$$

8/8

6

Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point($\bar{x}$, $\bar{y}$).

Answers:

The least squares line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1}$$

from (3.4):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

Then we replace the $\hat{\beta}_0$ in equation 1 using equation 2. We can get:

$$\hat{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \tag{3}$$

Then we input the point($\bar{x}$, $\bar{y}$) to equation 3:

$$\bar{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}$$
$$0 = 0$$

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

8/8

7

It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

Answers:

First of all, we assume $\bar{x} = \bar{y} = 0$.

From 3.17, we can get:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

From 3.18, we can get:

$$r = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \times \sum_{i=1}^n y_i^2}}$$

Then,

$$r^2 = \frac{\left(\sum_{i=1}^n x_i y_i\right)^2}{\sum_{i=1}^n x_i^2 \times \sum_{i=1}^n y_i^2}$$

Also,

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2 = \sum_{i=1}^n \left(y_i - \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) x_i\right)^2$$

Thus:

$$\frac{TSS - RSS}{TSS} = r^2$$

$$R^2 = [Cor\,(X, Y)]^2$$

9/9

8

This question involves the use of simple linear regression on the *Auto* dataset.

(a) Use the *lm()* function to perform a simple linear regression with *mpg* as the response and *horsepower* as the predictor. Use the *summary()* function to print the results. Comment on the output.

```
library(MASS)
library(ISLR)
attach(Auto)
names(Auto)
```

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

```
lm1.fit = lm(mpg ~ horsepower)
summary(lm1.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower)
##
## 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
```

Conclusion: We can see from the summary table that there is a statistically significant relationship between `horsepower` and `mpg` and that the relationship is pretty strong due to the very small p-value. And the coefficient is negative which means given other conditions constant, if `horsepower` increases, `mpg` would be very likely to decrease. R-squared can also be used to analyze here

```
predict(lm1.fit, data.frame(horsepower = 98), interval = 'confidence')
```

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

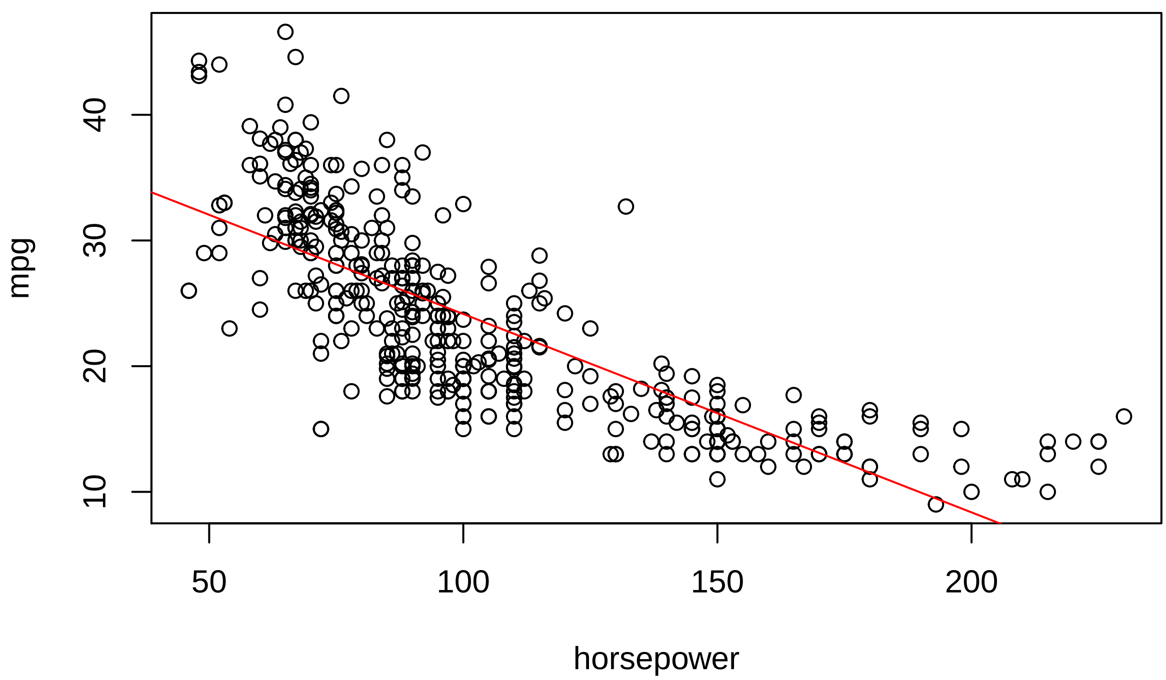
```r
predict(lm1.fit, data.frame(horsepower = 98), interval = 'prediction')
```

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

Conclusion: The predicted `mpg` associated with a `horsepower` of 98 is about 24.467. The associated 95% confidence interval is [23.973, 24.961] and prediction interval is [14.809, 34.125].

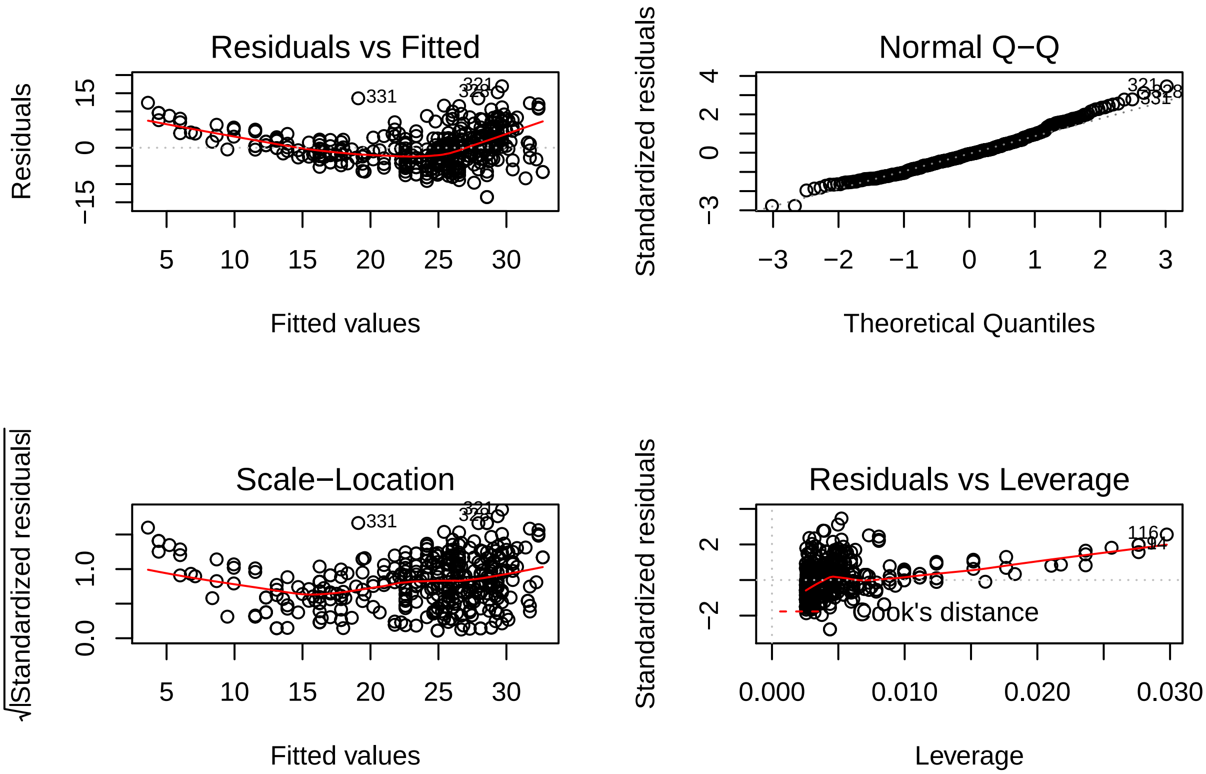
(b) Plot the response and the predictor. Use the *abline()* function to display the least squares regression line.

```r
plot(horsepower, mpg)
abline(lm1.fit, col = 'red')
```



(c) Use the *plot()* function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```r
par(mfrow=c(2,2))
plot(lm1.fit)
```



Conclusion: From the diagnostic plots, we can see the there is a curve pattern for Residuals vs Fitted. relationship between `horsepower` and `mpg` should not be linear. Also, we find that observation 116 and 94 are high leverage points and that observation 321, 328 and 331 are outliers.

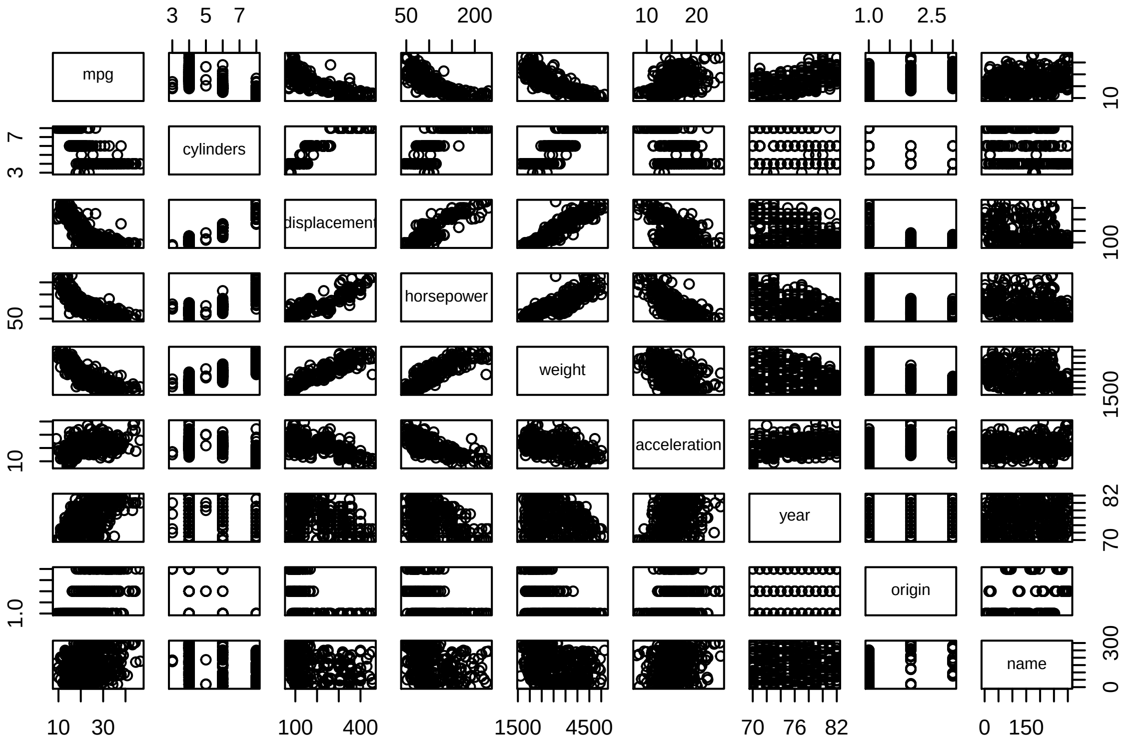
12/12

9

This question involves the use of multiple linear regression on the *Auto* dataset.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
library(ISLR)
data(Auto)
pairs(Auto)
```



(b) Compute the matrix of correlations between the variables using the function *cor()*. You will need to exclude the name variable, which is qualitative.

```
x = Auto[1:8]
y = Auto[1:8]
cor(x, y)
```

```
##                     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) Use the *lm()* function to perform a multiple linear regression with *mpg* as the response and all other variables except *name* as the predictors. Use the *summary()* function to print the results.

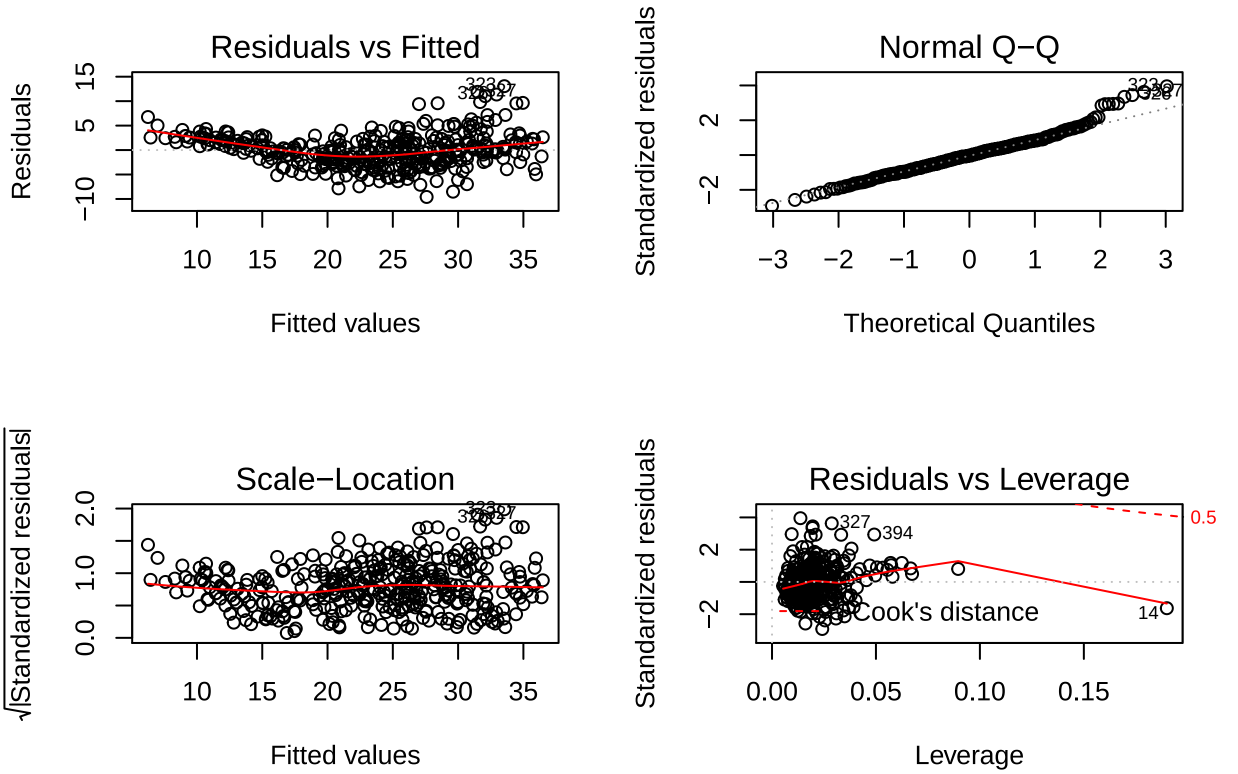
```r
lm2.fit = lm(mpg ~.-name, data=Auto)
summary(lm2.fit)
```

```
##
## 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
```

Conclusion: From the summary table, we can conclude that there is a relationship between the predictors and the response. Among the predictors, `displacement`, `weight`, `year` and `origin` have statistically significant relationships with mpg. The coefficient for `year` variable is positive which suggests that cars in the later years have higher mpg.

(d) Use the *plot()* function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```r
par(mfrow=c(2,2))
plot(lm2.fit)
```



Conclusion: From the diagnostic plots, we can see the there is a curve pattern for Residuals vs Fitted. The relation between the predictors and `mpg` should not be linear. Also, based on the Residuals vs Leverage graph, we can find out that observation 327 and 394 are outliers and observation 14 are high leverage points.

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
summary(lm(mpg ~ displacement*weight, data = Auto))
```

```
##
## Call:
## lm(formula = mpg ~ displacement * weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8664  -2.4801  -0.3355   1.8071  17.9429
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.372e+01  1.940e+00  27.697  < 2e-16 ***
## displacement        -7.831e-02  1.131e-02  -6.922 1.85e-11 ***
## weight              -8.931e-03  8.474e-04 -10.539  < 2e-16 ***
## displacement:weight  1.744e-05  2.789e-06   6.253 1.06e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.097 on 388 degrees of freedom
## Multiple R-squared:  0.7265, Adjusted R-squared:  0.7244
## F-statistic: 343.6 on 3 and 388 DF,  p-value: < 2.2e-16
```

```r
summary(lm(mpg ~ displacement:weight, data = Auto))
```

```
##
## Call:
## lm(formula = mpg ~ displacement:weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5066  -3.3499  -0.6626   2.7207  17.4808
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.126e+01  3.879e-01   80.59   <2e-16 ***
## displacement:weight -1.182e-05  4.600e-07  -25.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.763 on 390 degrees of freedom
## Multiple R-squared:  0.6285, Adjusted R-squared:  0.6276
## F-statistic: 659.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

Conclusion: These two operaters are different. We use two highly correlated variables to try these two. From the result we can see that the interaction term between `weight` and `displacement` is statistically significant. This means that the `weight` and `displacement` can affect the `mpg` of the car itself and they also have interactive effect to affect the `mpg` of the car.

(f) Try a few different transformations of the variables. Comment on your findings.

```r
summary(lm(mpg~ log(horsepower), data=Auto))
```

```
##
## Call:
## lm(formula = mpg ~ log(horsepower), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2299  -2.7818  -0.2322   2.6661  15.4695
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     108.6997     3.0496   35.64   <2e-16 ***
## log(horsepower) -18.5822     0.6629  -28.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.501 on 390 degrees of freedom
## Multiple R-squared:  0.6683, Adjusted R-squared:  0.6675
## F-statistic: 785.9 on 1 and 390 DF,  p-value: < 2.2e-16
```

```r
summary(lm(mpg~ horsepower+I(horsepower^(1/2)), data=Auto))
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + I(horsepower^(1/2)), data = Auto)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5479  -2.5677  -0.2663   2.2998  15.5098
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         105.31581    6.64657  15.845  < 2e-16 ***
## horsepower            0.41913    0.05867   7.144 4.49e-12 ***
## I(horsepower^(1/2)) -12.48574    1.26337  -9.883  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.392 on 389 degrees of freedom
## Multiple R-squared:  0.685,  Adjusted R-squared:  0.6834
## F-statistic:   423 on 2 and 389 DF,  p-value: < 2.2e-16
```

```r
summary(lm(mpg~ horsepower+I(horsepower^2), data=Auto))
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + I(horsepower^2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7135  -2.5943  -0.0859   2.2868  15.8961
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     56.9000997  1.8004268   31.60   <2e-16 ***
## horsepower      -0.4661896  0.0311246  -14.98   <2e-16 ***
## I(horsepower^2)  0.0012305  0.0001221   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.374 on 389 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.686
## F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

```r
summary(lm(mpg~ poly(horsepower, 5), data=Auto))
```

```
##
## Call:
## lm(formula = mpg ~ poly(horsepower, 5), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4326  -2.5285  -0.2925   2.1750  15.9730
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            23.4459     0.2185 107.308  < 2e-16 ***
## poly(horsepower, 5)1 -120.1377     4.3259 -27.772  < 2e-16 ***
## poly(horsepower, 5)2   44.0895     4.3259  10.192  < 2e-16 ***
## poly(horsepower, 5)3   -3.9488     4.3259  -0.913  0.36190
## poly(horsepower, 5)4   -5.1878     4.3259  -1.199  0.23117
```

```
## poly(horsepower, 5)5   13.2722     4.3259   3.068  0.00231 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.326 on 386 degrees of freedom
## Multiple R-squared:  0.6967, Adjusted R-squared:  0.6928
## F-statistic: 177.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

Conclusion: We can make the model fit better by overfitting it, for we can see that the R^2 and RSE in the last model are both the smallest among all of the four models. But by doing so, we are sacrificing on the dimensionality and the level of significance among our predictors.

16/16

10

This question should be answered using the Carseats dataset.

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

```r
library(ISLR)
attach(Carseats)
names(Carseats)
```

```
##  [1] "Sales"       "CompPrice"   "Income"      "Advertising" "Population"
##  [6] "Price"       "ShelveLoc"   "Age"         "Education"   "Urban"
## [11] "US"
```

```r
help('Carseats')
lm3.fit = lm(Sales ~ Price + Urban + US)
summary(lm3.fit)
```

```
##
## 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
```

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

- The relationships between `Price` and `Sales`, `US` and `Sales` are statistically significant. For every dollar increase in `Price`, `Sales` would decrease for 54.459 dollars. If a store is in the `US`, `Sales` would be 1200.573 dollars higher.
- Even thought the relationship between `Urban` and `Sales` is not statistically significant, we can still conclude from the coefficient that if a store is in an `Urban` location, there is a possibility that `Sales` would be 21.916 dollars lower than if it's in a rural location.

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

Let:

$$X_1 = Price \tag{1}$$

$$X_2 = \begin{cases} 1 & \text{if } Urban, \\ 0 & \text{if } Rural. \end{cases} \tag{2}$$

$$X_3 = \begin{cases} 1 & \text{if } US, \\ 0 & \text{if } Non-US. \end{cases} \tag{3}$$

So the model written in equation form would be:

$$Sales = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon = \begin{cases} 13.043 - 0.054X_1 - 0.022X_2 + 1.201X_3 + \epsilon & \text{if } Urban, US, \\ 13.043 - 0.054X_1 - 0.022X_2 + \epsilon & \text{if } Urban, Non-US, \\ 13.043 - 0.054X_1 + 1.201X_3 + \epsilon & \text{if } Rural, US, \\ 13.043 - 0.054X_1 + \epsilon & \text{if } Rural, Non-US. \end{cases} \tag{4}$$

(d) For which of the predictors can you reject the null hypothesis?

Answer: `Price` and `USYes`.

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

```r
lm4.fit = lm(Sales ~ Price + US, data = Carseats)
summary(lm4.fit)
```

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

Conclusion: They both have quite appropriate R^2 and RSE. The F-statistics prove statistically significant for two models. But for model 1, it is not significant for urbanicity and model 2 also have lower RSE. Thus, model 2 is better.

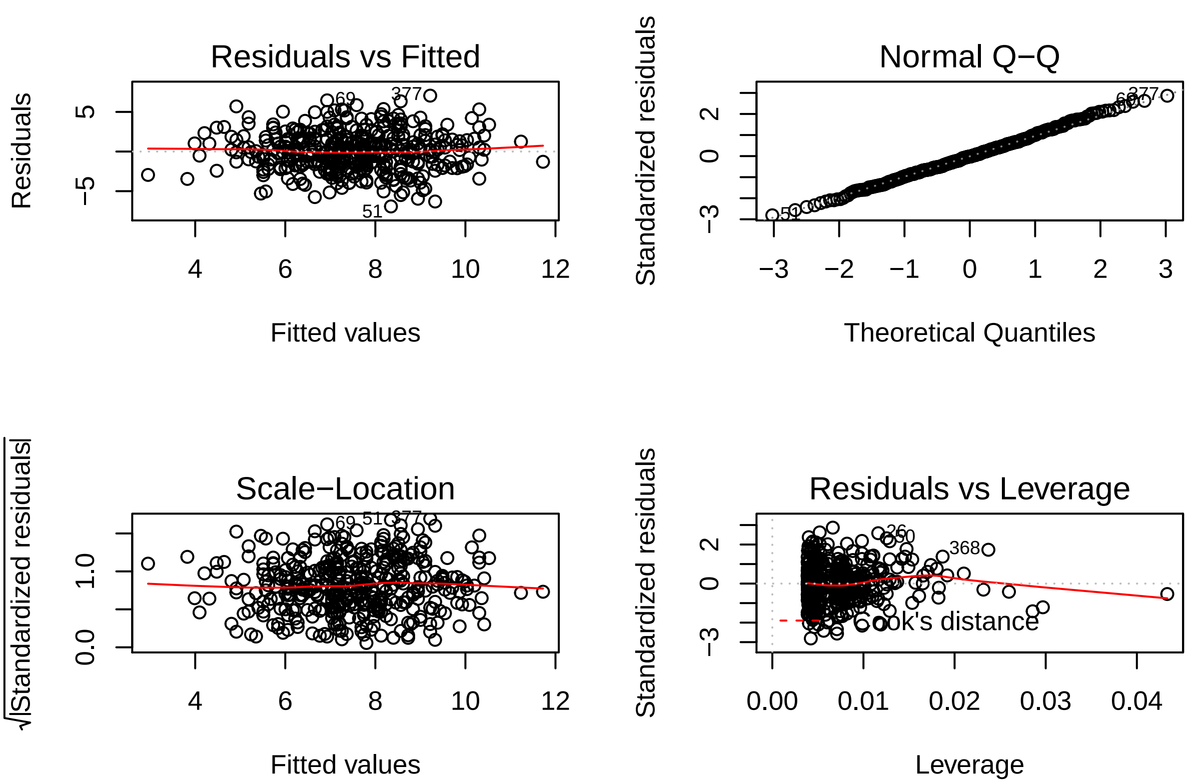
(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```r
confint(lm4.fit)
```

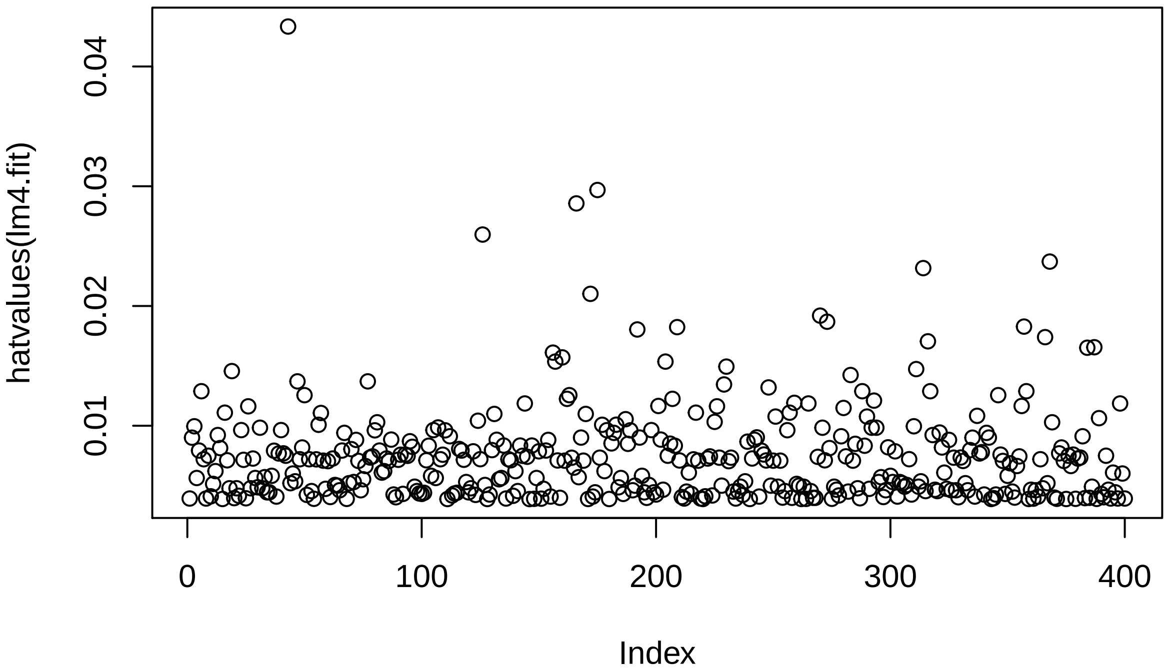
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

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

```r
par(mfrow=c(2,2))
plot(lm4.fit)
```



```r
plot(hatvalues(lm4.fit))
```




```r
which.max(hatvalues(lm4.fit))
```

```
## 43
## 43
```

Conclusion: From this diagnostic plots, we can see that observation 51, 69 and 377 are big outliers. We can see that observation 43 is the highest leverage point.