Chapter 3 Exercises

Adrien Osakwe

Conceptual

1)

Table 3.4 shows basic stats from the linear model for the estimation of the number of units sold given the advertising budgets for (the predictors): *TV, radio and newspaper*. The null hypotheses the p-values correspond to are that there is no link between the budgets of each form of advertisement and the number of units sold (== the coefficients of the predictors are equal to 0).

Based on this relationship, the p-values indicate that the null hypotheses for both the TV and radio budgets' relationship to sales can be rejected. However, the null hypothesis for newspaper must be accepted as its p-value is much larger than 0.05. We would therefore conclude that TV and radio budgets are two predictors worth using to estimate sales.

2)

The KNN classifier will be used to estimate a qualitative response, (a category). This functions as a conditional probability where we find the k nearest neighbors to our observation and determine what proportion of the neighbors are in a given class. i.e: If k-1/k of our observation's neig- hbors are in class Green, then we would classify our observation as green with a conditional probability of k-1/k.

The KNN regression method is used to estimate quantitative responses, where the estimation of f(x) will be the average response of the k nearest-neighbors . i.e: if the average of our observation's neighborhood response is 8, then our regression model with k neighbors will estimate the response as 8.

3)

The linear model can be written as follows:

y = 50 + 20GPA + 0.07IQ + 35Level + 0.01(GPA x IQ) -10(GPA x Level)

We can simplify the model as the following: Level = 1: College graduate y = 85 + 10GPA + 0.07IQ + 0.01(GPA x IQ) Level = 0: High School Graduate y = 50 + 20GPA + 0.07IQ +0.01(GPA X IQ)

–> These models differ at two coefficients: the intercept is 35 units larger for a college graduate and the GPA coefficient is twice as large for the high school graduate

A) Based on the different coefficients in the two instances in the model, the student with the higher salary is dependent on GPA. Because the intercept is larger for the college instance, the college graduate will have a higher salary for low GPAs. However, when GPA is 3.5, both students have the same salary (10GPA =35, which equals the difference in the intercept). Therefore, GPA > 3.5 will lead to a higher salary for a high school graduate.

   The correct answer is therefore (iii)

B) The salary will be 137100 dollars. What did they study? How is it so high?!! Who knows...

C) This is false. Evidence for an interaction effect would be based on the t-statistic and resulting p-value for the interaction's coefficient to know if we can accept the null hypothesis that there is no interaction effect.

4)

a. There is not enough information to know as the cubic regression may enable the model to overfit more to any noise in the data. On the other hand, since the linear model is closer to the truth, it may also have a lower RSS.
b. We expect it to be lower for the linear regression as its assumption are closest to the true trend, reducing reducible error. The cubic model is more likely to have overfit, increasing reducible error.
c. There is not enough information as the true trend may not be cubic either, in which case RSS will also be larger for the cubic regression. If anything, the size of the RSS for either may help give a clue of what kind of non-linear relationship is the truth.
d. As in c, little can be said here without knowing what the true relationship is. Therefore, it is most likely higher in the regression furthest from the truth.

5)

Note: I really need to review LaTex...

y(i) = x(i)sum(x(j)y(j))/sum(x^2(i'))

y(i) = sum(x(i)x(j)y(j))/sum(x^2(i'))

y(i) = sum(((x(i)x(j))/sum(x^2(i')))y(j))

y(i) = sum(a(j)y(j))

a(j) = x(i)(j)/sum(x^2(i'))

6)

y = b0 + b1x

x = x(^)

y = b0 + b1x(^)

b0 = y(^) - b1x(^)

y = y(^)

7)

```
setwd(getwd())
```

$$R^2 = 1 - \frac{RSS}{TSS} \qquad \bar{x} = \bar{y} = 0$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum y_i^2}$$

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

$$\bar{x} = \bar{y} = 0 \rightarrow \beta_0 = 0$$

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

$$\beta_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$R^2 = 1 - \frac{\sum \left(y_i - \frac{\sum x_j y_j}{\sum x_j^2} \cdot x_i\right)^2}{\sum y_i^2}$$

$$R^2 = 1 - \frac{\left(\sum y_i^2 - 2 \sum y_i \cdot \frac{(\sum x_j y_j)}{\sum x_j^2} \cdot x_i + \sum \left(\frac{(\sum x_j y_j)}{\sum x_j^2}\right)^2 x_i^2\right)}{\sum y_i^2}$$

$$R^2 = 2 \sum y_i \cdot \frac{(\sum x_j y_j)}{\sum x_j^2} \cdot x_i - \left(\frac{\sum (x_j y_j)^2}{\sum x_j^4} \cdot x_i^2\right) \cdot \frac{1}{\sum y_i^2}$$

$$\frac{2\frac{\sum(x_i y_i)^2}{\sum x_i^2} - \frac{\sum(x_i y_i)^2}{\sum x_i^2}}{\sum y_i^2}$$

$$\hookrightarrow R^2 = \frac{\sum(x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

Note: $\bar{x} = \bar{y} = 0$

$$Cor(X,Y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}}$$

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

Applied

Packages

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.1.2
```

8)

A)

i) The summary of lm.fit shows that there is a relationship between mpg and horsepower (very small p-value).

ii) According to the R^2 value, horsepower accounts for 60% of the variance in mpg. Note: I saw in the solutions manual afterwards that we can also check the percentage error (RSE/mean response) to evaluate relationship strength.

iii) Negative relationship

iv) Prediction: mpg ~ 24.47, CI = [23.97-24.96], PI = [14.81-34.12] (rounded to 2 d.p.)

```
# Part A
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
```

```
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
```

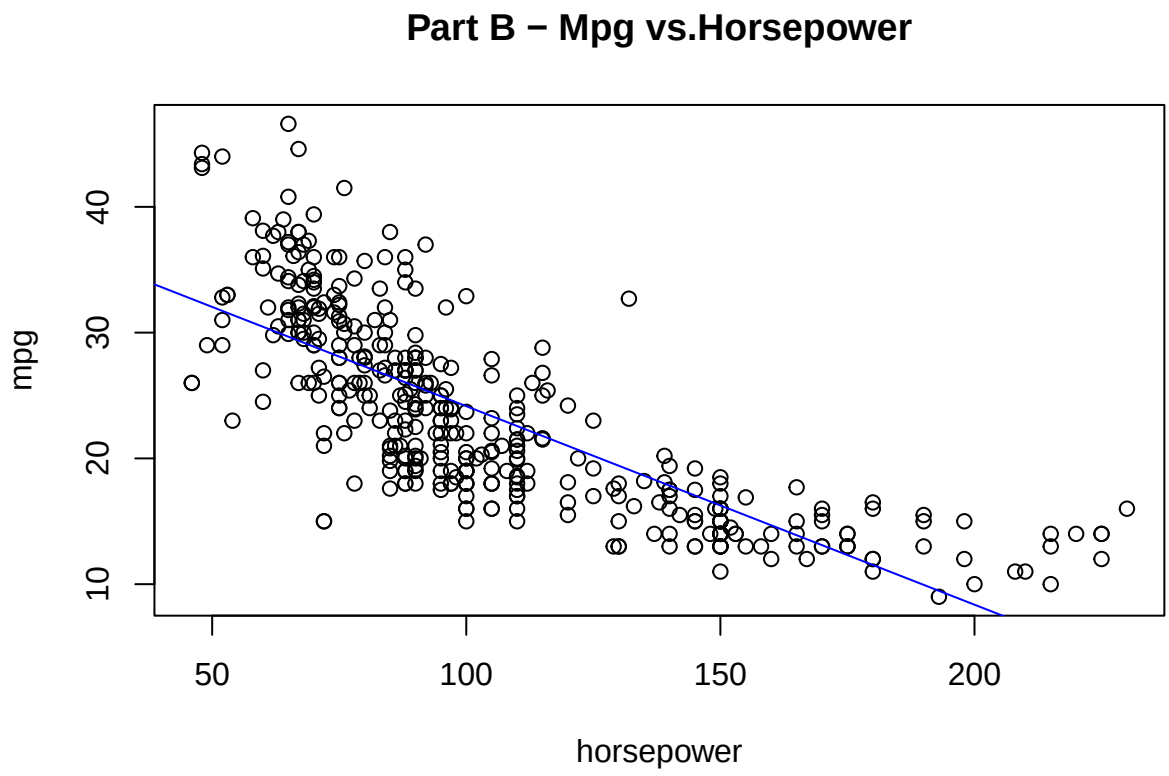
B)

See plot below

```
# Part B
attach(Auto)
plot(horsepower,mpg) + title("Part B - Mpg vs.Horsepower")
```

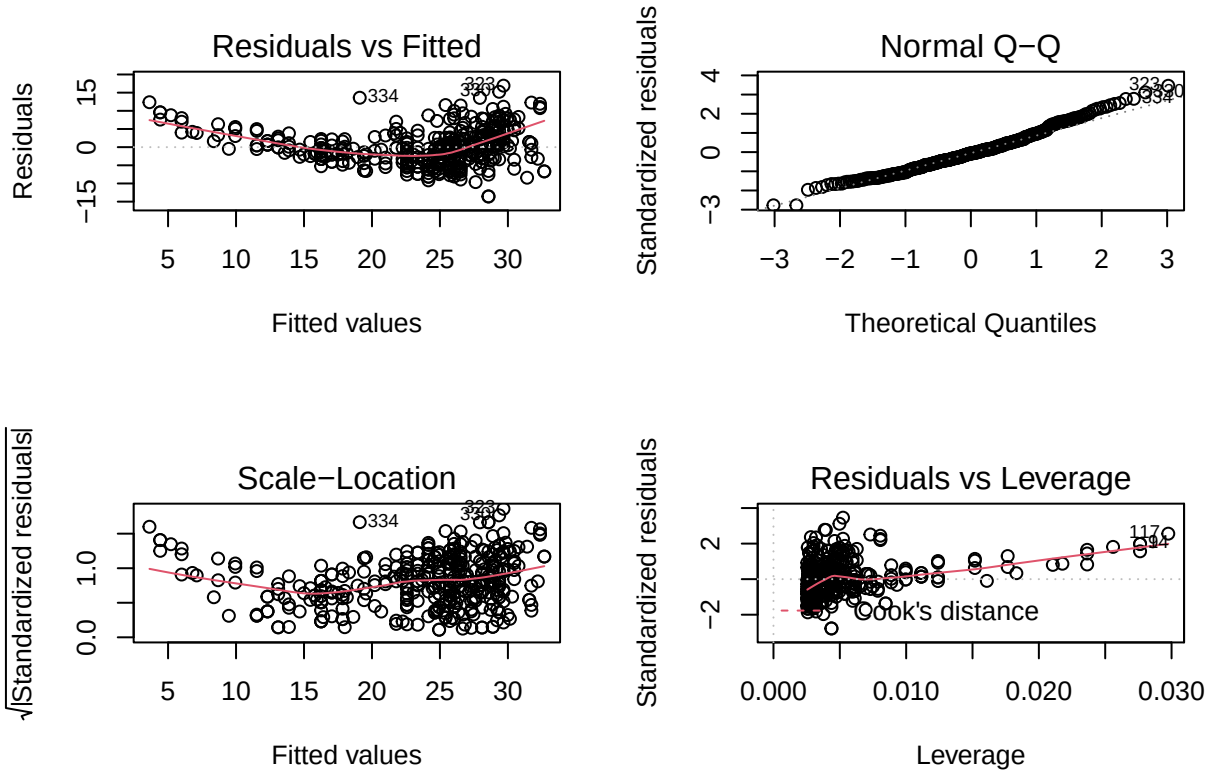
```
## integer(0)
```

```
abline(lm.fit, col = "blue")
```



C) Based on the residuals v. fitted value plot, we can see that the true relationship is not linear, indicating that some for of transformation will be required to best fit the data. There is not a large fat tail in the Q-Q plot which indicates that the responses are not very non-normal. The remaining plots highlight one or two observations that are outliers.

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

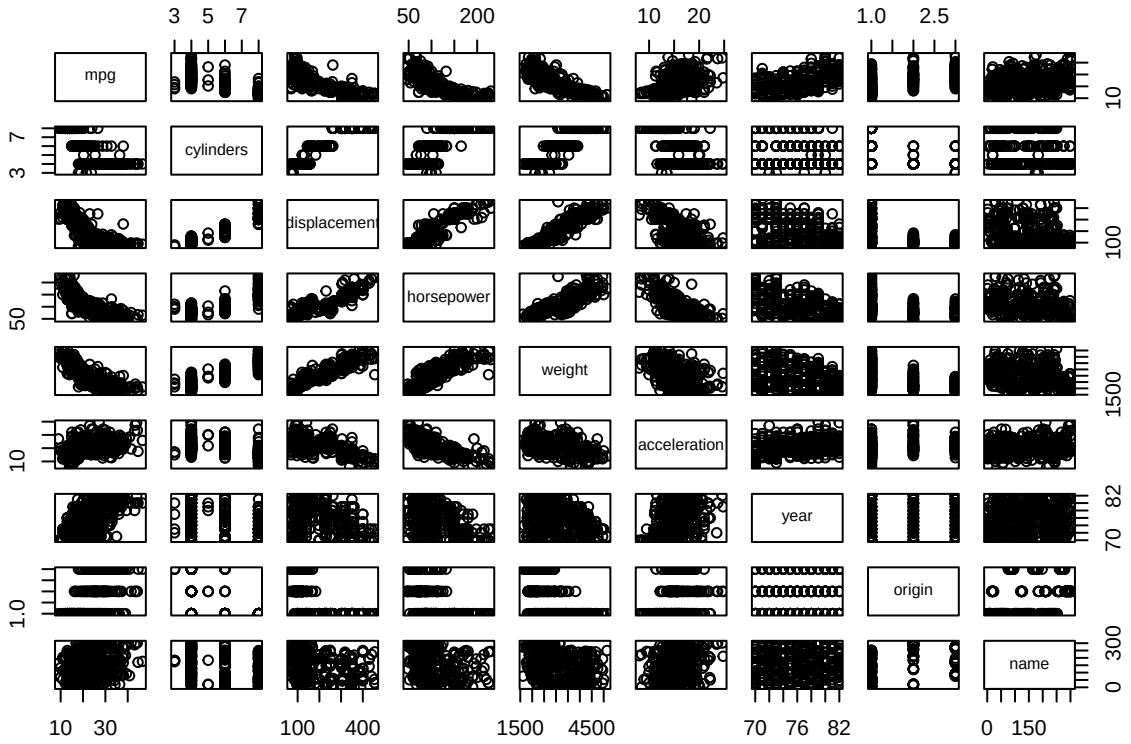


9)

A)

See plot below.

```
#Part A
pairs(Auto)
```



B) See matrix below.

```
#Part B
mat.cor <- cor(Auto[,!(names(Auto) == "name")])
data.frame(mat.cor)
```

```
##                     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) i) The F-statistic is below 0.05 which confirms we can reject the null hypothesis.

ii) Based on the p-values, only some of the predictors are shown to have a relationship with the response. These are: displacement, weight, year, and origin.

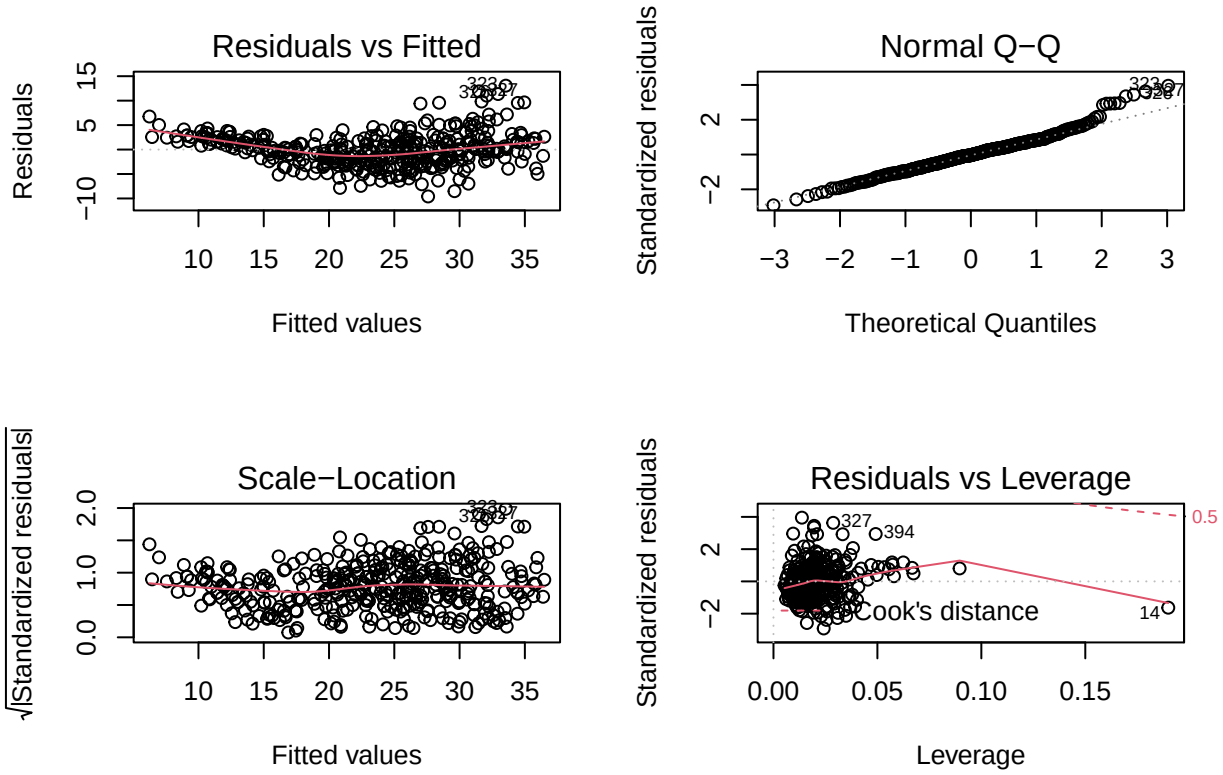
iii)The coefficient for year suggests that as the year increases, the mpg increases. This is reassuring to see as I would be concerned if more technologically advanced cars were incapable of a better mileage than a model T...

```
#Part C
lm.fit <- lm(mpg ~ . - name, data = Auto)
summary(lm.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
```

D) The residuals vs leverage plot shows observation 14 to be a considerable outlier compared to the rest of the dataset. Observations 323,326 and 327 are also shown to be outliers based on the other three plots. The residuals vs. fitted plot seems to have a very light non-linearity. Q-Q plot shows as fat tail indicating some of the data is non-normal.

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



E) Checked for interactions effects between cylinders and weight as well as horsepower and acceleration. Both interactions are shown to have a significant relationship. In addition this model reduces RSE and increases R^2. An anova analysis reveals that this model is a significant improvement of the previous one.

```
#Part E
lm.fit2 <- lm(mpg ~ . - name + cylinders:weight + horsepower:acceleration, data = Auto)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name + cylinders:weight + horsepower:acceleration,
##     data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3499  -1.6685  -0.0141   1.5009  12.0571
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -7.4837124  5.5164273  -1.357 0.175702
## cylinders              -4.0179263  0.5878598  -6.835 3.26e-11 ***
## displacement           -0.0048855  0.0075637  -0.646 0.518716
## horsepower              0.0806652  0.0236283   3.414 0.000709 ***
## weight                 -0.0120372  0.0012008 -10.024  < 2e-16 ***
```

```
## acceleration             0.7885834  0.1514762   5.206 3.16e-07 ***
## year                     0.7814074  0.0447175  17.474  < 2e-16 ***
## origin                   0.5833654  0.2552730   2.285 0.022845 *
## cylinders:weight         0.0013150  0.0001633   8.051 1.05e-14 ***
## horsepower:acceleration -0.0092849  0.0016782  -5.533 5.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.912 on 382 degrees of freedom
## Multiple R-squared:  0.864,  Adjusted R-squared:  0.8608
## F-statistic: 269.7 on 9 and 382 DF,  p-value: < 2.2e-16
```

```
anova(lm.fit,lm.fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ (cylinders + displacement + horsepower + weight + acceleration +
##     year + origin + name) - name
## Model 2: mpg ~ (cylinders + displacement + horsepower + weight + acceleration +
##     year + origin + name) - name + cylinders:weight + horsepower:acceleration
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    384 4252.2
## 2    382 3238.4  2    1013.8 59.794 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

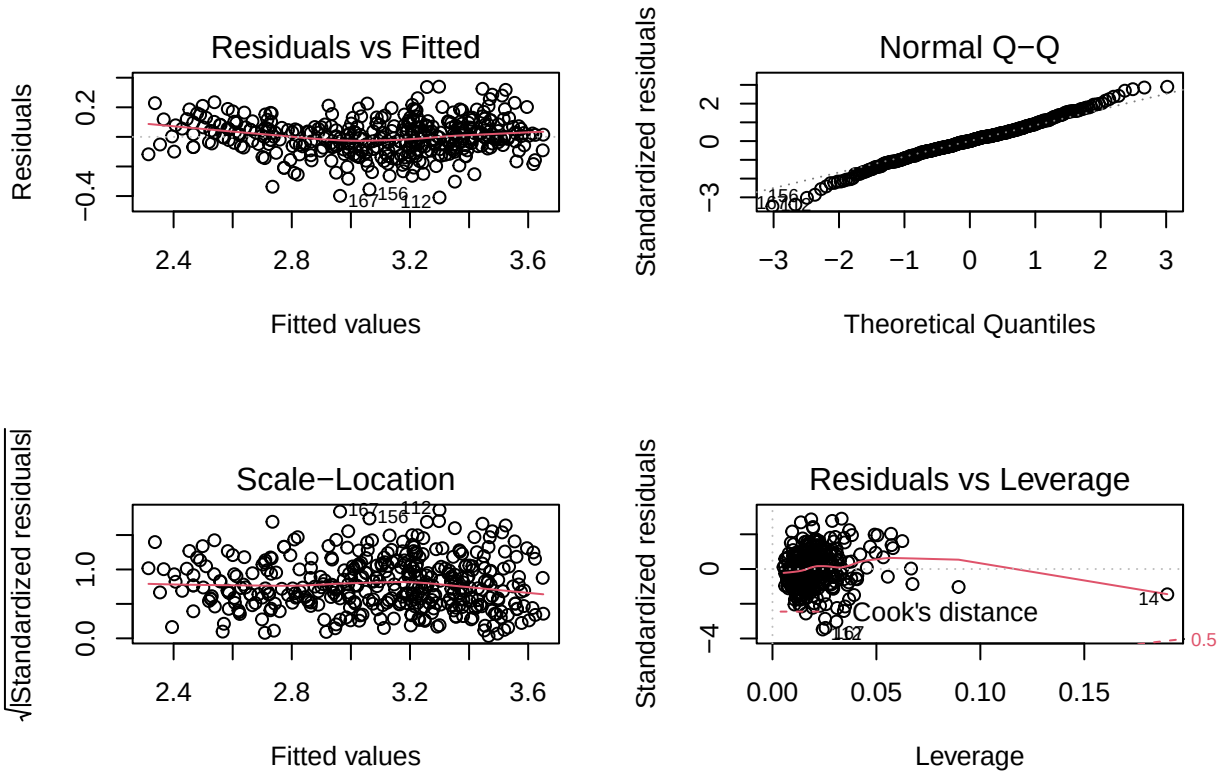
F) Log-transforming mpg seems ot better fit the data.RSE is very small and the residuals v. fitted plot shows no trend. A square root transform also reduces R^2, however the non-linear trend in the residuals vs. fitted plot shows the model does not fit the data as well as the log transformation. See plots below

```
#Part F
log.fit <- lm(log(mpg) ~ . - name, data = Auto)
summary(log.fit)
```

```
##
## Call:
## lm(formula = log(mpg) ~ . - name, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40955 -0.06533  0.00079  0.06785  0.33925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.751e+00  1.662e-01  10.533  < 2e-16 ***
## cylinders    -2.795e-02  1.157e-02  -2.415  0.01619 *
## displacement  6.362e-04  2.690e-04   2.365  0.01852 *
## horsepower   -1.475e-03  4.935e-04  -2.989  0.00298 **
## weight       -2.551e-04  2.334e-05 -10.931  < 2e-16 ***
## acceleration -1.348e-03  3.538e-03  -0.381  0.70339
## year          2.958e-02  1.824e-03  16.211  < 2e-16 ***
## origin        4.071e-02  9.955e-03   4.089 5.28e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1191 on 384 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8773
## F-statistic: 400.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

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

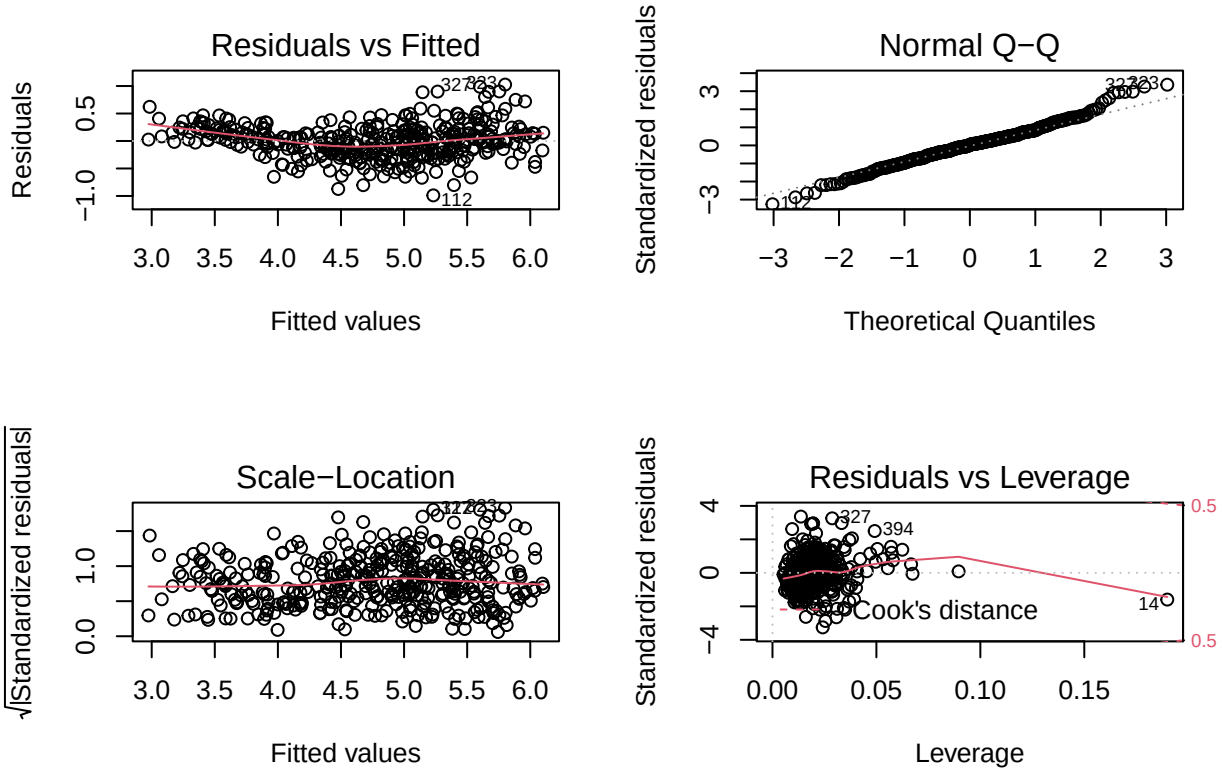



```
sqrt.fit <-lm(sqrt(mpg) ~ . - name, data = Auto)
summary(sqrt.fit)
```

```
##
## Call:
## lm(formula = sqrt(mpg) ~ . - name, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98891 -0.18946  0.00505  0.16947  1.02581
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.075e+00  4.290e-01   2.506   0.0126 *
## cylinders    -5.942e-02  2.986e-02  -1.990   0.0474 *
```

```
## displacement  1.752e-03  6.942e-04   2.524   0.0120 *
## horsepower   -2.512e-03  1.274e-03  -1.972   0.0493 *
## weight       -6.367e-04  6.024e-05 -10.570  < 2e-16 ***
## acceleration  2.738e-03  9.131e-03   0.300   0.7644
## year          7.381e-02  4.709e-03  15.675  < 2e-16 ***
## origin        1.217e-01  2.569e-02   4.735 3.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3074 on 384 degrees of freedom
## Multiple R-squared:  0.8561, Adjusted R-squared:  0.8535
## F-statistic: 326.3 on 7 and 384 DF,  p-value: < 2.2e-16
```

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



10)

A)

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

```
##
```

```
## 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) See above for the summary statistics of the model.

Based on the summary, the F-statistic indicates that we can reject then null hypothesis that all coefficients are 0. The individual p-values show that Urban is the only predictor to not be related to sales. The coefficient of Price tells us that a one unit increase in price will reduce sales by 0.05 units. The coefficient of US tells us that being in the US increases sales by 1.2 units.

C) If in the US: Sales = 14.2 -0.05Price Else: Sales = 13 - 0.05Price

D) As mentioned in B, the null hypothesis can be rejected for Price and US.

E)

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

```
##
## 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) The summary stats for both models tell us that the predictors in both cases only account for about 23% of the variance in sales. Based off this alone, we will likely have to use other predictors and/or interaction effects to have a better fit.

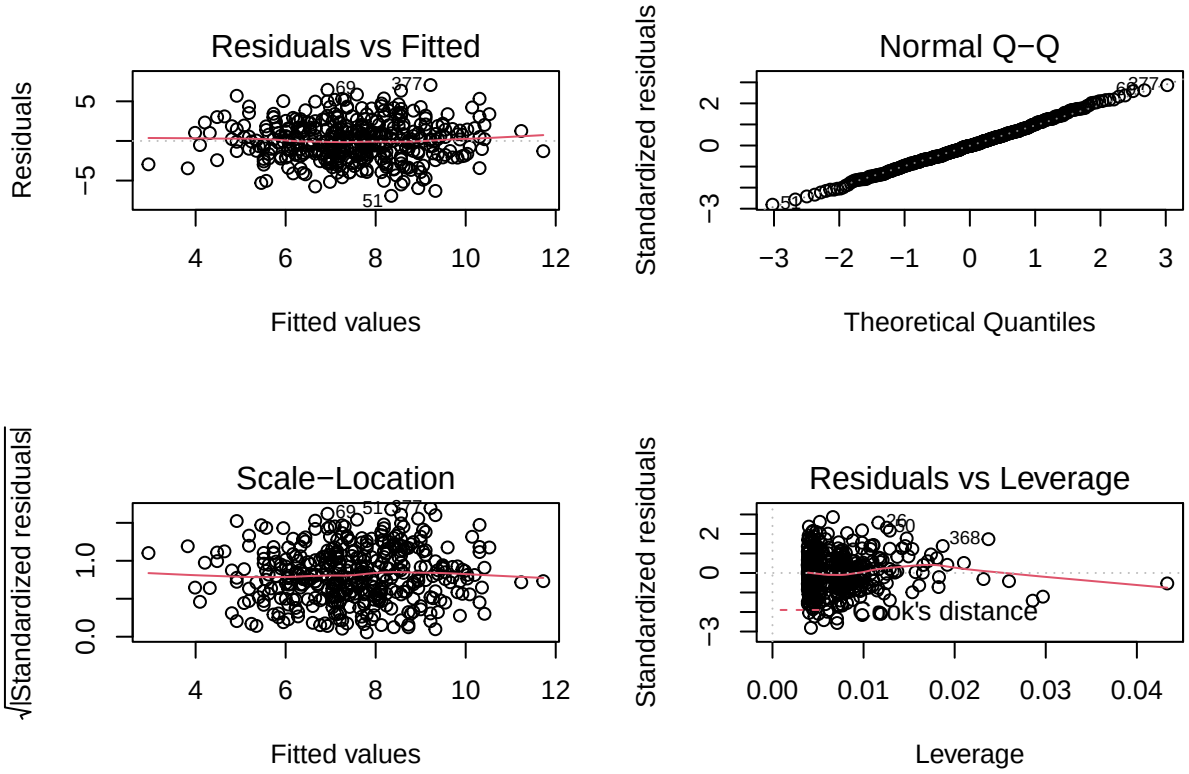
G)

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

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

H) The diagnostics show some evidence for one outlier (377) and one case of high leverage.

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



11)

A) Coefficients: x

Estimate SE t value Pr(>|t|)

1.9939 0.1065 18.73 <2e-16

These results show that the null hypothesis can be rejected as the p-value is below 0.05.

```
set.seed(1)
x <- rnorm(100)
y <- 2*x +rnorm(100)
lm.fit <- lm(y ~ x + 0)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

B) Coefficients: y

Estimate Std. Error t value Pr(>|t|)

0.39111 0.02089 18.73 <2e-16

Much like before, we can reject the null hypothesis as p-value is below 0.05.

```
set.seed(1)
x <- rnorm(100)
y <- 2*x +rnorm(100)
lm.fit <- lm(x ~ y + 0)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

C) Both results have the same t-statistic and p-value. This is expected as both models are producing the same line (we simply inversed the roles of predictor and response).