HOMEWORK 3

Andrew Guo

2/17/2022

PROBLEM 1

Using (3.4) on p. 62, show that in the case of simple linear regression, the least squares line always passes through the point ($\bar{x}$, $\bar{y}$).

If the LSRL always passes through the point (xbar, ybar), then this means that there is a point where x = xbar, and y = ybar.

b0(hat) = ybar - b1(hat)xbar

SLR: y(hat) = b0(hat) + b1(hat)x

substitute b0(hat):

y(hat) = ybar - b1(hat)xbar + b1(hat)x

yhat-ybar = b1(hat)[-xbar + x]

If xbar, ybar, is a point on the graph, then we would expect that if we were to plug in xbar for x in the SLR equation, then we should get that the expected value of y is ybar.

y(hat) - ybar = b1(hat)[-xbar + xbar]

yhat - ybar = b1(hat)[0]

yhat - ybar = 0

yhat = ybar

The expected value of y is ybar. Therefore, the least squares line always passes through the point (xbar, ybar).

PROBLEM 2

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

y = 50 + 20x1 + 0.07x2 + 35x3 + 0.01x1x2 - 10x1x3

x1 = GPA

x2 = IQ

x3 = Level (1 = college, 0 = high school)

x1x2: Interaction between GPA and IQ

x1x3: Interaction between GPA and Level

College Graduate model (x3 = 1): y = 50 + 20x1 + 0.07x2 + 35(1) + 0.01x1x2 -10x1(1); y = 50 + 20x1 + 0.07x2 + 35 + 0.01x1x2 - 10x1

High School Graduate model (x3 = 0): y = 50 + 20x1 + 0.07x2 + 35(0) + 0.01x1x2 -10x1(0); y = 50 + 20x1 + 0.07x2 + 0.01x1x2

College Grad Model, unique terms: 35 - 10x1

High School Grad model, unique terms: None

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

Not necessarily. If the GPA is lower than 3.5, then the model predicts that the starting salary will be somewhat greater for college students than high school students. However, if the GPA is higher than 3.5, then this would create a negative value for the college graduate model, meaning that high school students would have a higher starting salary.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

Same principle as i. If the GPA is higher than 3.5, then this would mean that high school graduates would be earning more than college graduates due to the negative value that is within the college graduate model but is not present in the high school graduate model.

iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

This is correct. If the GPA is higher than 3.5, then this would create a negative value for the college graduate model, meaning that high school students would have a higher starting salary. For instance, if the GPA was 4.0, college students would have a 35-10(4) = -5 decrease in their starting salary according to the model, while high school students do not.

iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

If the GPA is high enough, then this would mean that the college graduate model has a negative, subtracting value that is not present in the high school model. This would mean that we would expect high school graduates to earn MORE than college graduates if the GPA is high enough. This statement proclaims the opposite.

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

x1 = GPA, x2 = IQ, x3 = college student

x1 = 4, x2 = 110, x3 = 1

College Graduate model: y = 50 + 20x1 + 0.07x2 + 35 + 0.01x1x2 -10x1

y = 50 + 20(4) + 0.07(110) + 35 + 0.01(4*110) - 10(4) = 137.1

Based on the model, I predict that the salary of a college graduate with an IQ of 110 and a GPA of 4.0 will have a starting salary of 137.1 thousand dollars.

(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.

This is false. A small interaction term does not mean that the actual statistical significance is insignificant. Likewise, the size of the interaction term is not its statistical significance; these are two different statistics. Yes, it is possible for a small interaction term (for instance, a beta hat equal to 0.01) to not be statistically significant (for instance, having a p value of 0.09), but it is also possible for that small predictor term to also be statistically significant (by having a p value of less than 0.05).

PROBLEM 3

This question involves the use of simple linear regression on the `Auto` data set. Information can be found here: `https://rdrr.io/cran/ISLR/man/Auto.html`

(a) Read in the data and 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.

```
Auto <- read.csv ("Auto.csv", na.strings = "?", stringsAsFactors = T)
Auto = na.omit(Auto)
attach(Auto)

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

```
##
## 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. Is there a relationship between the predictor and the response?

Assume b1 = horsepower

Ho: b1 = 0

Ha: b1 != 0

t = -0.157845/0.006446 = -24.49

2[p(t<-24.49)] = approx. 2e-16

With a p-value less than a = 0.05, we reject the null. There is convincing evidence that a relationship exists between the predictor and the response.

ii. How strong is the relationship between the predictor and the response?

Because the number of observations in our data is significantly larger than the number of predictors in the model, it is safe to use r-squared and adjusted r-squared as interpreters for how good of a fit the model is for showcasing the relationship between the predictor and the response. Thus, with an adjusted r-squared value of 0.6049, there is a moderately strong relationship between the predictor and the response; approximately 60.49% of the variation in mpg can be explained by the LSRL on horsepower, taking into account both sample size n and the number of parameters in the model.

iii. Is the relationship between the predictor and the response positive or negative?

Because the estimated beta coefficient is negative, the relationship between the predictor and the response is negative.

iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

y = 39.935861 - 0.157845x y = 39.935861 - 0.157845(98) = 24.467051

The predicted mpg associated with a horsepower of 98 is 24.467051.

```
predict(lm.fit.mpg, data.frame(mpg = c(24.467051), horsepower = c(98)), interval = 'conf
```

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

```
predict(lm.fit.mpg, data.frame(mpg = c(24.467051), horsepower = c(98)), interval = 'pred
```

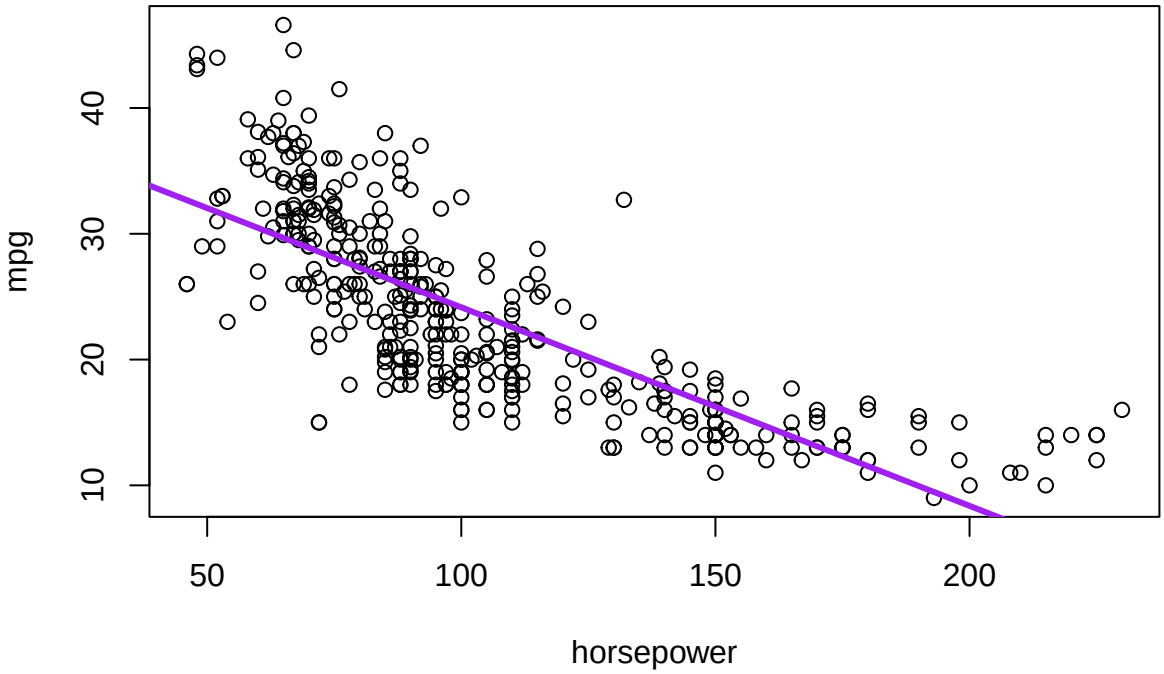
```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

The associated 95% confidence interval is [23.97308, 24.96108].

The associated 95% prediction interval is [14.8094 34.12476].

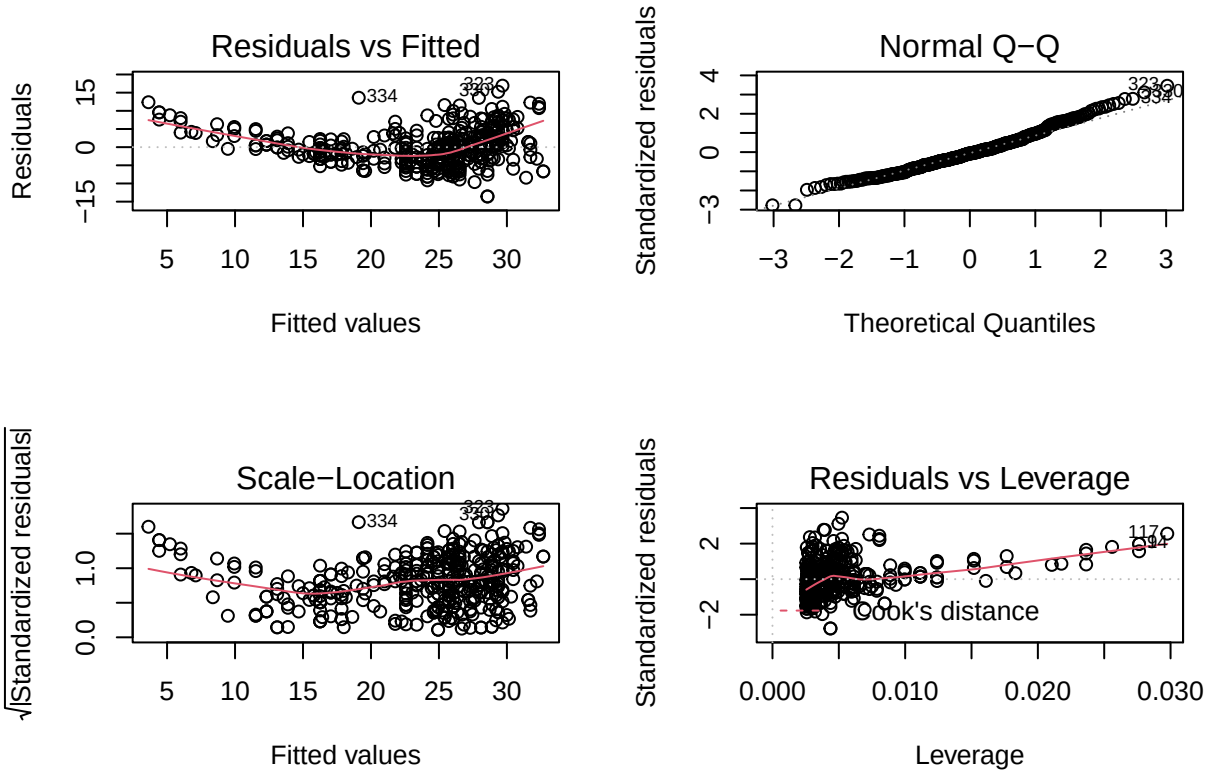
(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

```
plot(horsepower, mpg)
abline(lm.fit.mpg, lwd = 3, col = 'purple')
```



(c) Use the `parfor()` and `plot()` functions to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit in terms of the appropriateness of a linear model, normality, heteroscedasticity, outliers, and high leverage observations.

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



The residual vs. fitted values plot shows both a slight upward parabolic curvature as well as fanning outwards to the right, indicating that our data does not appear to be homoscedastic. The normal probability plot shows a relationship that is strongly linear, indicating that there does not appear to be any problems visually in terms of normality. The residuals vs leverage plot also does not contain many concerning points. A few observations, such those nearing the +4 standard deviations, as well as some observations such as 117 and 84 that are higher than the observations in terms of leverage, are noticable, but not too concerning. The observations that are on the higher ends of the standards of deviation do not have much leverage, and even those observations that are comparatively higher in leverage do not yield that high of leverage; for instance, observations 117 and 84 only have a leverage statistic of approximately 0.03 when the maximum value of a leverage statistic is 1.

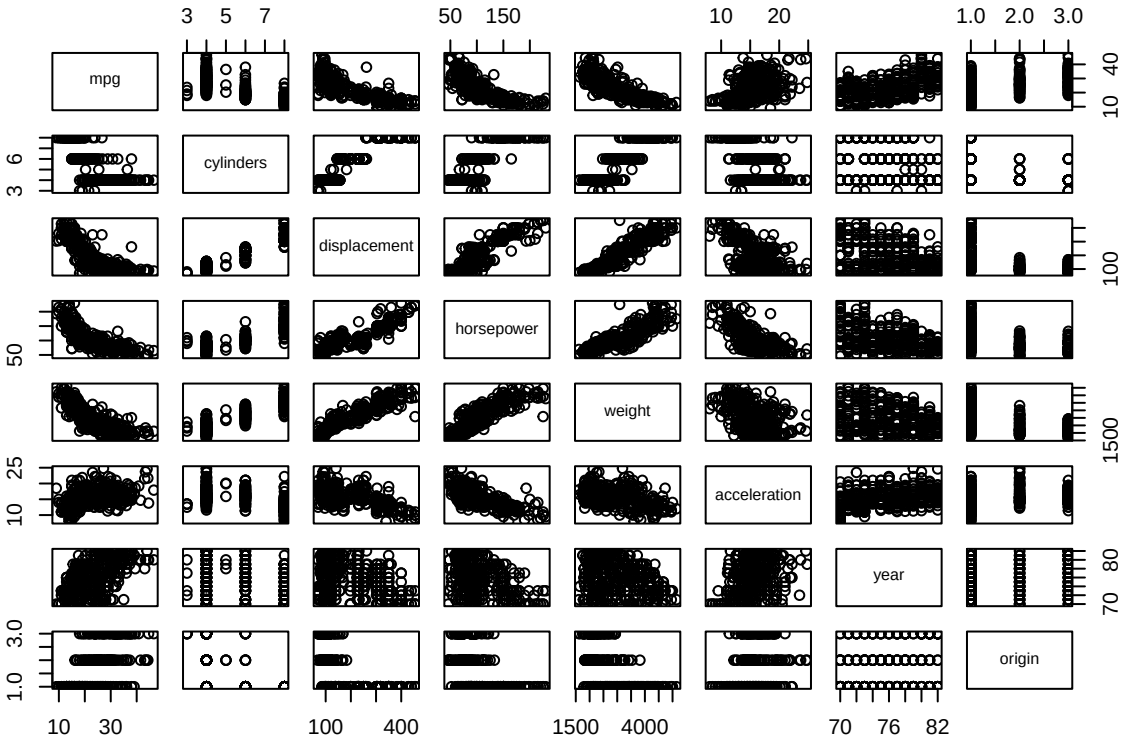
PROBLEM 4

This question involves the use of multiple linear regression on the `Auto` data set.

(a) Create a new data frame called `Auto2` which includes all of the variables in the data set except for `name` (you learned to remove a variable in HWK2). Then produce a scatterplot matrix of `Auto2`.

```
#Auto2 <- read.csv ("Auto.csv", na.strings = "?", stringsAsFactors = T)
#Auto2 = na.omit(Auto2)
#Auto2$origin = as.factor(Auto2$origin)
#attach(Auto2)

Auto2 <- Auto[, c(1:8)]
pairs(Auto2)
```



(b) Compute the matrix of correlations between the variables in `Auto2` using the function `cor()`. You will need to exclude `origin`, which is qualitative.

```
#Auto2 <- Auto2[, c(1:7)]
cor(Auto2[ ,c(1:7)])
```

```
##                     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
##              acceleration       year
## mpg             0.4233285  0.5805410
## cylinders      -0.5046834 -0.3456474
## displacement   -0.5438005 -0.3698552
## horsepower     -0.6891955 -0.4163615
## weight         -0.4168392 -0.3091199
## acceleration    1.0000000  0.2903161
## year            0.2903161  1.0000000
```

(c) Recode `origin` as follows:

```
```{r}
library(dplyr)
Auto2$origin <- recode(Auto2$origin, "1" = "American", "2" = "European", "3" = "Japanese")
attach(Auto2)
```
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Auto2$origin <- recode(Auto2$origin, "1" = "American", "2" = "European", "3" = "Japanese
attach(Auto2)
```

```
## The following objects are masked from Auto:
##
##     acceleration, cylinders, displacement, horsepower, mpg, origin,
##     weight, year
```

```
#Auto2 <- read.csv('Auto.csv')
#attach(Auto2)
#Auto2$horsepower <- as.integer(Auto2$horsepower)
```

```
#Auto2 = na.omit(Auto2)
#Auto2 <- Auto2[, c(1:8)]
```

Then 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. Comment on the output.

```
lm.fit.Auto2 = lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin,
summary(lm.fit.Auto2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin, data = Auto2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders      -4.897e-01  3.212e-01  -1.524 0.128215
## displacement    2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower     -1.818e-02  1.371e-02  -1.326 0.185488
## weight         -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration    7.910e-02  9.822e-02   0.805 0.421101
## year            7.770e-01  5.178e-02  15.005  < 2e-16 ***
## originEuropean  2.630e+00  5.664e-01   4.643 4.72e-06 ***
## originJapanese  2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

i. Is there a relationship between the predictors and the response?

Based on the general omnibus test, we have a p-value of less than a = 0.05. This indicates that at least one of our predictors is statistically significant.

ii. Which predictors appear to have a statistically significant relationship to the response?

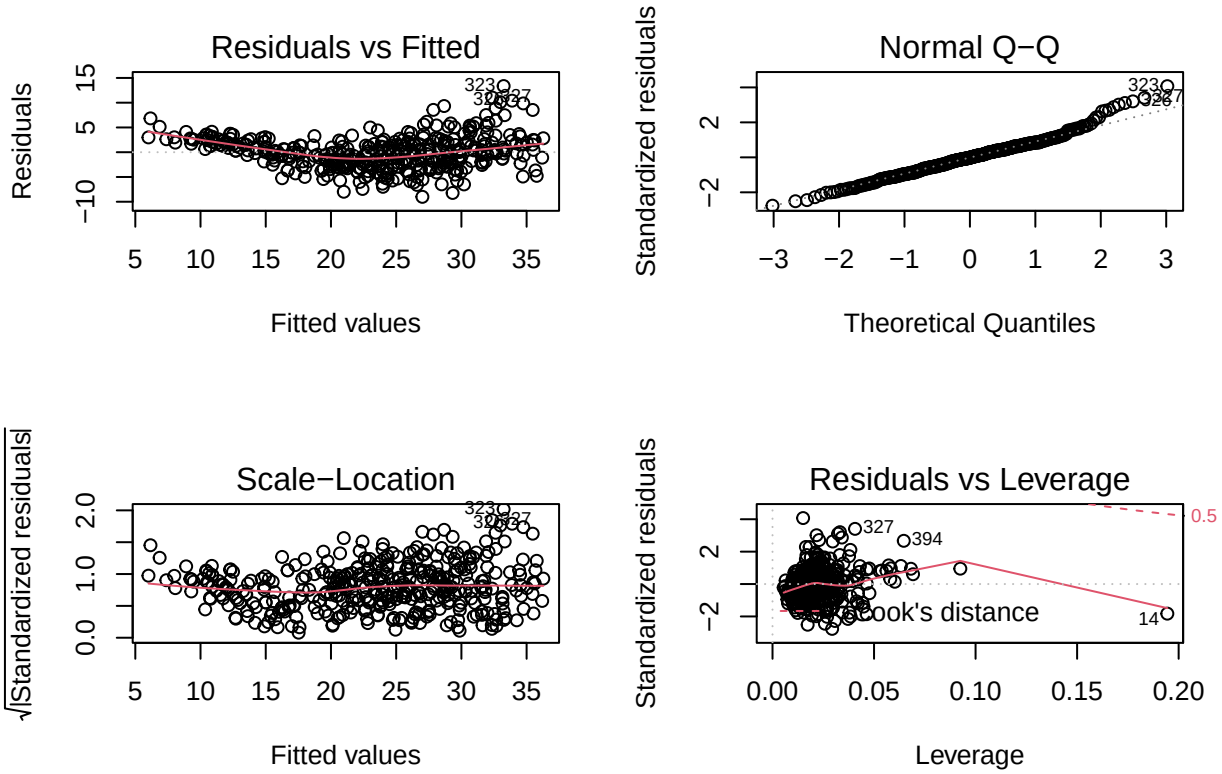
A predictor with a p-value of less than a = 0.05 is considered statistically significant. These predictors are: originEuropean, originJapanese, year, weight, displacement, and originAmerican.

iii. What does the coefficient for the year variable suggest?

A one unit change in year will lead to a change of 7.770e-01 units for mpg, holding all other predictors constant.

(d) Use the `parfor()` and `plot()` functions to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit in terms of the appropriateness of a linear model, normality, heteroscedasticity, outliers, and high leverage observations.

```
par(mfrow = c(2, 2))
plot(lm.fit.Auto2)
```



The residual vs. fitted values plot shows no clear or definitively strong curvature, but it does show visible fanning outwards to the right, indicating that our data does not appear to be homoscedastic. The normal Q-Q plot shows a relationship that is strongly linear,

indicating that there does not appear to be any problems in terms of normality. Although, the observations at the end of the plot are not as linear as the observations in the other points of the graph, because the overall shape of the plot is linear, we can say that our data is approximately normal. The residuals vs leverage plot also does not contain many concerning points. A few observations may be concerning, such 327 and 394, near +4 standard deviations, and observation 14 is higher than the observations in terms of leverage. But, the observations that are on the higher ends of the standards of deviation do not have much leverage, and even those observations that are comparatively higher in leverage do not yield that high of leverage; observation 14 only have a leverage statistic of approximately 0.18 when the maximum value of a leverage statistic is 1.

(e) The plots seem to indicate heteroscedasticity. Try three different transformations of `mpg`, such as `log(mpg)`, `sqrt{mpg}`, `mpg^2` and determine which one seems best in eliminating this issue. Only show the results for the best of the three.

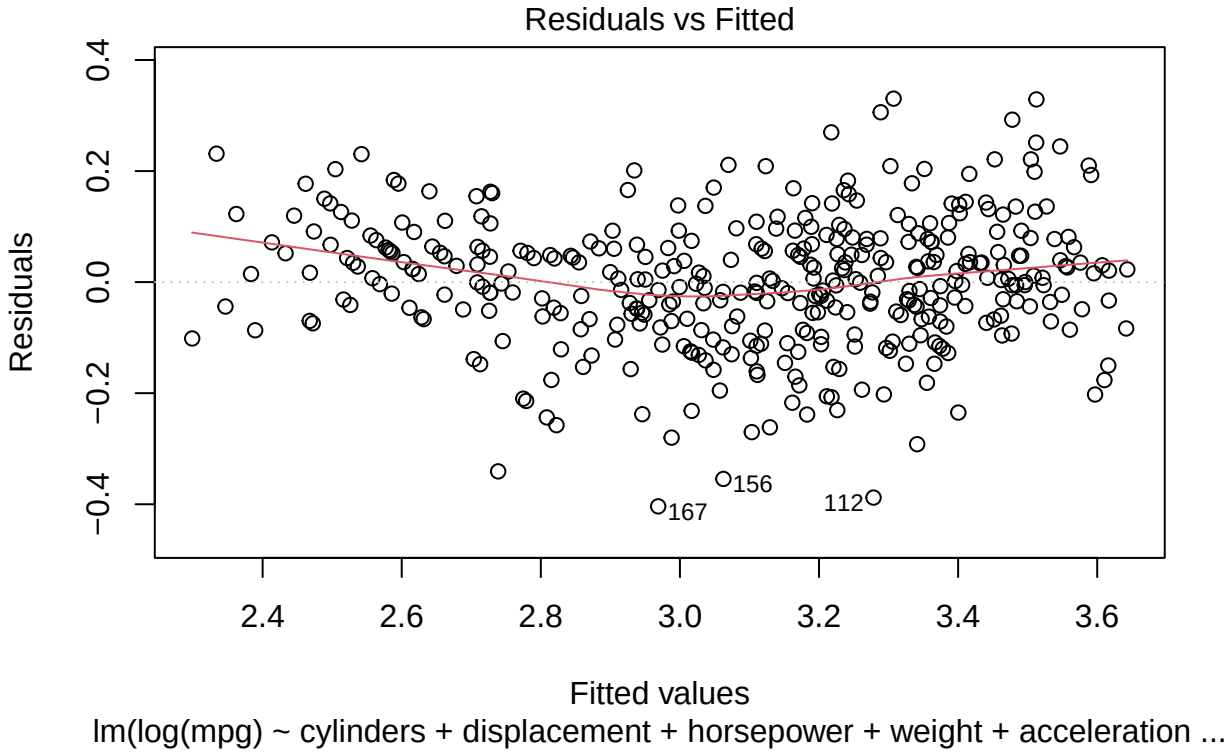
$$Y' = \log Y$$

```
lm.fit.Auto2 = lm(log(mpg)~cylinders+displacement+horsepower+weight+acceleration+year+or
summary(lm.fit.Auto2)
```

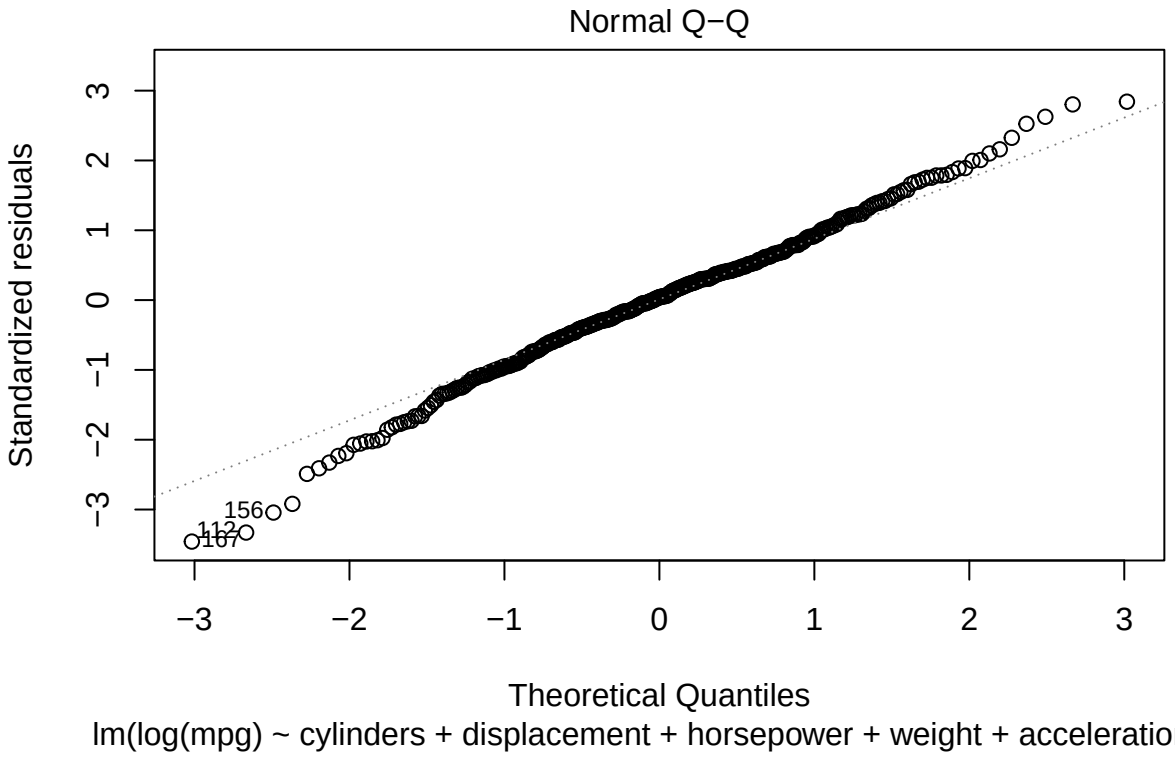
```
##
## Call:
## lm(formula = log(mpg) ~ cylinders + displacement + horsepower +
##     weight + acceleration + year + origin, data = Auto2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40380 -0.06679  0.00493  0.06913  0.33036
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.712e+00  1.673e-01  10.230  < 2e-16 ***
## cylinders      -2.781e-02  1.149e-02  -2.420  0.01598 *
## displacement    7.874e-04  2.738e-04   2.876  0.00425 **
## horsepower     -1.520e-03  4.904e-04  -3.100  0.00208 **
## weight         -2.639e-04  2.344e-05 -11.260  < 2e-16 ***
## acceleration   -1.403e-03  3.513e-03  -0.399  0.68996
## year            3.055e-02  1.852e-03  16.491  < 2e-16 ***
## originEuropean  8.531e-02  2.026e-02   4.210 3.18e-05 ***
## originJapanese  8.145e-02  1.977e-02   4.119 4.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1183 on 383 degrees of freedom
```

```
## Multiple R-squared:  0.8815, Adjusted R-squared:  0.879
## F-statistic: 356.1 on 8 and 383 DF,  p-value: < 2.2e-16
```

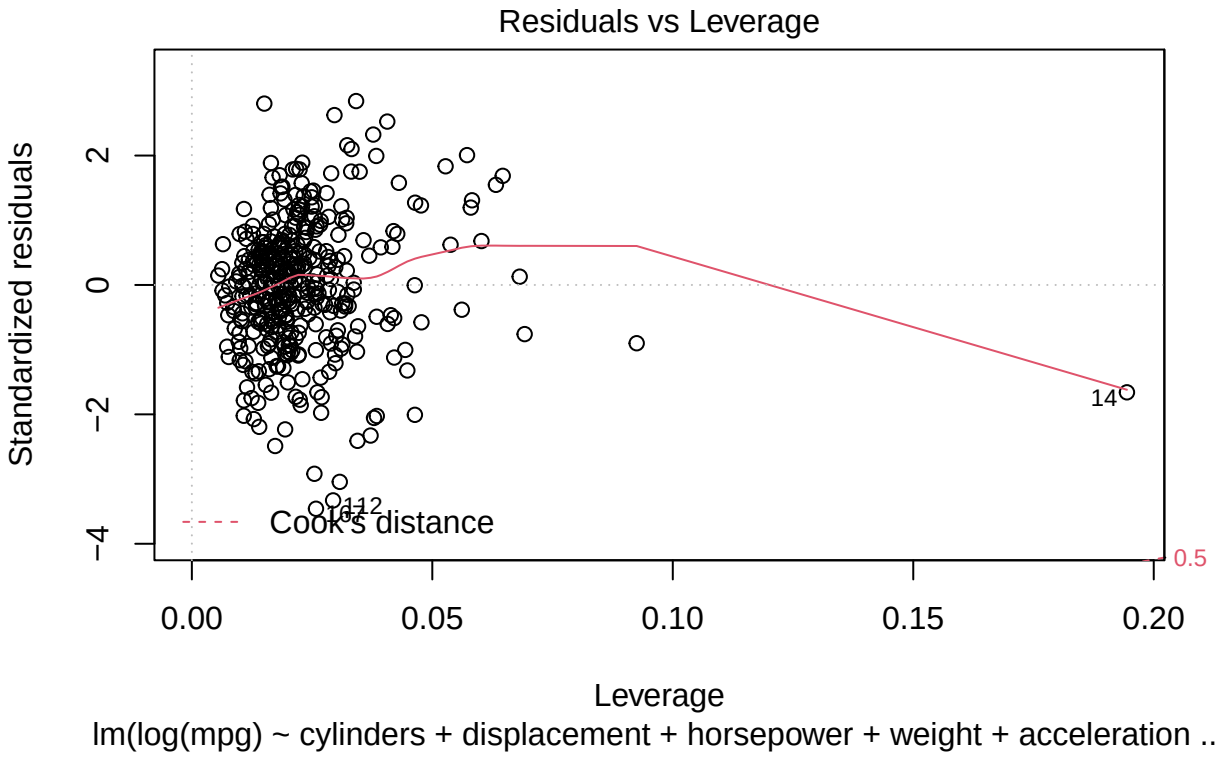
```
par(mfrow=c(1,1))
plot(lm.fit.Auto2, which = c(1))
```



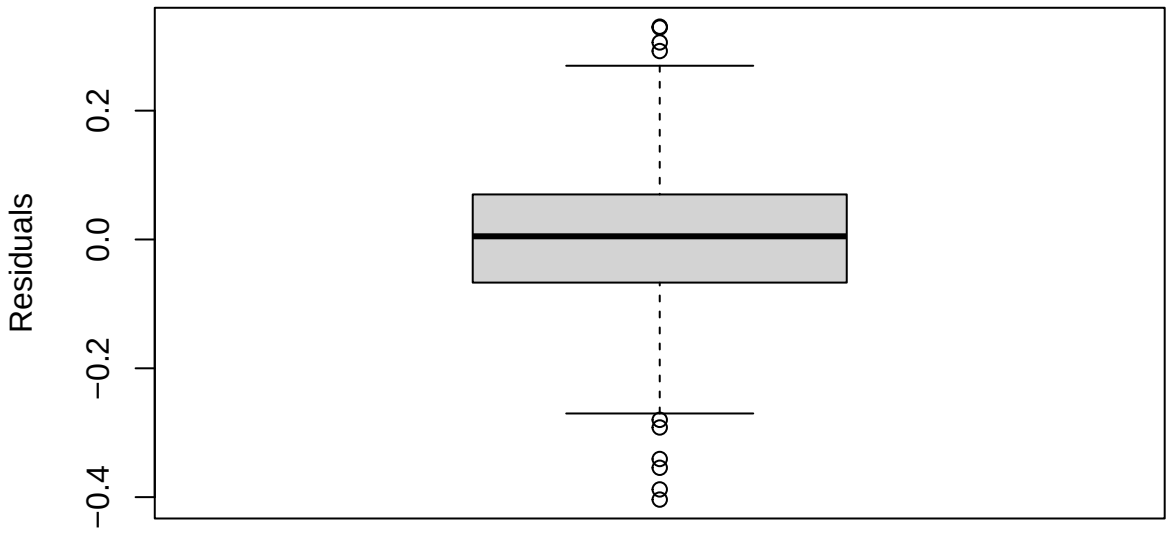

```
plot(lm.fit.Auto2, which = c(2))
```



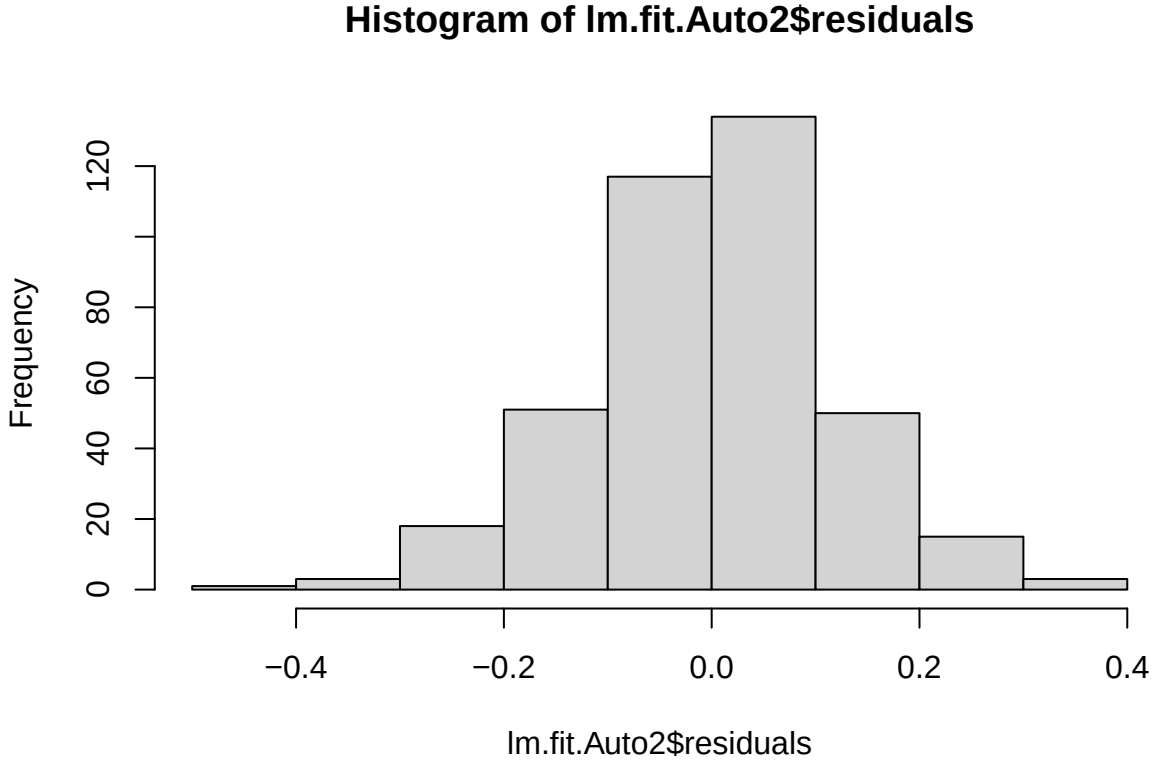

```
plot(lm.fit.Auto2, which = c(5))
```




```
boxplot(lm.fit.Auto2$residuals, ylab="Residuals")
```




```
hist(lm.fit.Auto2$residuals)
```




```
shapiro.test(lm.fit.Auto2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.fit.Auto2$residuals
## W = 0.9923, p-value = 0.04071
```

The boxplot is fairly symmetrical with the median located roughly in the center. The

histogram is roughly symmetric, contains no observable outliers nor gaps and is unimodal. Thus, our data appears to be visibly normal. However, the results of our Shapiro test suggest otherwise. The Shapiro test yielded a p-value less than a = 0.05, means that we would reject the null hypothesis that the data is normally distributed.

There doesn't appear to be any curvature or a definite shape by the residuals, nor does there appear to be any fanning. Thus, we are confident that this transformation does not have strong levels of heteroscedasticity, if any. The values in the Q-Q plot also follows a strong linear pattern, indicating that our data may be normally distributed. Additionally, the standardized residuals versus leverage plot do not indicate many outliers that fall outside of +-3 standard deviations. There are a few observations that are near -4 standard deviation, but these points are minimal in leverage and have little influence in distorting the shape of the data's relationship.

Thus, while our data appears to visibly pass homoscedasticity, it still has not passed normality. Additional transformations on the predictors may be needed in order to meet this condition.

(f) Use the * or : symbols to fit a multiple linear regression model which includes the two most significant predictors from part (c) and their interaction effect. Comment on the statistical significance of the result.

The two most significant predictors from part (c) would be those that have the lowest p value. Those predictors would be: weight and year, both of which have p-values < 2e-16.

```
lm.fit.Auto3 = lm(mpg~weight + year + weight:year, data = Auto2)
summary(lm.fit.Auto3)
```

```
##
## Call:
## lm(formula = mpg ~ weight + year + weight:year, data = Auto2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0397 -1.9956 -0.0983  1.6525 12.9896
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
## weight       2.755e-02  4.413e-03   6.242 1.14e-09 ***
## year         2.040e+00  1.718e-01  11.876  < 2e-16 ***
## weight:year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.193 on 388 degrees of freedom
## Multiple R-squared:  0.8339, Adjusted R-squared:  0.8326
```

```
## F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

The p-value of our global omnibus test yields a p-value less than 0.05, which means that at least one of our predictors is statistically significant. If we look at the summary table, we see that both main predictors and the interaction term all have p-values less than 0.05, indicating that all of these terms are statistically significant.