IDA LAB-4

SOLUTION FOR Practical-1, Practical-2, and Practical-3

Practical-1:

```
library(MASS)
library(ISLR)
```

1a Simple linear regression model

Name of the columns

```
names(Auto)
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

Fit Model: mpg ~ horsepower

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

Model Summary

```
summary(auto.lm)
##
## 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?
(Answers obtained using summary(auto.lm)) There is a relationship between horsepower (predictor) and mpg (response) because the p-value is extremely below 0.05, which means that chances that this relationship occurred, when there is no relationship at all , is extremely slim, therefore there has to be a relationship

(ii) How strong is the relationship between the predictor and the response?
(Answers obtained using summary(auto.lm)) The relationship is strong, about 60%, because the R^2 = .6059. This statistic measures the proportion of variability in response that can be explained using the predictor.

(iii) Is the relationship between the predictor and the response positive or negative?
(Answers obtained using summary(auto.lm)) The relationship between mpg and horsepower has a negative relationship because the coefficient of horsepower (predictor) is negative

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

```
predict(auto.lm, data.frame(horsepower=c(98)), interval="prediction")
##       fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

1b Plot Regression Line

```
attach(Auto)
plot(horsepower, mpg) # Plot points
abline(auto.lm) # Add Least Squares Regression Line
```

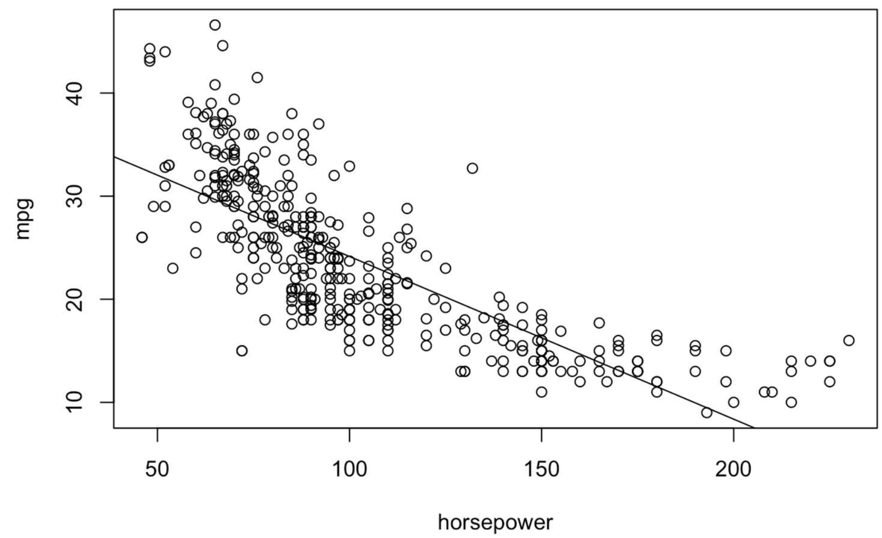

Figure: 1-b

1c Diagnostic Plots

```
par(mfrow = c(2,2)) # 4 plots in same picture
plot(auto.lm)
```

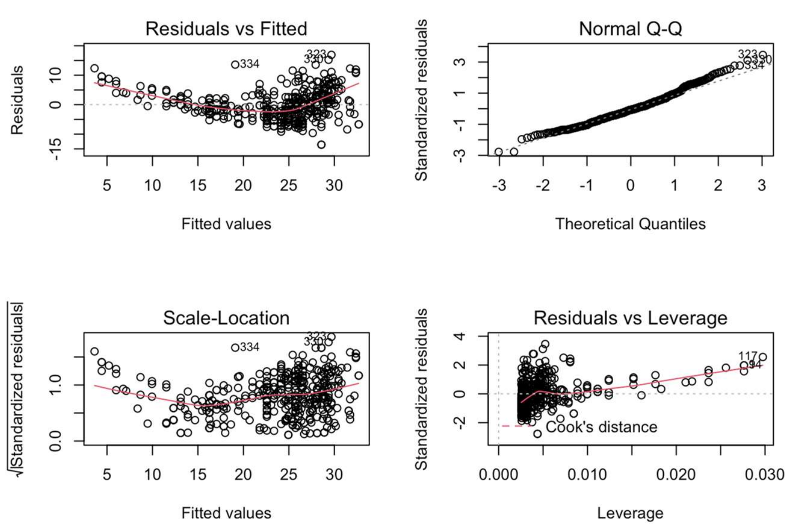

Figure: 1-c

Practical-2:

```
library(ISLR)
library(tidyverse)
library(GGally)
library(car) # scatterplotMatrix
```

2a Scatterplot Matrix
Produce a scatterplot matrix which includes all of the variables in the data set.

Basic Scatterplot Matrix

```
pairs(Auto)
```

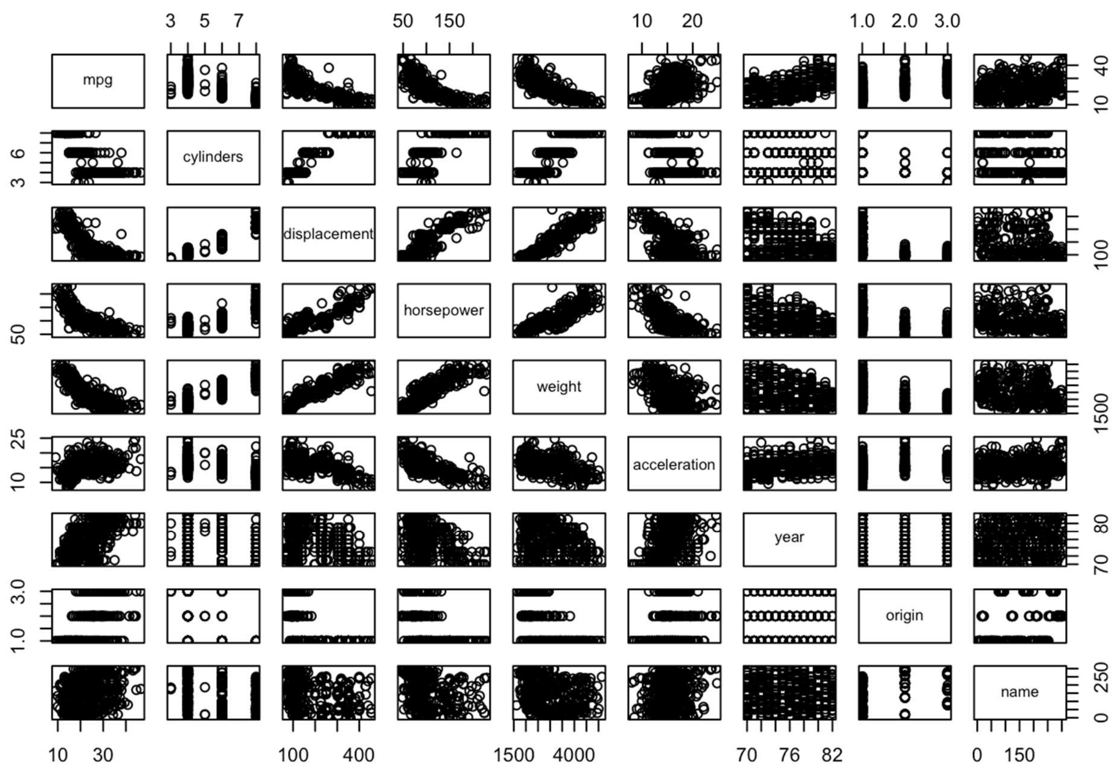

Figure: 2-a-i

Enhanced Pairs Plot

```
auto <- as_tibble(Auto)
auto <- select(auto, -name)
colnames(auto)

## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"
```

Rename a few columns

```
names(auto)[names(auto) == "displacement"] <- "displ"
names(auto)[names(auto) == "horsepower"] <- "hp"
names(auto)[names(auto) == "acceleration"] <- "accel"
ggpairs(auto)
```

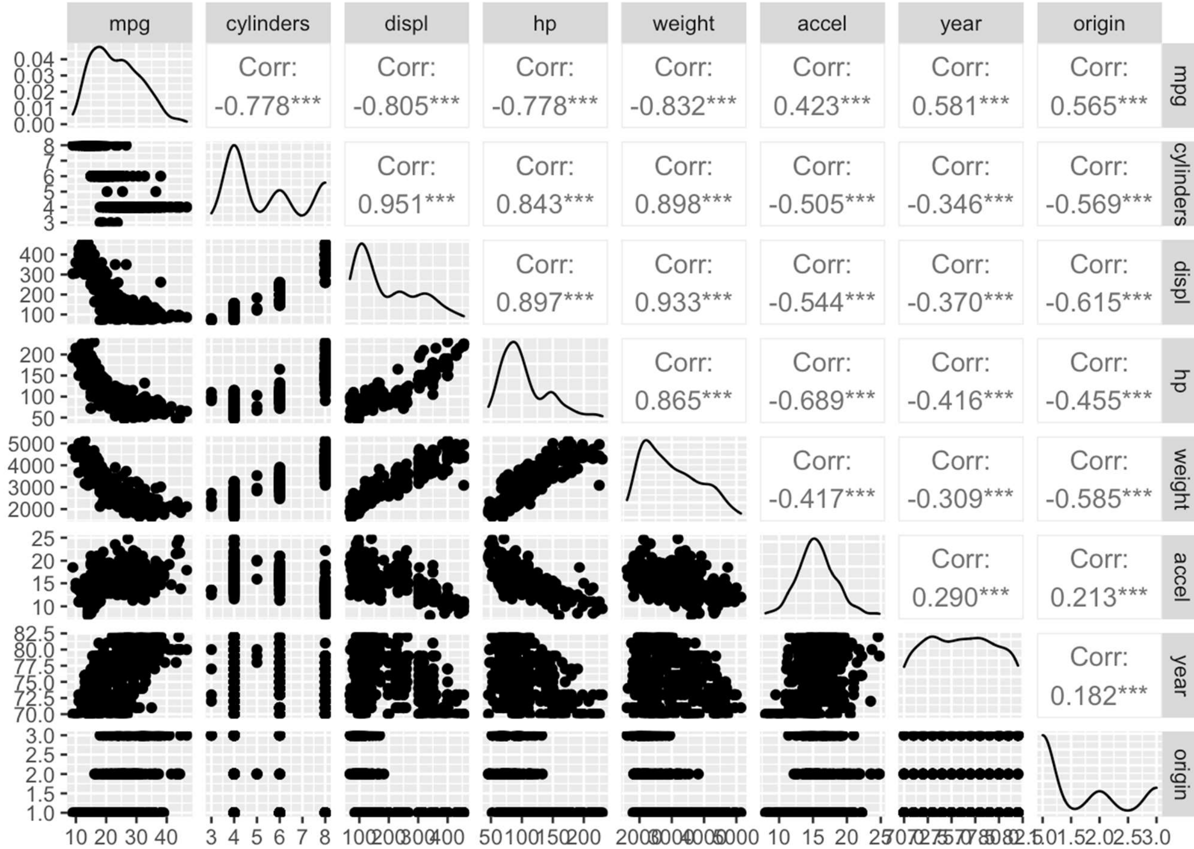

Figure: 2-a-ii

car Package scatterplotMatrix

scatterplotMatrix(auto, smooth = FALSE, main="Scatter Plot Matrix")

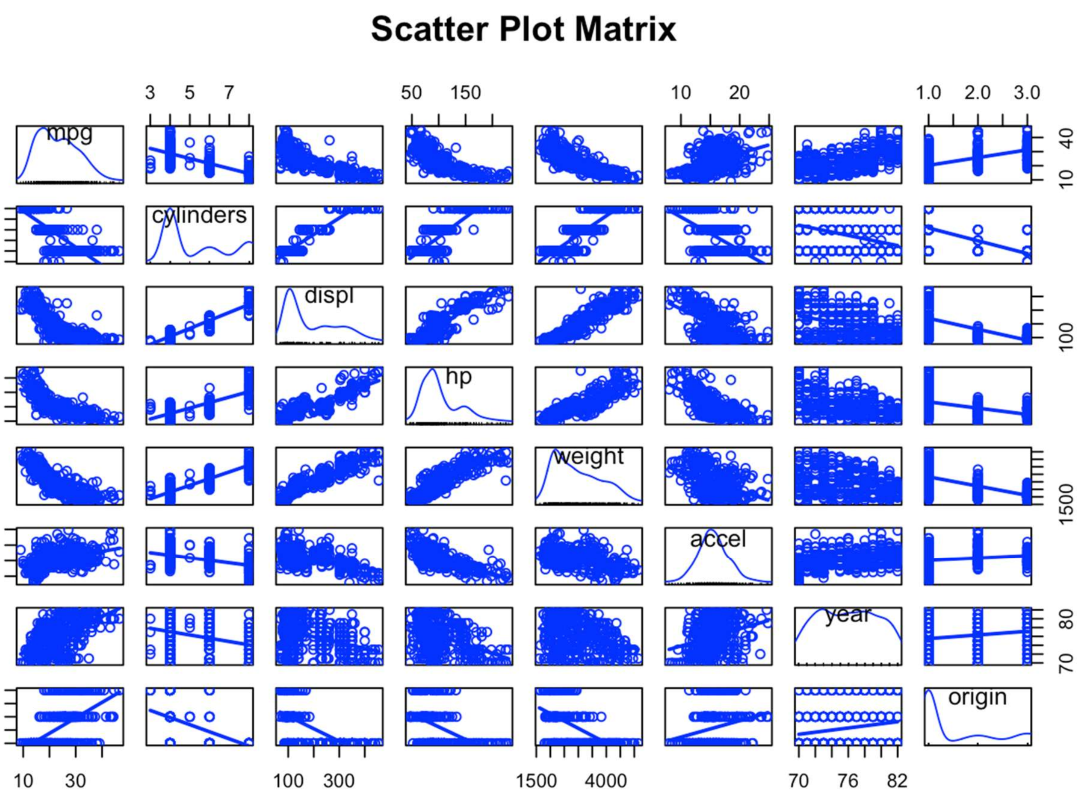

Figure: 2-a-iii

2b Correlations Matrix
Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

```
options(digits=2)
cor(auto[,!colnames(auto) %in% c("name")]) # Skip name column
```

```
##           mpg cylinders displ    hp weight accel  year origin
## mpg       1.00    -0.78 -0.81 -0.78  -0.83  0.42  0.58   0.57
## cylinders -0.78    1.00  0.95  0.84   0.90 -0.50 -0.35  -0.57
## displ     -0.81    0.95  1.00  0.90   0.93 -0.54 -0.37  -0.61
## hp        -0.78    0.84  0.90  1.00   0.86 -0.69 -0.42  -0.46
## weight    -0.83    0.90  0.93  0.86   1.00 -0.42 -0.31  -0.59
## accel      0.42   -0.50 -0.54 -0.69  -0.42  1.00  0.29   0.21
## year       0.58   -0.35 -0.37 -0.42  -0.31  0.29  1.00   0.18
## origin     0.57   -0.57 -0.61 -0.46  -0.59  0.21  0.18   1.00
```

Enhanced Correlation Plot

```
ggcorr(auto)
```

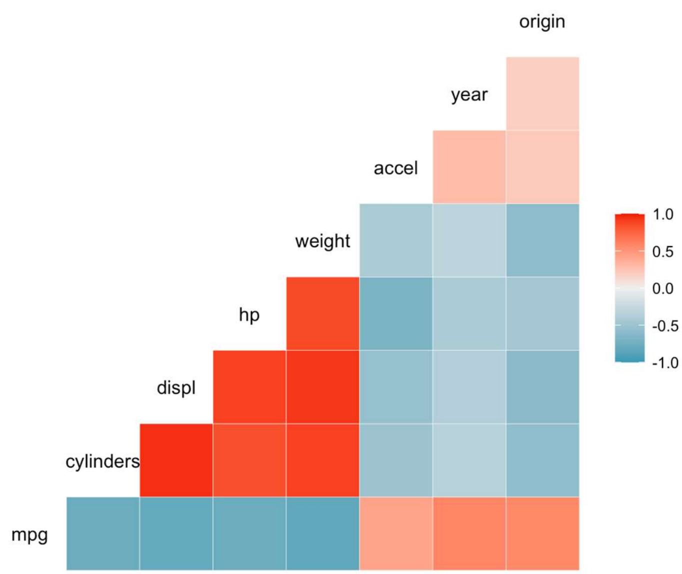

Figure: 2-b

2c Multiple Linear Regression: mpg ~ .
Running a MLR on all predictors except for name

Model: mpg ~ . -name

```
auto.mlr = lm(mpg ~ . -name, data=Auto)
```

Model Summary

```
summary(auto.mlr)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.590 -2.157 -0.117  1.869 13.060
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.72e+01   4.64e+00   -3.71  0.00024 ***
## cylinders    -4.93e-01   3.23e-01   -1.53  0.12780
## displacement  1.99e-02   7.51e-03    2.65  0.00844 **
## horsepower   -1.70e-02   1.38e-02   -1.23  0.21963
## weight       -6.47e-03   6.52e-04   -9.93  < 2e-16 ***
## acceleration  8.06e-02   9.88e-02    0.82  0.41548
## year          7.51e-01   5.10e-02   14.73  < 2e-16 ***
## origin        1.43e+00   2.78e-01    5.13  4.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.3 on 384 degrees of freedom
## Multiple R-squared:  0.821,  Adjusted R-squared:  0.818
## F-statistic:  252 on 7 and 384 DF,  p-value: <2e-16
```

(i) Is there a relationship between the predictors and the response?
There are multiple predictors that have relationship with the response because their associated p-value is significant

(ii) Which predictors appear to have a statistically significant relationship to the response?
The predictors: displacement, weight, year, and origin have a statistically significant relationship.

(iii) What does the coefficient for the year variable suggest?
The coefficient of year suggests that every 4 years, the mpg goes up by 3

2d Diagnostic Plots

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?

Diagnostic Plots

```
par(mfrow=c(2,2))
plot(auto.mlr)
```

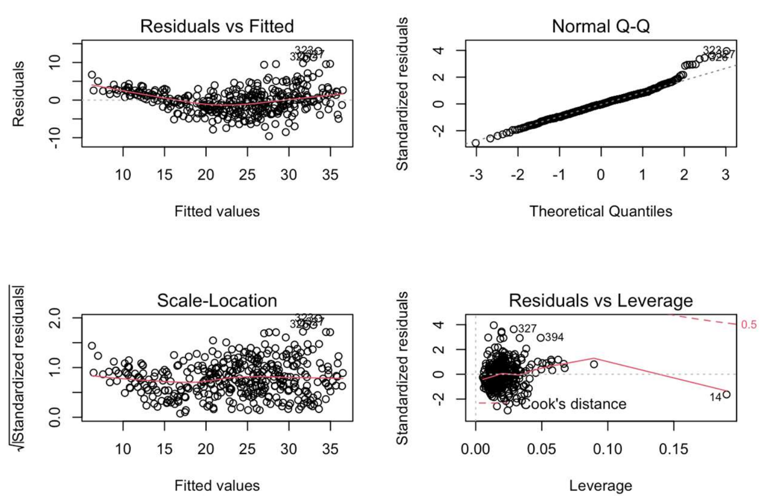

Figure: 2-d

Better Plots

#qplot(auto.mlr)

Non-Linearity: The residual plot shows that there is a U-shape pattern in the residuals which might indicate that the data is non-linear.

Non-constant Variance: The residual plot also shows that the variance is not constant. There is a funnel shape appearing at the end which indicates heteroscedasticity (non-constant variance)

Outliers: There seems to not be any outliers because in the Scale-Location, all values are within the range of [-2,2]. It will only be an outlier if standardized residual is outside the range of [-3, 3].

High Leverage Points: Based on the Residuals vs. Leverage graph, there is no observations that provides a high leverage

2e Interaction Effects

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

```
names(Auto)

## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"

interact.fit = lm(mpg ~ . -name + horsepower*displacement, data=Auto)
origin.hp = lm(mpg ~ . -name + horsepower*origin, data=Auto)
summary(origin.hp)

##
## Call:
## lm(formula = mpg ~ . - name + horsepower * origin, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.277 -1.875 -0.225  1.570 12.080
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.20e+01   4.40e+00   -5.00  8.9e-07 ***
## cylinders         -5.28e-01   3.03e-01   -1.74    0.082 .
## displacement      -1.49e-03   7.61e-03   -0.20    0.845
## horsepower         8.17e-02   1.86e-02    4.40  1.4e-05 ***
## weight            -4.71e-03   6.55e-04   -7.19  3.5e-12 ***
## acceleration      -1.12e-01   9.62e-02   -1.17    0.243
## year               7.33e-01   4.78e-02   15.33  < 2e-16 ***
## origin             7.70e+00   8.86e-01    8.69  < 2e-16 ***
## horsepower:origin -7.95e-02   1.07e-02   -7.40  8.4e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.1 on 383 degrees of freedom
## Multiple R-squared:  0.844,  Adjusted R-squared:  0.841
## F-statistic:  259 on 8 and 383 DF,  p-value: <2e-16
```

Statistically Significant Interaction Terms:
- displacement and horsepower
- horsepower and origin

```
inter.fit = lm(mpg ~ . -name + horsepower:origin + horsepower:
          + horsepower:displacement, data=Auto)
summary(inter.fit)

##
## Call:
## lm(formula = mpg ~ . - name + horsepower:origin +
horsepower:+horsepower:displacement,
##     data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.722 -1.525 -0.097  1.355 12.842
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -4.71e+00   4.69e+00   -1.00    0.316
## cylinders                5.14e-01   3.14e-01    1.64    0.102
## displacement            -6.97e-02   1.14e-02   -6.10  2.6e-09 ***
## horsepower              -1.54e-01   3.55e-02   -4.34  1.8e-05 ***
## weight                  -3.08e-03   6.48e-04   -4.76  2.7e-06 ***
## acceleration            -2.28e-01   9.10e-02   -2.50    0.013 *
## year                     7.35e-01   4.46e-02   16.48  < 2e-16 ***
## origin                   2.28e+00   1.09e+00    2.09    0.037 *
## horsepower:origin       -1.92e-02   1.28e-02   -1.50    0.134
## displacement:horsepower  4.67e-04   6.13e-05    7.61  2.1e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.9 on 382 degrees of freedom
## Multiple R-squared:  0.864,  Adjusted R-squared:  0.861
## F-statistic:  271 on 9 and 382 DF,  p-value: <2e-16
```

Adding more interactions, decreases the significance of previous significant values

Practical-3:

3a. Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
data(Carseats)
summary(Carseats)
     Sales          CompPrice       Income        Advertising
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
   Population        Price        ShelveLoc        Age
 Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00
 1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75
```

```
 Median :272.0   Median :117.0   Medium:219   Median :54.50
 Mean   :264.8   Mean   :115.8               Mean   :53.32
 3rd Qu.:398.5   3rd Qu.:131.0               3rd Qu.:66.00
 Max.   :509.0   Max.   :191.0               Max.   :80.00
  Education    Urban      US
 Min.   :10.0   No :118   No :142
 1st Qu.:12.0   Yes:282   Yes:258
 Median :14.0
 Mean   :13.9
 3rd Qu.:16.0
 Max.   :18.0
model1 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(model1)

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
```

3b. Provide an interpretation of each coefficient in the model. Be careful, some of the variables in the model are qualitative!

Price: suggests a relationship between price and sales given the low p-value of the t-statistic. The coefficient states a negative relationship between Price and Sales: as Price increases, Sales decreases.

UrbanYes: The linear regression suggests that there is not enough evidence for arelationship between the location of the store and the number of sales based.

USYes: Suggests there is a relationship between whether the store is in the US or not and the amount of sales. A positive relationship between USYes and Sales: if the store is in the US, the sales will increase by approximately 1201 units.

3c. Write out the model in equation form, being careful to handle the qualitative variables properly.

Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

3d. For which of the predictors can you reject the null hypothesis H0:βj=0?

Price and USYes, based on the p-values, F-statistic, and p-value of the F-statistic.

3e. 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.

```
model2 <-  lm(Sales ~ Price + US, data = Carseats)
summary(model2)

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
```

3f. How well do the models in (a) and (e) fit the data?

Based on the RSE and R^2 of the linear regressions, they both fit the data similarly, with linear regression from (e) fitting the data slightly better.

3g. Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
confint(model2)
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
```

3h. Is there evidence of outliers or high leverage observations in the model from (e)?

```
plot(predict(model2), rstudent(model2))
```

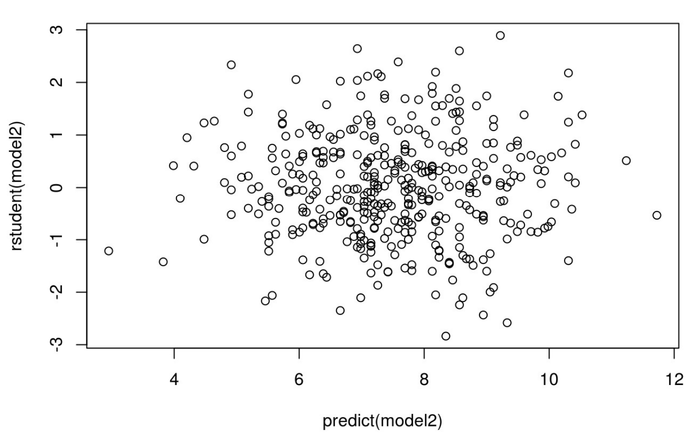

Figure: 3-h-i

All studentized residuals appear to be bounded by -3 to 3, so no potential outliers are suggested from the linear regression.

```
par(mfrow=c(2,2))
plot(model2)
```

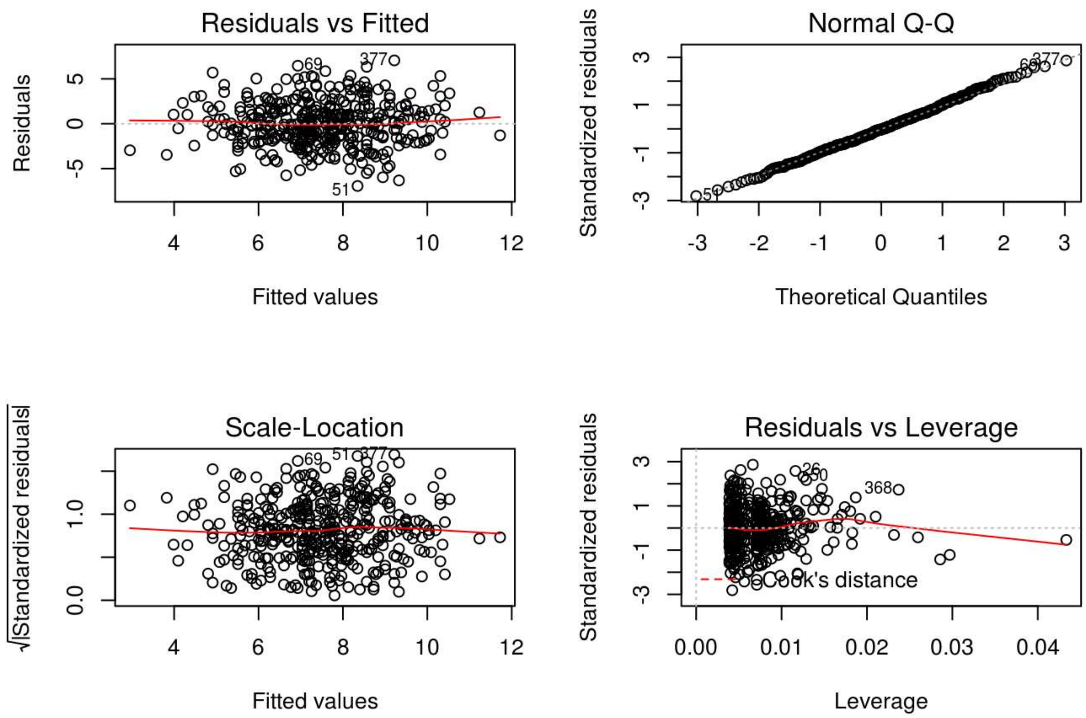

Figure: 3-h-ii

There are a few observations that greatly exceed (p+1)/n (0.0075567) on the leverage-statistic plot that suggest that the corresponding points have high leverage