3

Sai Ananthula - emw832

```
library(MASS)
library(ISLR2)
```

```
##
## Attaching package: 'ISLR2'
```

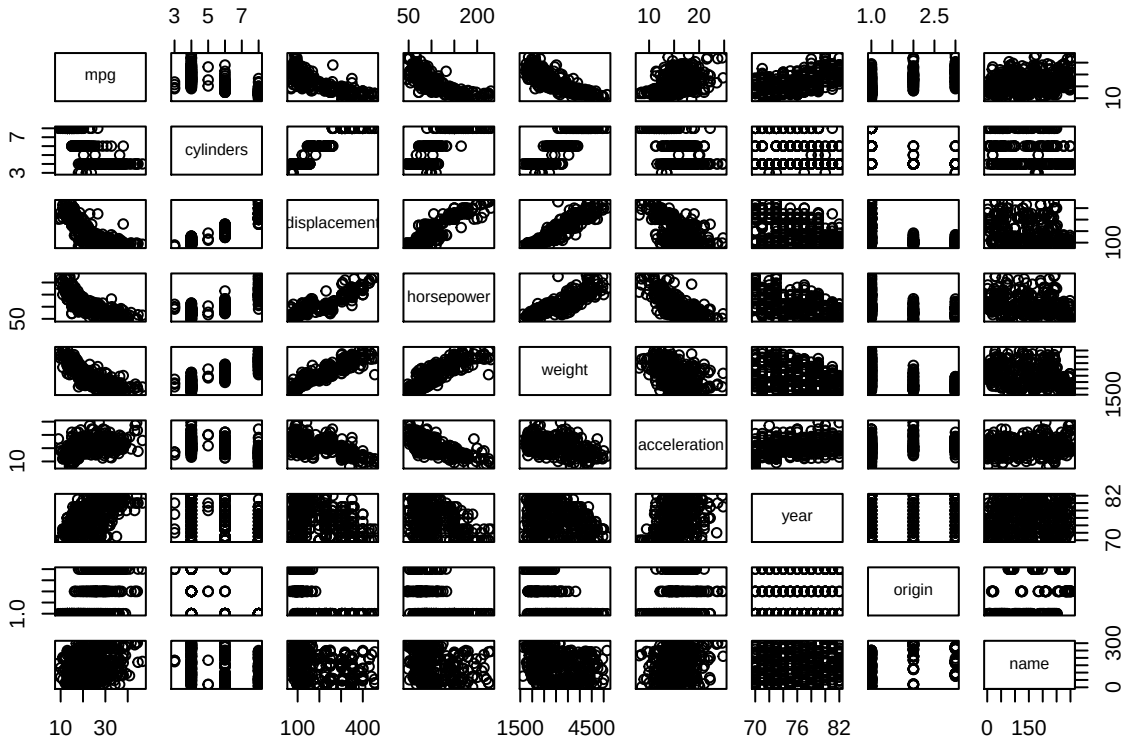
```
## The following object is masked from 'package:MASS':
##
##     Boston
```

2. The KNN classifier when given a point h uses n number of points near it to classify it while the KNN regression when given predictor h uses n number of points near it to calculate the observation for it. The classifer is qualitiative and the regression quantitative.

3.

a.

```
pairs(Auto)
```




```
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

b.

```
cor(subset(Auto, select=-name))
```

```
##                     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.

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

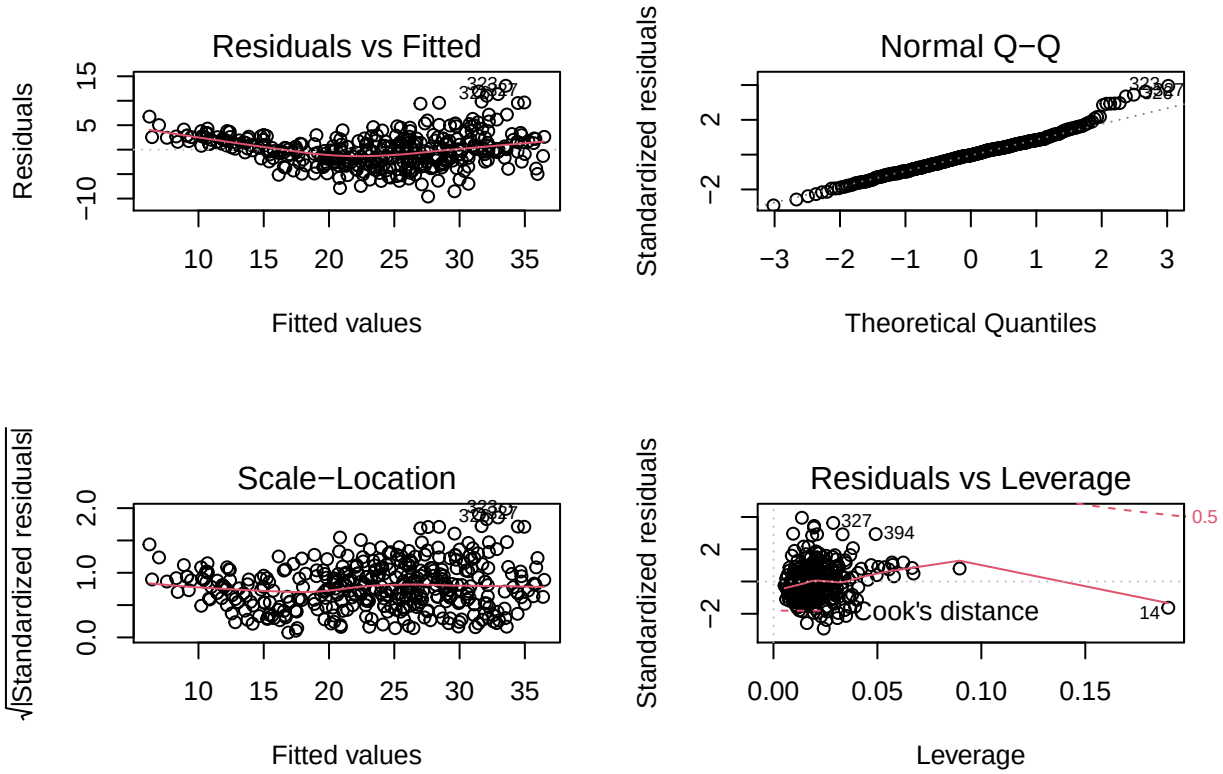
c1. There is a relationship between the predictos and the response variable which in this case is mpg. Tge F-Stat is 252.4 which is far from 1 so there is reasonable evidence to reject the null hypothesis.

c2. The predictors that seemed to be the most correlated with MPG are weight, year, origin, and displacement in that order.

c3.It mean every year that passes the mpg goes up .75.

d.

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



14 is an outlier with high leverage but a normalish standardized residual. Also, linear might not be the best fit for this due to the curve in residuals vs fitted chart.

e.

```
mpg2.fit <- lm(mpg ~ weight * displacement + year * origin, data = Auto )
summary(mpg2.fit)
```

```
##
## Call:
## lm(formula = mpg ~ weight * displacement + year * origin, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5758 -1.6211 -0.0537  1.3264 13.3266
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.793e+01  8.044e+00   2.229 0.026394 *
## weight               -1.035e-02  6.450e-04 -16.053  < 2e-16 ***
```

```
## displacement        -7.519e-02  9.091e-03  -8.271 2.19e-15 ***
## year                 4.864e-01  1.017e-01   4.782 2.47e-06 ***
## origin              -1.503e+01  4.232e+00  -3.551 0.000432 ***
## weight:displacement  2.098e-05  2.179e-06   9.625  < 2e-16 ***
## year:origin          1.980e-01  5.436e-02   3.642 0.000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.969 on 385 degrees of freedom
## Multiple R-squared:  0.8575, Adjusted R-squared:  0.8553
## F-statistic: 386.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

There seems to be a interaction effect between weight and displacement and another interaction effect year:origin.

f.

```
mpg3.fit <- lm(mpg ~ sqrt(weight) * displacement + year * origin, data = Auto )
summary(mpg3.fit)
```

```
##
## Call:
## lm(formula = mpg ~ sqrt(weight) * displacement + year * origin,
##     data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4965 -1.5769 -0.1341  1.3570 13.2981
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.340e+01  8.309e+00   5.223 2.89e-07 ***
## sqrt(weight)               -1.047e+00  6.517e-02 -16.061  < 2e-16 ***
## displacement               -1.181e-01  1.672e-02  -7.063 7.66e-12 ***
## year                        4.933e-01  1.015e-01   4.859 1.72e-06 ***
## origin                     -1.481e+01  4.224e+00  -3.506 0.000508 ***
## sqrt(weight):displacement   1.952e-03  2.573e-04   7.587 2.49e-13 ***
## year:origin                 1.951e-01  5.426e-02   3.596 0.000365 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.962 on 385 degrees of freedom
## Multiple R-squared:  0.8582, Adjusted R-squared:  0.856
## F-statistic: 388.4 on 6 and 385 DF,  p-value: < 2.2e-16
```

By taking the square root of weight the r^2 value increase by .007.

10.

a.

```
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          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

```
car.fit <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(car.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. There is most likely a relationship between price ans sales due to the p-value. The negative coefficient suggests as price increases sales decrease.

There is not a relationship between urban location and sales due to the high p value.

There is a relationship between USYes and sales due to the low p value. Since this variable is qualitative it mean stores in the US make more sales.

c. Sales = 13.04 - .054Price - .022UrbanYes + 1.201USYes

d. Price and USYes due their low p-values.

e.

```
car2.fit <- lm(Sales ~ Price + US, data = Carseats)
summary(car2.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. Equally bad due to R^2 that are hovering around .2393 and F-statistics in the 40-65 range. The model in e is marginally better but not by much.

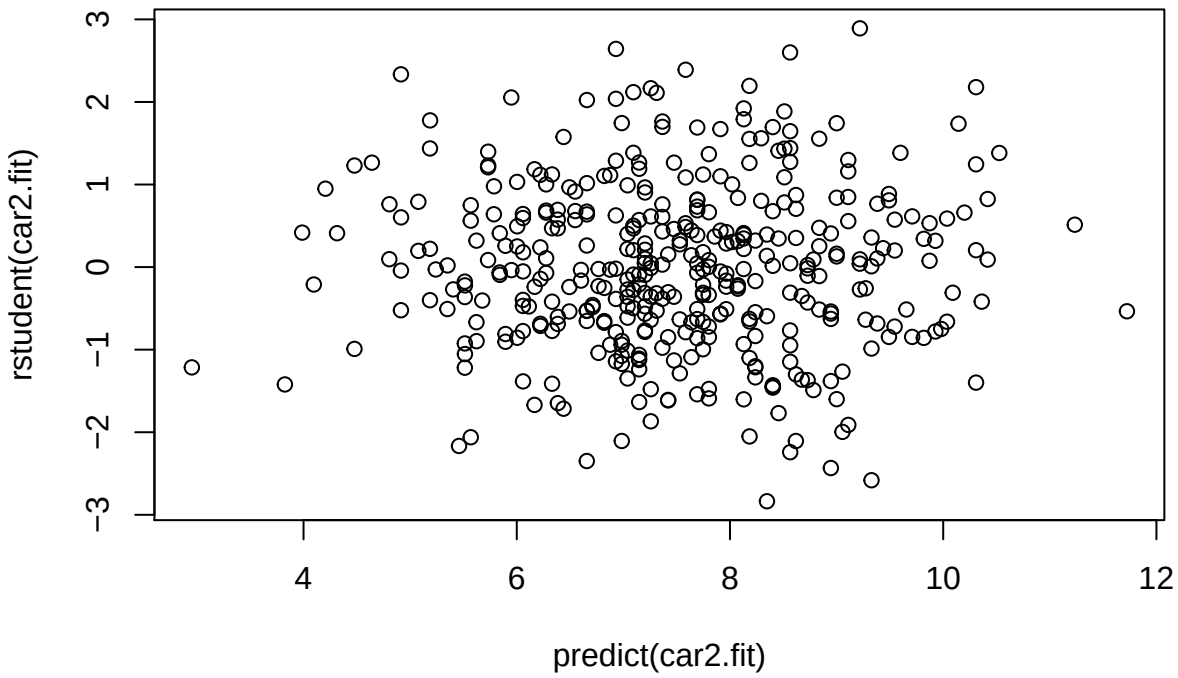
g.

```
confint(car2.fit)
```

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

h.

```
plot(predict(car2.fit), rstudent(car2.fit))
```



Everything is bounded from -3 to so there seems to not be any potenital outliers.

12a. They are the same when the coefficient is the same and these is a lack of noise.

12b.

```
x = rnorm(100)
y = 0.5 * x + rnorm(100)
coefficients(lm(x ~ y + 0))
```

```
##         y
## 0.5534561
```

```
coefficients(lm(y ~ x + 0))
```

```
##         x
## 0.4857814
```

12c.

```
x  =rnorm(100)
y = 1*x
coefficients(lm(x ~ y + 0))
```

```
## y
## 1
```

```r
coefficients(lm(y ~ x + 0))
```

```
## x
## 1
```