Assignment 4: Linear Regression

Use the Auto data set from the ISLR package to answer the following questions:

(a) Perform a simple linear regression with mpg as the response and cylinders, displacement, horsepower, weight, and acceleration as the predictors.

i. Is there a relationship between each predictor and the response?

Assuming the critical value is 0.05

Coding:

```
library(ISLR)
# call the library
attach(Auto)
lm_fit = lm(mpg~cylinders+displacement+horsepower+ weight+acceleration)
# fit the data as multiple linear regression, let mpg as y, others as predictor
summary(lm_fit)
```

Output:

```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5816  -2.8618  -0.3404   2.2438  16.3416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.626e+01  2.669e+00  17.331   <2e-16 ***
cylinders    -3.979e-01  4.105e-01  -0.969   0.3330
displacement -8.313e-05  9.072e-03  -0.009   0.9927
horsepower   -4.526e-02  1.666e-02  -2.716   0.0069 **
weight       -5.187e-03  8.167e-04  -6.351    6e-10 ***
acceleration -2.910e-02  1.258e-01  -0.231   0.8171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 386 degrees of freedom
Multiple R-squared:  0.7077,    Adjusted R-squared:  0.7039
F-statistic: 186.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

From the above output we can see that F-statistic is 186.9, so that means at

least one indicator has relationship with the response. And there are some indicator's p values is smaller than 0.05. So, there is a relationship between each predictor and the response.

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

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.626e+01  2.669e+00  17.331   <2e-16 ***
cylinders   -3.979e-01  4.105e-01  -0.969   0.3330
displacement -8.313e-05  9.072e-03  -0.009   0.9927
horsepower  -4.526e-02  1.666e-02  -2.716   0.0069 **
weight      -5.187e-03  8.167e-04  -6.351    6e-10 ***
acceleration -2.910e-02  1.258e-01  -0.231   0.8171
```

From above output, we can see cylinder's p value is 0.333 which is larger than 0.05, so based on multiple linear regression, there is no relationship between cylinders and MPG.

For displacement, its' p value is 0.9927 which is larger than 0.05, so based on multiple linear regression, there is no relationship between displacement and MPG.

For horsepower, its' p value is 0.0069 which is smaller than 0.05, so based on multiple linear regression, there is a strong relationship between horsepower and MPG.

For weight, its' p value is 6*10^(-10) which is significant smaller than 0.05, so based on multiple linear regression, there is a significant relationship between weight and MPG.

For acceleration, its' p value is 0.8171 which is larger than 0.05, so based on multiple linear regression, there is no relationship between acceleration and MPG.

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

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.626e+01  2.669e+00  17.331   <2e-16 ***
cylinders    -3.979e-01  4.105e-01  -0.969   0.3330
displacement -8.313e-05  9.072e-03  -0.009   0.9927
horsepower   -4.526e-02  1.666e-02  -2.716   0.0069 **
weight       -5.187e-03  8.167e-04  -6.351    6e-10 ***
acceleration -2.910e-02  1.258e-01  -0.231   0.8171
```

Based on the multiple linear regression output, every indicator's coefficient is smaller than 0. So, it is negative relationship between each predictor and the response.

iv. Characterize the fitness of the developed model.

```
Residual standard error: 4.247 on 386 degrees of freedom
Multiple R-squared:  0.7077,    Adjusted R-squared:  0.7039
F-statistic: 186.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

The residual standard error (RSE) is 4.247 on 386 degrees of freedom, which suggests that the average difference between the observed values of the response variable and the predicted values from the model is around 4.247 units.

The multiple R-squared value is 0.7077, which indicates that the model

explains 70.77% of the variation in the response variable. This is a relatively high value, suggesting that the model is a good fit for the data.

The adjusted R-squared value is 0.7039, which takes into account the number of predictor variables in the model. This value is slightly lower than the multiple R-squared value, which suggests that adding more predictor variables to the model may not substantially improve its fit.

The F-statistic is 186.9 on 5 and 386 degrees of freedom, with a p-value of less than 2.2e-16. This indicates that the overall fit of the model is statistically significant.

With a high R-squared value and a significant F-statistic, we can see that the model appears to be a good fit to the data.

v. What is the predicted mpg associated cylinders = 8, displacement = 307, horsepower = 130, weight = 3504, and acceleration = 12? What are the associated 95% confidence and prediction intervals?

Code:

```
new_data = data.frame(cylinders = c(8), displacement = c(307), horsepower = c(130), weight = c
# insert new data set
pre_con = predict(lm_fit,newdata = new_data,interval = "confidence", level = 0.95)
pre_con
pre_pre = predict(lm_fit,newdata = new_data,interval = "prediction", level = 0.95)
pre_pre
# check prediction value, confidence interval and prediction interval.
```

Output:

```
> new_data = data.frame(cylinders = c(8), displacement = c(307), horsepower = c(130), weight = c(3504),acceleration = c(12))
> # insert new data set
> pre_con = predict(lm_fit,newdata = new_data,interval = "confidence", level = 0.95)
> pre_con
       fit      lwr      upr
1 18.64772 17.52881 19.76663
> pre_pre = predict(lm_fit,newdata = new_data,interval = "prediction", level = 0.95)
> pre_pre
       fit      lwr      upr
1 18.64772 10.22283 27.07261
>
```

So, the prediction value of cylinders = 8, displacement = 307, horsepower = 130, weight = 3504, and acceleration = 12 is **18.64772,** and confidence interval is (17.52881,19.7663) while prediction interval is (10.22283,27.07261).

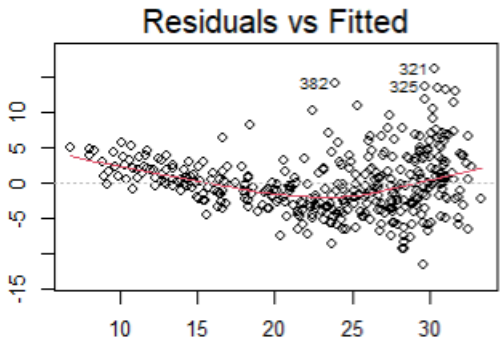
(b) Produce the diagnostic plots of the least squares regression fit.

Comment on each plot.

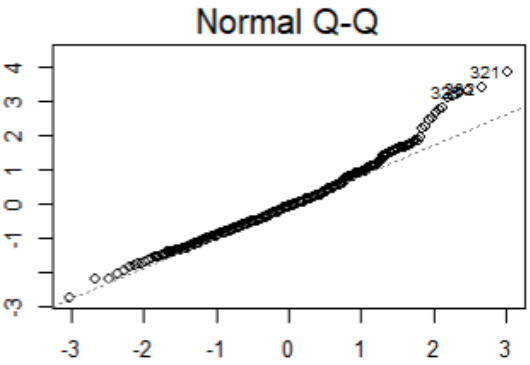
Code:

```
# diagnostic plots
par(mfrow = c(3, 2)) # Set plot layout to 3x2
plot(lm_fit, which = 1) # Residuals vs Fitted
plot(lm_fit, which = 2) # Normal Q-Q
plot(lm_fit, which = 3) # Scale-Location
plot(lm_fit, which = 4) # Cook's distance
plot(lm_fit, which = 5) # Residuals vs Leverage
```

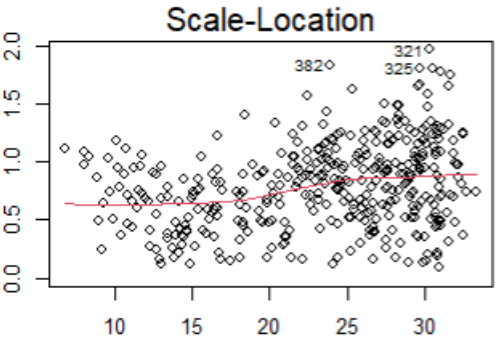
Output:



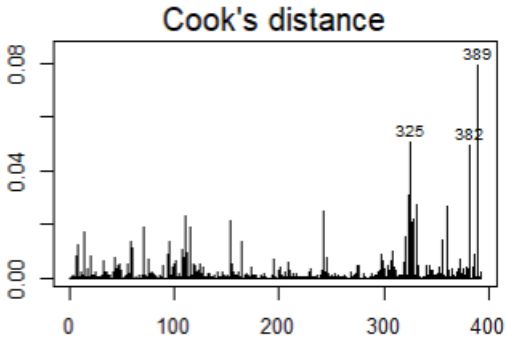
From above we can see that there is slight non-linearity.



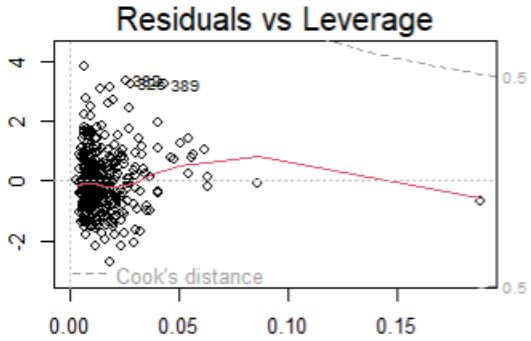
From above we can see that most of residuals follows normal distribution, until the end set of the data which are not on the line.



There are not very obvious trend shows up for this picture, so it may indicate constant variance.



From above picture we can see that there are 3 points which are influential points. Because it is significantly higher than other points.



From above we can see that there is one point which is out off center. So, it means there is unusual value for Xi, we need limit the values of x.

(c) Try one transformation of a select predictor variable (such as log(X), $\sqrt{X}$, X^2). Comment on your findings. Did it improve model fit?

Code:

```
#transforming  of a select predictor variable
lm_fit = lm(mpg~cylinders+I(displacement^2)+horsepower+ weight+acceleration)
summary(lm_fit)
```

Output:

```
> lm_fit = lm(mpg~cylinders+I(displacement^2)+horsepower+ weight+acceleration)
> summary(lm_fit)

Call:
lm(formula = mpg ~ cylinders + I(displacement^2) + horsepower +
    weight + acceleration)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6062  -2.6891  -0.2676   2.1902  16.0375

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.121e+01  2.749e+00  18.629  < 2e-16 ***
cylinders         -1.009e+00  3.401e-01  -2.967 0.003198 **
I(displacement^2)  5.200e-05  1.396e-05   3.726 0.000223 ***
horsepower        -7.270e-02  1.743e-02  -4.172 3.74e-05 ***
weight            -5.699e-03  7.349e-04  -7.754 7.99e-14 ***
acceleration      -1.402e-02  1.229e-01  -0.114 0.909285
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.173 on 386 degrees of freedom
Multiple R-squared:  0.7178,    Adjusted R-squared:  0.7142
F-statistic: 196.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

So, after transforming displacement to ^2, the R^2 and adjusted R^2 value is better than before. It means that there might be a quadratic relationship between MGP and displacement and it improves model fit.

(c) Try one transformation of the response variable. Comment on your findings. Did it improve model fit?

Code and output:

```
> lm_fit = lm(I(mpg^2)~cylinders+displacement+horsepower+ weight+acceleration)
> summary(lm_fit)

Call:
lm(formula = I(mpg^2) ~ cylinders + displacement + horsepower +
    weight + acceleration)

Residuals:
    Min      1Q  Median      3Q     Max
-554.75 -169.59  -48.58  119.54 1237.31

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1608.68435  161.10062   9.986  < 2e-16 ***
cylinders      -7.15937   24.77665  -0.289   0.7728
displacement    0.09086    0.54749   0.166   0.8683
horsepower     -1.69521    1.00556  -1.686   0.0926 .
weight         -0.28612    0.04929  -5.805 1.34e-08 ***
acceleration    3.36896    7.58978   0.444   0.6574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 256.3 on 386 degrees of freedom
Multiple R-squared:  0.5939,    Adjusted R-squared:  0.5886
F-statistic: 112.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

After transforming dependent variable to square, we can see that R^2 value and adjusted R^2 value is much lower than before. It does not improve model fit.

(d) Try one interaction effect between any two predictors. Comment on your findings. Did it improve model fit?

Code and output:

```
> lm_fit = lm(mpg~cylinders*displacement+horsepower+ weight+acceleration)
> summary(lm_fit)

Call:
lm(formula = mpg ~ cylinders * displacement + horsepower + weight +
    acceleration)

Residuals:
     Min       1Q   Median       3Q      Max
-14.2613  -2.4353  -0.4322   1.9954  16.2509

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         61.0206448  3.2996017  18.493  < 2e-16 ***
cylinders           -2.8502621  0.5248865  -5.430 9.99e-08 ***
displacement        -0.0980427  0.0165296  -5.931 6.69e-09 ***
horsepower          -0.0875506  0.0168740  -5.188 3.44e-07 ***
weight              -0.0036665  0.0008017  -4.574 6.47e-06 ***
acceleration        -0.0551309  0.1187976  -0.464    0.643
cylinders:displacement  0.0145998  0.0021070   6.929 1.79e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.01 on 385 degrees of freedom
Multiple R-squared:  0.7401,    Adjusted R-squared:  0.7361
F-statistic: 182.7 on 6 and 385 DF,  p-value: < 2.2e-16
```

So, after interaction independent variables between cylinders and displacement, the R^2 value and adjusted R^2 value improved a lot than before. So, that means that including the interaction term in the regression model has increased the amount of variance in the dependent variable that is explained by the independent variables. And it improves the model fit.