Apurva Shrivastava
925009508
ISEN-613 600

Assignment 2

Problem 1

Use the Auto data set to answer the following questions:

(a) Perform a simple linear regression with mpg as the response and horsepower as the predictor. Comment on the output. For example

```
> library(ISLR)
> data(Auto)
> fit <- lm(mpg~horsepower, data = Auto)
> summary(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
```

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

If we look at the p- value of the F – statistic, which is, 2.2e-16, very small. When we have a small p-value for the F-statistic it indicates that the predictor is sufficient for the response. In this case, as it is very small it shows there is some relationship between horsepower and mpg.

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

R-squared is a statistical measure of how close the data are to the fitted regression line. If we look at the value of R square, it is .6059 which means, almost 60.59% of variability in mpg can be explained using horsepower. Hence, there is a moderately strong relationship between the predictor and the response.

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

Apurva Shrivastava
925009508
ISEN-613 600

Whether relationship between predictor and response is positive or negative is judges by the estimated coefficient. As the coefficient of horsepower is negative (-.1578545), it shows that the relationship between mpg and horsepower is negative. It means, the more horsepower an automobile has the less mpg fuel efficiency it will have.

iv. How to interpret the estimate of the slope?

The estimate of the slope ($\beta 1$) which is represented by the coefficient indicates whether there is a relationship between the predictors and the response or not. If the value of $\beta 1$ is positive, this signifies a positive relation, if it is negative it implies a negative relation as in this case. If $\beta 1$ is 0, this indicates that there is no relationship between the predictor and the response. In this case, it is -0.15 which means there is negative relationship between the response and the predictor.

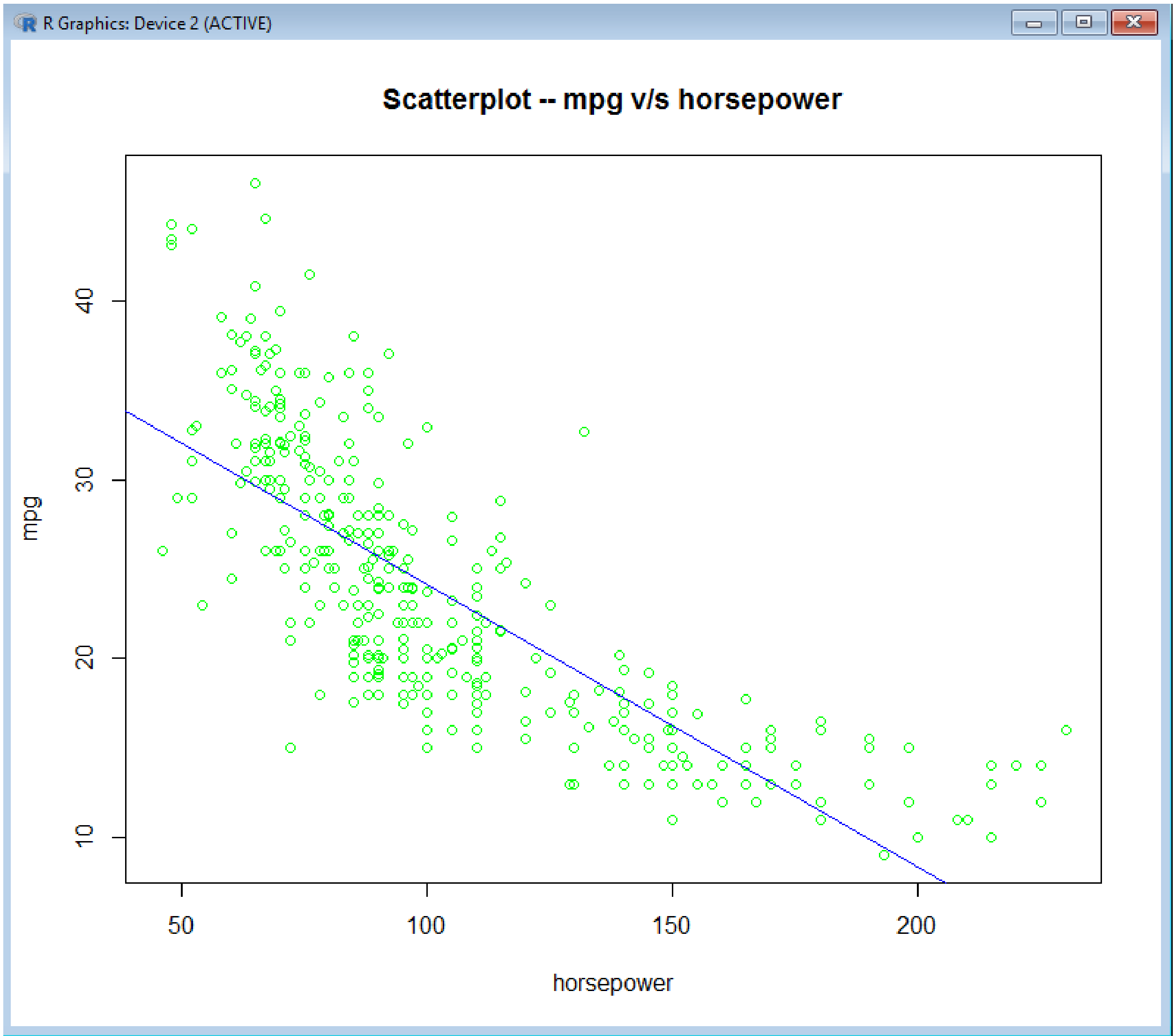
v. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
> 
```

(b) Plot the response and the predictor. Display the least squares regression line in the plot.

```
> plot(Auto$horsepower, Auto$mpg, main = "Scatterplot -- mpg v/s horsepower", xlab = "horsepower", ylab ="mpg", col = "green")
> abline(fit, col ="blue")
> 
```

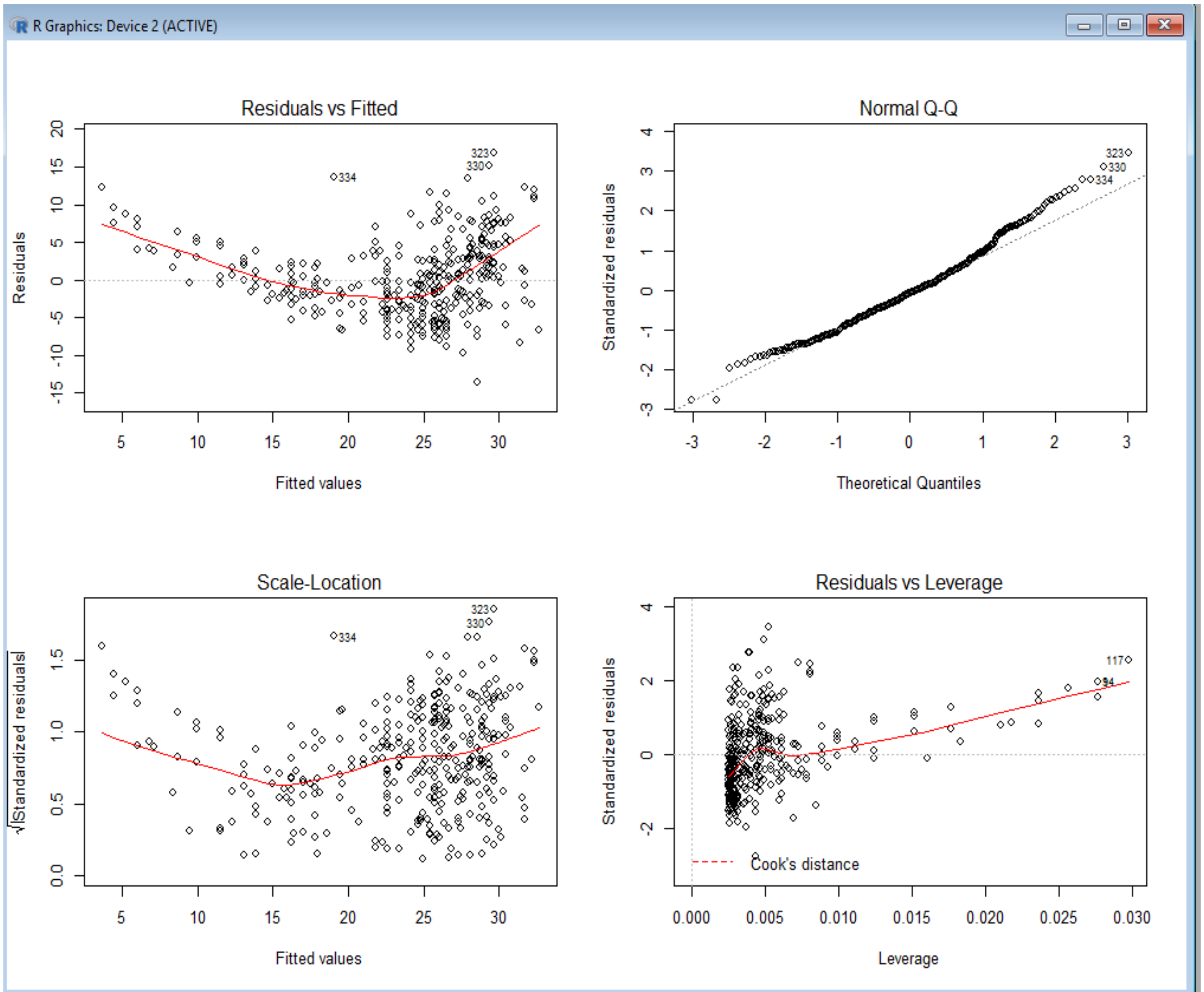
Apurva Shrivastava
925009508
ISEN-613 600



(c) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.

```
> par(mfrow = c(2,2))
> plot(fit)
>
```

Apurva Shrivastava
925009508
ISEN-613 600



Graph	Comment
Residual vs Fitted	This indicates the presence of non-linearity in the data
Normal Q-Q	This indicates that as it is a straight line the data is normally distributed
Scale-Location	It is almost horizontal line which indicates equally (randomly) spread points and verify the assumption of equal variance (homoscedasticity)
Residuals vs Leverage	This indicates the presence of few outliers (higher than 2 and lower than -2) and a few high leverage points

(d) Try a few different transformations of the predictor, such as log(X) , √X,X 2 , and repeat (a)-(c). Comment on your findings.

Apurva Shrivastava
925009508
ISEN-613 600

Log (Horsepower)

(a)

```
> fit <- lm(mpg ~ log(horsepower), data = Auto)
> summary(fit)

Call:
lm(formula = mpg ~ log(horsepower), data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.2299  -2.7818  -0.2322   2.6661  15.4695 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     108.6997     3.0496   35.64   <2e-16 ***
log(horsepower) -18.5822     0.6629  -28.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.501 on 390 degrees of freedom
Multiple R-squared:  0.6683,	Adjusted R-squared:  0.6675 
F-statistic: 785.9 on 1 and 390 DF,  p-value: < 2.2e-16

> 
```

(i) Very small p-value indicates there exists a relationship between the response and the predictor

(ii) R square is .6683 which means 66.83% of variability in mpg can be explained. Hence, it is strongly fitting model

(iii) As the coefficient of estimation is negative, there is a negative relation between mpg and log(horsepower)

(iv) The estimated slope ($\beta 1$) is very negative in this case, -18.58 which means there is a highly negative relation between the response and the predictors.

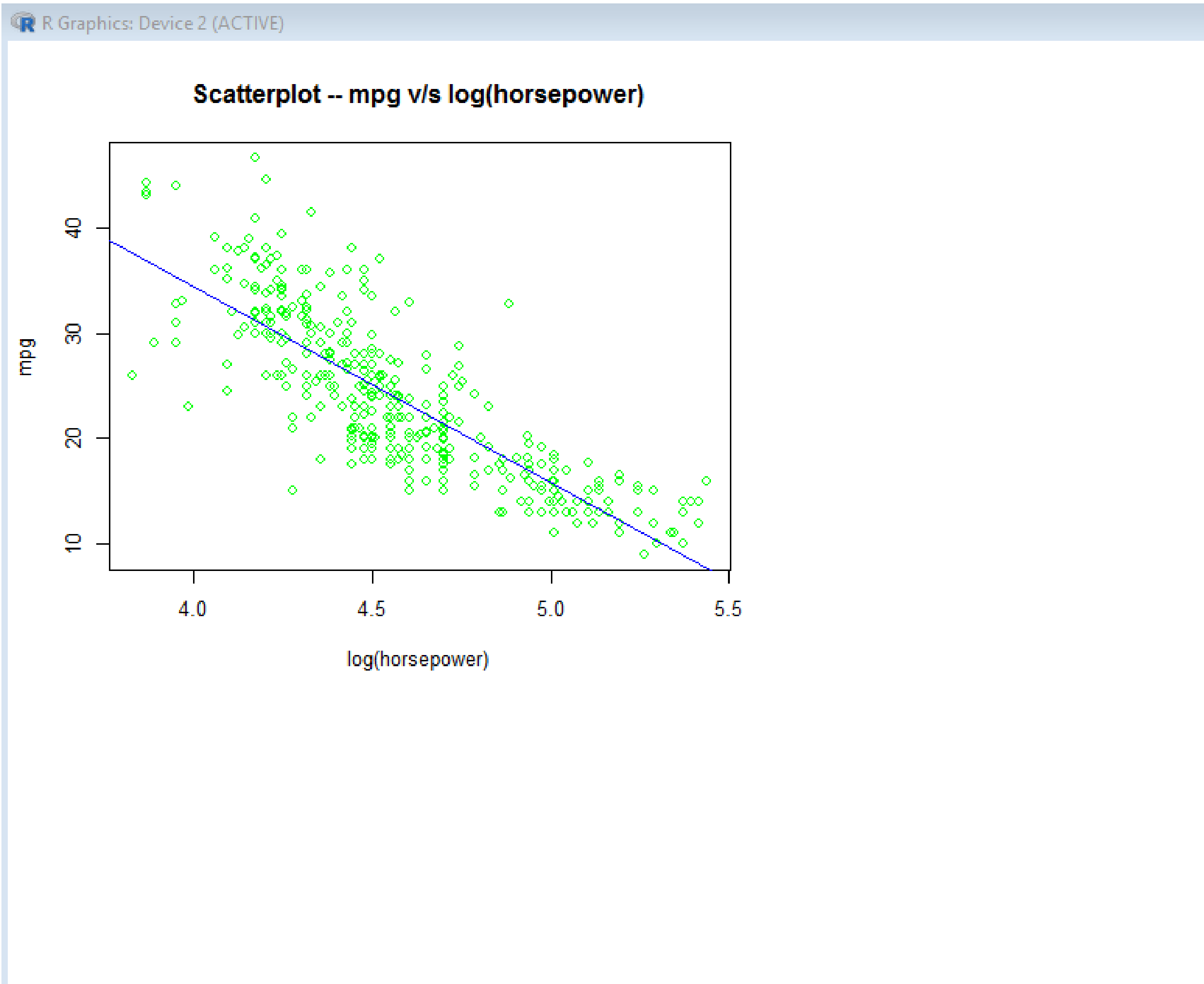
(v)

```
> x <- log(horsepower)
> fit<- lm(mpg ~ x, data = Auto)
> predict(fit, data.frame(x = 98), interval = "confidence")
        fit       lwr       upr
1 -1712.354 -1834.091 -1590.618
> predict(fit, data.frame(x = 98), interval = "predict")
        fit       lwr       upr
1 -1712.354 -1834.412 -1590.297
> 
```

(b)

Apurva Shrivastava
925009508
ISEN-613 600

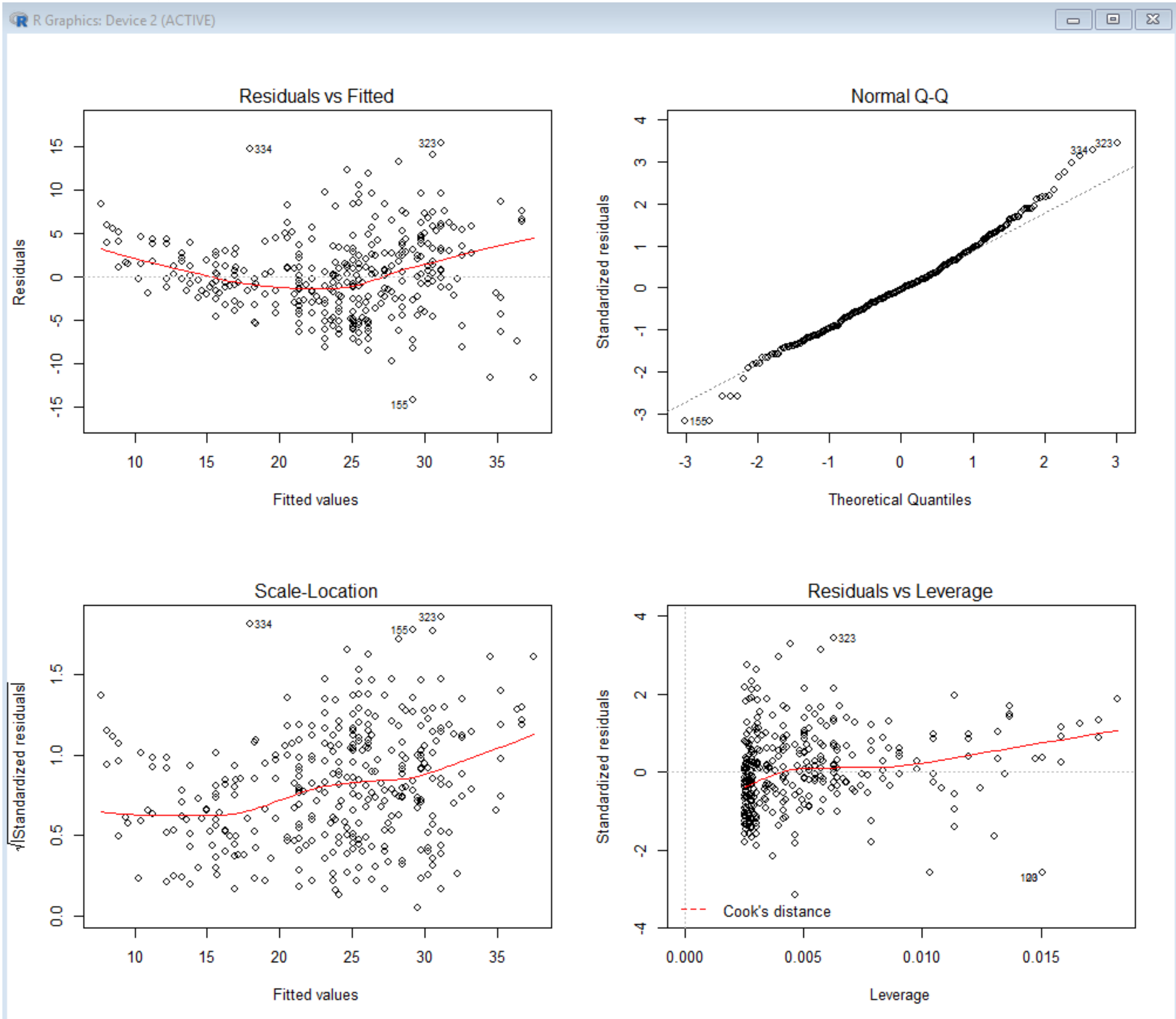
```
> x <- log(Auto$horsepower)
> plot(x, Auto$mpg, main = "Scatterplot -- mpg v/s log(horsepower)", xlab = "log(horsepower)", ylab ="mpg", col = "green")
> abline(fit, color ="blue")
Warning message:
In int_abline(a = a, b = b, h = h, v = v, untf = untf, ...) :
  "color" is not a graphical parameter
> abline(fit, col = "blue")
>
```



(c)

```
> par(mfrow = c(2,2))
> plot(fit)
>
```

Apurva Shrivastava
925009508
ISEN-613 600



Graph	Comment
Residual vs Fitted	This indicates the presence of non-linearity in the data
Normal Q-Q	This indicates that as it is a straight line the data is normally distributed
Scale-Location	It is skewed horizontal line which indicates not equally (randomly) spread points
Residuals vs Leverage	This indicates the presence of few outliers (higher than 2 and lower than -2) and a few high leverage points

Apurva Shrivastava
925009508
ISEN-613 600

Sqrt(Horsepower)

(a)

```
> fit <- lm(mpg ~ sqrt(horsepower), data = Auto)
> summary(fit)

Call:
lm(formula = mpg ~ sqrt(horsepower), data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.9768  -3.2239  -0.2252   2.6881  16.1411 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        58.705      1.349   43.52   <2e-16 ***
sqrt(horsepower)   -3.503      0.132  -26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.665 on 390 degrees of freedom
Multiple R-squared:  0.6437,	Adjusted R-squared:  0.6428 
F-statistic: 704.6 on 1 and 390 DF,  p-value: < 2.2e-16
```

(i) Very small p-value indicates there exists a relationship between the response and the predictor

(ii) R square is .6437 which means 64.37% of variability in mpg can be explained. Hence, it is strongly fitting model

(iii) As the coefficient of estimation is negative, there is a negative relation between mpg and log(horsepower)

(iv) The estimated slope ($\beta 1$) is negative in this case, -3.50 which means there is a negative relation between the response and the predictors.

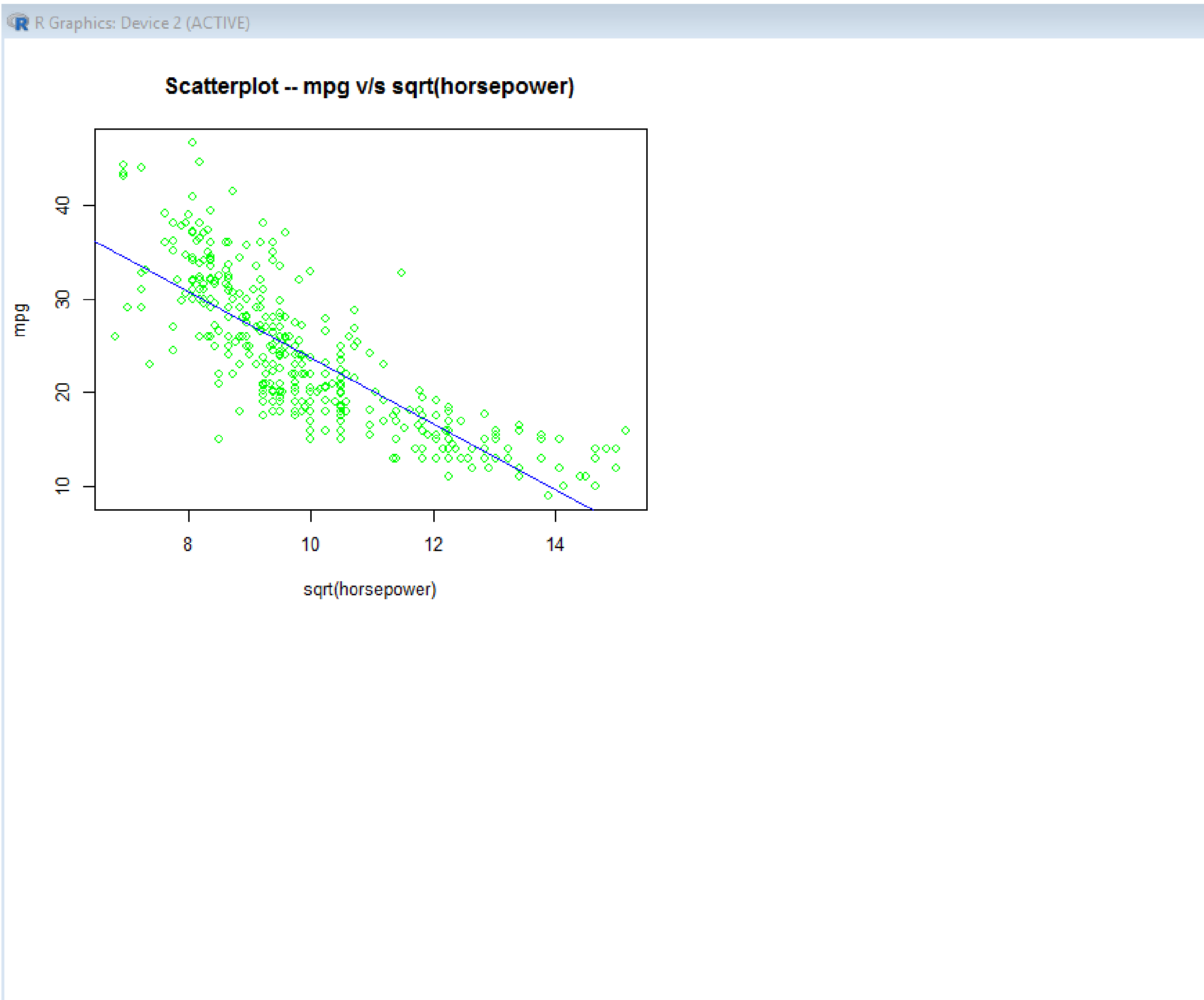
(v)

```
> x <- sqrt(horsepower)
> fit<- lm(mpg ~ x, data = Auto)
> predict(fit, data.frame(x = 98), interval = "predict")
        fit       lwr       upr
1 -284.6402 -309.2378 -260.0425
> predict(fit, data.frame(x = 98), interval = "confidence")
        fit       lwr       upr
1 -284.6402 -307.4641 -261.8163
> 
```

(b)

Apurva Shrivastava
925009508
ISEN-613 600

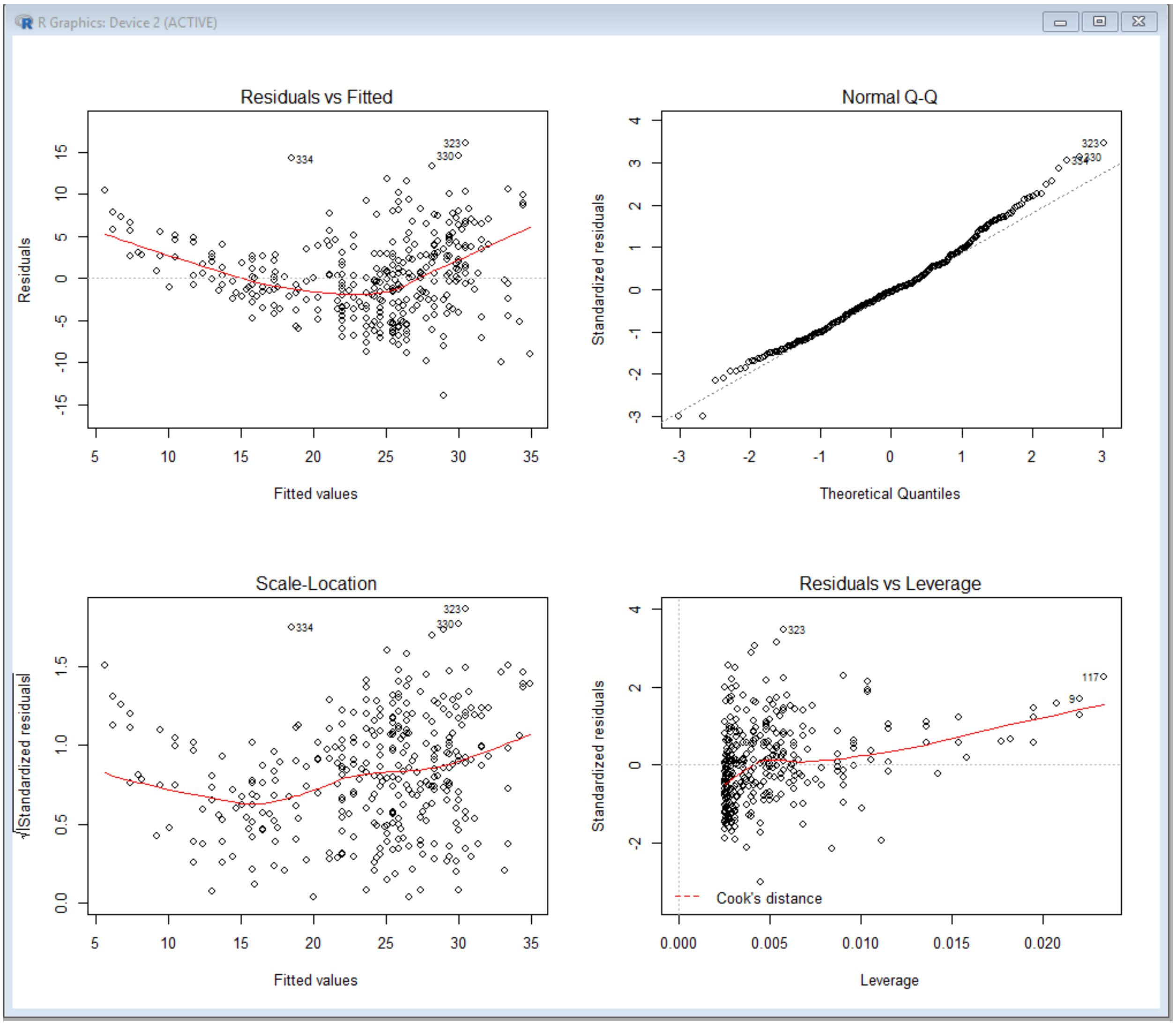
```
> x <- sqrt(Auto$horsepower)
> plot(x, Auto$mpg, main = "Scatterplot -- mpg v/s sqrt(horsepower)", xlab = "sqrt(horsepower)", ylab ="mpg", col = "green")
> abline(fit, col = "blue")
>
```



(c)

```
> par(mfrow = c(2,2))
> plot(fit)
>
```

Apurva Shrivastava
925009508
ISEN-613 600



Graph	Comment
Residual vs Fitted	This indicates the presence of non-linearity in the data
Normal Q-Q	This indicates that as it is a straight line the data is normally distributed
Scale-Location	It is skewed horizontal line which indicates not equally (randomly) spread points
Residuals vs Leverage	This indicates the presence of few outliers (higher than 2 and lower than -2) and a few high leverage points

Apurva Shrivastava
925009508
ISEN-613 600

Square(horsepower)

(a)

```
> fit <- lm(mpg ~ (horsepower)*(horsepower), data = Auto)
> summary(fit)

Call:
lm(formula = mpg ~ (horsepower) * (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

>
```

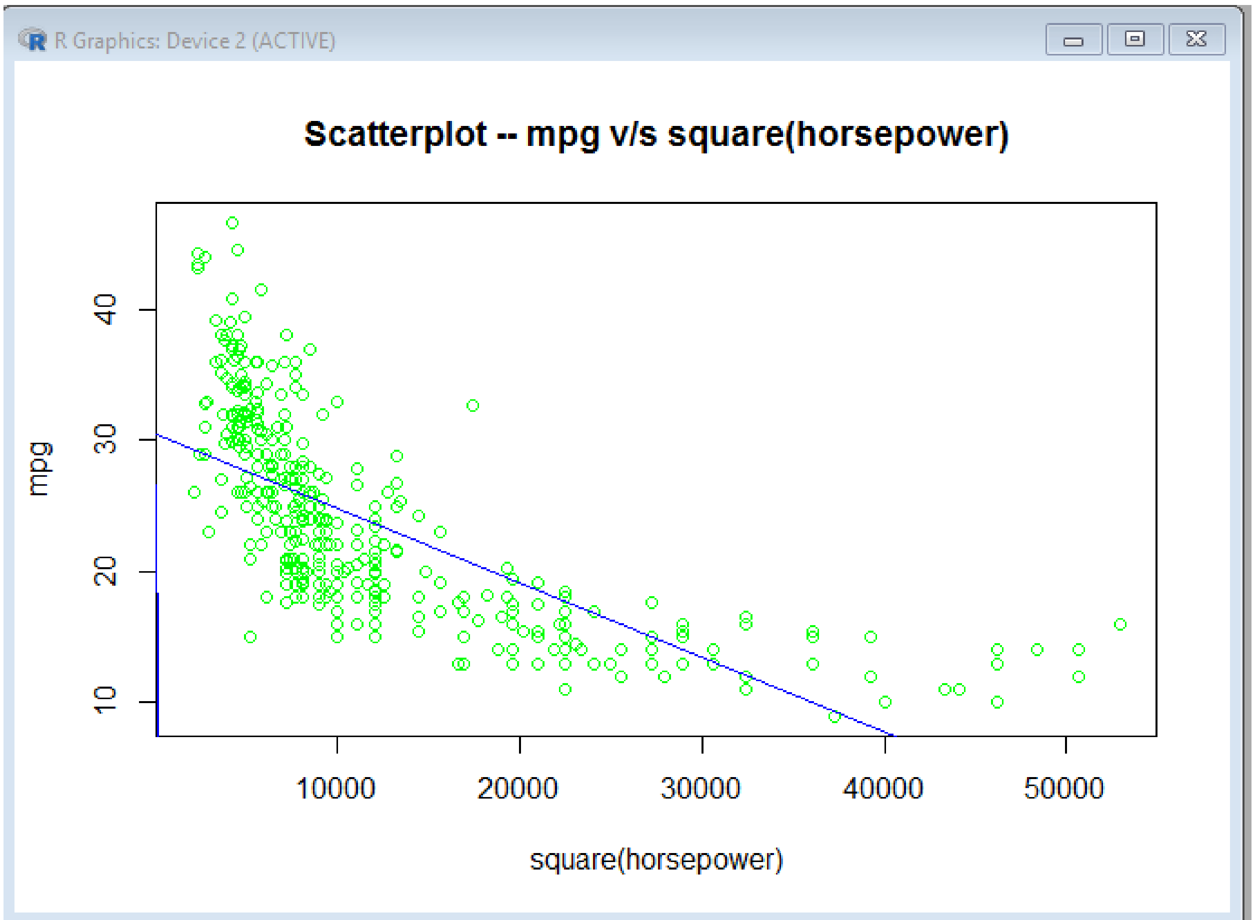
(i) Very small p-value indicates there exists a relationship between the response and the predictor
(ii) R square is .6059 which means 60.59% of variability in mpg can be explained. Hence, it is strongly fitting model
(iii) As the coefficient of estimation is negative, there is a negative relation between mpg and log(horsepower)
(iv) The estimated slope ($\beta1$) is negative in this case, -0.15 which means there is a negative relation between the response and the predictors.
(v)

```
> attach(Auto)
> x <- (horsepower)^2
> fit<- lm(mpg ~ (horsepower)^2, data = Auto)
> fit<- lm(mpg ~ x, data = Auto)
> predict(fit, data.frame(x = 98), interval = "confidence")
       fit     lwr      upr
1 30.41026 29.5365 31.28401
> predict(fit, data.frame(x = 98), interval = "predict")
       fit      lwr      upr
1 30.41026 19.59069 41.22982
>
```

Apurva Shrivastava
925009508
ISEN-613 600

(b)

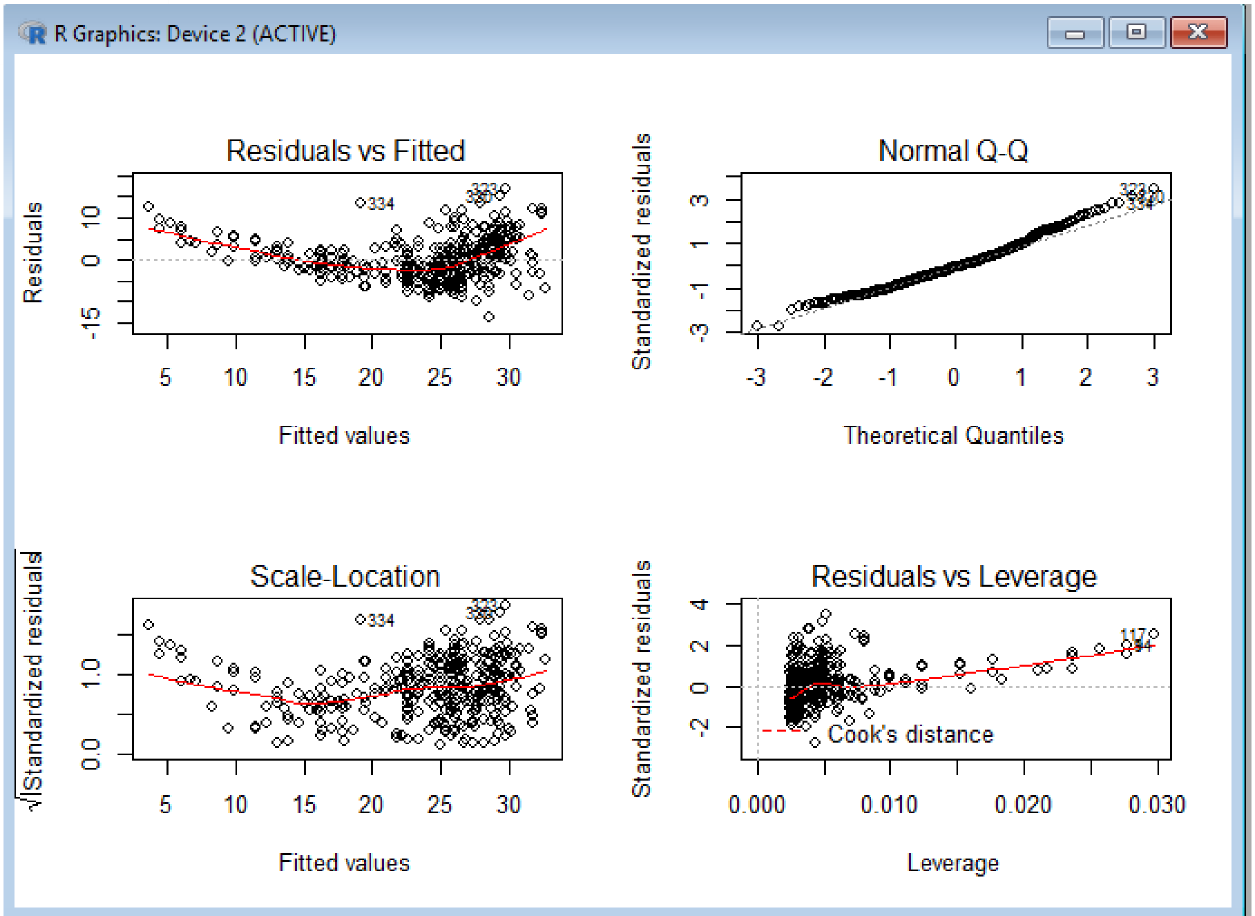
```
> x <- (Auto$horsepower)^2
> plot(x, Auto$mpg, main = "Scatterplot -- mpg v/s square(horsepower)", xlab = "square(horsepower)", ylab ="mpg", col = "green")
> abline(fit, col ="blue")
>
```



(c)

```
> par(mfrow = c(2,2))
> plot(fit)
>
```

Apurva Shrivastava
925009508
ISEN-613 600



Graph	Comment
Residual vs Fitted	This indicates the presence of non-linearity in the data
Normal Q-Q	This indicates that as it is a straight line the data is normally distributed
Scale-Location	It is almost horizontal line which indicates equally (randomly) spread points and verify the assumption of equal variance (homoscedasticity)
Residuals vs Leverage	This indicates the presence of few outliers (higher than 2 and lower than -2) and a few high leverage points

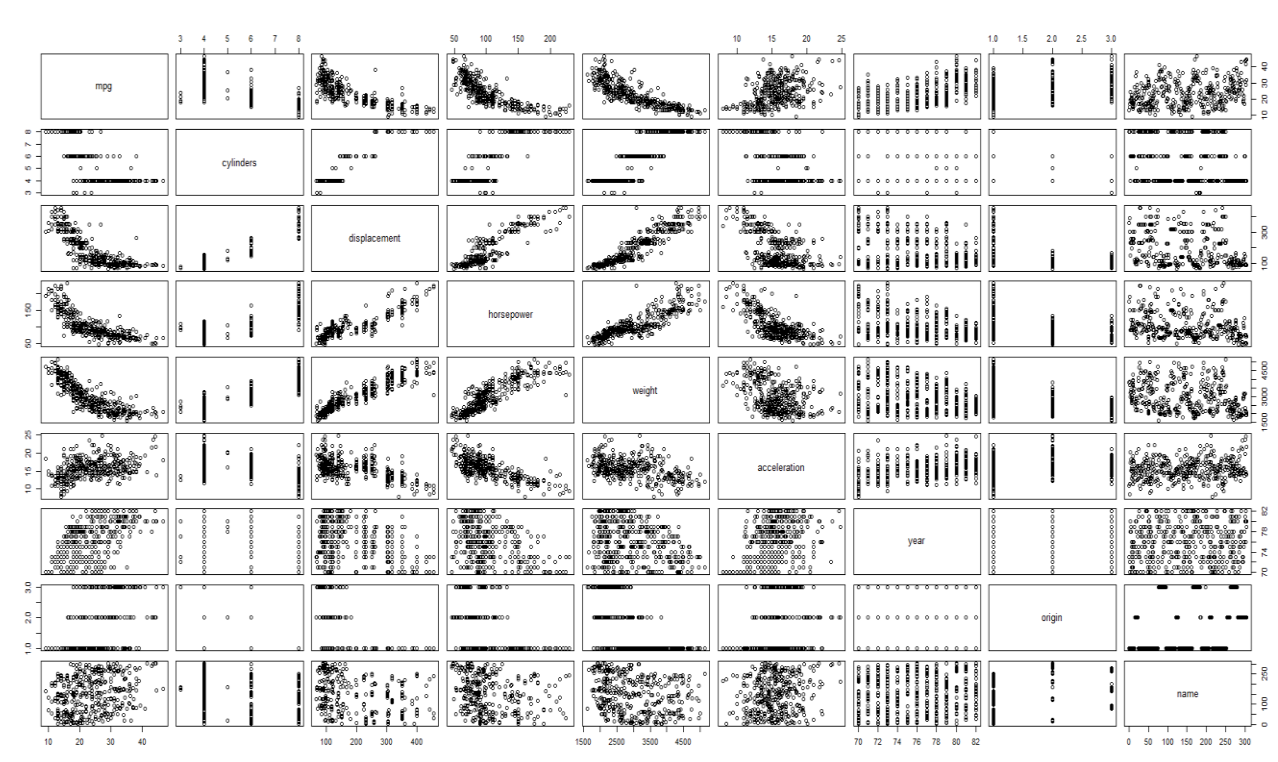
Apurva Shrivastava
925009508
ISEN-613 600

Problem 2

Use the Auto data set to answer the following questions:

(a) Produce a scatterplot matrix which includes all of the variables in the data set. Which predictors appear to have an association with the response?

```
>
> pairs(Auto)
>
```



Based on the scatterplot, the predictors which appear to have an association with the response (mpg) are:

Displacement	Slightly Linearly decreasing
Horsepower	Linearly decreasing but skewed
Weight	Linearly decreasing but skewed

Rest all the predictors are scattered and no definite association can be inferred.

Apurva Shrivastava
925009508
ISEN-613 600

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

```
> names(Auto)
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"       "acceleration" "year"         "origin"       "name"
> cor(Auto[1:8])
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
> 
```

(c) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Comment on the output. For example,

Apurva Shrivastava
925009508
ISEN-613 600

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

> |
```

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

If we look at the p- value of the F – statistic, which is, 2.2e-16, very small. It shows there is some relationship between mpg and other predictors.

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

We can answer this question by checking the p-values associated with each predictor's t-statistic. We may conclude that all predictors are statistically significant except "cylinders", "horsepower" and "acceleration" as their p-values are greater than 0.05 which is higher than 95% confidence interval.

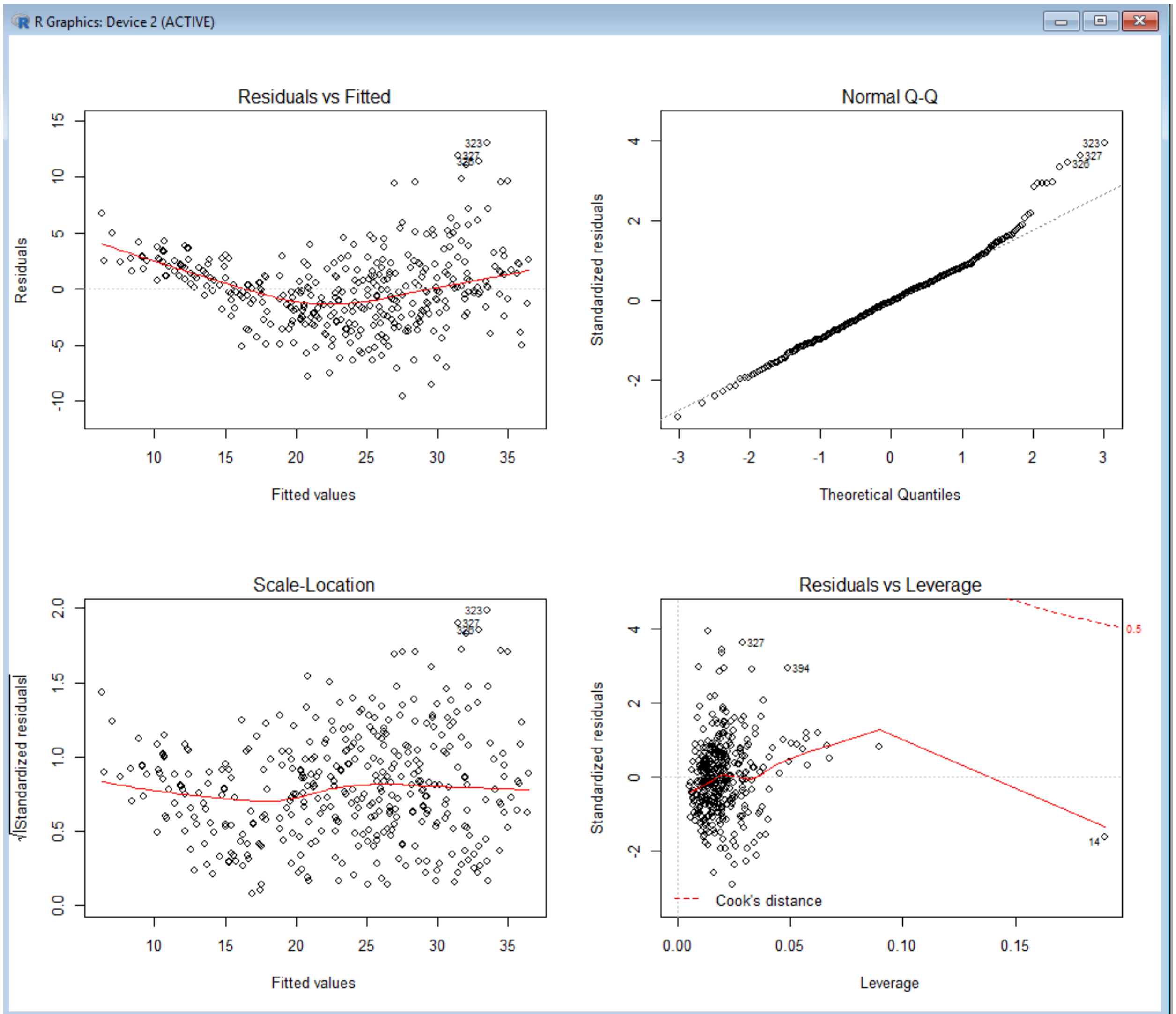
iii. What does the coefficient for the year variable suggest?

The coefficient of the year variable is positive, which indicates that the average effect of an increase of one year corresponds to an increase of 0.7507 in mpg, considering all other predictors being constant. Hence, we can say that cars become more fuel efficient every year by almost 1.4 mpg/year.

(d) Produce diagnostic plots of the linear regression fit. Comment on each plot.

Apurva Shrivastava
925009508
ISEN-613 600

```
>
> par(mfrow = c(2,2))
> plot(fit)
>
```



Graph	Comment
Residual vs Fitted	This indicates the presence of mild non-linearity in the data
Normal Q-Q	This indicates that as it is a straight line the data is normally distributed and right skewed
Scale-Location	It is almost horizontal line which indicates equally (randomly) spread points and verify the assumption of equal variance (homoscedasticity)
Residuals vs Leverage	This indicates the presence of few outliers (higher than 2 and lower than -2) and one high leverage point (point 14)

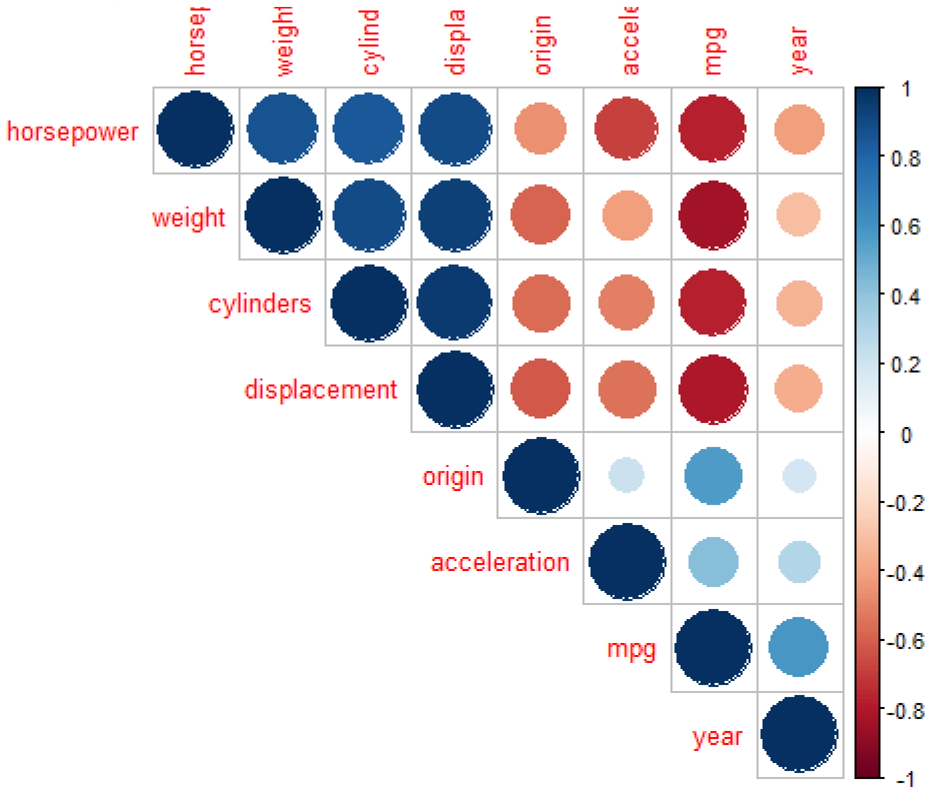
Apurva Shrivastava
925009508
ISEN-613 600

(e) Is there serious collinearity problem in the model? Which predictors are collinear?

In regression, "collinearity" refers to predictors that are correlated with other predictors. Collinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. Hence, finding the correlation amongst predictors:

```
> library(corrplot)
> data("Auto")
> my_data <- Auto[, c(1,2,3,4,5,6,7,8)]
> res <- cor(my_data)
> corrplot(my_data, method ="circle")
Error in matrix(unlist(value, recursive = FALSE, use.names = FALSE), nrow = nr,  :
  length of 'dimnames' [2] not equal to array extent
>
>
> library(corrplot)
> data("Auto")
> M <- cor(Auto)
Error in cor(Auto) : 'x' must be numeric
> res <- cor(my_data)
> round(res,2)
               mpg cylinders displacement horsepower weight acceleration  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
displacement -0.81      0.95         1.00       0.90   0.93        -0.54 -0.37  -0.61
horsepower   -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
acceleration  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

> corrplot(res, type ="upper", order = "hclust")
>
```

Apurva Shrivastava
925009508
ISEN-613 600



Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. In the right side of the correlogram, the legend color shows the correlation coefficients and the corresponding colors.

To identify which predictor has significant collinearity, we find the VIF for all the predictors:

```
> library(car)

Attaching package: 'car'

The following object is masked from 'package:VIF':

    vif

> attach(Auto)
The following objects are masked from Auto (pos = 4):

    acceleration, cylinders, displacement, horsepower, mpg, name, origin, weight, year

The following objects are masked from Auto (pos = 5):

    acceleration, cylinders, displacement, horsepower, mpg, name, origin, weight, year

> fit <- lm(mpg ~ . - name , data = Auto)
> vif(lm.fit)
Error: object of type 'closure' is not subsettable
> vif(fit)
   cylinders displacement   horsepower       weight acceleration         year       origin
   10.737535    21.836792     9.943693    10.831260     2.625806     1.244952     1.772386
> |
```

Apurva Shrivastava
925009508
ISEN-613 600

As the VIF for Cylinders, displacement, horsepower and weight are more than 5, these predictors are highly collinear.

(f) Fit linear regression models with interactions. Are any interactions statistically significant?

```
> fit3 <- lm(mpg ~ cylinders * displacement+displacement * weight, data = Auto[, 1:8])
There were 50 or more warnings (use warnings() to see the first 50)
> summary(fit3)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight, data = Auto[, 1:8])

Residuals:
     Min       1Q   Median       3Q      Max
-13.2934  -2.5184  -0.3476   1.8399  17.7723

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders               7.606e-01  7.669e-01   0.992    0.322
displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,    Adjusted R-squared:  0.7237
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

From the p-values, we can see that the interaction between displacement and weight is statistically significant, while the interaction between cylinders and displacement is not.

Apurva Shrivastava
925009508
ISEN-613 600

Problem 3
Use the Carseats data set to answer the following questions:
(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
> data(Carseats)
> fit <- lm(Sales ~ Price + Urban + US, data = Carseats)
> summary(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) Provide an interpretation of each coefficient in the model (note: some of the variables are qualitative).

The coefficient of the US variable: This may be interpreted by saying that on average the unit sales in a US store are 1200.57 units more than in a non-US store all other predictor remaining fixed.

The coefficient of the Price variable: This may be interpreted by saying that the average effect of a price increase of 1 dollar is a decrease of 54.45 units in sales all other predictors remaining fixed.

The coefficient of the Urban variable: This may be interpreted by saying that on average the unit sales in urban location are 21.91 units less than in rural location all other predictors remaining fixed.

(c) Write out the model in equation form.

$$Sales=13.0434689+(-0.0544) \times Price+(-0.0219) \times Urban+(1.2005) \times US+ \varepsilon$$

Urban =1 if the store is in an urban location and 0 if not,
and US=1 if the store is in the US and 0 if not.

Apurva Shrivastava
925009508
ISEN-613 600

(d) For which of the predictors can you reject the null hypothesis $H0$: βj = 0 ?

We can reject the null Hypothesis for Price and USYes predictors as their p-values are significantly low.

(e) On the basis of your answer to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the response.

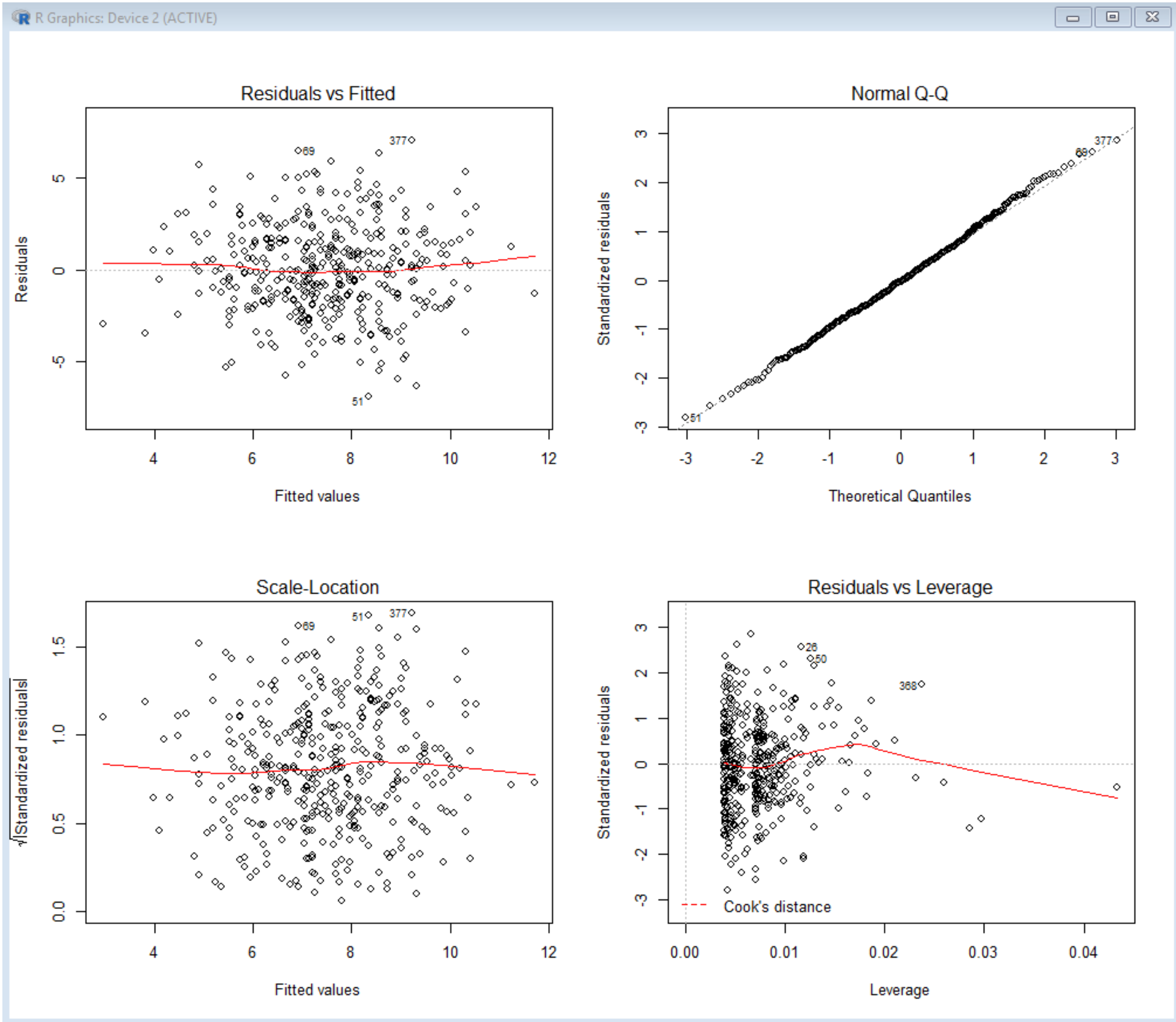
```
> fit1 <- lm(Sales ~ Price + US, data = Carseats)
> summary(fit1)

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) How well do the models in (a) and (e) fit the data?
The F statistic value for the smaller model is higher than the previous one leading to small p values indicating that the predictors are sufficient for the response. However, the p-values are very small for both models indicating definite relationship between response and predictors. Also, both models have R square value of 0.2393 which means 23.93% of variability is explained by these models. This could be interpreted as a slightly loosely fitting model.

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

```
> par(mfrow = c(2,2))
> plot(fit1)
> 
```

Apurva Shrivastava
925009508
ISEN-613 600



The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and some leverage points.