**Assignment #2: Linear Regression**

**Submit through link: eCampus -> Assignments->Assignment 2 Submission**

**Deadline: September 20 (Friday) @11:59pm**

**# YOUR NAME:** Joel Dsouza

**#UIN :** 327003425

**# RULES:**

**#**

**# 1) You can discuss the homework with each other in general terms,**

**# but you must write your own solutions and not copy from anyone.**

**#**

**# 2) Edit your answers into this file following each problem. Write the answer Problem 1(b)**

**# right after Problem 1 part (b).**

**#**

**# 3) For problems 1&2 you do not need to use R.**

**#**

**# 4) Post your solutions to e-campus**

**#**

**# The filename should have this format: LastName-FirstName-hw02.doc**

**# An example would be: Eksin-Ceyhun-hw02.doc**

**Problem 1 (1pt)**

Assume that the mean height for women at a large university (to be viewed as a population) is 65 inches with a standard deviation of 3 inches.

1. If the women are placed randomly into classes of 36 each, what will be the standard deviation of the class means for height (i.e. the standard error of the mean)?

Standard deviation: 3 inches

Sample size = N = 36

**Standard error** =  3/sqrt(36) = 3/6 = **0.5**

1. Using your answer to part (a), what is the z score for a class whose average height is 64.2 inches? What is the two-tailed p-value for this class?

1) Z score = (sample mean – population mean) / standard error

**Z score** = 64.2-65/0.5 = **-1.6**

2) P value for 2 tailed test is ( 1- p value) \* 2

Two-tailed p-value : (1-0.9452)\*2 = 0.1096

1. If you were testing the null hypothesis (i.e. μ=65), would you reject H0 for this class at the 0.05 level for a two-tailed test?

H0:μ = 65

H1: μ <> 65

Cutoff value = =  0.05

Since it is a two-tailed test /2 = 0.025

z alpha = z table(1- 0.025)  =  z table(0.975) = 1.96

Rejection region =  z score = 1.96

**Z** =  64.2-65/0.5 = **-1.6**

p value from the chart is 0.0548

For a two tailed test, p value is 0.1096

**Zalpha** is **1.0398** from the chart

**1.6 < 1.96  the null hypothesis cannot be rejected for alpha=0.05 for a two tailed test**

1. Repeat part (b) for a class whose mean height is 67.4 inches. Would you reject the null hypothesis for this class with a two-tailed test at the 0.05 level? At the 0.01 level?

**Z** = 67.4-65/0.5 = 2.4/0.5 = **4.8**

**P value** is **0.0001** from the Z-table

**For both 0.05 and 0.01 the p value 0.0001 is much smaller. Therefore we reject the null hypothesis**

**Problem 2 (1pt)**

This exercise relies on the following data set: 1, 3, 6, 0, 1, 1, 2, 1, 4.

1. Perform a t test in order to decide whether you can reject the null hypothesis of μ=2.5 at the 0.05 level (two-tailed) for these data.

H0:μ = 2.5

H1: μ <> 2.5

sample mean = 2.1111

sample standard deviation = 1.90029

standard error = 0.63343

**T** =  (2.11-2.5)/0.633 = **-0.61611**

For two-tailed test /2 = 0.025

Critical value from the t table = **talpha**= **2.306**

**Since 0.616<2.306, we cannot reject the null hypothesis**

1. Redo your t-test in part (a) for a null hypothesis of μ=6.0

**T** = (sample mean – hypothesized mean value)/standard error

=  (2.11-6)/0.633 = **-6.145**

Degree of freedom = n -1 = 9 -1 = 8

Significance level = =  0.05

For two-tailed test /2 = 0.025

Critical value from the t table = **talpha= 2.306**

**Since 6.145>2.306, we reject the null hypothesis**

1. Compute the 95% confidence interval (CI) for the population mean form which these data were drawn. Explain how this CI could be used to draw conclusions about the null hypothesis in parts (a) and (b).

**95% confidence interval (CI)** for the population mean = (sample mean – tn-1, ⍺/2\* standard error, sample mean + tn-1, ⍺/2\* standard error)

= (sample mean – t8, 0.025\* standard error, sample mean + t8, 0.025\* standard error)

= (2.11 – (2.306 \* 0.633), 2.11 + (2.306 \* 0.633))

= (2.11 - 1.45, 2.11 + 1.45)

= **(0.650, 3.57)**

Since μ=2.5 falls in the confidence interval, we have t accept the hypothesis. However, for μ=6.0 we will **reject the hypothesis** because it does not fall into the range.

**Problem 3 (2pt)**

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.

library(ISLR)

head(Auto)

fit=lm(formula=mpg~horsepower, data=Auto)

Call:

lm(formula = mpg ~ horsepower, data = Auto)

Coefficients:

(Intercept) horsepower

39.9359 -0.1578

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

Comment on the output. For example

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

Yes there is a linear relationship between the predictor and response

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

The t-value of the predictor is large and the p value is very small and clearly indicates that the null hypothesis can be rejected and a strong relationship is established

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

Since the slope of the regressed line is negative, the relationship is inversely proportional

iv. **How to interpret the estimate of the slope?**

From the summary function, we can find the slope to be -0.1578 that is negative

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

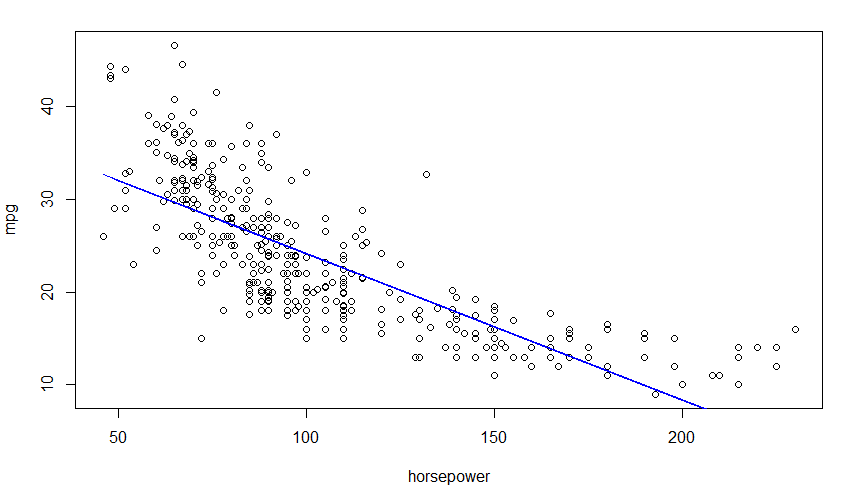
1 24.46708 23.97308 24.96108

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

fitted(fit)

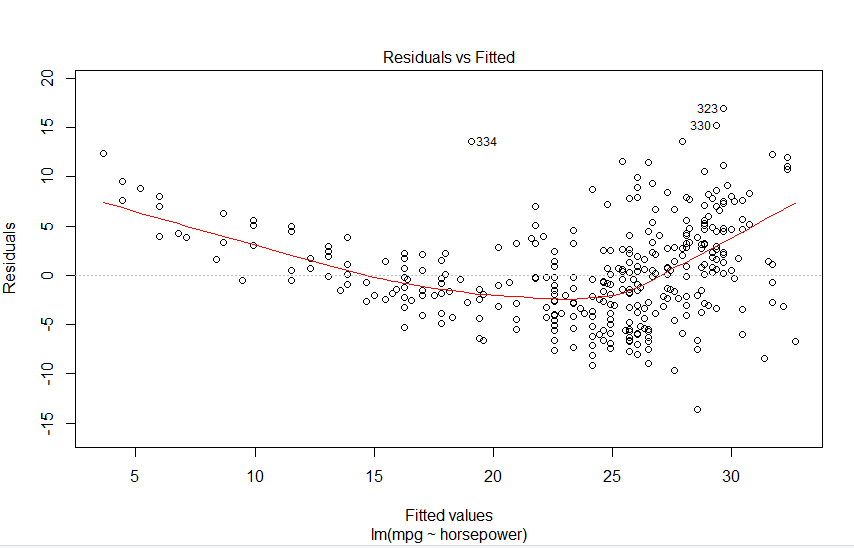
plot(mpg~horsepower,Auto)

lines(Auto$horsepower,predict(fit),col="blue")

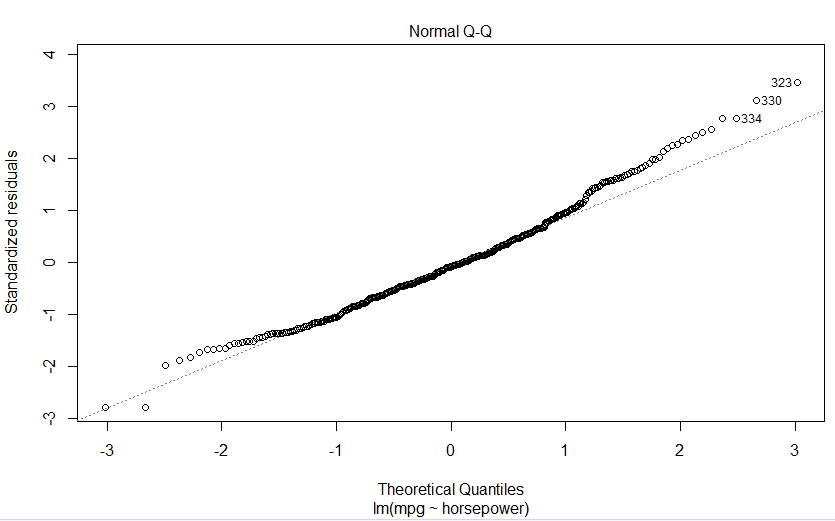


(c) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.

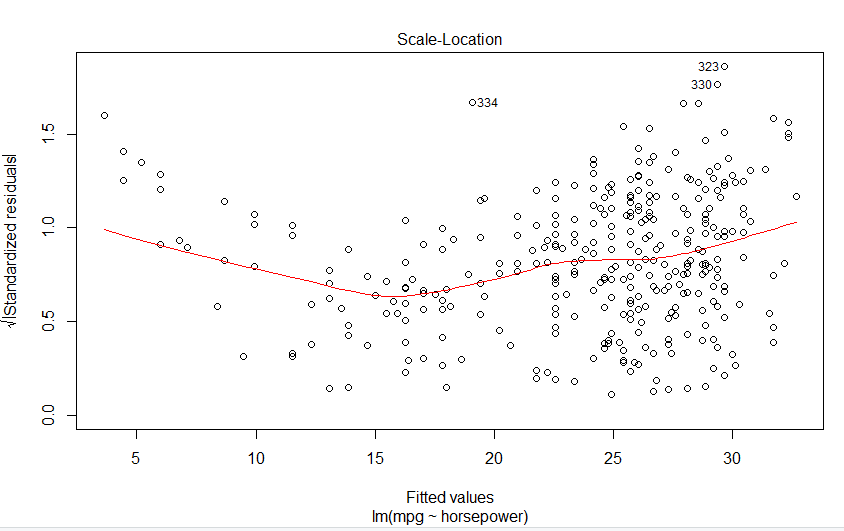
The residual vs fitted plot shows that the data is slightly parabolic with a pattern of non linearity which was not explained by the model. However, this non linearity is very small and the model mostly gives us a linear relationship between the response and predictor.



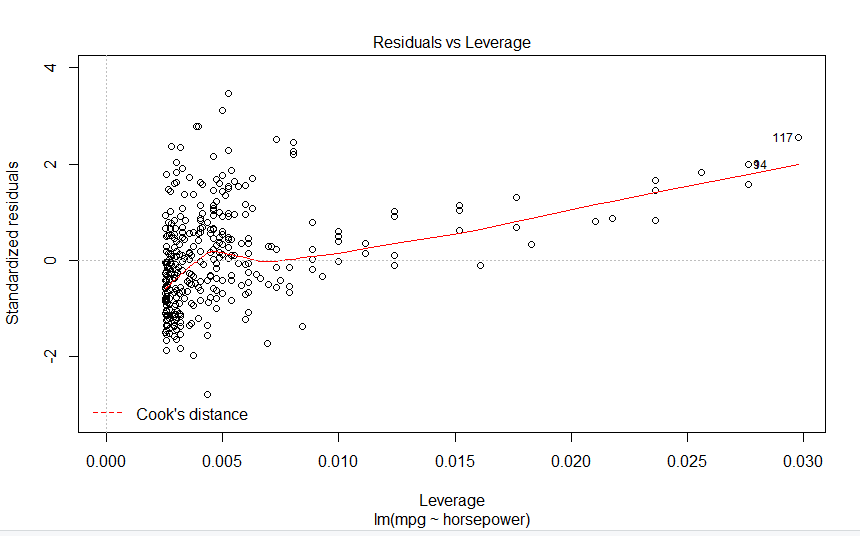
The Normal Q-Q plot shows 3 points that slightly deviate from the dotted line that may be problematic but mostly all points are along the line.



The points are equally spread except 3 points in the chart below. These points define the slight non linear behavior of our regression but it can be ignored.



Since we are not able to view the cooks distance lines, we can conclude that the relationship is fairly linear and can be modeled with a linear regression well.



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

# Log(X)

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

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

Call:

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

Coefficients:

(Intercept) log(horsepower)

108.70 -18.58

> summary(fitlog)

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

There is a relation between the log of horsepower and mpg based on the high t-value and low p-value

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

T >> 1 and p << 0.001 which shows that the relation is very strong

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

The relation is negative because of the slope is negative

iv. How to interpret the estimate of the slope?

The slope is interpreted from the summary data to be -18.5822

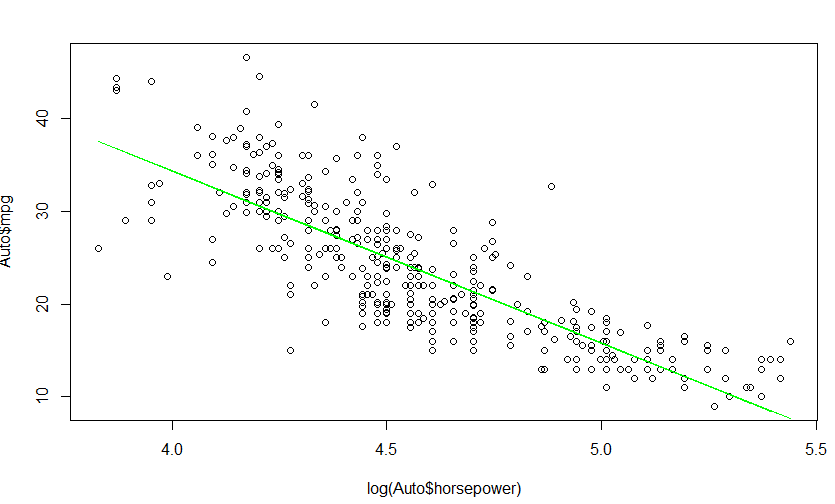
v. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

> predict(fitlog,data.frame(horsepower=98),interval="confidence")

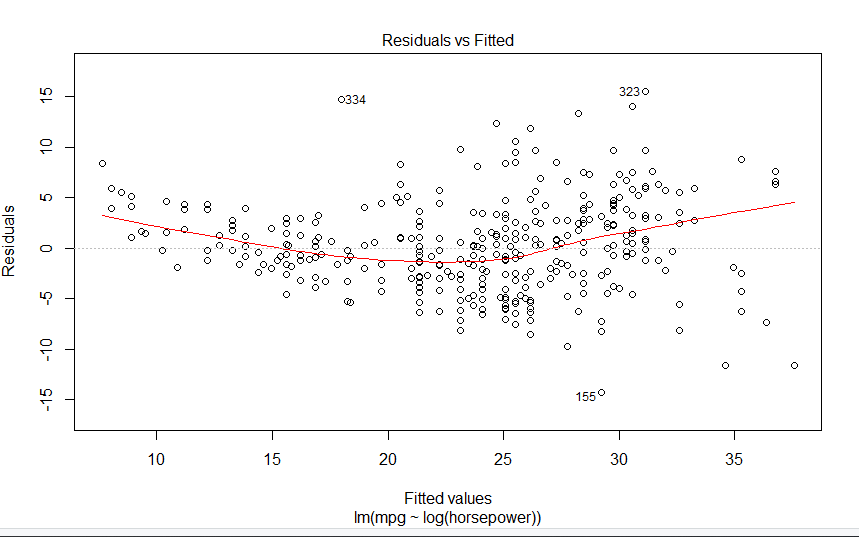
fit lwr upr

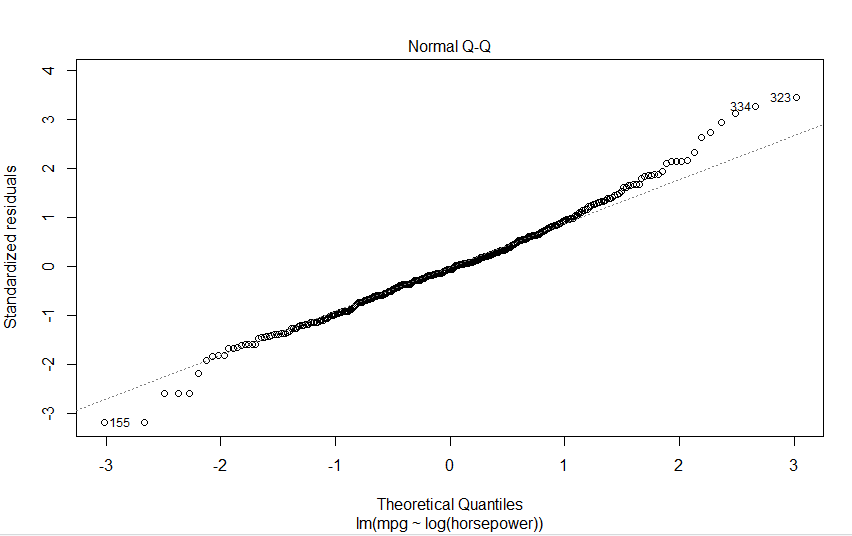
1 23.50099 23.05405 23.94794

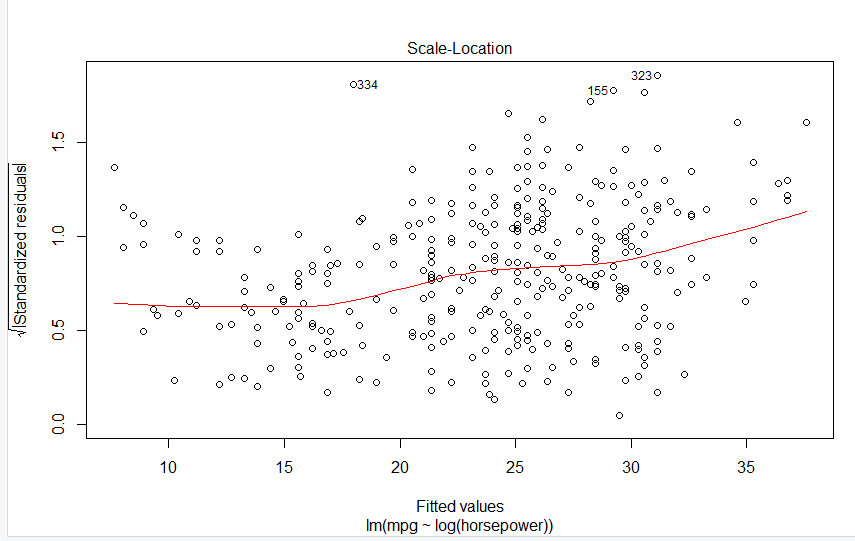
(b) Plot the response and the predictor. Display the least squares regression line in the plot.

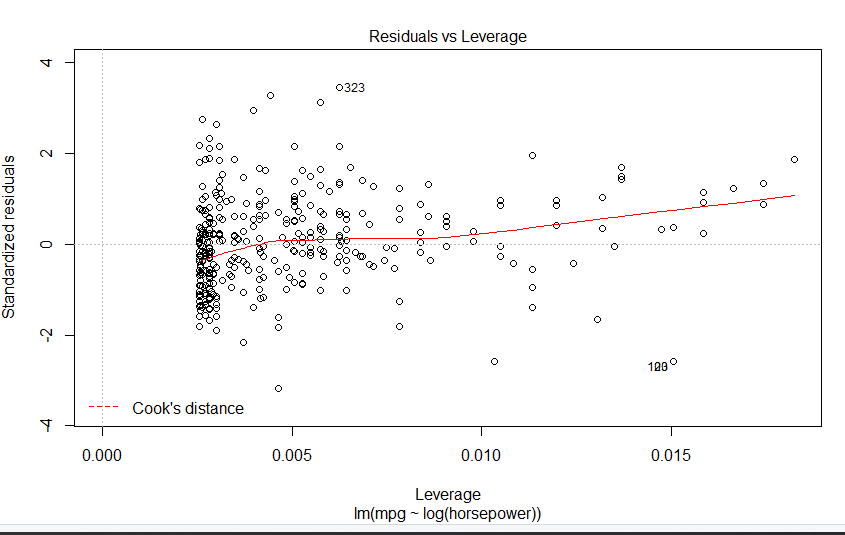


(c) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.









**As seen from the above plots, the log(x) acts similar to the predictor without log transformation but has a more linear behavior as compared to just x because the same sample points are more densely concentrated.**

# Sqrt(X)

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

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

Call:

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

Coefficients:

(Intercept) sqrt(horsepower)

58.705 -3.504

summary(fitroot)

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

There is a relationship between response and predictor because of the t and p value

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

T >>1 and p << 0.001 shows that the relationship is strong

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

Response is negative because of negative slope

iv. How to interpret the estimate of the slope?

We see the slope is -3.2239 from the summary and fit information

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

> predict(fitroot,data.frame(horsepower=98),interval="confidence")

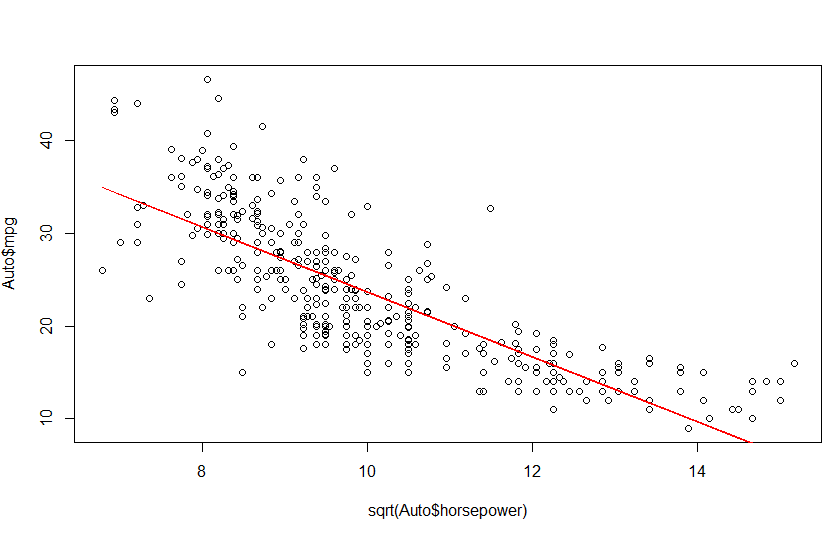
fit lwr upr

1 24.02206 23.55687 24.48724

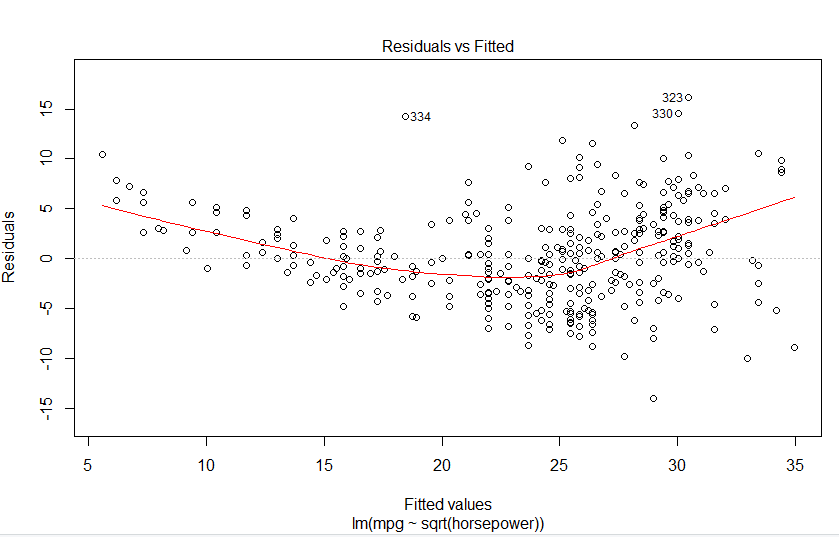
(b) Plot the response and the predictor. Display the least squares regression line in the plot.

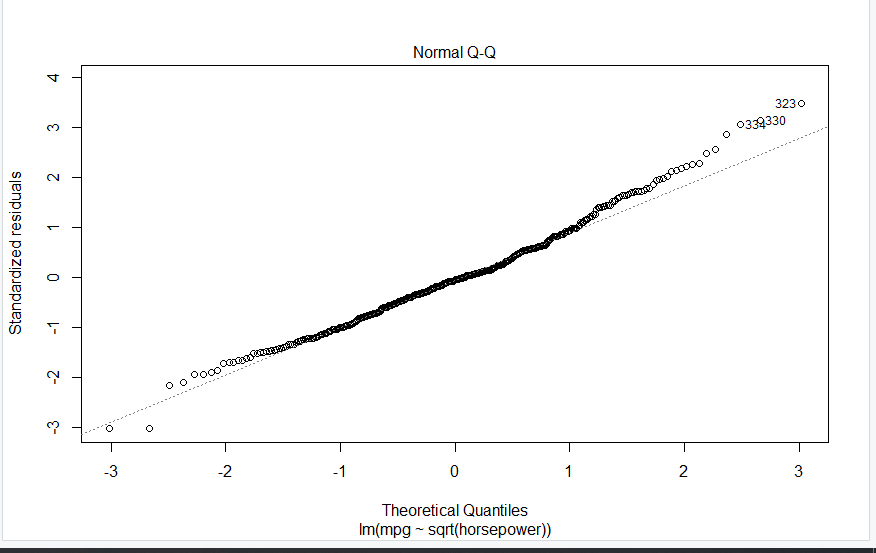
> plot(Auto$mpg~sqrt(Auto$horsepower))

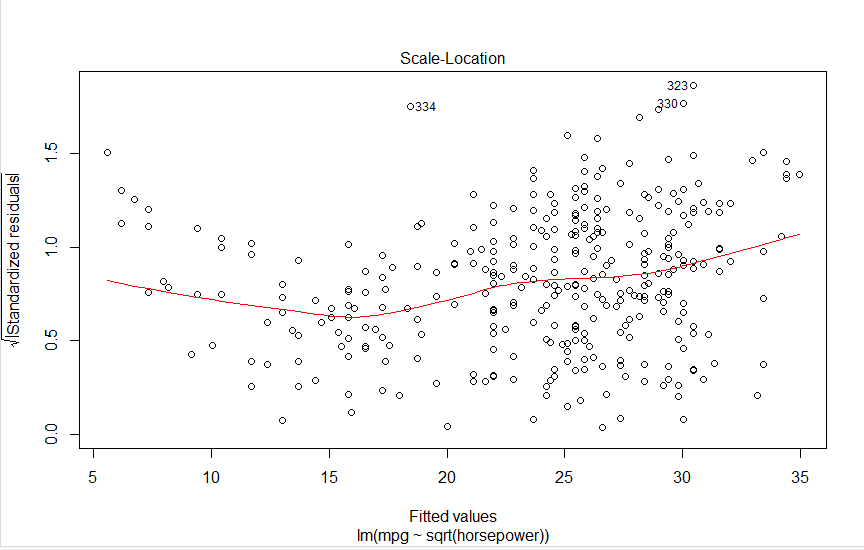
> lines(sqrt(Auto$horsepower),predict(fitroot),col="red")

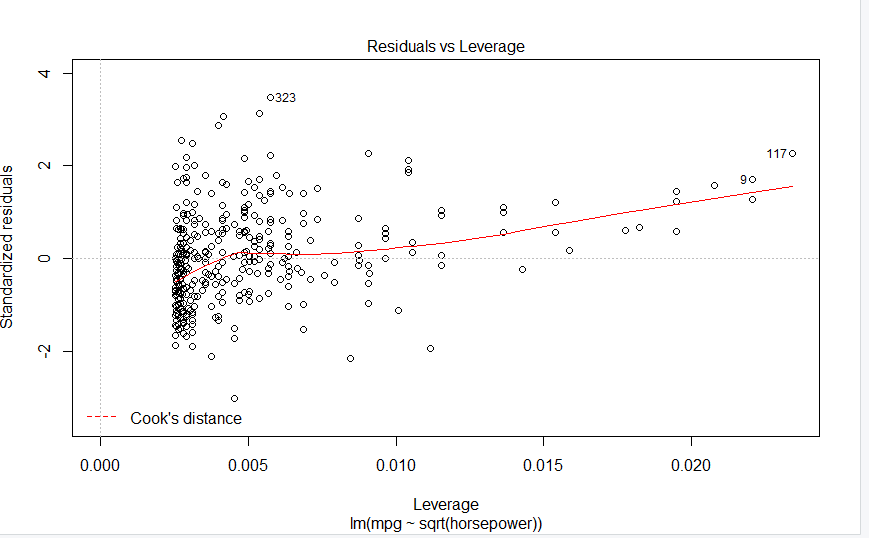


(c) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.









**The above plots show that the root x is slightly more linear than the log x and just x since it typically normalizes the values closer to each other.**

# X^2

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

fitsq=lm(formula=mpg~sapply(horsepower, function(x) x^2), data=Auto)

lm(formula = mpg ~ sapply(horsepower, function(x) x^2), data = Auto)

Coefficients:

(Intercept) sapply(horsepower, function(x) x^2)

30.4657729 -0.0005665

summary(fitsq)

Call:

lm(formula = mpg ~ sapply(horsepower, function(x) x^2), data = Auto)

Residuals:

Min 1Q Median 3Q Max

-12.529 -3.798 -1.049 3.240 18.528

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.047e+01 4.466e-01 68.22 <2e-16 \*\*\*

sapply(horsepower, function(x) x^2) -5.665e-04 2.827e-05 -20.04 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.485 on 390 degrees of freedom

Multiple R-squared: 0.5074, Adjusted R-squared: 0.5061

F-statistic: 401.7 on 1 and 390 DF, p-value: < 2.2e-16

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

There is a clear relationship between predictor and response because of t and p value

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

t>>1 and p<<1 so we see the relationship is strong

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

Slope is negative so the relationship is inverse

iv. How to interpret the estimate of the slope?

The slope is very low and we can see this information from the summary

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

> predict(fitsq,data.frame(horsepower=98),interval="confidence")

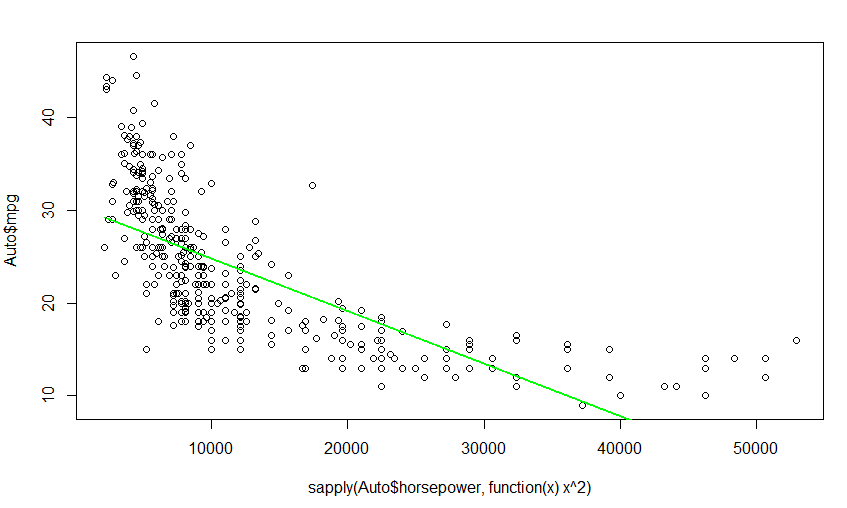
fit lwr upr

1 25.02512 24.45883 25.5914

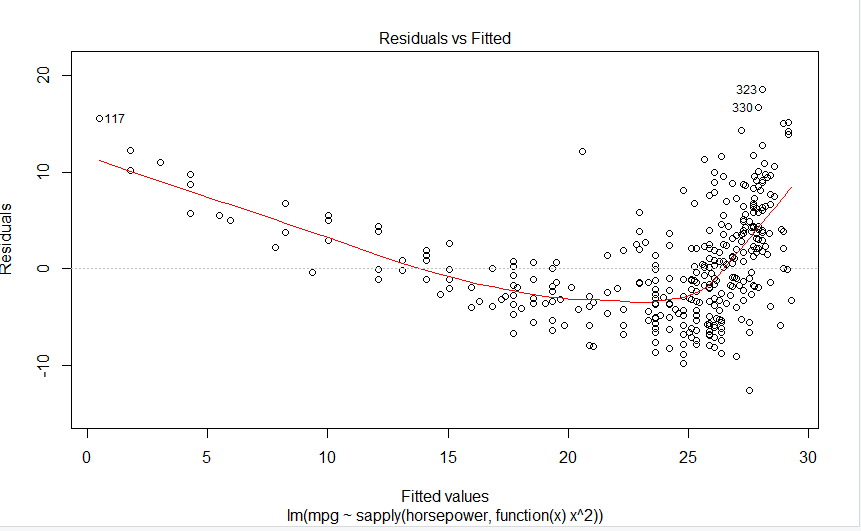
(b) Plot the response and the predictor. Display the least squares regression line in the plot.

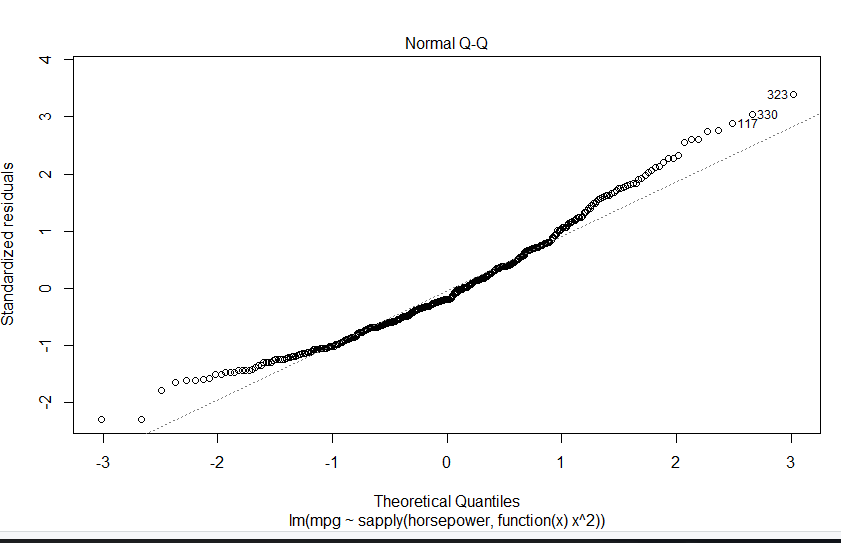
> plot(Auto$mpg~sapply(Auto$horsepower,function(x) x^2))

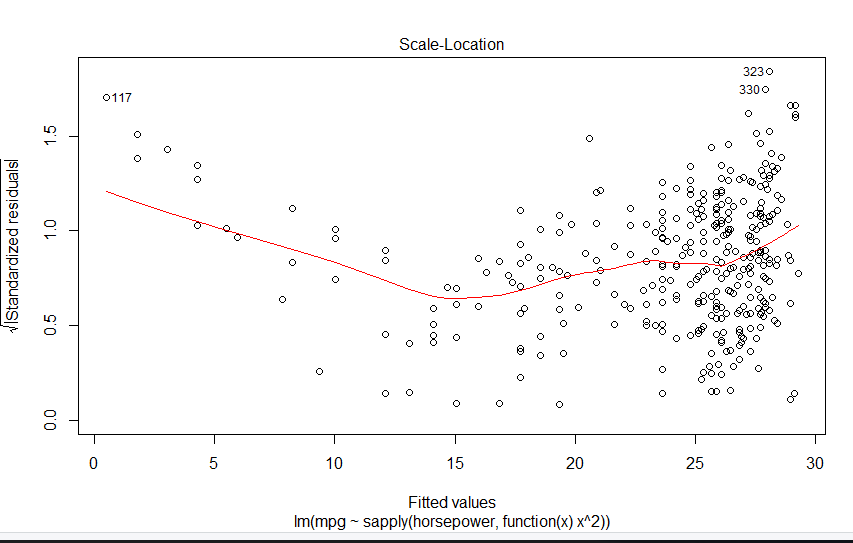
> lines(sapply(Auto$horsepower,function(x) x^2),predict(fitsq),col="green")

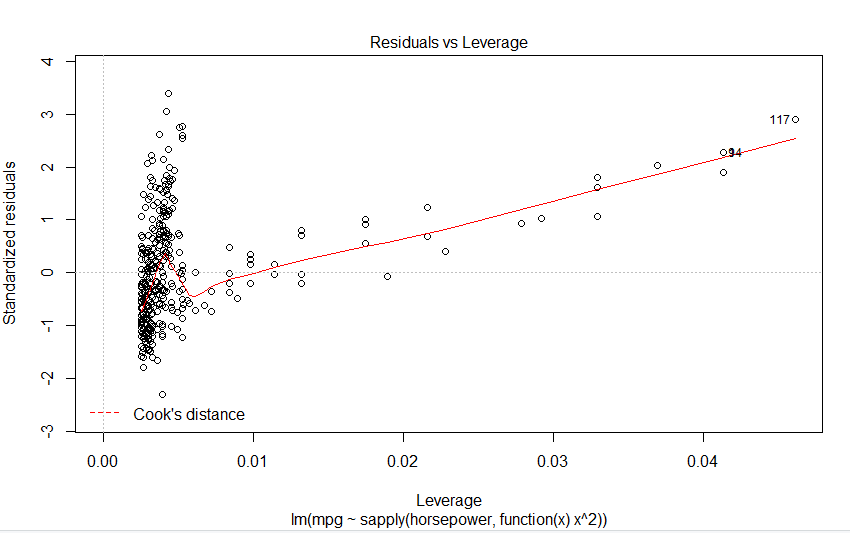


(c) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.









**The above 4 plots show that squaring the predictor will result in more sparse data here and is resulting in an increased nonlinear relationship as compared to the other approaches.**