Homework 2

Richard Albright
ISYE6414
Spring 2020

Background

The Motor Trend Car Road Tests data set was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). Here, we will perform some analysis on this data set to understand better the effect of hp, wt and disp on mpg.

Data Description

The data consists of a data frame with 32 observations on the following 11 variables:

1. mpg: Miles/(US) gallon
2. cyl: Number of cylinders
3. disp: Displacement (cu.in.)
4. hp: Gross horsepower
5. drat: Rear axle ratio
6. wt: Weight (1000 lbs)
7. qsec: 1/4 mile time
8. vs: Engine (0 = V-shaped, 1 = straight)
9. am: Transmission (0 = automatic, 1 = manual)
10. gear: Number of forward gears
11. carb: Number of carburetors

We will focus on the the effect of hp, wt and disp on mpg.

Instructions on reading the data

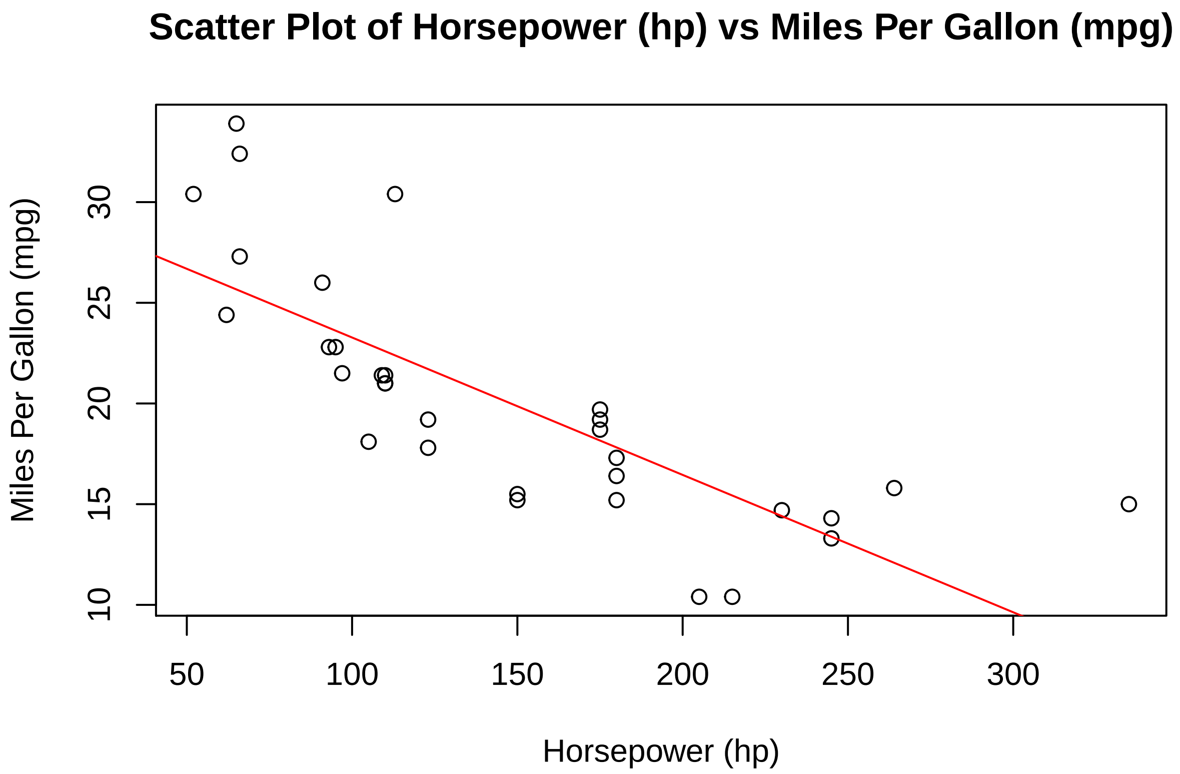
```
data(mtcars)
```

This loads the data set into your workspace.

Question 1: Exploratory Data Analysis - 12 points

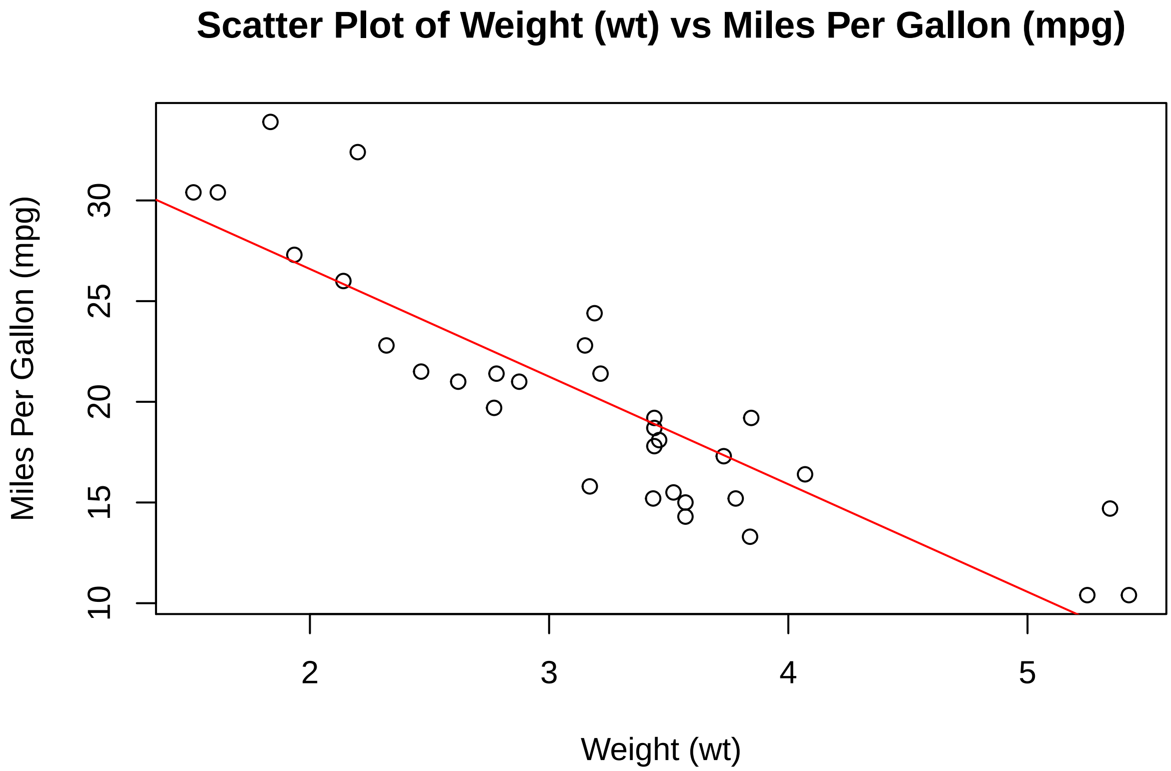
a. **3 pts** Plot the data (scatterplot) to observe and report the relationship between the response and each of the three predictors hp, wt and disp (there should be 3 plots reported). Comment on the general trend (direction and form).

```
plot(mtcars$hp,
     mtcars$mpg,
     xlab='Horsepower (hp)',
     ylab='Miles Per Gallon (mpg)',
     main='Scatter Plot of Horsepower (hp) vs Miles Per Gallon (mpg)')
abline(lm(mpg ~ hp, data=mtcars), col='red')
```



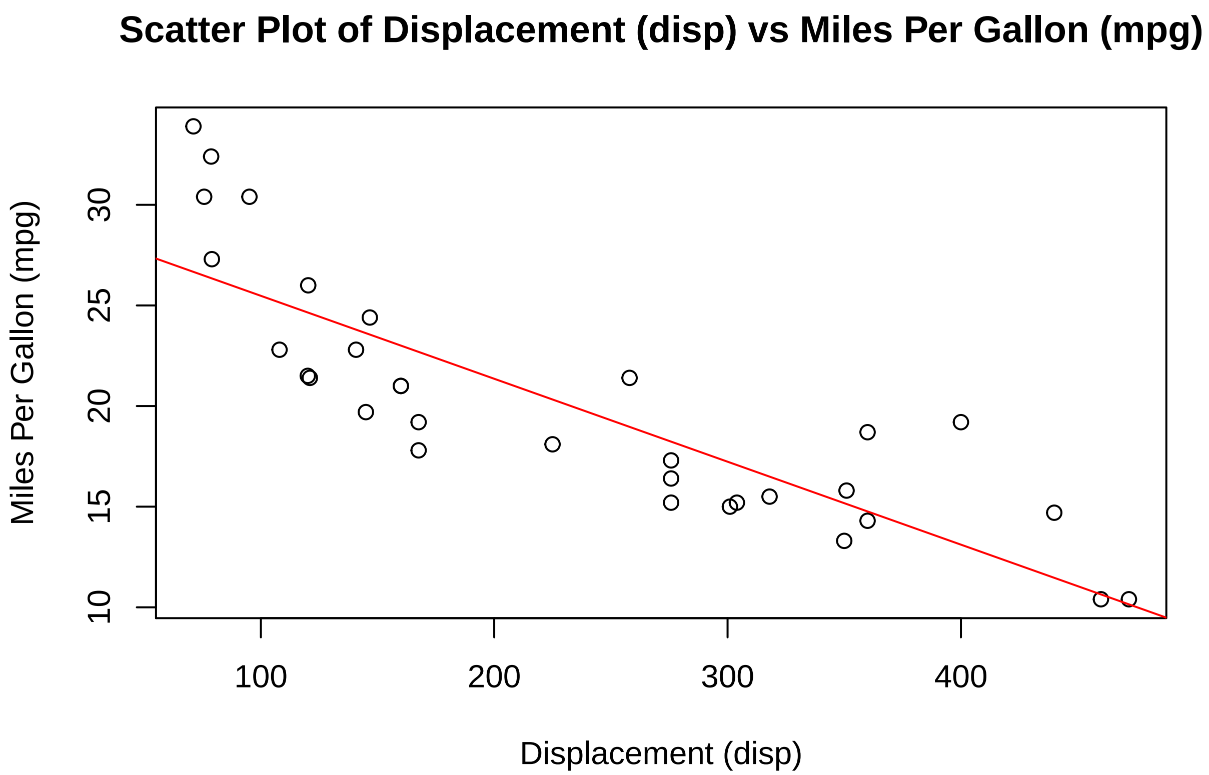
There is a moderate negative linear relationship between horsepower (hp) and miles per gallon (mpg). There appears to be outliers on both the low and high end of the horsepower (hp) range that are highly above the AB line.

```
plot(mtcars$wt,
     mtcars$mpg,
     xlab='Weight (wt)',
     ylab='Miles Per Gallon (mpg)',
     main='Scatter Plot of Weight (wt) vs Miles Per Gallon (mpg)')
abline(lm(mpg ~ wt, data=mtcars), col='red')
```



There is a strong negative linear relationship between weight (wt) and miles per gallon (mpg). There appears to be outliers on both the low and high end of the weight range that are highly above the abline.

```r
plot(mtcars$disp,
     mtcars$mpg,
     xlab='Displacement (disp)',
     ylab='Miles Per Gallon (mpg)',
     main='Scatter Plot of Displacement (disp) vs Miles Per Gallon (mpg)')
abline(lm(mpg ~ disp, data=mtcars), col='red')
```



There is a strong negative linear associaion between engine displacement (disp) and miles per gallon (mpg). There appears to be outliers in the low end of the displacement (disp) range that are above the AB line. There may be a tail on the low end of the distribution.

b. **3 pts** What is the value of the correlation coefficient for each of the above pair of response and predictor variables? What does it tell you about your comments in part (a).

Lets derive the correlations of the variables vs mpg.

```
variable <- c('hp', 'wt', 'disp')
correlation = c(cor(mtcars$hp,mtcars$mpg),
  cor(mtcars$wt,mtcars$mpg),
  cor(mtcars$disp,mtcars$mpg))
xt <- xtable(data.frame(variable, correlation), caption='Correlation of Variables vs MPG')
print(xt, caption.placement='top')
```

Table 1: Correlation of Variables vs MPG

	variable	correlation
1	hp	-0.78
2	wt	-0.87
3	disp	-0.85

The correlations of the predictor variables vs miles per gallon (mpg) verify the strength of the linear relationships as noted in the above scatterplot analysis.

c. **3 pts** Based on this exploratory analysis, is it reasonable to assume a multiple linear regression model for the relationship between mpg and all the predictor variables (hp, wt and disp)? Did you note anything unusual?

All the variables are highly correlated to miles per galon (mpg), I suspect that we have variables that are not independent and exhibit colinearity in our model. In order to assume a multiple linear regression model, the variables should be independent and not be highy correlated with each other. In order to reasonably assume the variables are not highly correlated, a variance inflation factor test should be performed on the model to ensure there is no multicolinearity.

```r
model <- lm(mpg ~ hp + wt + disp, data=mtcars)
xt <- xtable(data.frame(car::vif(model)), caption='Variance Inflation Factors of Model')
print(xt, caption.placement='top')
```

Table 2: Variance Inflation Factors of Model

	car..vif.model.
hp	2.74
wt	4.84
disp	7.32

Both miles per gallon (mpg) and weight (wt) having a variance inflation factor above 2.5 shows cause for concern that the variables are not independent. Also, using a variance inflation factor threshold of 5, we may want to exclude displacement (disp) altogether from our model.

d. **3 pts** Based on the analysis above, would you pursue a transformation of the data? Based on the analysis above I would consider a log transformation of the horsepower (hp) and weight (wt) variables.

Please work on non-transformed data for all of the following questions.

Question 2: Fitting the Multiple Linear Regression Model - 8 points

Build a multiple linear regression model using the response and all the three predictors and then answer the questions that follow:

```r
summ(model, confint = TRUE, digits = 6)
```

Observations	32
Dependent variable	mpg
Type	OLS linear regression

F(3,28)	44.565520
R^2	0.826836
Adj. R^2	0.808283

	Est.	2.5%	97.5%	t val.	p
(Intercept)	37.105505	32.781696	41.429314	17.578756	0.000000
hp	-0.031157	-0.054582	-0.007731	-2.724476	0.010971
wt	-3.800891	-5.984883	-1.616898	-3.564926	0.001331
disp	-0.000937	-0.022138	0.020263	-0.090535	0.928507

Standard errors: OLS

a. **4 pts** Report the coefficient of determination for the model and give a single line interpretation of this value.

The coefficient of determination (R^2) = 0.8268. 82.68% of the variance of miles per gallon (mpg) is explained by the total variation of the model inputs (horsepower (hp) + weight (wt) + displacement (disp)).

b. **4 pts** Is the model of any use in predicting mpg? Conduct a test of overall adequacy of the model, using $\alpha = 0.05$. Provide the following elements of the test: null hypothesis H_0, alternative hypothesis H_a, F- statistic or p-value, and conclusion.

H_0: $\beta = \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$

H_a: $\beta \neq 0$

F(3, 28) = 44.5655

p-value = 8.65e-11

The p value of 8.65e-11 $<$ the α of 0.05. We reject the null hypothesis that $\beta = 0$. The model is useful in predicting miles per gallon (mpg).

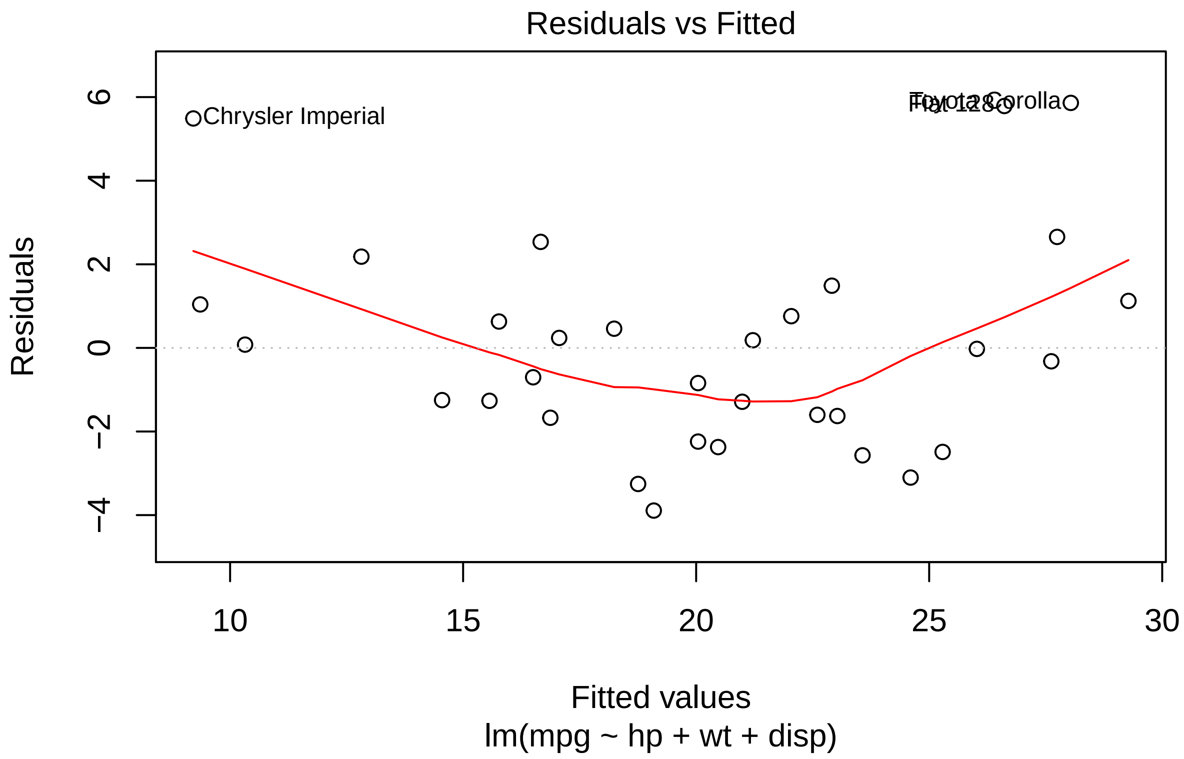
Question 3: Checking Assumptions of Model - 15 points

Provide plots to check for Linearity, Constant Variance and Normality assumptions of the model (use your knowledge from Homework 1 Peer Assessment). Provide your interpretations (i.e. whether the assumptions hold) for each plot.

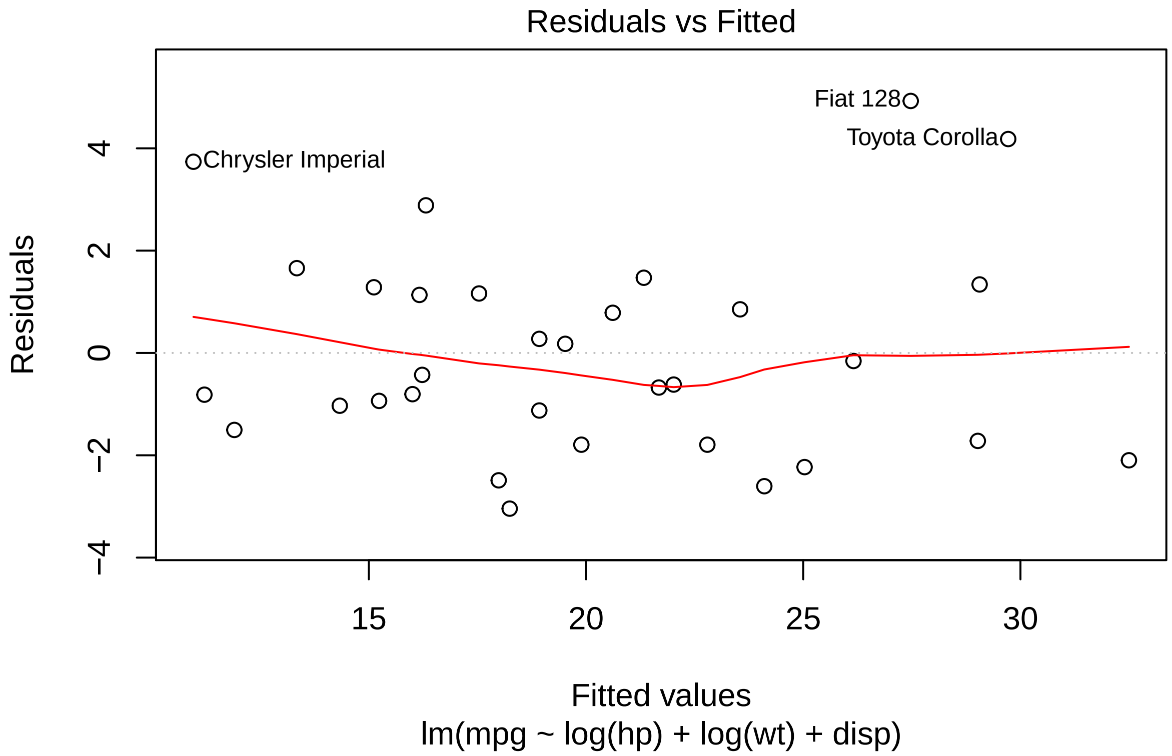
a. **5 pts** Linearity Assumption

Plot(s):

```r
plot(model, 1)
```



```r
logmodel <- lm(mpg ~ log(hp) + log(wt) + disp, data=mtcars)
plot(logmodel, 1)
```



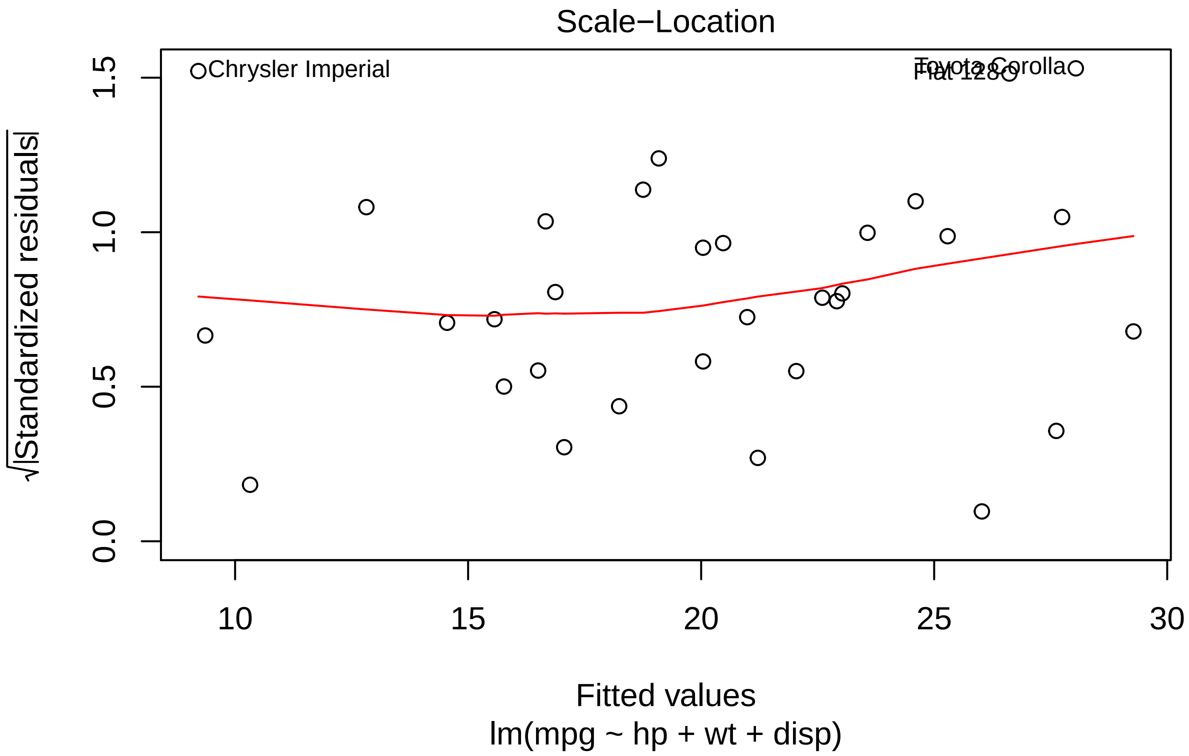
Interpretation:

A plot of the residuals vs the fitted values of the original model shows that the relationship is curvilinear. The assumption of linearity of the model does not hold. Doing a log transformation of horsepower(hp) and weight(wt) would transform the model into a linear relationship.

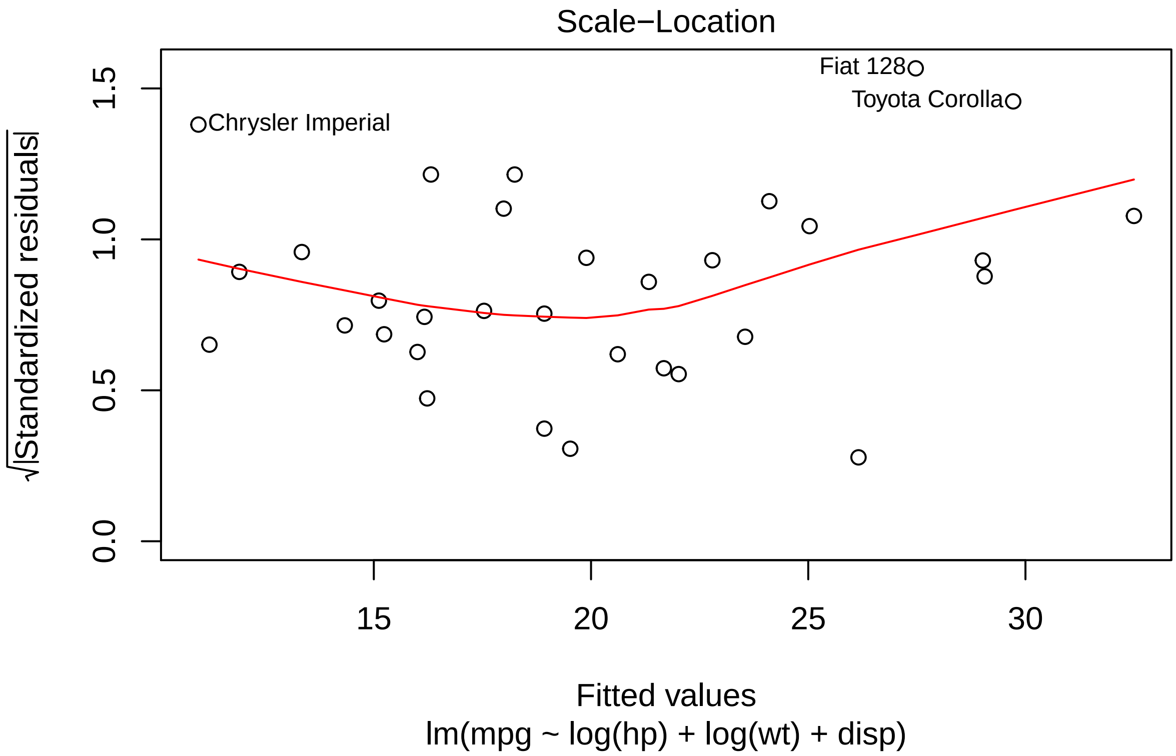
b. **5 pts** Constant Variance Assumption

Plot(s):

```
plot(model, 3)
```



```
plot(logmodel, 3)
```



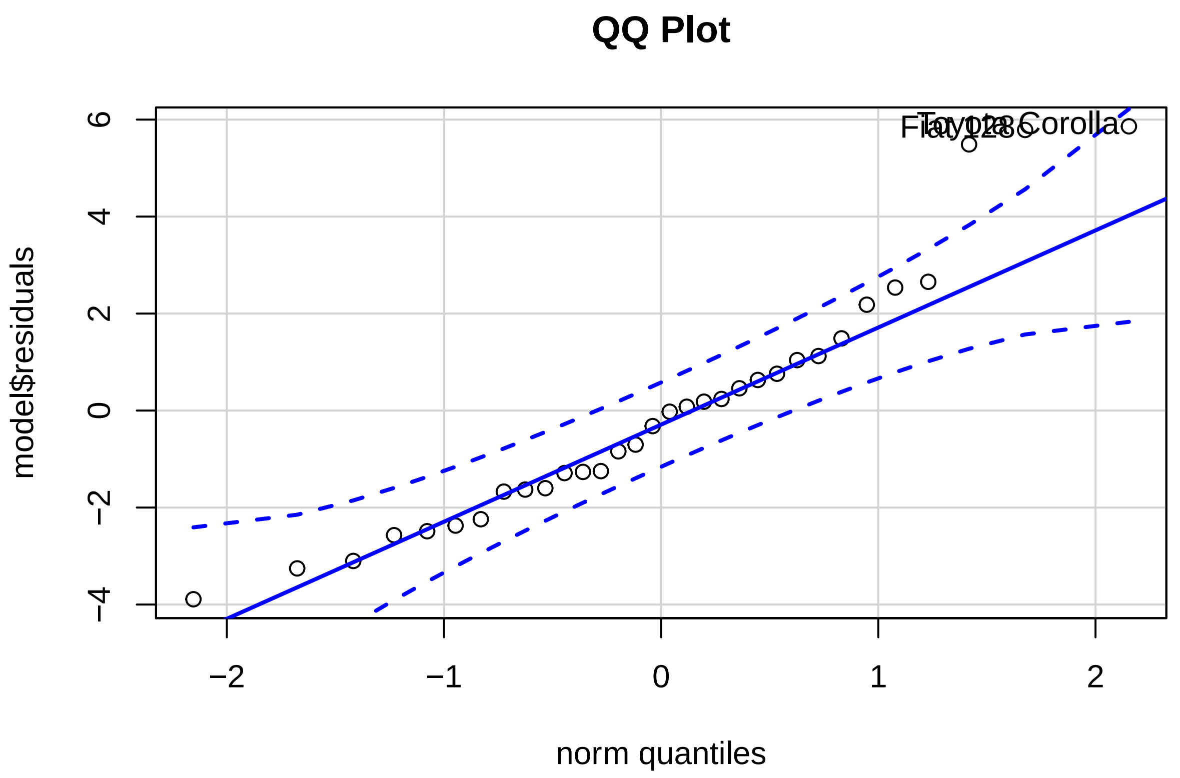
Interpretation:

The original model has appears to have constant variance. The assumption of constant variance for the original model holds. while the proposed log transformation of horsepower (hp) and weight (wt) does not have constant variance. The residuals of miles per gallon (mpg) is more heteroscedastic in the lower and higher end of the range for the proposed log transformation.

c. **5 pts** Normality Assumption

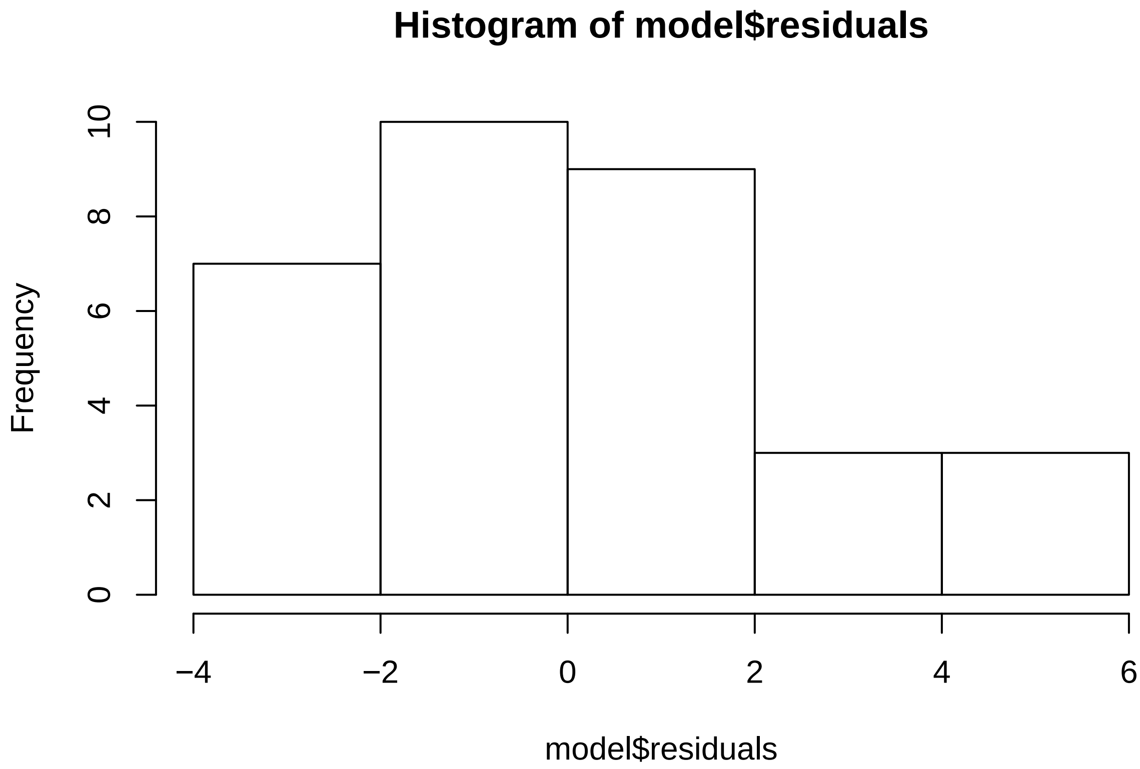
Plot(s):

```
qqPlot(model$residuals, main='QQ Plot', pch=1)
```

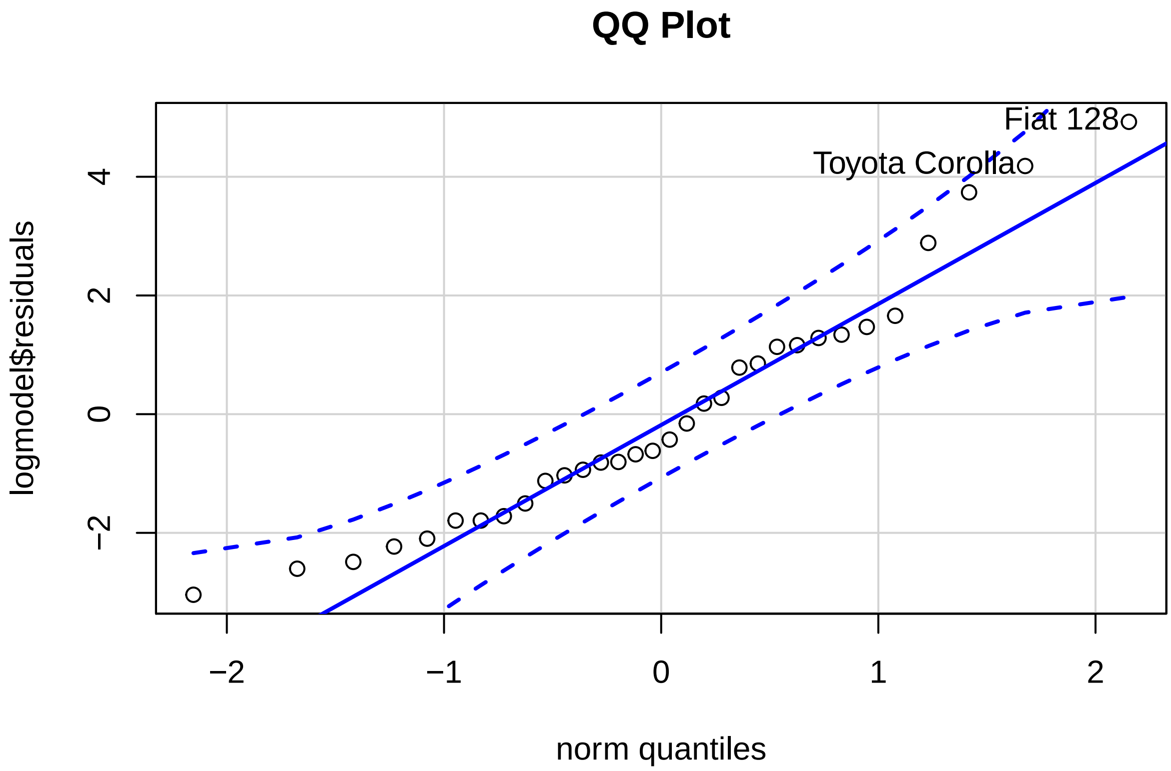


Toyota Corolla Fiat 128 20 18

```
hist(model$residuals)
```

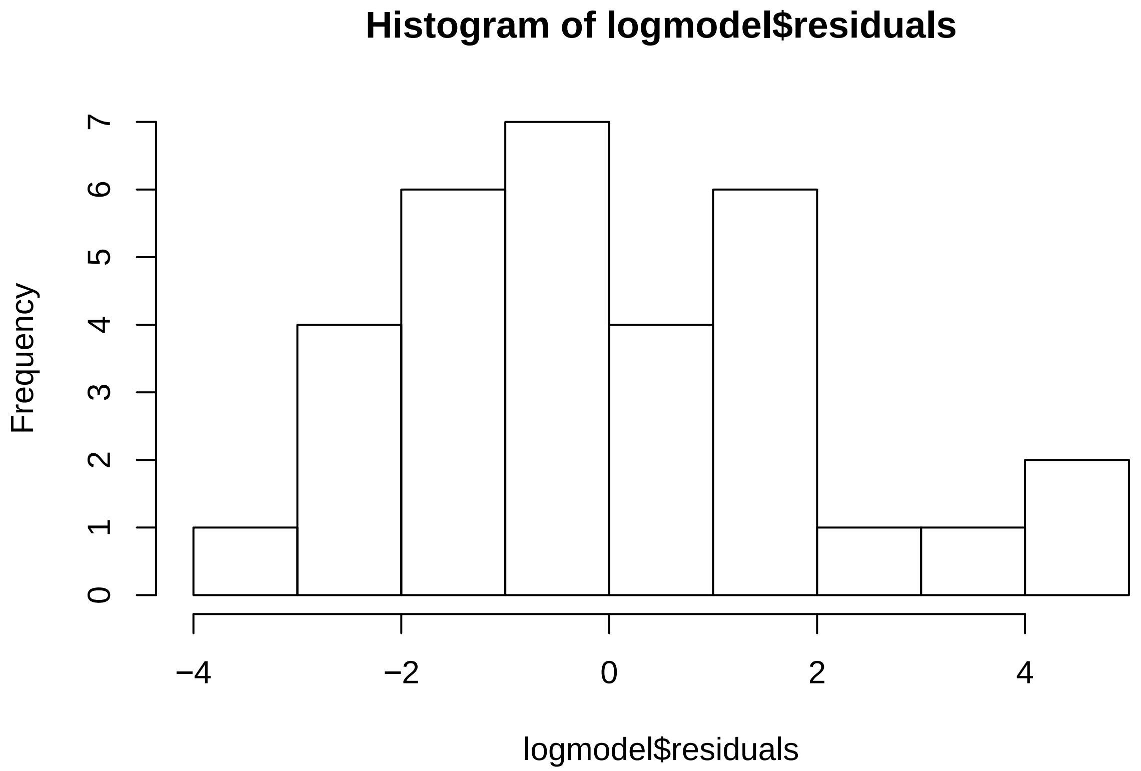


```
qqPlot(logmodel$residuals, main='QQ Plot', pch=1)
```



```
Fiat 128 Toyota Corolla 18 20
```

```r
hist(logmodel$residuals)
```



Interpretation:

The distribution of our error terms for the model is mostly normal with 2 cars as outliers, the Toyota Corolla & Fiat 128. The distribution if fat tailed on the right side. Persuing a log transformation of the horsepower (hp) and weight (wt) variables normalizes the distribution (although there is still some fatness in the right tail) and the 2 cars above are no longer outliers.

Question 4: Coefficient Interpretation - 6 points

a. **3 pts** Interpret the coefficient of wt (mention any assumption you make about other predictors clearly when stating the interpretation).

For every 1000 lbs of weight (wt) added to a car results in a -3.8 miles per gallon (mpg) drop in fuel economy, assuming horsepower hP and displacement (disp) are held constant at their current values in the model.

b. **3 pts** If value of predictor wt in the above model is increased by 0.01 keeping other predictors constant, what change in the response would be expected?

Given 1000 * 0.01 = 10 lbs of weght added, we would see a -0.38 mpg drop in fuel economy, keeping horsepower and displacement constant.

Question 4: Confidence Intervals and Interpretation - 9 points

a. **4 pts** Compute 90% and 95% confidence intervals (CIs) for the parameter associated with disp (β_3) for the model in Question 2.

The 90% confidence interval for dispacement (disp):

```
print(confint(model, "disp", level = 0.90))
```

```
          5 %       95 %
disp -0.01854328 0.01666926
```

The 95% confidence interval for displacement (disp):

```
print(confint(model, "disp", level = 0.95))
```

```
       2.5 %     97.5 %
disp -0.0221375 0.02026348
```

b. **5 pts** Using just these intervals, what could you deduce about the range (Upper Bound or Lower Bound or both) of the p-value for testing H_0: $\beta_3 = 0$ in the model in Question 2?

Both confidence intervals contain the null hypthesis of $\beta_3 = 0$ in their range. This means that the results are not stastistically significant. We can accept the null hypothesis. The lower bound of the p value is at least $>$ 0.10.