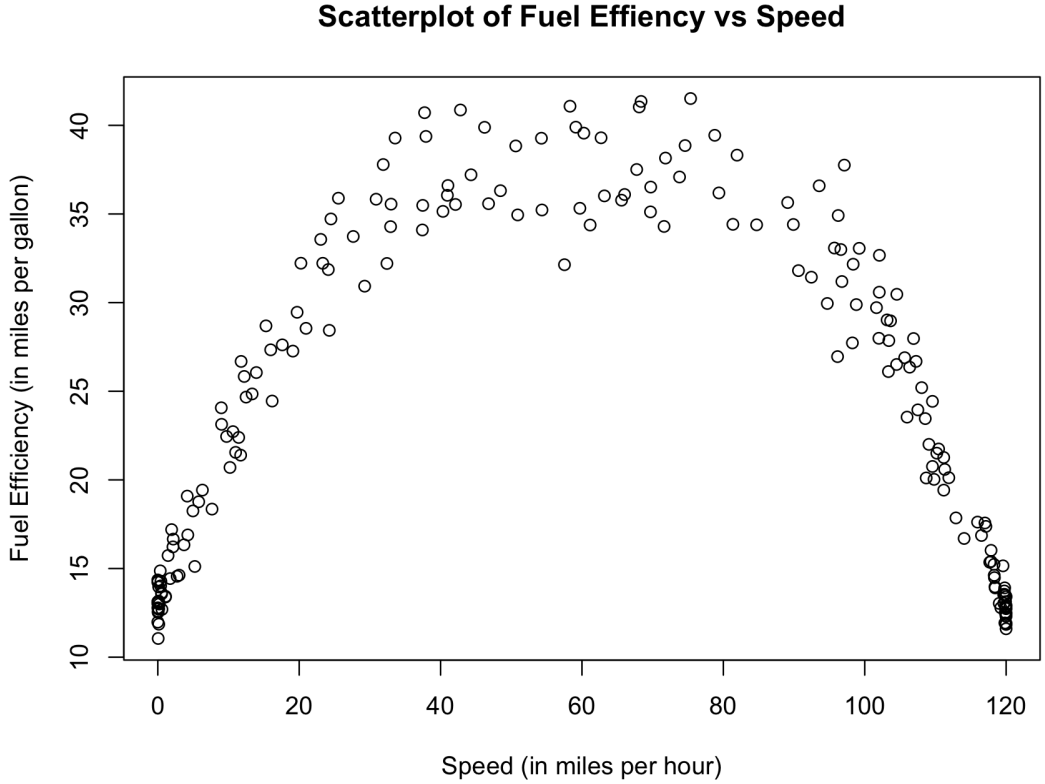
STAT 463 - Assignment 4

Due Data: Saturday, April 1st, 2023

Data Description and Background
Rocket Motors, manufacturer of high-end sport bikes, just released its newest bike line, the Speed Demon. To get an understanding of the motorcycle's performance, the bike was driven on a closed course by a professional driver at a constant speed for 5 minutes and various data points recorded, among them being the fuel efficiency (measured in miles per gallon (MPG)). These calibration runs were performed 200 times at a variety of speeds.

1) Create a scatterplot of the data from the calibration runs, plotting the MPG on the vertical axis and speed on the horizontal axis (be sure to properly label your plot). Does there appear to be an association between the speed the bike is driven at and the MPG? If so, explain what the nature of the relationship seems to be. – 1.5 point



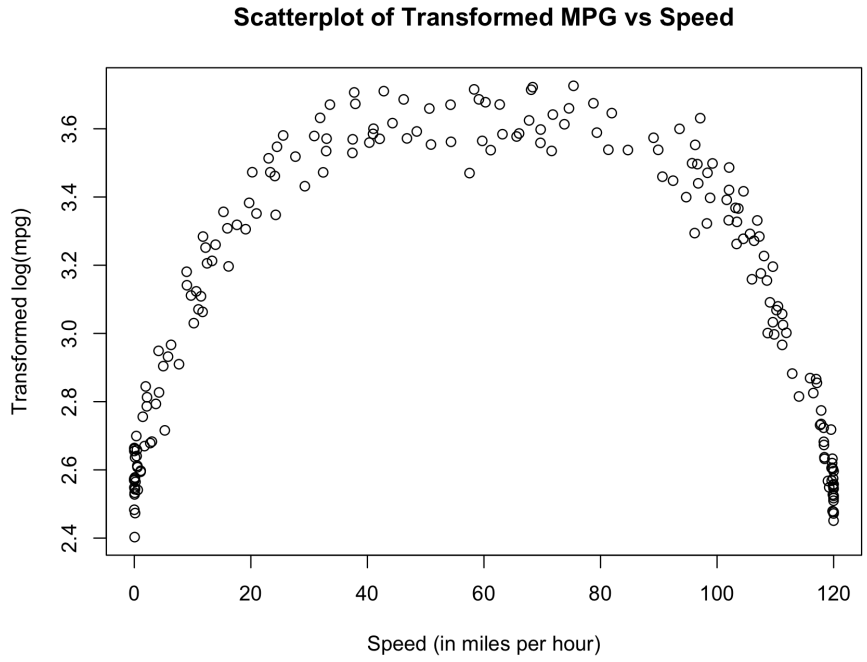
Based on the scatterplot there appears to be an association between the speed the bike is driven at and the mpg. The relationship between fuel efficiency (in mpg) and speed (in mph) seems to be concave down quadratic and not simple or monotonic as it changes direction and convexity. Therefore it cannot be corrected by simply applying a linearizing transformation.

2) Treating MPG as the response variable and speed as the explanatory variable, are enough of the model assumptions satisfied in order to fit a polynomial model to this data towards the prior

purpose? If not, explain what must be done to address the deviations from the needed model assumptions (if necessary).– 2 points

In order to fit a polynomial model to this data, the behavior of a response variable (mpg) can be explained by a curvilinear relationship between the response variable (mpg) and the explanatory variable (speed). That assumption has been met because when looking at the scatter plot a very obvious quadratic relationship exists between speed and mpg. And the curve is not simple or monotonic as it changes direction and convexity. Therefore since it doesn't pass the linearity assumption, it cannot be corrected by simply applying a linearizing transformation and so polynomial fit would be the best model. The assumption of errors is justified because for any given speed, the variation of mpg mostly appears to be the same, being distributed evenly above and below where the polynomial regression line would fit through the data. There are no apparent outliers.

But the homoscedasticity assumption is violated for this dataset because the variation of the observations of mpg around the polynomial regression 'line' is greater for values at the middle of the spread of speed values. To address the heteroscedasticity, I apply a transformation to the data of mpg. I tested 1/mpg, log(mpg) and sqrt(mpg). The log function happened to correct for heteroscedasticity the best.




```
> var(mpg.data[,2])
[1] 93.63014
> var(log(mpg.data[,2]))
[1] 0.1755884
```

The variance dropped drastically after the transformation and based off the scatterplot above, the variation of the observations of mpg around the polynomial regression 'line' is no longer greater for values at the middle of the spread of speed values. The homoscedasticity assumption is justified and the polynomial model can be fit to the data.

3) After addressing any issues in part 2, fit a polynomial model to the data. Clearly explain the process with which you went about arriving at the order of the polynomial model you fit (you will need to fit several polynomial models and compare them). Explicitly write out the estimated model equation for the polynomial model you decided upon (on the transformed scales if data transformations were needed). – 2 points

I began by fitting several polynomial models by increasing order and looking for the highest adjusted R^2 value found among the summaries to select the best polynomial model order.

```
> mod2 = lm(log(MPG) ~ Speed + I(Speed^2) )
> mod3 = lm(log(MPG) ~ Speed + I(Speed^2) + I(Speed^3) )
> mod4 = lm(log(MPG) ~ Speed + I(Speed^2) + I(Speed^3) + I(Speed^4) )
> mod5 = lm(log(MPG) ~ Speed + I(Speed^2) + I(Speed^3) + I(Speed^4) + I(Speed^5) )
> summary(mod2)

Call:
lm(formula = log(MPG) ~ Speed + I(Speed^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.303201 -0.089438 -0.005311  0.091005  0.281194

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.649e+00  1.721e-02  153.97   <2e-16
Speed        3.745e-02  8.096e-04   46.26   <2e-16
I(Speed^2)  -3.113e-04  6.577e-06  -47.33   <2e-16

(Intercept) ***
Speed       ***
I(Speed^2)  ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1197 on 197 degrees of freedom
Multiple R-squared:  0.9192,  Adjusted R-squared:  0.9184
F-statistic:  1121 on 2 and 197 DF,  p-value: < 2.2e-16

> summary(mod3)

Call:
lm(formula = log(MPG) ~ Speed + I(Speed^2) + I(Speed^3))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.301422 -0.089994 -0.008313  0.089468  0.265627

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.660e+00  1.920e-02 138.547  < 2e-16 ***
Speed        3.526e-02  1.864e-03  18.918  < 2e-16 ***
I(Speed^2)  -2.611e-04  3.915e-05  -6.669 2.56e-10 ***
I(Speed^3)  -2.804e-07  2.156e-07  -1.300    0.195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1195 on 196 degrees of freedom
Multiple R-squared:  0.9199,  Adjusted R-squared:  0.9187
F-statistic: 750.2 on 3 and 196 DF,  p-value: < 2.2e-16

> summary(mod4)

Call:
lm(formula = log(MPG) ~ Speed + I(Speed^2) + I(Speed^3) + I(Speed^4))

Residuals:
     Min       1Q   Median       3Q      Max
-0.183479 -0.046792 -0.002348  0.051251  0.169324

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.572e+00  1.212e-02  212.11   <2e-16 ***
Speed        6.831e-02  2.019e-03   33.83   <2e-16 ***
I(Speed^2)  -1.689e-03  7.693e-05  -21.96   <2e-16 ***
I(Speed^3)   1.876e-05  9.873e-07   19.00   <2e-16 ***
I(Speed^4)  -7.865e-08  4.045e-09  -19.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06989 on 195 degrees of freedom
Multiple R-squared:  0.9727,  Adjusted R-squared:  0.9722
F-statistic:  1740 on 4 and 195 DF,  p-value: < 2.2e-16

> summary(mod5)

Call:
lm(formula = log(MPG) ~ Speed + I(Speed^2) + I(Speed^3) + I(Speed^4) +
```

```
   I(Speed^5))

Residuals:
      Min        1Q    Median        3Q       Max
-0.182939 -0.047550 -0.003123  0.050619  0.169658

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.573e+00  1.286e-02 200.127  < 2e-16 ***
Speed        6.756e-02  3.383e-03  19.972  < 2e-16 ***
I(Speed^2)  -1.636e-03  2.078e-04  -7.872  2.4e-13 ***
I(Speed^3)   1.750e-05  4.689e-06   3.732 0.000249 ***
I(Speed^4)  -6.667e-08  4.387e-08  -1.520 0.130216
I(Speed^5)  -3.967e-11  1.446e-10  -0.274 0.784172
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07006 on 194 degrees of freedom
Multiple R-squared:  0.9728,  Adjusted R-squared:  0.972
F-statistic:  1385 on 5 and 194 DF,  p-value: < 2.2e-16
```

The p-value from the F-test for the significance of the regression is highly significant for all 4 models.Thus we can conclude that there is a relationship between the response & explanatory variable.

The 4th order had the highest Adjusted R^2 value of (0.9722) which we are using as the criteria to find the best model order because R^2 is the proportion of the total variance of the response variable explained by the polynomial model on the explanatory variable but adding higher order polynomial terms will only increase it, therefore we use Adjusted R^2 which takes into account model complexity.

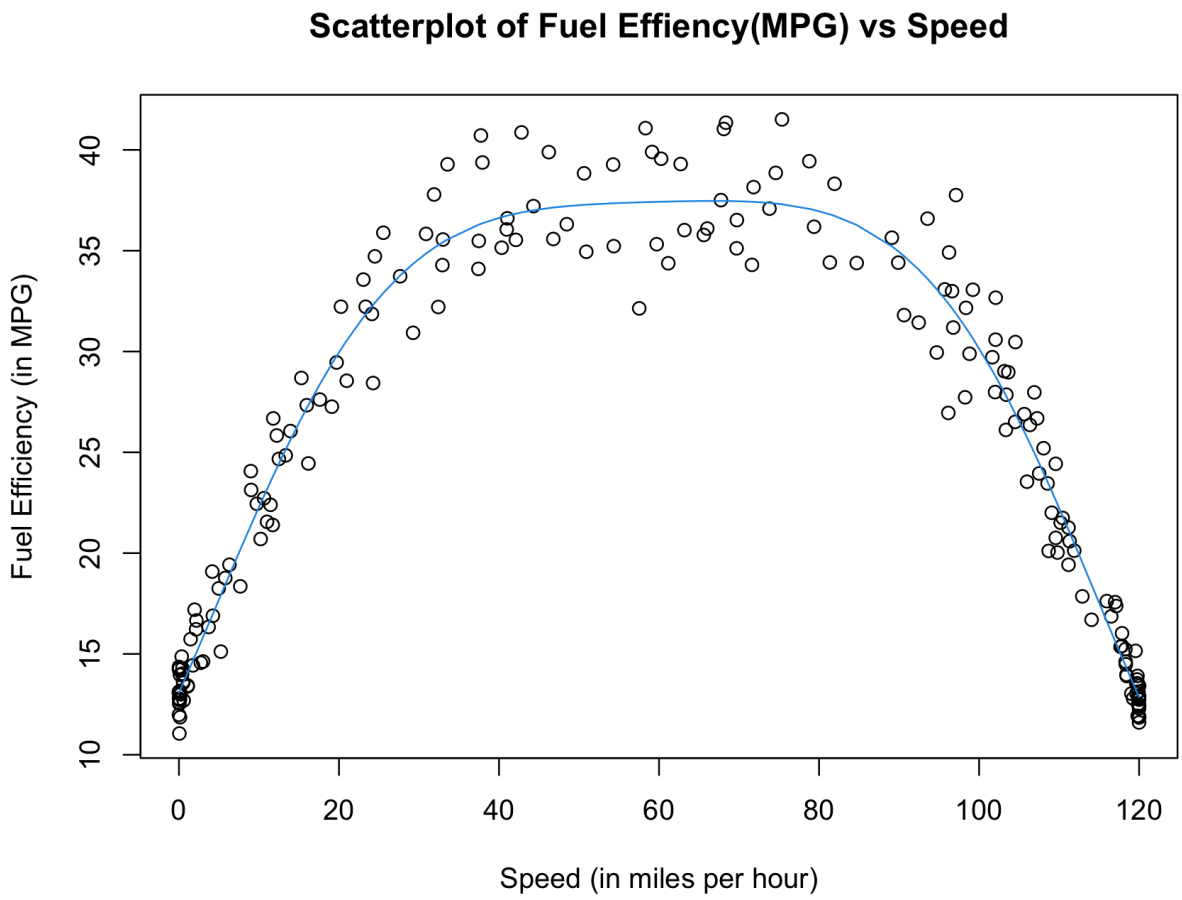
The estimated model equation:

$y = 2.572 + 0.06831x - 0.001689x^2 + 0.00001876x^3 - 0.00000007865x^4$

4) On a scatter plot depicting the MPG on the vertical axis and speed on the horizontal axis (on their original, untransformed measurement scales), overlay the estimated model on the plot (in the event you transformed any of your variables, this may necessitate back transforming the polynomial model that was constructed on the transformed data). – 2 points

```
> CI4 = predict(mod4, interval = "confidence", level = 0.9)
> CI4 = cbind(Speed, CI4)
> CI4 = CI4[order(CI4[,1]),]
```

```
> plot(mpg.data[,1],mpg.data[,2], main="Scatterplot of Fuel Efficiency(MPG) vs
Speed",xlab="Speed (in miles per hour)",ylab="Fuel Efficiency (in MPG)")
> points(CI4[,1], exp(CI4[,2]), type = "l", col = 4)
```



5) From the model constructed in part 3, can one conclude that there is a statistically significant relationship between MPG and the speed? Explain what procedure you used to determine so and why you arrived at your conclusion. – 1 points

Conduct hypothesis test:
H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$
H_1: At Least one β is not equal to zero

At the significance level of 0.05, the p value is 2e-16 so we reject the null hypothesis, so a statistically significant relationship does exist between speed and mpg. We have chosen a good model for prediction.

6) Calculate the coefficient of determination for the model on the original measurement scale (if transformations were applied to the data, calculations of the various sums of squares requires

back transforming the fitted values from the polynomial model on the transformed data to get the fitted values and residuals on the original scale). – 2 points

We must first find the fitted and residual values on the original scale by applying the inverse of log function which is exp() to the regression model.

```
> fitted = exp(predict(mod4))
> resid = MPG - fitted
```

We can then go ahead and calculate R^2 (coefficient of determination) by doing intermediate calculations of TSS RSS and ESS. TSS is the sum of the squared differences between the observed MPG values and their mean. RSS is the sum of the squared residuals.

```
> mean = mean(MPG)
> TSS = sum((MPG-mean)^2)
> RSS = sum(resid^2)
> ESS = TSS-RSS
> Rsq = ESS/TSS
```

$R^2 = 0.9639837$, which gives the proportion of the total variance in the MPG data that is explained by the polynomial regression model.

7) According to the model constructed in part 3, at what speed is the engine most fuel efficient (i.e. what speed does it have the highest MPG on average). Explain how you arrived at this value (this can certainly be ascertained analytically, but providing a numerically approximated value is also acceptable as well). – 2.5 points

#First I used the predicted values already calculated from the observed speed values prior that had been stored along the upper and lower confidence interval values in the object CI4.

```
> CI4 = predict(mod4, interval = "confidence", level = 0.9)
> CI4 = cbind(Speed, CI4)
> CI4 = CI4[order(CI4[,1]),]
```

#then i stored the speed column which is the original speed data with the transformed (by log() function) predicted fit data for mpg.

```
> s=CI4[,1]
> fit = CI4[,2]
```

then I found the index of the maximum value of the predicted mpg values when back transformed with exp() function.

```
> ind <- which.max(exp(fit))
```

#then found the speed at which it has the highest MPG with the index

```
> corspspeed <- s[ind]
```

#then i find the maxMPG

```
> maxMPG <- exp(fit[ind])
```

I print out the max MPG and the speed at which it has this max MPG for the model equation
> maxMPG
    12
37.46651
> corspspeed
    12
66.02823

The speed at which the maximum MPG is achieved is 66.028 miles per hour with a fuel efficiency of 37.4665 Miles Per Gallon

8) On a scatter plot depicting the MPG on the vertical axis and speed on the horizontal axis (on their original, untransformed measurement scales), overlay 90% confidence bands for the mean MPG as functions of the speed – 3 points
> CI4 = predict(mod4, interval = "confidence", level = 0.9)
> CI4 = cbind(Speed, CI4)
> CI4 = CI4[order(CI4[,1]),]
> plot(mpg.data[,1],mpg.data[,2], main="Scatterplot of Fuel Effiency(MPG) vs Speed with 90% > confidence bands",xlab="Speed (in miles per hour)",ylab="Fuel Efficiency (in MPG)")
> points(CI4[,1], exp(CI4[,2]), type = "l", col = 4)
> points(CI4[,1], exp(CI4[,3]), type = "l", lty = 2, col = 3)
> points(CI4[,1], exp(CI4[,4]), type = "l", lty = 2, col = 3)