Homework4

Benjamin Wang

2023-03-22

Assignment 4

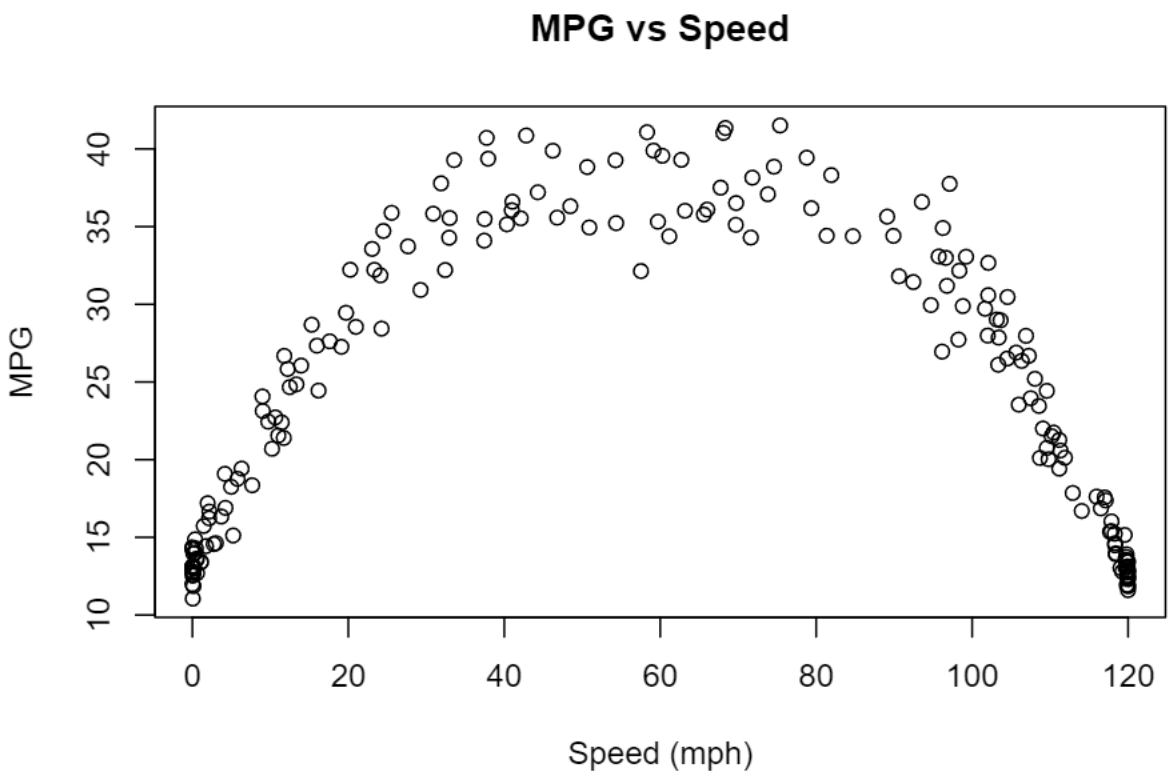
Rocket Motors, manufacturer of high-end sport bikes, just released its newest bike line, the Speed Demon. To get an understanding of the motorcycle's performance, the bike was driven on a closed course by a professional driver at a constant speed for 5 minutes and various data points recorded, among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

```
mpg = read.csv("./mpg_data.txt",sep="\t")
mpg$speed = mpg$Speed_.mph.
```

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.

There appears to be a quadratic relationship between MPG and Speed. As speed gets nearer to 60 mph, MPG increases. As speed gets further from 60 mph, MPG decreases.

```
plot(MPG ~ Speed_.mph., data=mpg,
     xlab="Speed (mph)",
     ylab="MPG",
     main = "MPG vs Speed"
  )
```



2) The National Highway Traffic Safety Administration (NHTSA) requires all vehicles marketed in the US to provide ranges for what the mean MPG is at a variety of speeds. 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).

a. The linearity assumption is satisfied - the model appears as if it could be modeled closely by a polynomial model (quadratic).

b. The p-value is less than 0.05 (obtained by performing a breusch-pagan test, assuming a quadratic model) and therefore **there is heteroscedasticity** in this model. We may address this by applying a log transformation to the model.

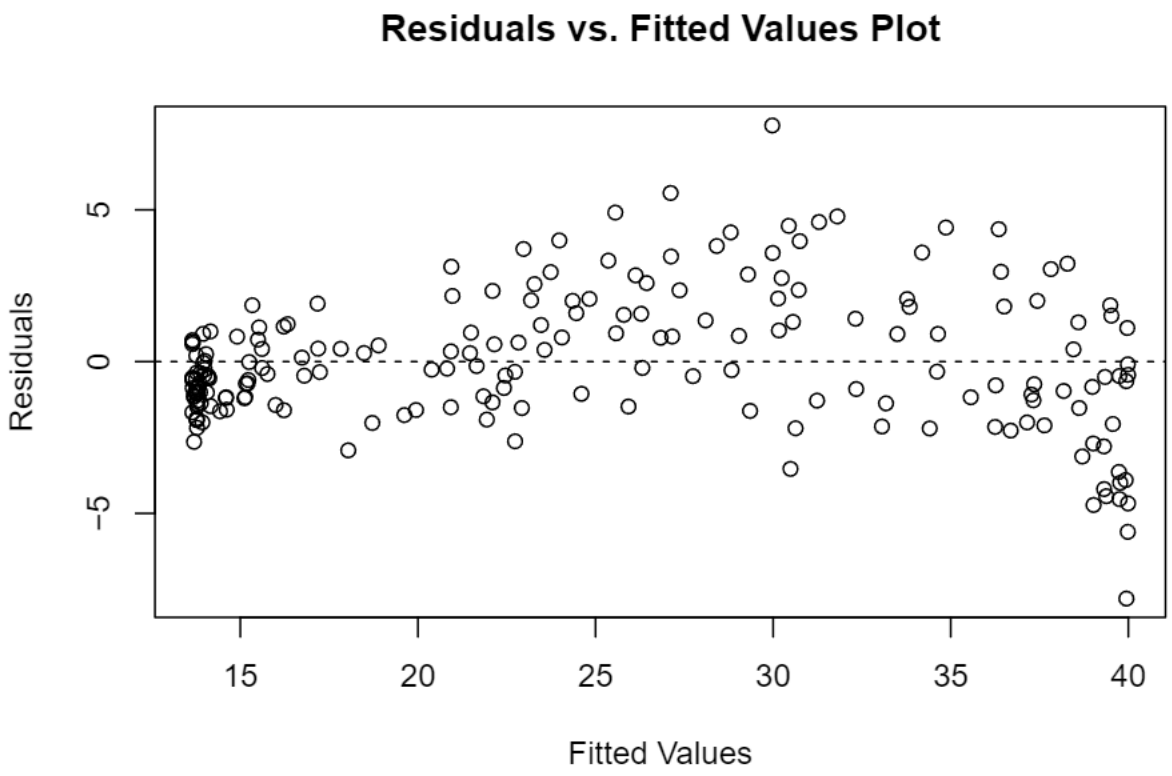
```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
speedValues = mpg[order(mpg[,1]),1]
quadraticModel = lm(MPG ~ speed + I(speed^2), data=mpg)

# Display residuals vs fitted values plot
plot(fitted(quadraticModel), resid(quadraticModel), main = "Residuals vs. Fitted Values Plot",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, lty = "dashed")
```

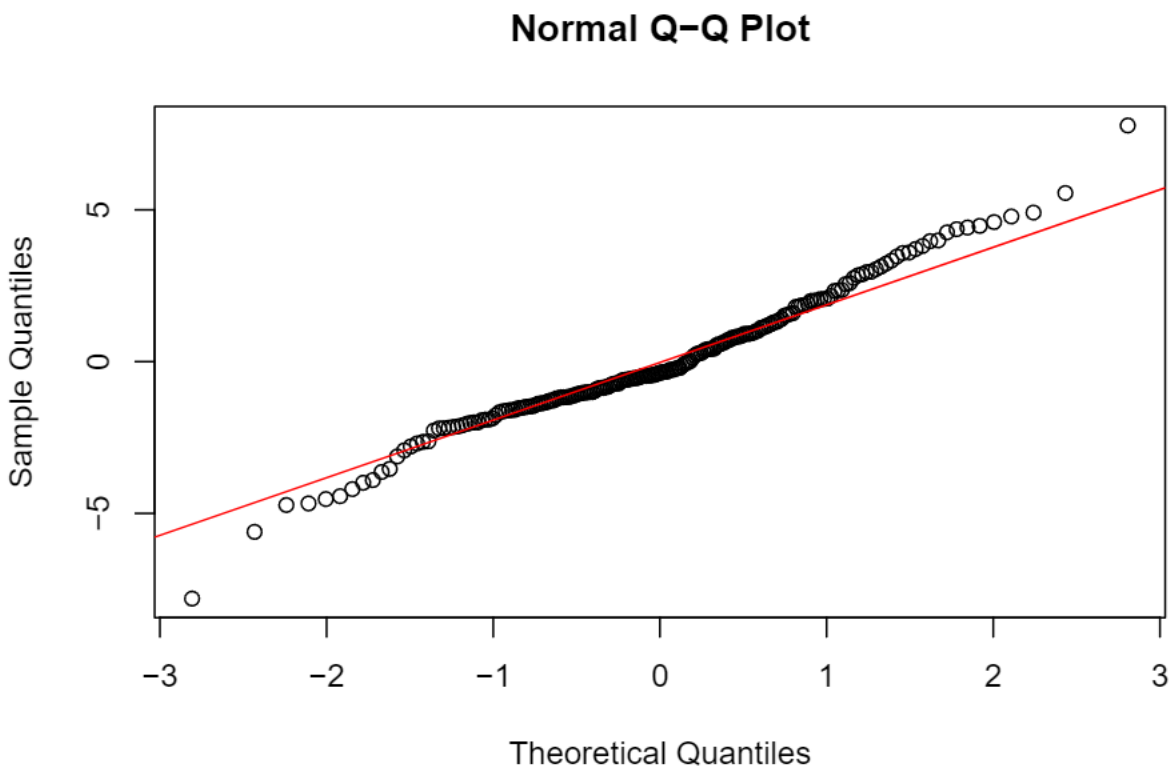



```
# Breusch Pagan test
bptest(quadraticModel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  quadraticModel
## BP = 30.147, df = 2, p-value = 2.842e-07
```

c. Generally, errors are normal. However at the ends of the line, there is an outlier and a slight deviation from the line (negative on left end and positive on right end). To address this, we could remove the two outliers.

```
qqnorm(resid(quadraticModel))
qqline(resid(quadraticModel),col="red")
```



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).

```
# Apply log transformation to both MPG and speed variables
mpg$logMPG <- log(mpg$MPG)

# Fit quadratic model with log-transformed variables
logQuadraticModel <- lm(logMPG ~ speed + I(speed^2), data=mpg)
logMpgPredict = predict(logQuadraticModel, list(speed=speedValues))

bptest(logQuadraticModel)
```

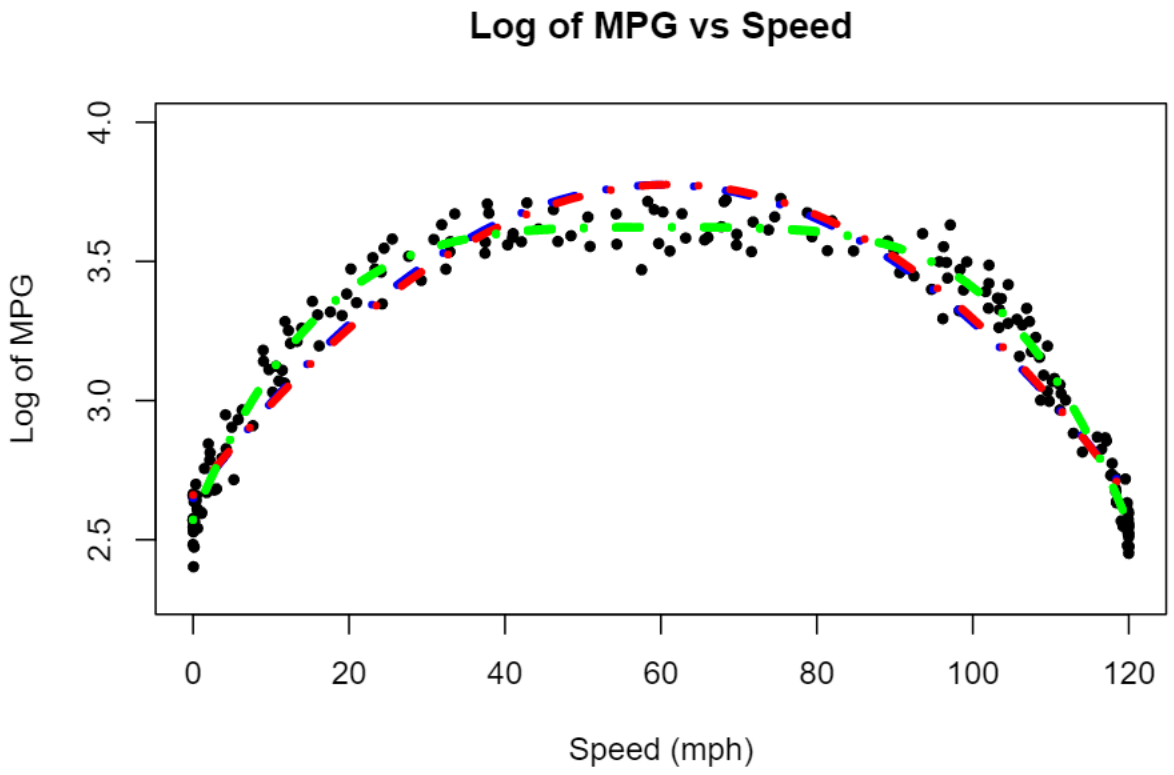
```
##
##  studentized Breusch-Pagan test
##
## data:  logQuadraticModel
## BP = 5.1553, df = 2, p-value = 0.07595
```

Performing a log transformation followed by another Breusch-Pagan test on the transformed model yields a p-value of 0.07595, which is greater than our significance level, 0.05.

Although the data appears to be closely modeled by the quadratic model, it is important to fit polynomial models of other orders to the data as well. We test polynomial models of increasing degrees and take the one with the highest R-squared value.

```
# Get line of fit
logCubicModel = lm(log(MPG) ~ speed + I(speed^2) + I(speed^3), data=mpg)
cubicPredict = predict(logCubicModel, list(speed=speedValues))
```

```
logQuarticModel = lm(log(MPG) ~ speed + I(speed^2) + I(speed^3) + I(speed^4), data=mpg)
quarticPredict = predict(logQuarticModel, list(speed=speedValues))

logQuinticModel = lm(log(MPG) ~ speed + I(speed^2) + I(speed^3) + I(speed^4) + I(speed^5), data=mpg)
quinticPredict = predict(logQuinticModel, list(speed=speedValues))

logSexticModel = lm(log(MPG) ~ speed + I(speed^2) + I(speed^3) + I(speed^4) + I(speed^5) + I(speed^6),
sexticPredict = predict(logSexticModel, list(speed=speedValues))

logSepticModel = lm(log(MPG) ~ speed + I(speed^2) + I(speed^3) + I(speed^4) + I(speed^5) + I(speed^6) +
septicPredict = predict(logSepticModel, list(speed=speedValues))

plot(mpg$Speed_.mph., log(mpg$MPG), pch=20, cex=1, ylim=c(2.3,4), main = "Log of MPG vs Speed", ylab="L
lines(speedValues, logMpgPredict, col='blue', lwd=4, lty=4)
lines(speedValues, cubicPredict, col='red', lwd=4, lty=4)
lines(speedValues, quarticPredict, col='green', lwd=4, lty=4)
lines(speedValues, sexticPredict, col='green', lwd=4, lty=4)
lines(speedValues, septicPredict, col='green', lwd=4, lty=4)
```




```
cat("Quadratic R^2:", summary(logQuadraticModel)$r.squared, "\n")
```

```
## Quadratic R^2: 0.9191992
```

```
cat("Cubic R^2:", summary(logCubicModel)$r.squared, "\n")
```

```
## Cubic R^2: 0.9198904
```

```
cat("Quartic R^2:", summary(logQuarticModel)$r.squared, "\n")
```

```
## Quartic R^2: 0.9727395
```

```
cat("Sextic R^2:", summary(logSexticModel)$r.squared, "\n")
```

```
## Sextic R^2: 0.9727511
```

```
cat("Septic R^2:", summary(logSexticModel)$r.squared, "\n")
```

```
## Septic R^2: 0.9727511
```

```
summary(logSexticModel)
```

```
##
## Call:
## lm(formula = log(MPG) ~ speed + I(speed^2) + I(speed^3) + I(speed^4) +
##     I(speed^5) + I(speed^6), data = mpg)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.182343 -0.047475 -0.003195  0.050774  0.170197
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.573e+00  1.340e-02 191.990  < 2e-16 ***
## speed        6.787e-02  5.039e-03  13.469  < 2e-16 ***
## I(speed^2)  -1.669e-03  4.463e-04  -3.739 0.000243 ***
## I(speed^3)   1.870e-05  1.529e-05   1.223 0.222793
## I(speed^4)  -8.634e-08  2.425e-07  -0.356 0.722165
## I(speed^5)   1.076e-10  1.791e-09   0.060 0.952152
## I(speed^6)  -4.115e-13  4.988e-12  -0.082 0.934337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07024 on 193 degrees of freedom
## Multiple R-squared:  0.9728, Adjusted R-squared:  0.9719
## F-statistic:  1148 on 6 and 193 DF,  p-value: < 2.2e-16
```

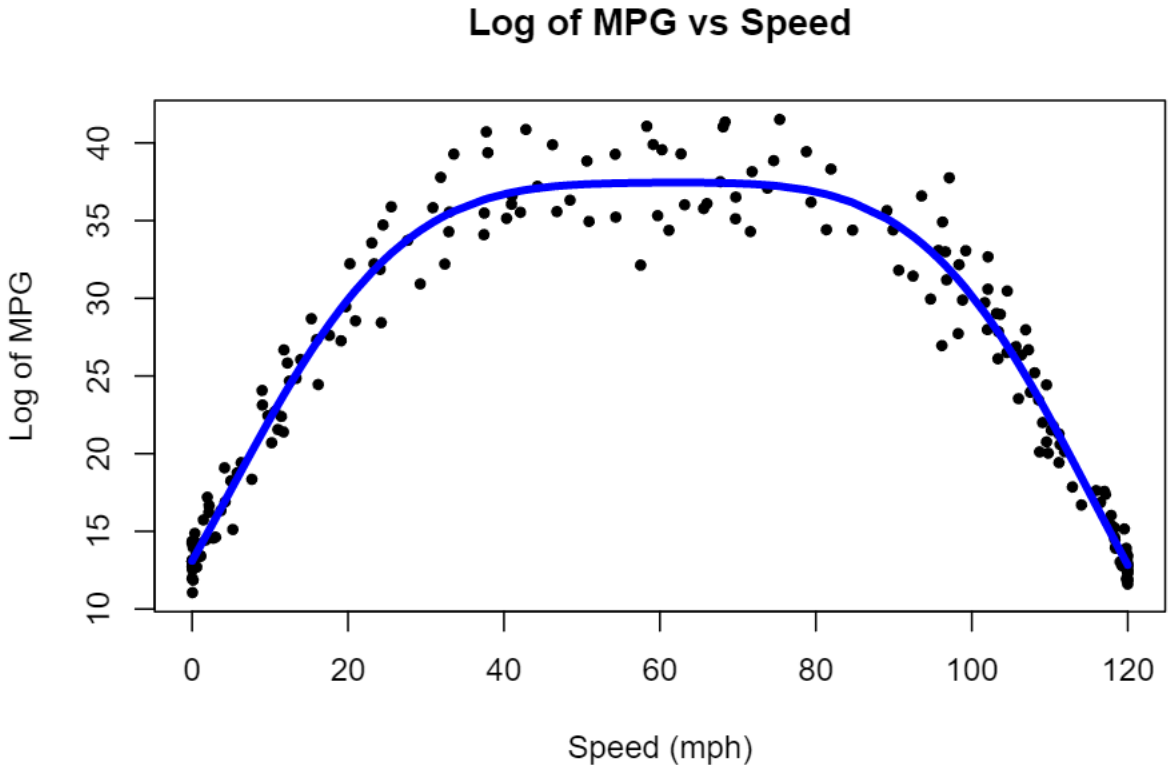
Examining the different models shows that the blue line, or the sextic model best fits the line. It also has the highest R-squared value of all the models. The septic model has a similar R^2 (significant figures) to the sextic, but has one more degree so it's unnecessary to use the septic. For the sake of simplicity, we use the sextic model.

The formula of this model is:

$$y = 2.572 + (6.787 * 10^{-02})x + (-1.669 * 10^{-3})x^2 + (-1.870 * 10^{-5})x^3 + (-8.634 * 10^{-8})x^4 + (1.076 * 10^{-10})x^5 + (-4.115 * 10^{-13})x^6$$

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)

```
plot(mpg$Speed_.mph., mpg$MPG, pch=20, cex=1, main = "Log of MPG vs Speed", ylab="Log of MPG", xlab="Sp
lines(speedValues, exp(sexticPredict), col='blue', lwd=4, lty=1)
```



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.

```
summary(logSexticModel)
```

```
##
## Call:
## lm(formula = log(MPG) ~ speed + I(speed^2) + I(speed^3) + I(speed^4) +
##     I(speed^5) + I(speed^6), data = mpg)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.182343 -0.047475 -0.003195  0.050774  0.170197
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.573e+00  1.340e-02 191.990  < 2e-16 ***
## speed        6.787e-02  5.039e-03  13.469  < 2e-16 ***
## I(speed^2)  -1.669e-03  4.463e-04  -3.739 0.000243 ***
## I(speed^3)   1.870e-05  1.529e-05   1.223 0.222793
## I(speed^4)  -8.634e-08  2.425e-07  -0.356 0.722165
## I(speed^5)   1.076e-10  1.791e-09   0.060 0.952152
## I(speed^6)  -4.115e-13  4.988e-12  -0.082 0.934337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07024 on 193 degrees of freedom
## Multiple R-squared:  0.9728, Adjusted R-squared:  0.9719
## F-statistic:  1148 on 6 and 193 DF,  p-value: < 2.2e-16
```

To determine if this model indicates that there is a statistically significant relation between MPG and speed, we examine the F-statistic of the sextic model. The F-statistic is 1148 with a p-value of less than $2.2 * 10^{-16}$. Therefore, it is statistically significant.

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)

```
logSexticPredict = predict(logSexticModel)
logResiduals = logSexticPredict - log(mpg$MPG)
backResid = exp(logResiduals)
RSq = 1 - (sum(backResid^2)/ sum( (mpg$MPG - mean(mpg$MPG))^2 ) )
cat("R^2: ",RSq,"\n")
```

```
## R^2:  0.989163
```

The coefficient of determination (R^2 value) after backtransforming is 0.989.

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).

The engine is most fuel efficient at 62.70145 mph with a fuel efficiency of 39.29755 MPG

```
cat("Maximum fuel efficiency: ",mpg$MPG[which.max(logSexticPredict)],"\n")
```

```
## Maximum fuel efficiency:  39.29755
```

```
cat("Speed at maximum fuel efficiency: ",mpg$Speed_.mph.[which.max(logSexticPredict)],"\n")
```

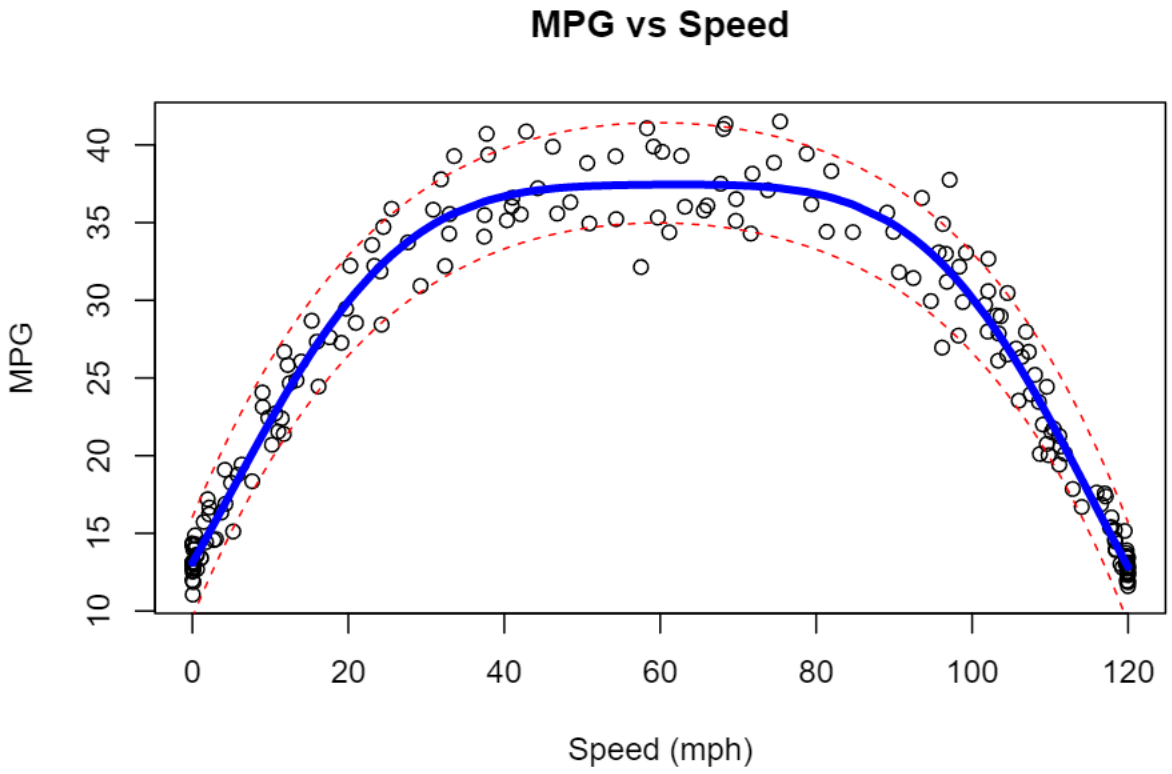
```
## Speed at maximum fuel efficiency:  62.70145
```

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 (in the event you transformed any of your variables, this will necessitate back transforming the 90% confidence bands for the polynomial model that were constructed on the transformed data).

```
plot(mpg$MPG ~ mpg$speed,
     xlab="Speed (mph)",
     ylab="MPG",
     main = "MPG vs Speed"
  )

backTransformed = data.frame(x = mpg$Speed_.mph., y = exp(logSexticPredict))
backTransformed = backTransformed[order(backTransformed$x), ]

backQuintic = lm(MPG ~ speed + I(speed^2) + I(speed^3) + I(speed^4) + I(speed^5), data=mpg)
lines(sort(backTransformed$x), backTransformed$y, col='blue', lwd=4, lty=1)

PI = predict(backQuintic, interval = "prediction", level = 0.9)
```

```
## Warning in predict.lm(backQuintic, interval = "prediction", level = 0.9): predictions on current dat
```

```
PI = cbind(mpg$speed, PI)
PI = PI[order(PI[,1]),]
points(PI[,1], PI[,3], type = "l", lty = 2, col = "red")
points(PI[,1], PI[,4], type = "l", lty = 2, col = "red")
```


MPG vs Speed
MPG
10
15
20
25
30
35
40
0
20
40
60
80
100
120
Speed (mph)