Homework 2

Simple Linear Regression Model Assumptions

AUTHOR

Zeb Sorenson

Data and Description

One key component of determining appropriate speed limits is the amount of distance that is required to stop at a given speed. For example, in residential neighborhoods, when pedestrians are commonly in the roadways, it is important to be able to stop in a very short distance to ensure pedestrian safety. The speed of vehicles may be useful for determining the distance required to stop at that given speed, which can aid public officials in determining speed limits.

The Stopping Distance data set compares the **distance (column 2)** (in feet) required for a car to stop on a certain rural road against the **speed (column 1)** (MPH) of the car. Download the StoppingDistance.txt file from Canvas, and put it in the same folder as this quarto file.

0. Replace the text "< PUT YOUR NAME HERE >" (above next to "author:") with your full name.

1. Read in the data set, and call the data frame "stop".

```
stop_data <- read.csv("~/Desktop/Stat 330/StoppingDistance.txt", sep="")

head(stop_data)
```

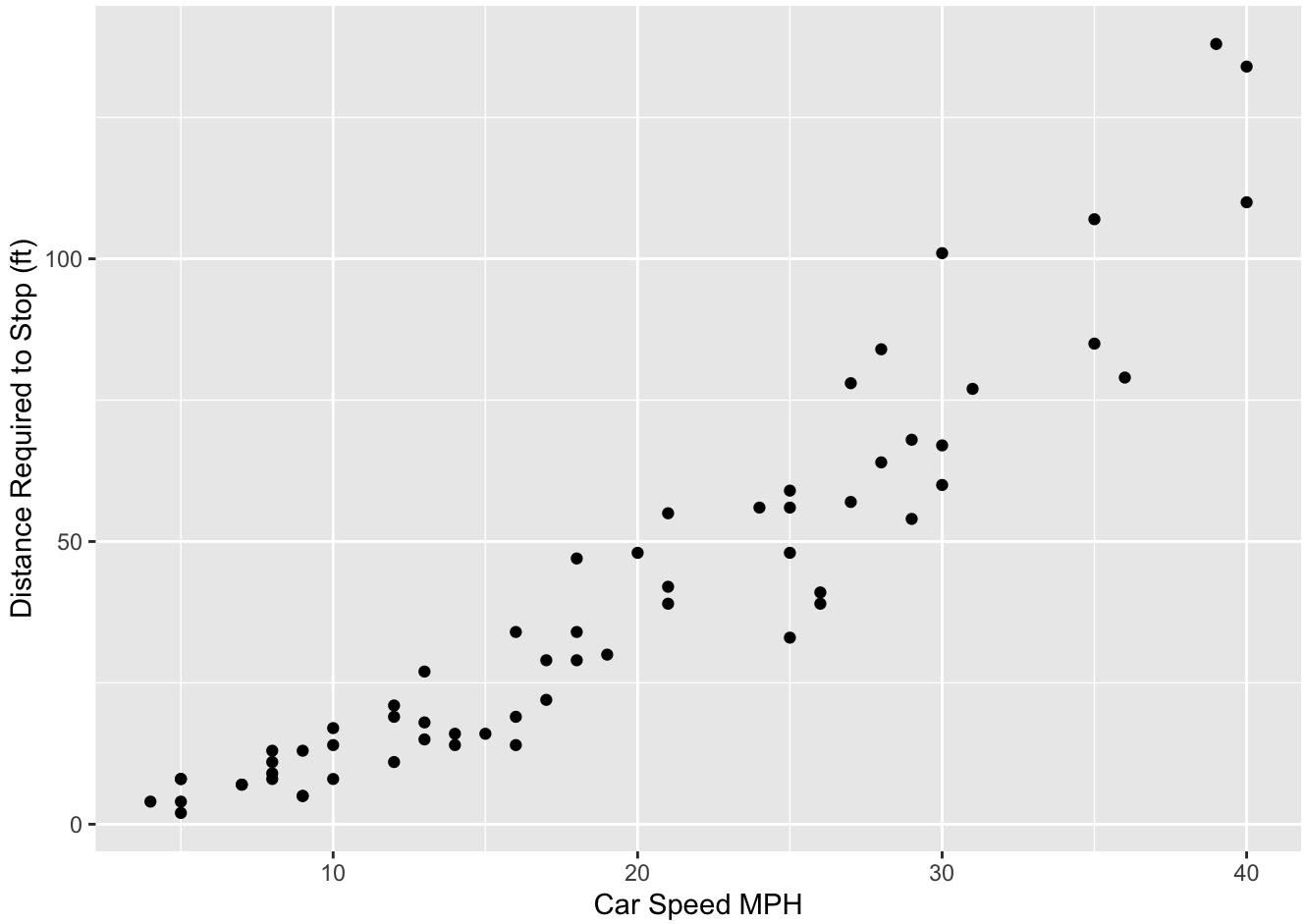
```
  Speed Distance
1     4        4
2     5        2
3     5        4
4     5        8
5     5        8
6     7        7
```

```
summary(stop_data)
```

```
     Speed          Distance
 Min.   : 4.00   Min.   :  2.00
 1st Qu.:10.00   1st Qu.: 13.25
 Median :17.50   Median : 29.50
 Mean   :18.92   Mean   : 39.31
 3rd Qu.:26.75   3rd Qu.: 56.75
 Max.   :40.00   Max.   :138.00
```

2. Create a scatterplot of the data with variables on the appropriate axes (think about which variable makes the most sense to be the response). Make your plot look professional (make sure the axis labels are descriptive).

```r
#speed will be our x and distance will be our y

speed_scatter_plot <- ggplot(data = stop_data, mapping = aes(x = Speed, y = Distance)) +
  geom_point() +

  labs(x = "Car Speed MPH", y = "Distance Required to Stop (ft)")

print(speed_scatter_plot)
```



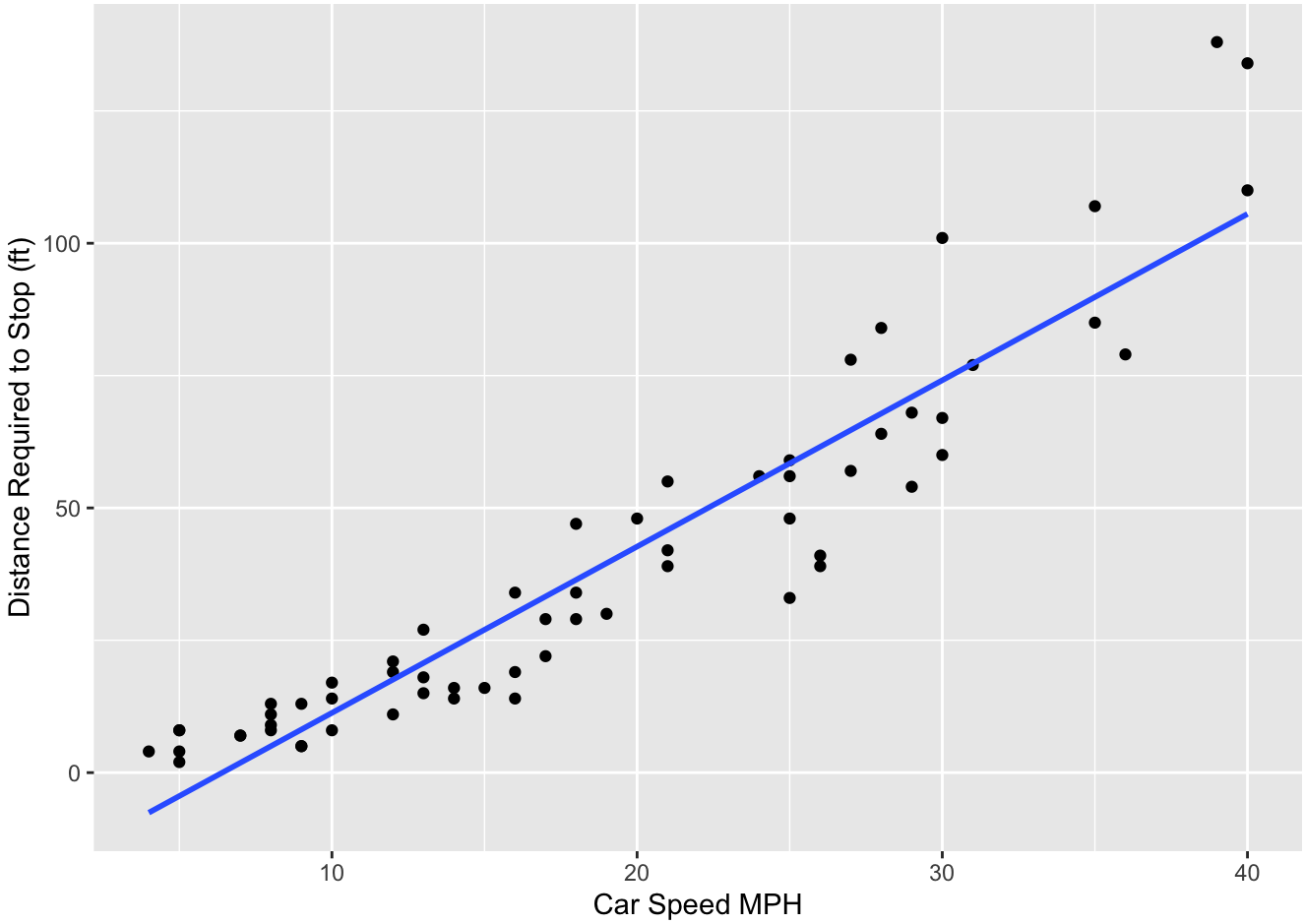
3. Briefly describe the relationship between Speed and Distance. (Hint: you should use 2 or 3 key words.)

There is a moderately positive linear relation, however, we do see strong curvature towards the end at about the 30 MPH mark, which we will keep an eye on.

4. Add the OLS regression line to the scatterplot you created in question 2 (note: if you receive a warning about rows with missing values, you may need to adjust an axis limit using `scale_y_continuous(limits = c(###, ###))`).

```
speed_scatter_plot +  geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



5. (a) Apply linear regression to the data (no transformations). (b) Print out a summary of the results from the `lm` function. (c) Save the residuals and fitted values to the `stop` dataframe.

```
#A
car_speed_LM <- lm(Distance~Speed, data = stop_data) #Are my X and Y values in the correct sp
#B
summary(car_speed_LM)
```

```
Call:
```

```
lm(formula = Distance ~ Speed, data = stop_data)

Residuals:
    Min      1Q  Median      3Q     Max
-25.410  -7.343  -1.334   5.927  35.608

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.1309     3.2308  -6.231 5.04e-08 ***
Speed         3.1416     0.1514  20.751  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.77 on 60 degrees of freedom
Multiple R-squared:  0.8777,    Adjusted R-squared:  0.8757
F-statistic: 430.6 on 1 and 60 DF,  p-value: < 2.2e-16
```

```
#C
stop_data$residuals <- car_speed_LM$residuals
stop_data$fits <- car_speed_LM$fitted.values

head(stop_data)
```

```
  Speed Distance residuals      fits
1     4        4 11.564466 -7.564466
2     5        2  6.422847 -4.422847
3     5        4  8.422847 -4.422847
4     5        8 12.422847 -4.422847
5     5        8 12.422847 -4.422847
6     7        7  5.139611  1.860389
```

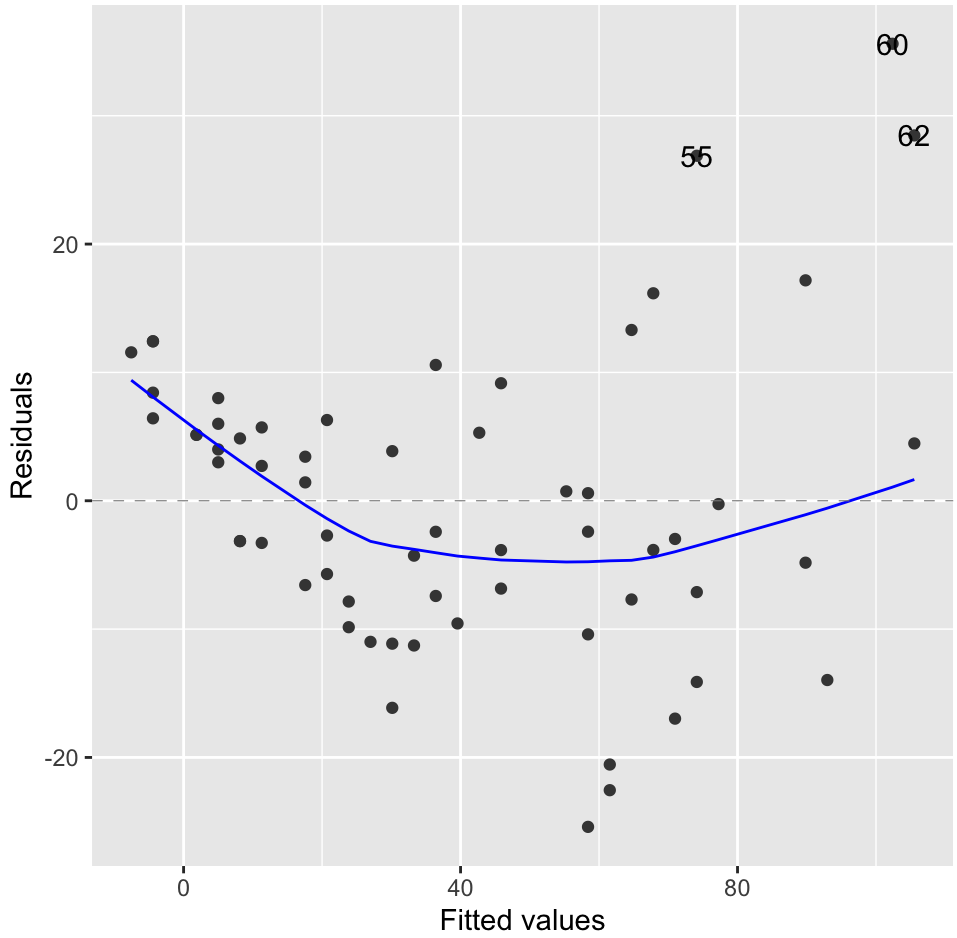
6. Mathematically write out the **fitted** simple linear regression model for this data set using the coefficients you found above. Do not use “x” and “y” in your model - use variable names that are fairly descriptive.

Estimated_StopDistance_Given_CarSpeed = -20.1309 + 3.1416×Car_Speed

Questions 7-11 involve using diagnostics to determine if the linear regression assumptions are met. For each assumption, (1) perform appropriate diagnostics to determine if the assumption is violated, and (2) explain whether or not you think the assumption is violated and why you think that.

7. (L) X vs Y is linear

```
autoplot(car_speed_LM,which =1, ncol = 1, nrow = 1)+ theme(aspect.ratio = 1) +
labs(title = NULL)
```




```
#residuals vs fitted plot may be better to check for linearity. If it's not close to 0 then.
#will give you 4 but just want 1st
```

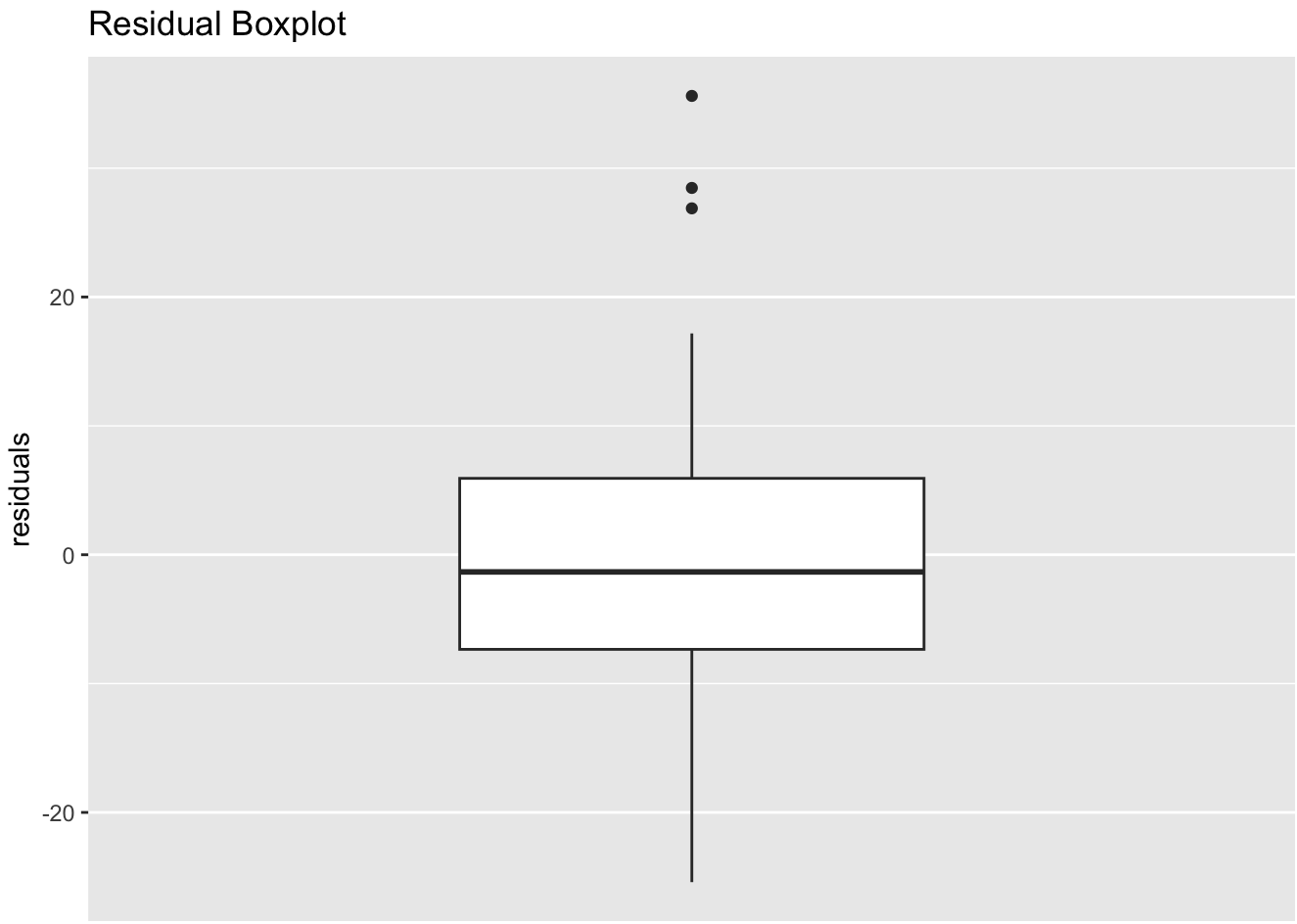
Looking at the residuals vs Fitted plot given from the autoplot function, we see that our data is still following a type of curvature. We want our data to uniform in randomness and our fitted line to follow along zero.

8. (I) The residuals are independent (no diagnostic tools required in this particular instance - just think about how the data was collected and briefly write your thoughts)

If we were to just think about plotting our residuals in the order they occurred in (having time as the x axis), we would be able to see that they are independent

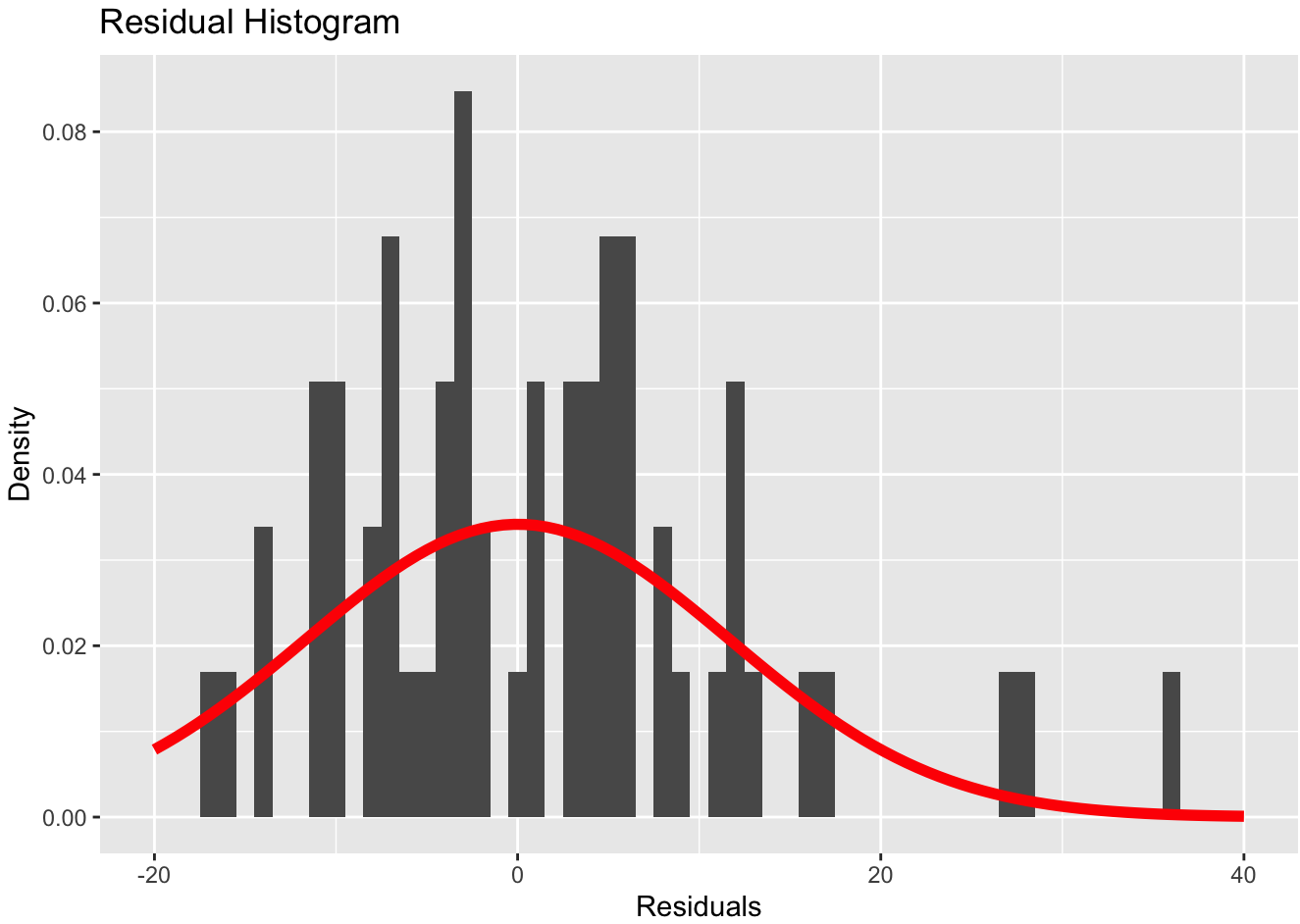
9. (N) The residuals are normally distributed. Use at least two diagnostic tools.

```
# <your code here>

stop_data$residuals <- car_speed_LM$residuals
# The code below produces a basic boxplot...Taken from the completed in class activity 2
ggplot(data = stop_data, mapping = aes(y = residuals)) +
  geom_boxplot() +
  scale_x_discrete() +
  labs(title = "Residual Boxplot")
```




```
#Now for a Histogram

ggplot(data = stop_data, mapping = aes(x = residuals)) +
  geom_histogram(mapping = aes(y = after_stat(density)),
                 binwidth = 1) +
  stat_function(fun = dnorm, color = "red", linewidth = 2,
                args = list(mean = mean(stop_data$residuals),
                            sd = sd(stop_data$residuals))) +
  labs(x = "Residuals", y = "Density", title = "Residual Histogram") +
  scale_x_continuous(limits = c(-20, 40))
```

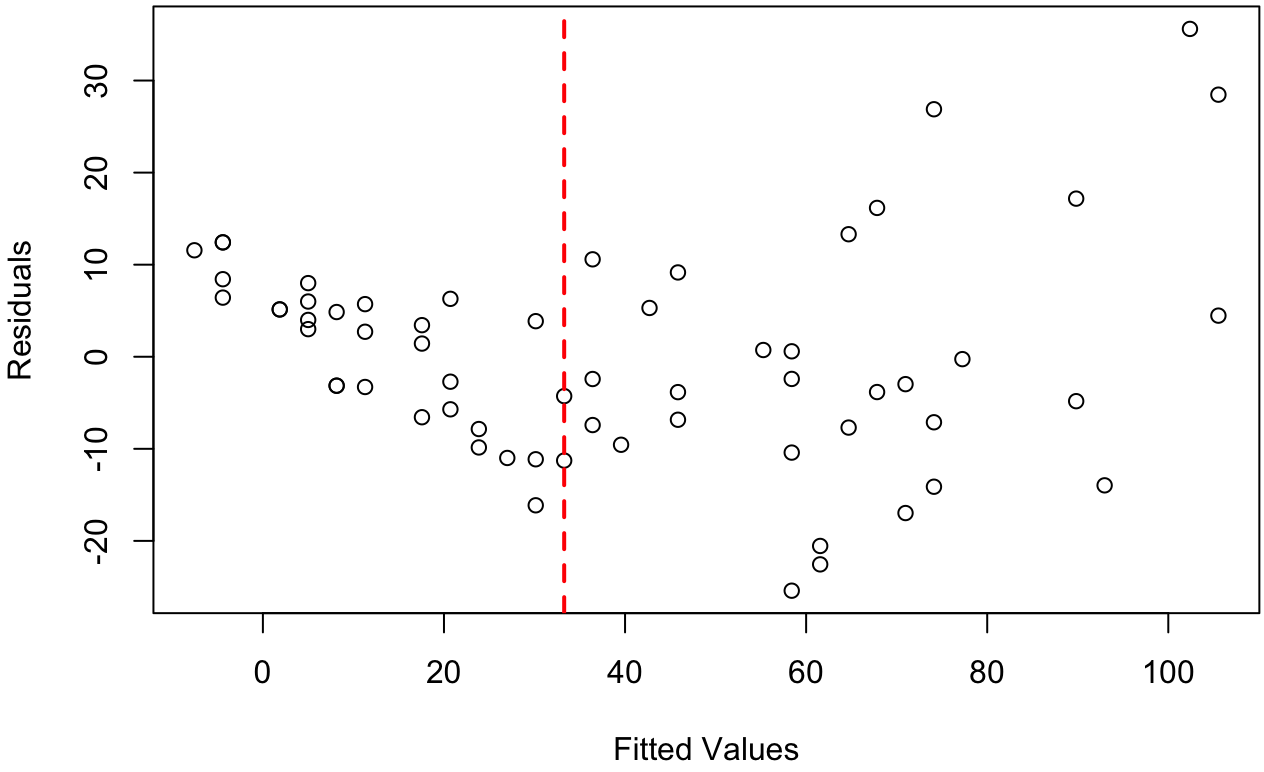
```
Warning: Removed 3 rows containing non-finite values (`stat_bin()`).

Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



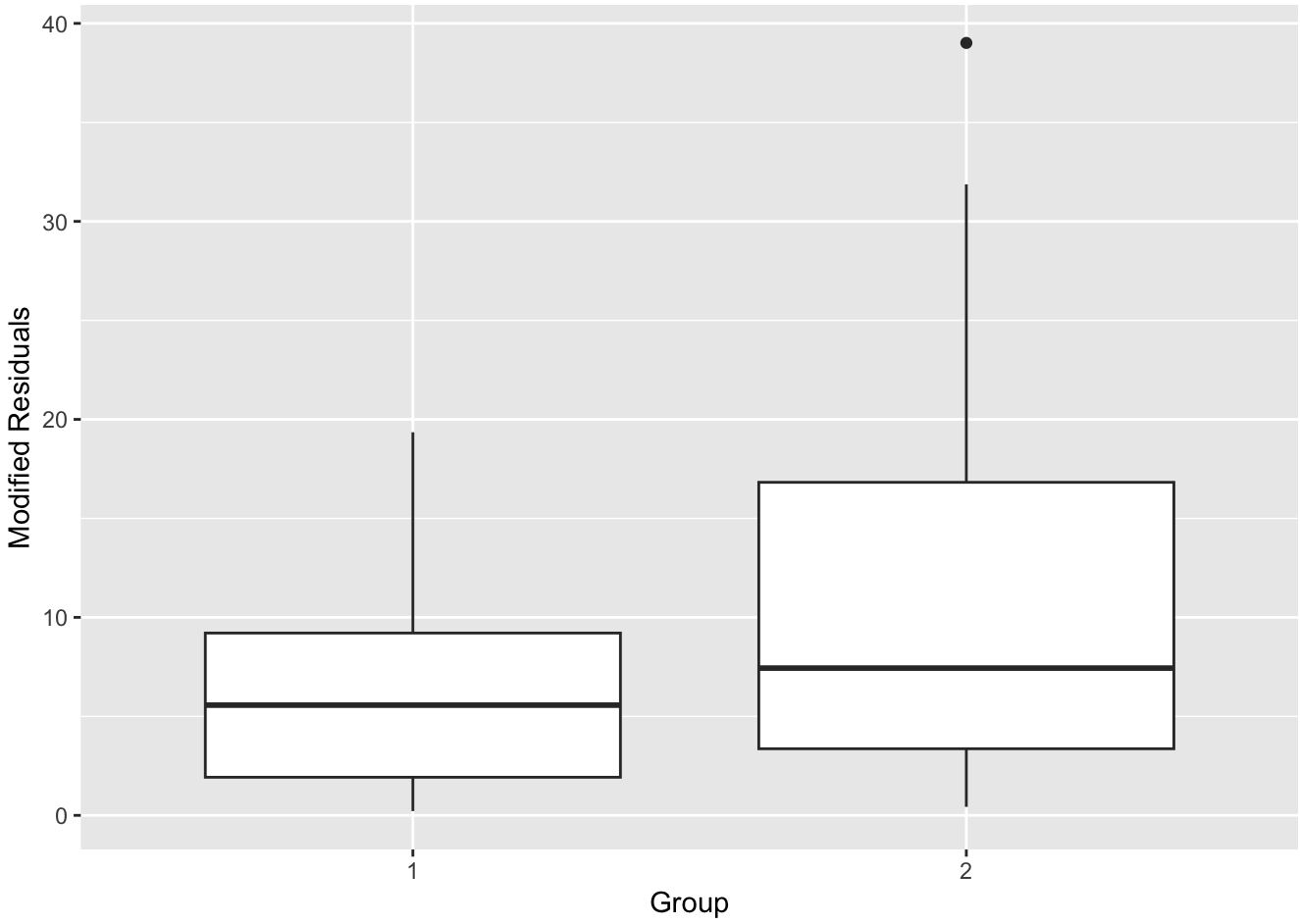
Our histogram shows us a right skew in our residuals and our box plot shows us outliers that correspond to what histogram shows us. Violating our assumption that the residuals are normally distributed.

10. (E) The residuals have equal/constant variance across all values of X.

```
#Brown-Forsythe Test
#This code is created by modifying the given code in the lecture slides

stop_LM <- lm(Distance~Speed, data = stop_data)
stop_data$resids <- stop_LM$residuals
stop_data$fits <- stop_LM$fitted.values
stop_bw <- stop_data |>
arrange(fits) |>
mutate(index = 1:nrow(stop_data),
e_group = ifelse(index < nrow(stop_data)/2, 1, 2) |>
as.factor()) |>
mutate(vals = abs(resids - median(resids)), .by = "e_group")
plot(stop_data$fits, stop_data$resids, xlab = "Fitted Values", ylab = "Residuals")
abline(v = stop_data$fits[round(nrow(stop_bw)/2)],
col = "red", lty = 2, lwd = 2)
```




```
stop_bw |>
ggplot(aes(e_group, vals)) +
geom_boxplot() +
labs(x = "Group", y = "Modified Residuals")
```




```
bf.test(vals ~ e_group, stop_bw)
```

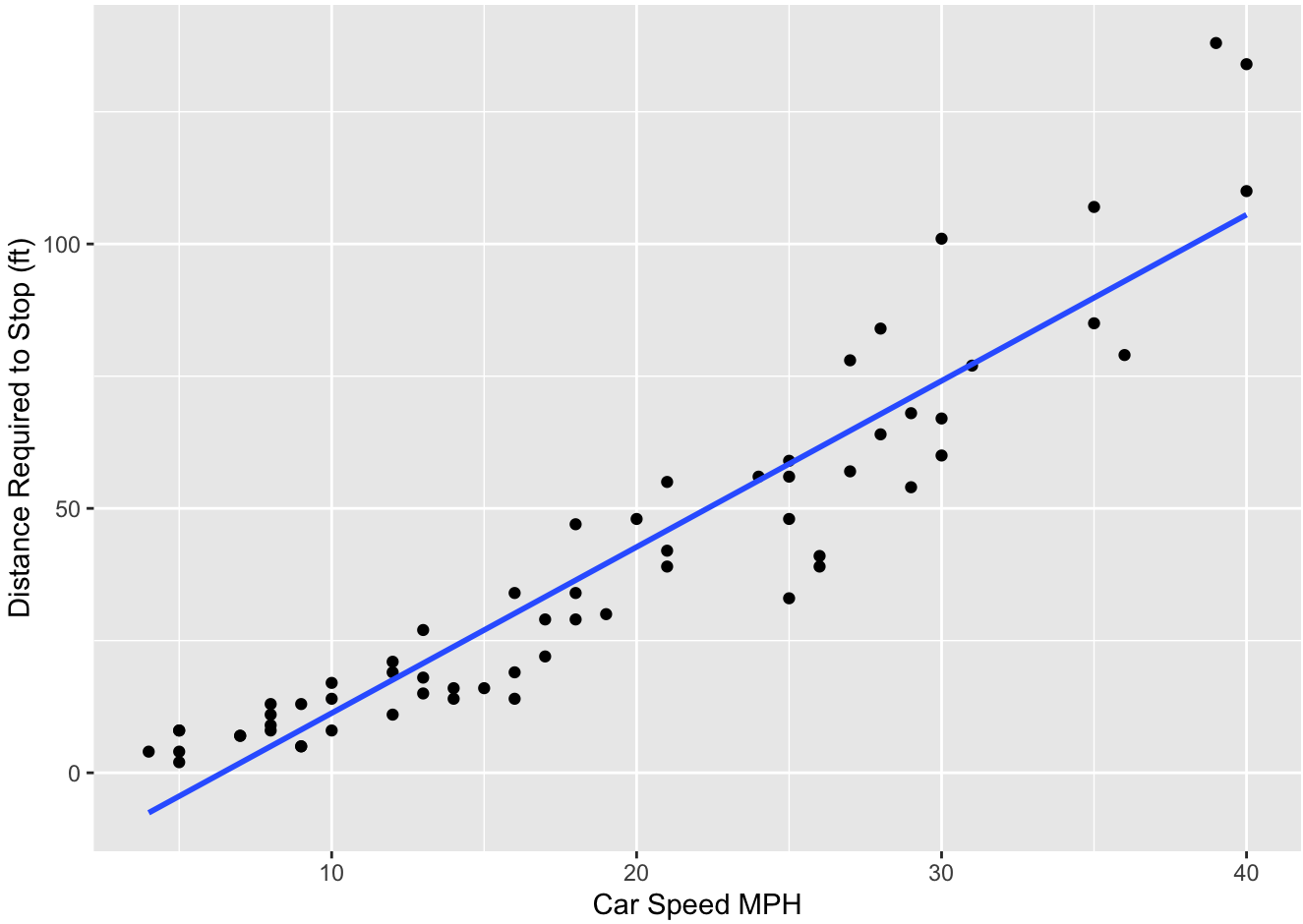
```
  Brown-Forsythe Test (alpha = 0.05)
------------------------------------------------------------
  data : vals and e_group

  statistic  : 4.602905
  num df     : 1
  denom df   : 46.76867
  p.value    : 0.03714113

  Result     : Difference is statistically significant.
------------------------------------------------------------
```

After running the Brown-Forsythe Test of Constant Variance, our results indicate that the difference is significant at an alpha of 0.05 we get a p value of 0.03714113. Causing us to reject our null hypothesis, therefore constant variance is violated.

11. Check if there are any influential points deserving of attention.

```
speed_scatter_plot +  geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



We're looking for outliers and leverage points. An outlier would be a high Y value that doesn't follow the trend of our data and a leverage point would be an extreme X value that actually does follow our trend. Without transforming the data, we can look at our model we can plot our current linear model and see two outliers at the 40 mark for car MPH speed. We will look to address these data points in our transformations.

12. Based on your analysis of the diagnostic measures, briefly discuss why this simple linear regression model on the raw data (not transformed) is not appropriate.

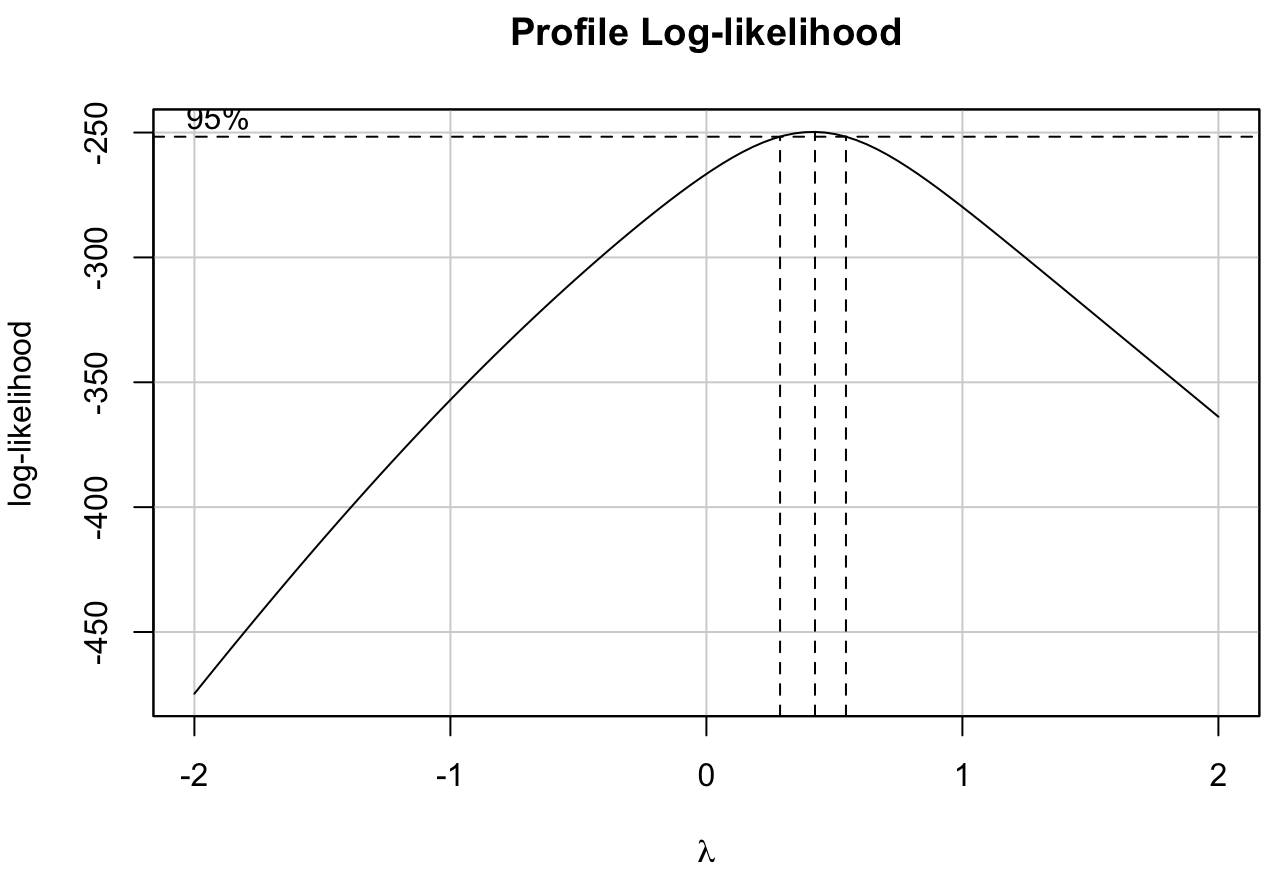
There are a number of linear regression assumptions that are violated in our current model that need to addressed in order to give a fair analysis of the data. Some of those violeted assumptions are, The residuals have equal/constant variance across all values of X, The residuals are normally distributed and X vs Y is linear. These are items that we will work to address in our transformations.

13. Fix the model by making any necessary transformations. Justify the transformation(s) you chose in words. (Note: if boxCox(mod) throws an error, replace mod with the formula for the linear model, y ~ x.) (Note: you will most likely need to repeat questions 13 and 14 until you are satisfied with the transformation(s) you chose. Only then should you fill out this section - I only want to see the model you end up choosing, not all of your attempted models.)

< your response here >

```
# <your code here>

boxCox_Car <- car_speed_LM
```

```r
boxCox(boxCox_Car)
```

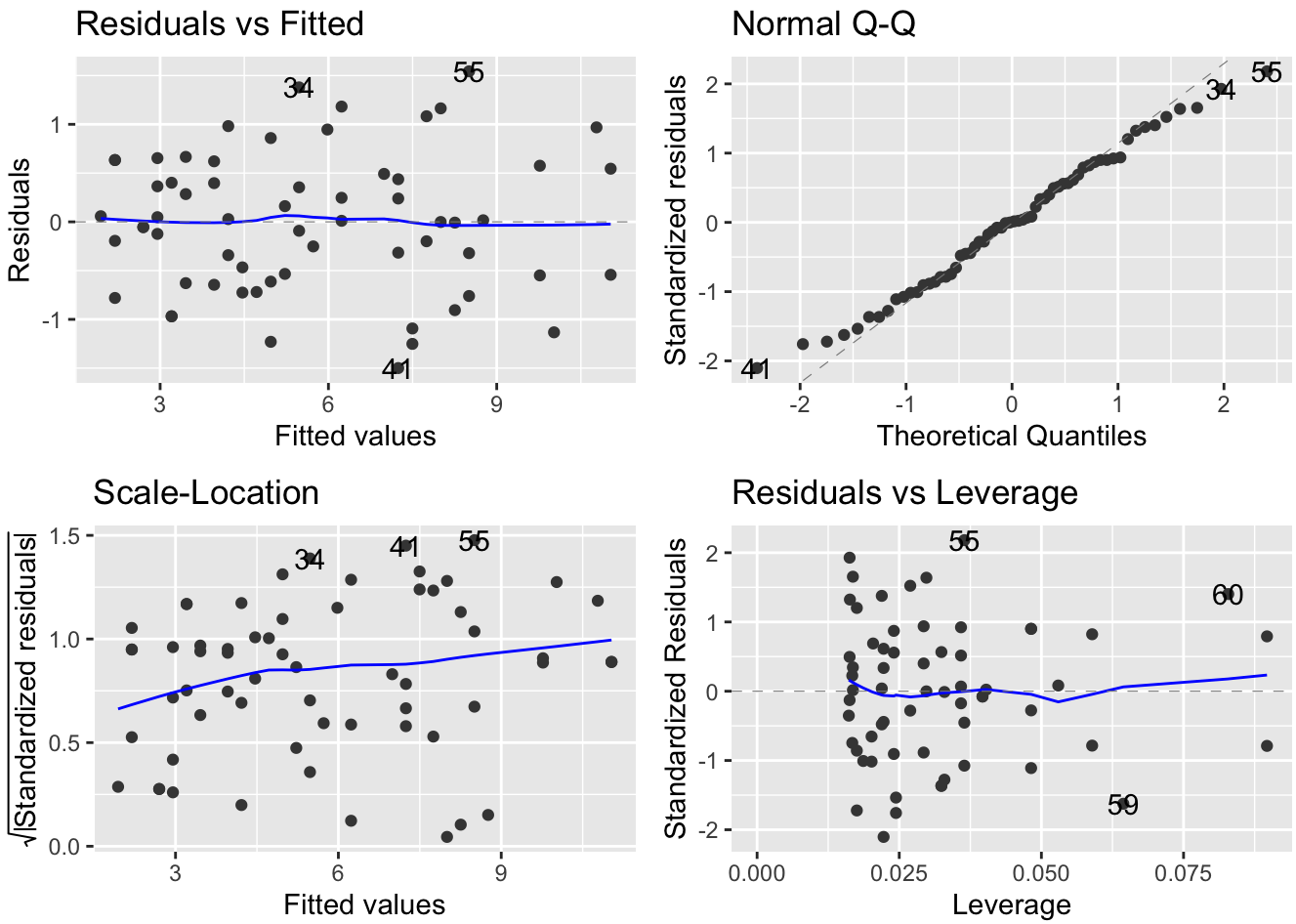



```r
new_linear_Model <- lm(sqrt(Distance)~((Speed)), data = stop_data)

summary(new_linear_Model)
```

```
Call:
lm(formula = sqrt(Distance) ~ ((Speed)), data = stop_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49948 -0.54761  0.00469  0.53153  1.54350

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.932396   0.197909   4.711  1.5e-05 ***
Speed       0.252466   0.009274  27.223  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7209 on 60 degrees of freedom
Multiple R-squared:  0.9251,    Adjusted R-squared:  0.9239
F-statistic: 741.1 on 1 and 60 DF,  p-value: < 2.2e-16
```

We began our transformation by running the boxBox test. Which will inform us of the lamda values that should work best for us. Although the result do not reach zero, they are extremely close, allowing us to use zero for our transformations. This means that we can experiment with taking the log of our data. I began by first taking only the log of x, then y, and finally both x and y together. On their own, just x or y was not promising, however, taking the log of both looked promising when. creating a plot with the new model. However, after checking our regression assumption against the new model with both the log of x and y, there were a number that failed, mainly residuals vs fitted values to check for linearity. After this, I began experimenting with transforming the data with the square roots of our x's and y's. I found that taking the square root of our x, being distance, provided the best solution to our model to fit our assumptions after running a number of tests to recheck if our assumptions had been violeted. The option of taking the square root of distance proved the most effective. #### 14. Now, re-check your transformed model and verify that the assumptions (the assumptions that were addressed in questions 7 to 11 above) are met. Provide a brief discussion about how each of the previously violated assumptions are now satisfied. Also, provide the code you used to assess adherence to the assumptions. (Note that transforming will not change your responses about (I) the residuals being independent, so you can skip that assumption here.

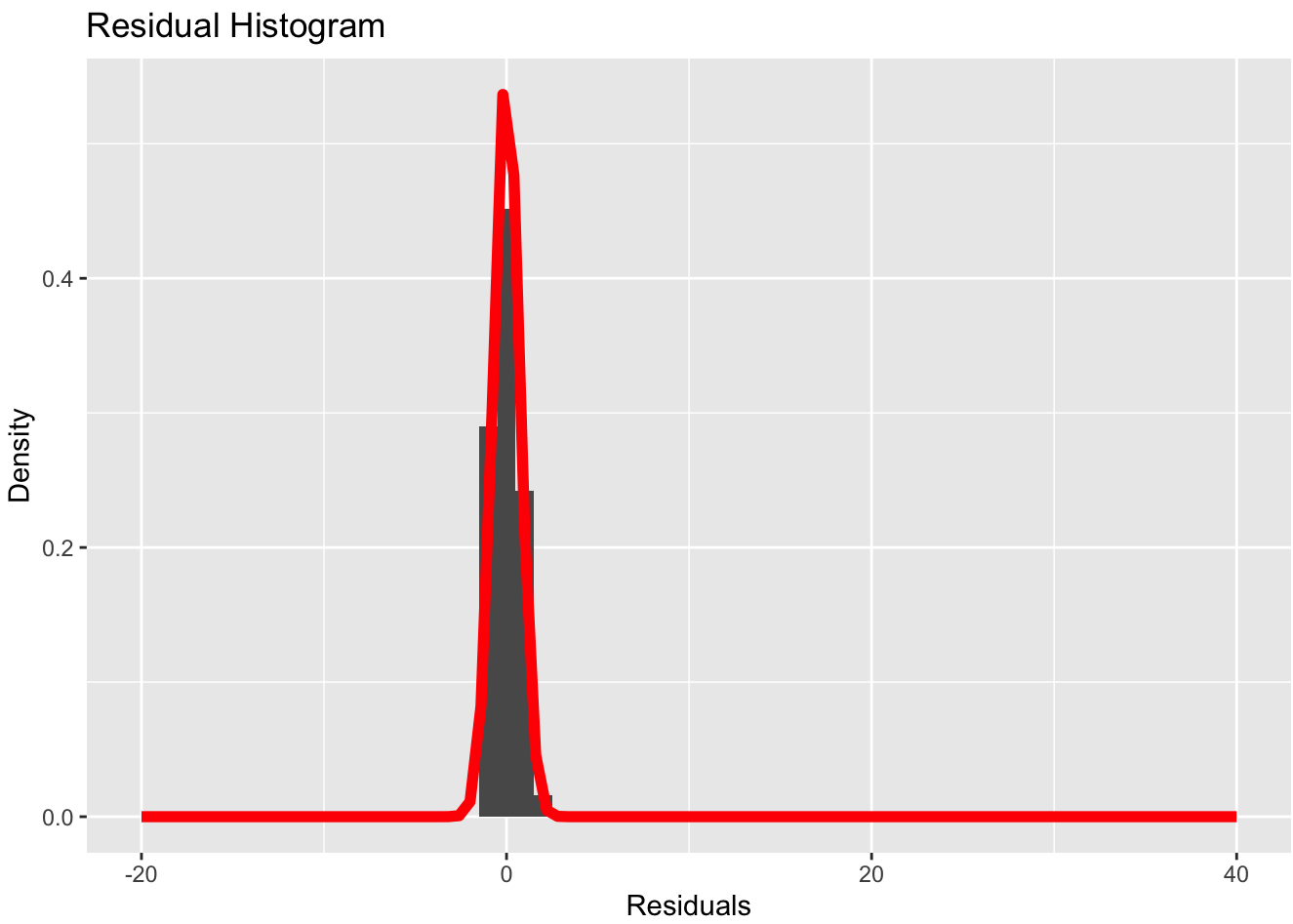
```
#7 X vs Y is linear


autoplot(new_linear_Model)
```



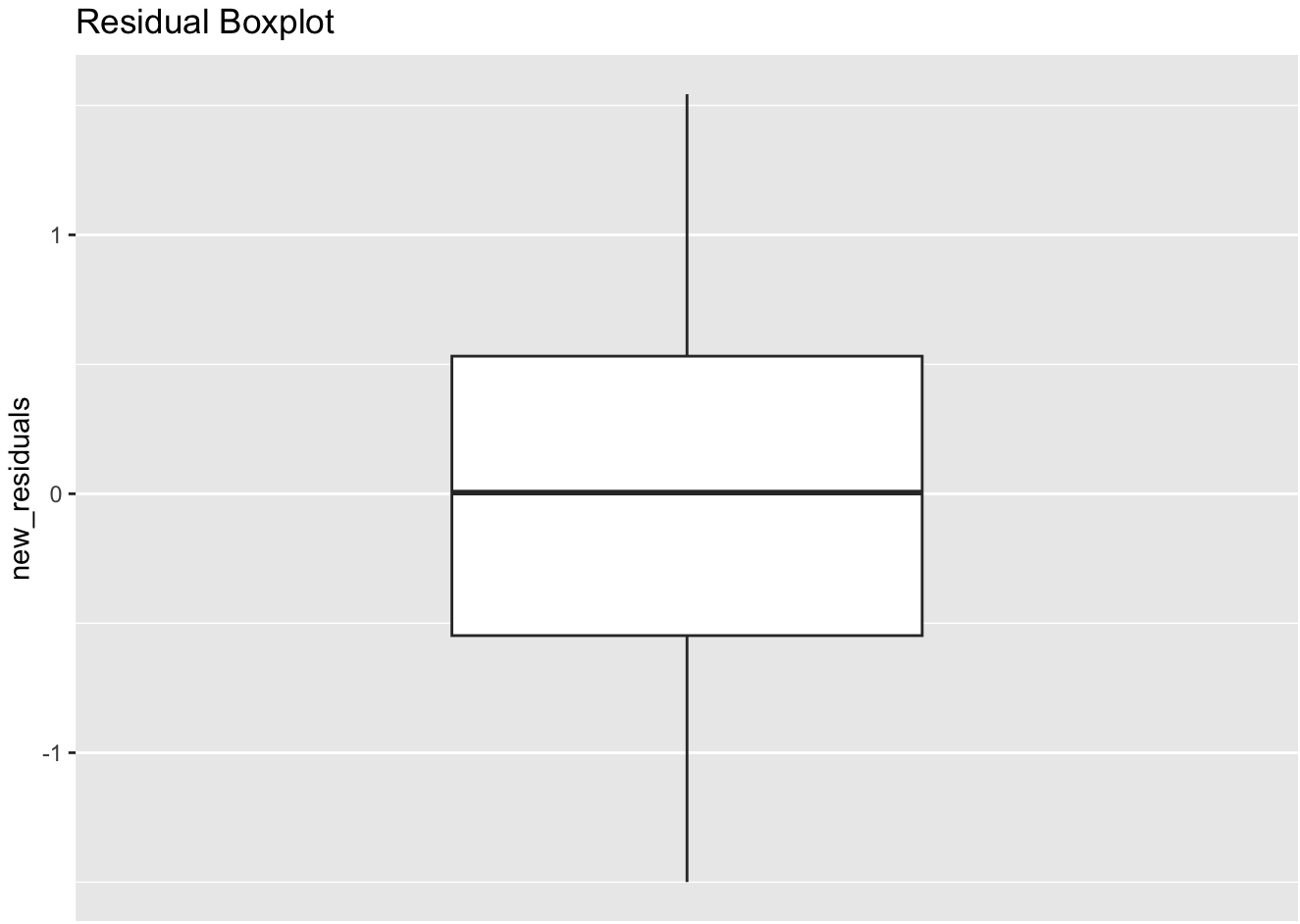

```
#9
```

```
#our histogram

new_residuals <- new_linear_Model$residuals
new_fits <- new_linear_Model$fitted.values

ggplot(data = stop_data, mapping = aes(x = new_residuals)) +
  geom_histogram(mapping = aes(y = after_stat(density)),
                 binwidth = 1) +
  stat_function(fun = dnorm, color = "red", linewidth = 2,
                args = list(mean = mean(new_residuals),
                            sd = sd(new_residuals))) +
  labs(x = "Residuals", y = "Density", title = "Residual Histogram") +
  scale_x_continuous(limits = c(-20, 40))
```

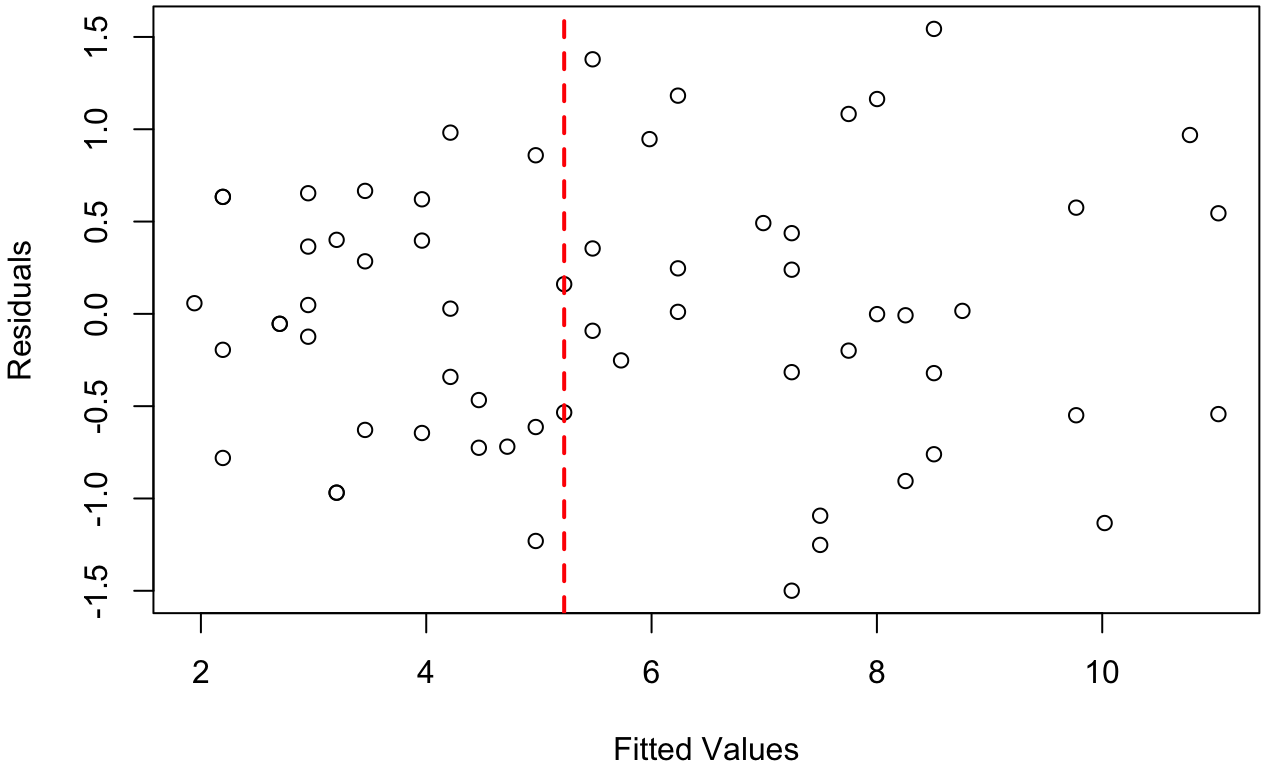
```
Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



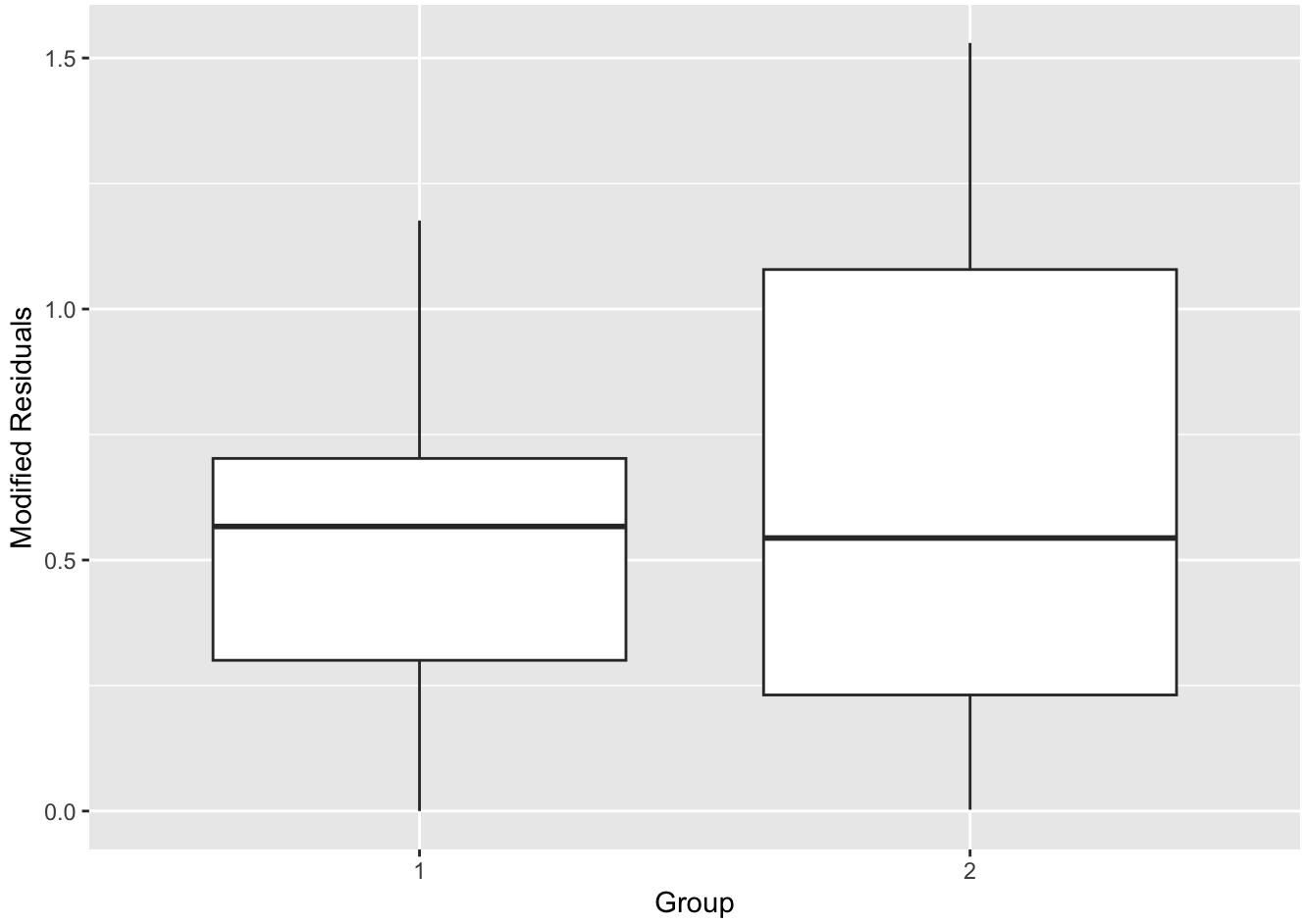

```
#our Residual Box plot

ggplot(data = stop_data, mapping = aes(y = new_residuals)) +
  geom_boxplot() +
  scale_x_discrete() +
  labs(title = "Residual Boxplot")
```




```
#10 the residuals have equal/constant variance across all values of x



stop_data$resids <- new_linear_Model$residuals
stop_data$fits <- new_linear_Model$fitted.values
stop_bw <- stop_data |>
arrange(fits) |>
mutate(index = 1:nrow(stop_data),
e_group = ifelse(index < nrow(stop_data)/2, 1, 2) |>
as.factor()) |>
mutate(vals = abs(resids - median(resids)), .by = "e_group")
plot(new_fits, new_residuals, xlab = "Fitted Values", ylab = "Residuals")
abline(v = new_fits[round(nrow(stop_bw)/2)],
col = "red", lty = 2, lwd = 2)
```




```
stop_bw |>
ggplot(aes(e_group, vals)) +
geom_boxplot() +
labs(x = "Group", y = "Modified Residuals")
```



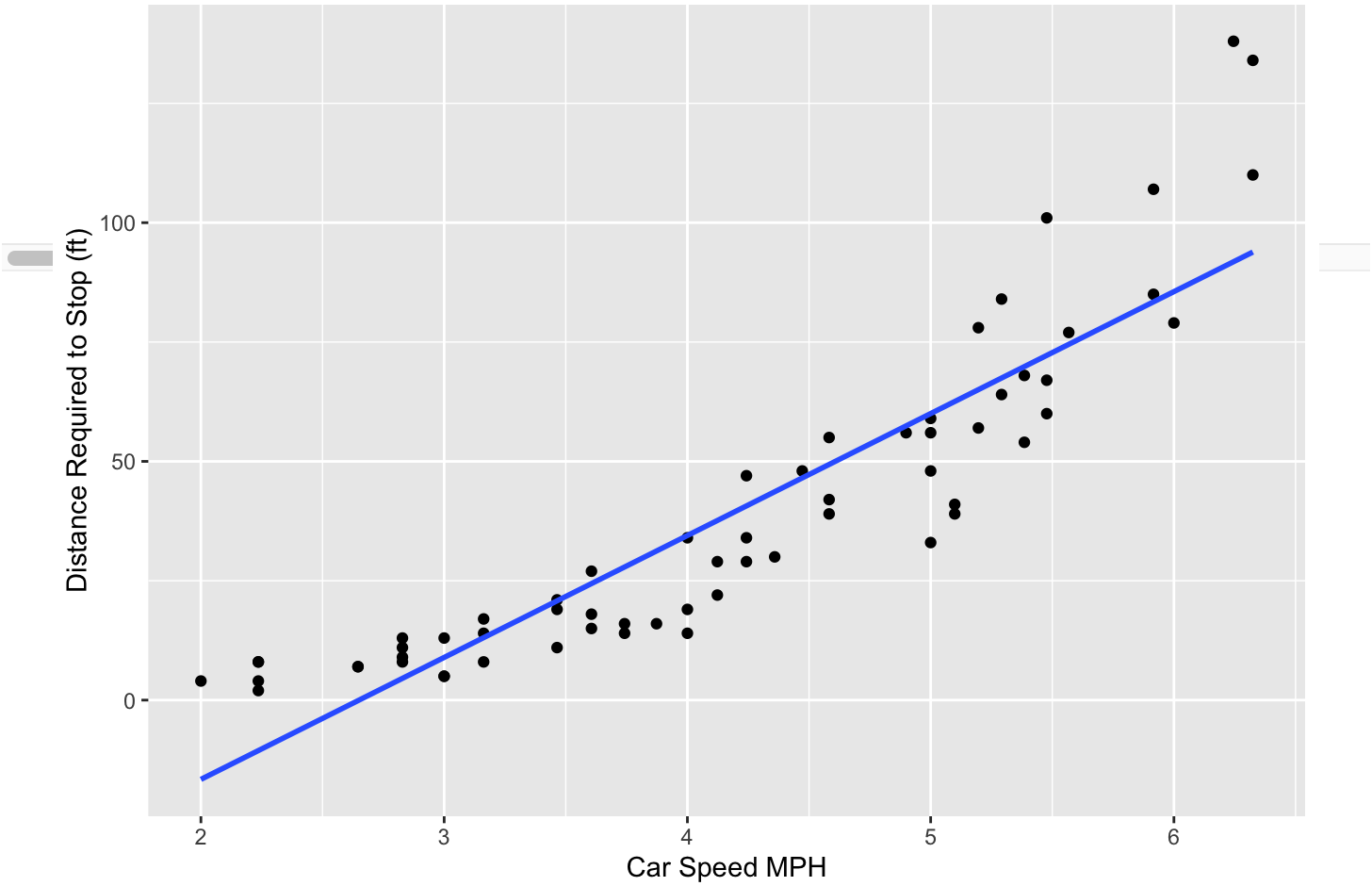

```
bf.test(vals ~ e_group, stop_bw)
```

```
  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : vals and e_group

  statistic  : 1.224225
  num df     : 1
  denom df   : 54.30738
  p.value    : 0.2734103

  Result     : Difference is not statistically significant.
-------------------------------------------------------------
```

```
#11

#Any additional infuential points deserving of attention


transformed_Scatter_plot<- ggplot(data = stop_data, mapping = aes(x = sqrt(Speed), y = (Dista
  geom_point() +

  labs(x = "Car Speed MPH", y = "Distance Required to Stop (ft)")
```

```
transformed_Scatter_plot + geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



For assumption on question 7, X vs Y is linear, we can use the autplot function and see a now beautiful Fitted Values Vs Residuals plot showing evenly distributed data and our fitted line following nicely along 0. Showing that they are now linear.

We can skip 8 based off of our instructions.

For number 9, we create our histogram and boxplot with the residuals from our new model and can see that the residuals are now nicely normally distributed with no major skew to the left or right.Fullfiling our needed assumption that the residuals be normally distributed.

For number 10, we can re run the Brown-Forsythe Test with our newly transformed data and see that our residuals now have constant variance across all values of x. If we compare with our previous Brown-Forsythe Test, we see that there is much more clustering around the 0-25 mark, but with the new data, it is much more even. We can even now see hard evidence for this in our new P value of 0.2734103 which allows us to accept the null hypothesis that variance is constant accross all values of x for our residuals.

For 11, our new model is very similar to our original model when we look at it at face value, the resolving of violated assumptions cannot necessarily be seen by just looking at the new scatter plot and OLS line. There is unfortunately still a few outliers new the 40 MPH mark, which we were not able to resolve. However,

considering the vast improvements from our transformations, this model now serves our purpose if we are mindful of the outliers.

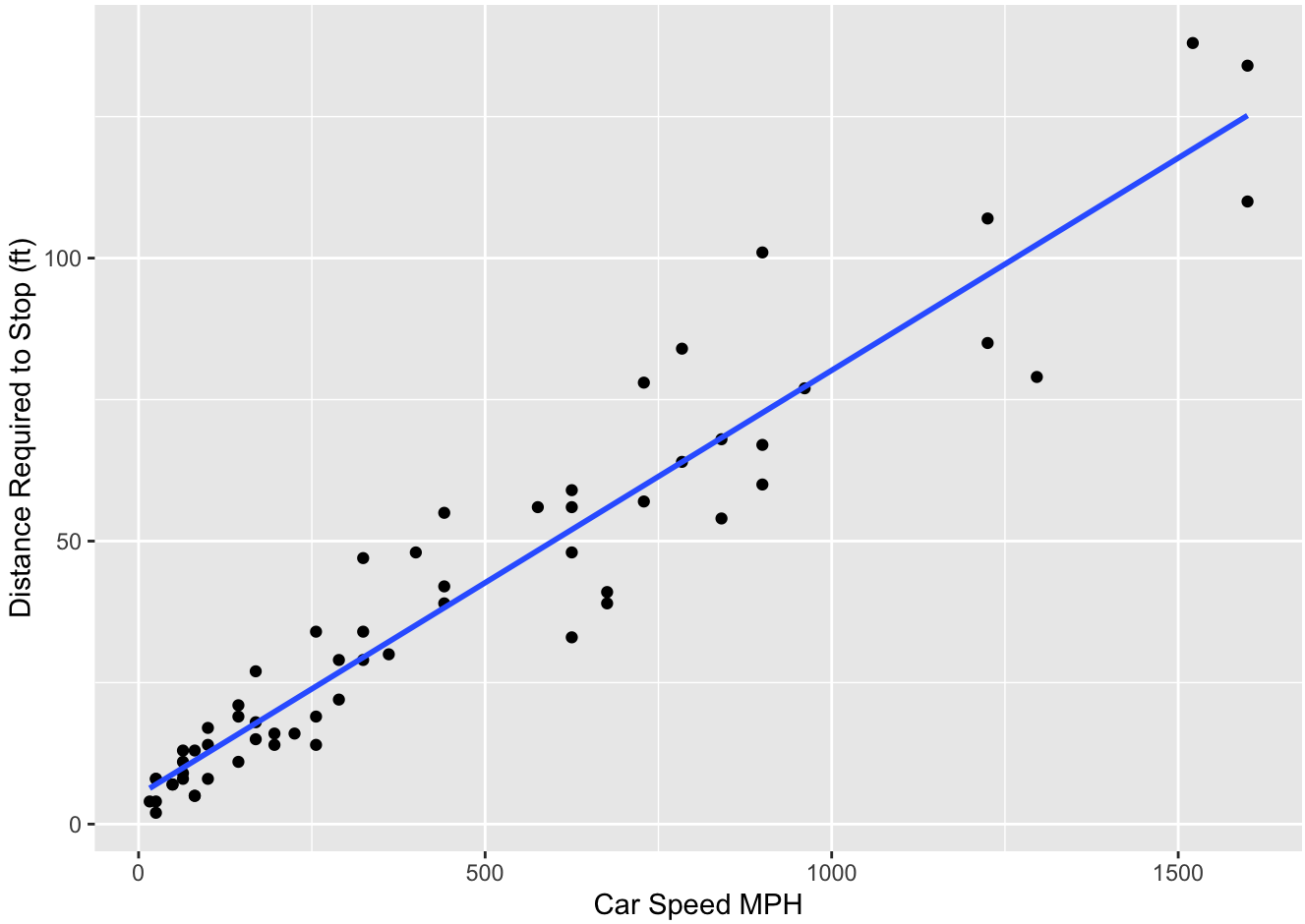
15. Mathematically write out the fitted simple linear regression model for this data set using the coefficients you found above from your transformed model. Do not use “x” and “y” in your model - use variable names that are fairly descriptive.

Estimated_StopDistance_Given_CarSpeed = 0.932396 + 0.252466 × Car_Speed

16. Plot your new fitted *curve* on the scatterplot of the original data (on the original scale - not the transformed scale). Do you think this model fits the data better than the original model?

```
# Tip: To get the fitted curve for your transformed model on the original scale, you have to
# <your code here>


original_Scatte_with_new_OLS <- ggplot(data = stop_data, mapping = aes(x = (Speed)^2, y = Dis
  geom_point() +

  labs(x = "Car Speed MPH", y = "Distance Required to Stop (ft)") + geom_smooth(method = "lm"

original_Scatte_with_new_OLS
```

```
`geom_smooth()` using formula = 'y ~ x'
```



This model most definitely fits the data better. It's a much straighter cut through our data and we no longer have as much curvature. Allowing a much more straight forward prediction at any x value within our range. This along with the fixed assumptions that were previously violeted, this transformed model serves us much better.

17. Briefly summarize (1) the purpose of this data set and analysis and (2) what you learned about this data set from your analysis. Write your response as if you were addressing a non-statistician (do not include any numbers or software output).

#1 Sometimes, our data is spread out in such a way that it is easy to perform the type of analysis we wish in order to find the types of relationships in the data we're looking for. However, sometimes the data is spread out in such a way that we cannot immediately performing our analysis, we have to make adjustments to our data. Now, we are not necessarily "changing" our data, rather, reformatting it in such a way that we can confidently perform our tests and draw conclusions. That is what happened with this dataset. #2. I learned from this dataset that although the data may look good on face value, i.e, linearity with just minor curvature, we still need to inspect our data to ensure assumptions aren't met. That was the biggest thing for me with this homework, without doing the violation checks, I would have just assumed we could go straight to the lm function because the scatter plot looked good to me. We have to be very critical of the steps we take before we draw conclusions. From a more statistical analysis, for every unit of change in speed of a car, there is a change of 0.252466 in the distance needed to stop the car.

Thanks!