DATA 605 - Homework #11

Zach Alexander

4/18/2020

> Using the "cars" dataset in R, build a linear model for stopping distance as a function of speed and replicate the analysis of your textbook chapter 3 (visualization, quality evaluation of the model, and residual analysis)

Downloading the Cars Dataset from R

Since R has this preloaded dataset ready to use, we can simply do the following to take a quick look at the first 10 rows of the `cars` dataset:

```
kable(head(cars, n = 10L)) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

speed	dist
4	2
4	10
7	4
7	22
8	16
9	10
10	18
10	26
10	34
11	17

For more information about this dataset, you can read some of the R documentation here. It's interesting to read that this data is from the 1920s, and mentions that it documents the speed of cars and the distances taken to stop.

Initial Analysis

Since the data is already cleaned and tidy'd, we can take an initial look at the dataset dimensions:

```
dim(cars)
```

```
## [1] 50  2
```

Our dataframe consists of 50 rows of speed and distance measurements, with two columns `speed` and the corresponding `dist`.

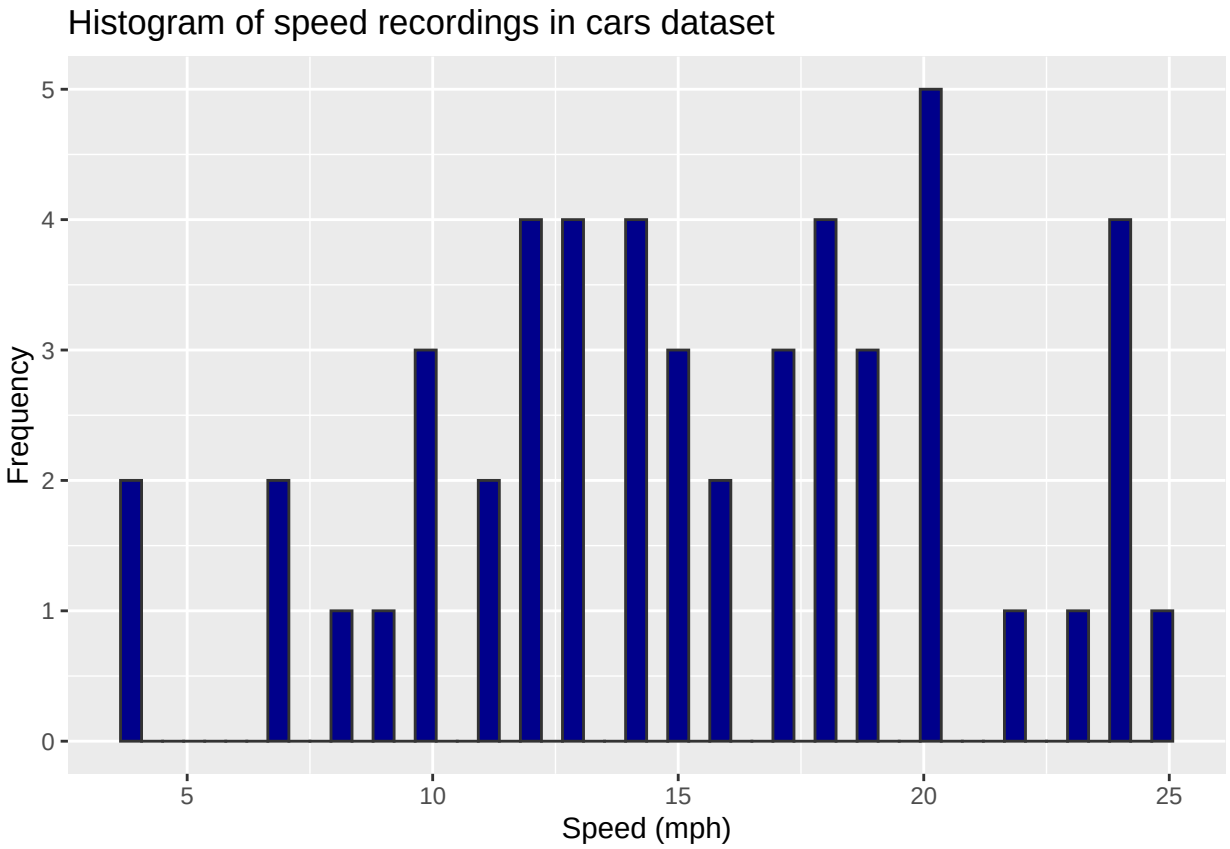
Next, we can take some summary statistics of the speed measurements:

```
summary(cars$speed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    12.0    15.0    15.4    19.0    25.0
```

We can see that the mean speed recording in our dataset is about 15.4 mph, with a median of 15 mph. From the summary statistics, this distribution seems pretty normal, with a pretty proportional range between the first and third quartiles and the mean and median being roughly equal.

We can plot the speeds on a histogram to better visualize the distribution:



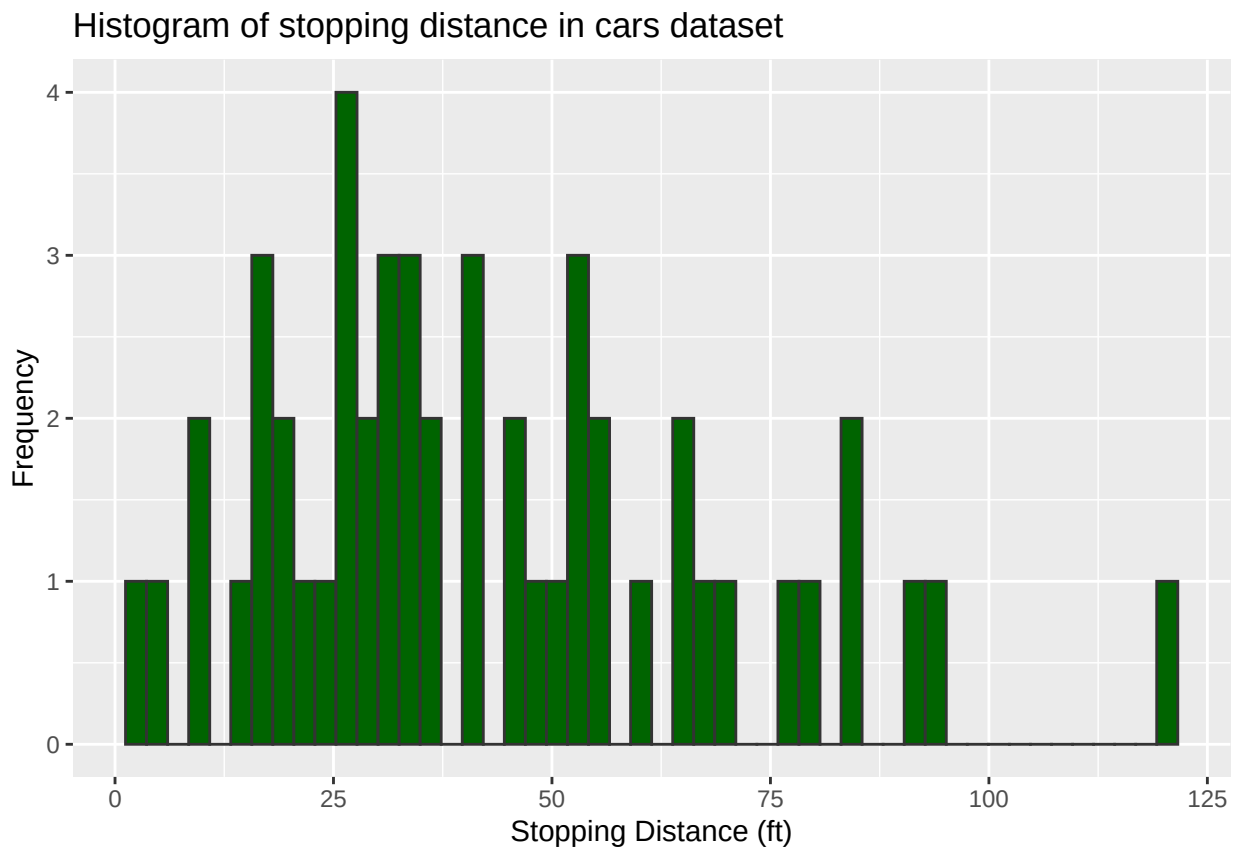
We'll do the same for the distance measurements. Here's the summary:

```
summary(cars$dist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   26.00   36.00   42.98   56.00  120.00
```

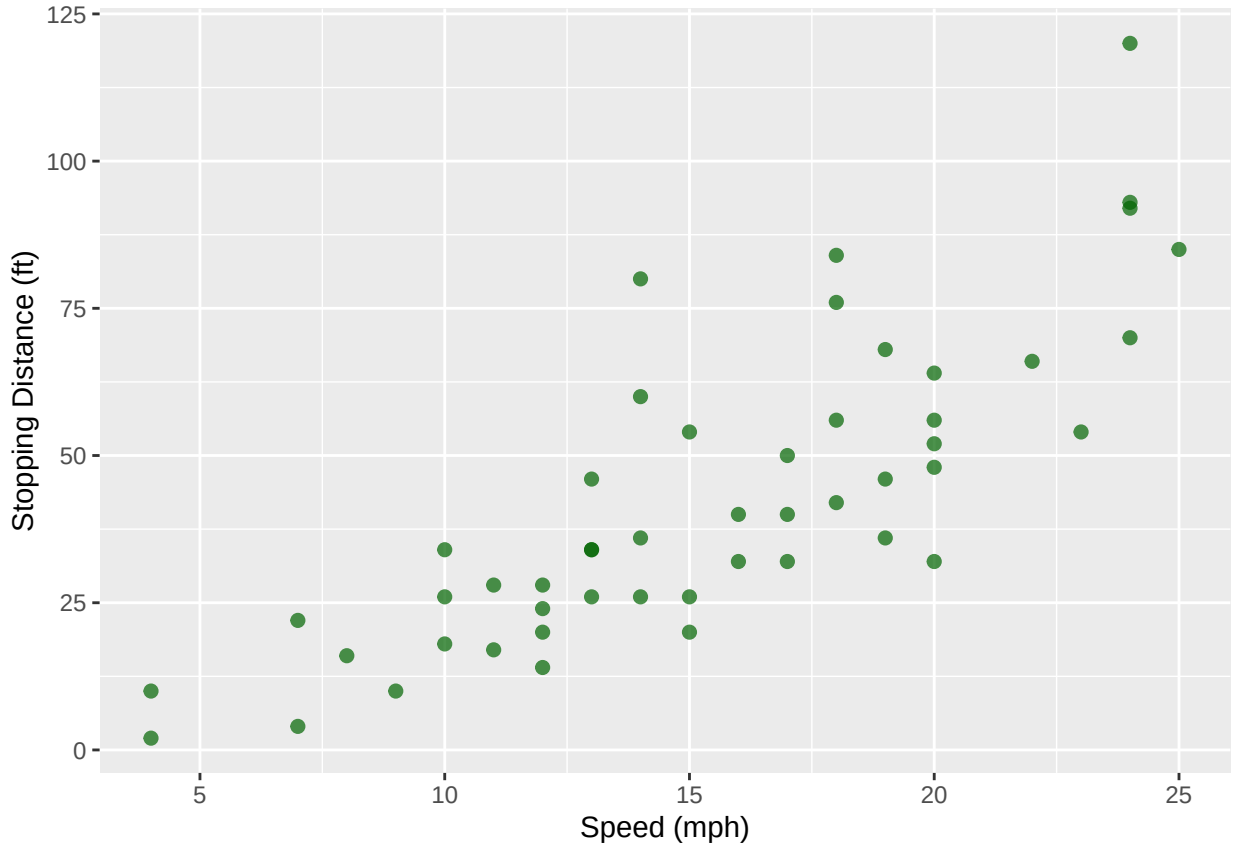
We can see that the mean distance recording in our dataset is about 42 ft, with a median of 36 ft. It does appear that we may have a few outliers, since 75% of the distance measurements fall below 57 ft and we have a maximum distance of 120 mph.

This can also be seen when we plot a histogram of stopping distance measurements:



We can also visualize both variables on the same plot:

```
ggplot(cars, aes(x=speed, y=dist)) +
  geom_point(size=2, color="darkgreen", alpha=0.7) +
  xlab("Speed (mph)") +
  ylab("Stopping Distance (ft)")
```



Initially, it looks like there may be a positive, linear relationship between speed and stopping distance, however we can investigate further by creating a linear regression.

Building the One-Factor Linear Regression

With our factors ready to go, we can create a linear regression model.

```
cars_lm <- lm(cars$dist ~ cars$speed)
cars_lm
```

```
##
## Call:
## lm(formula = cars$dist ~ cars$speed)
##
## Coefficients:
## (Intercept)   cars$speed
##     -17.579        3.932
```

Above, we can see the intercept and slope of our linear regression (-17.579 and 3.932 respectively). To get a more detailed outlook of the performance of our model, we can use the `summary()` function in R:

```
summary(cars_lm)
```

```
##
## Call:
```

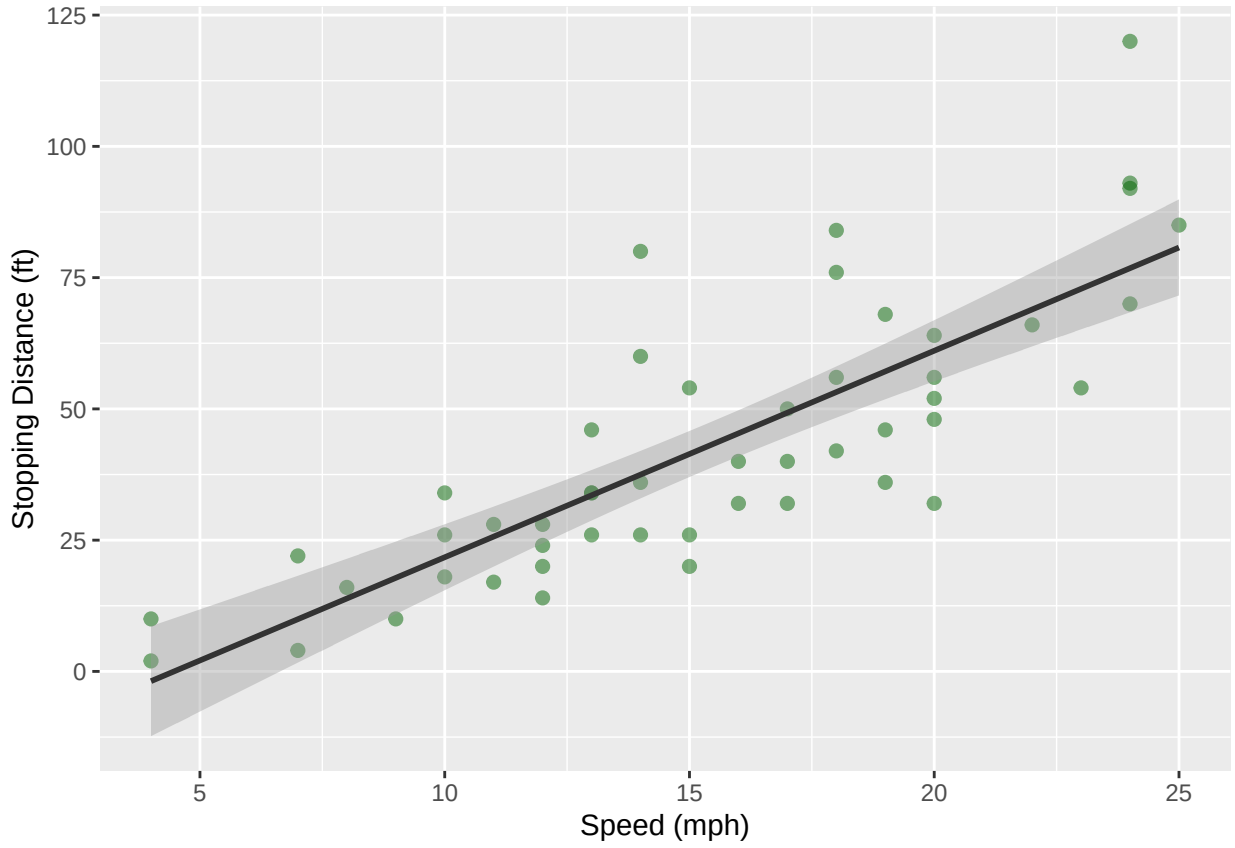
```
## lm(formula = cars$dist ~ cars$speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Evaluating the Quality of the Model and Residual Analysis

After running the summary above, we can see a few things:

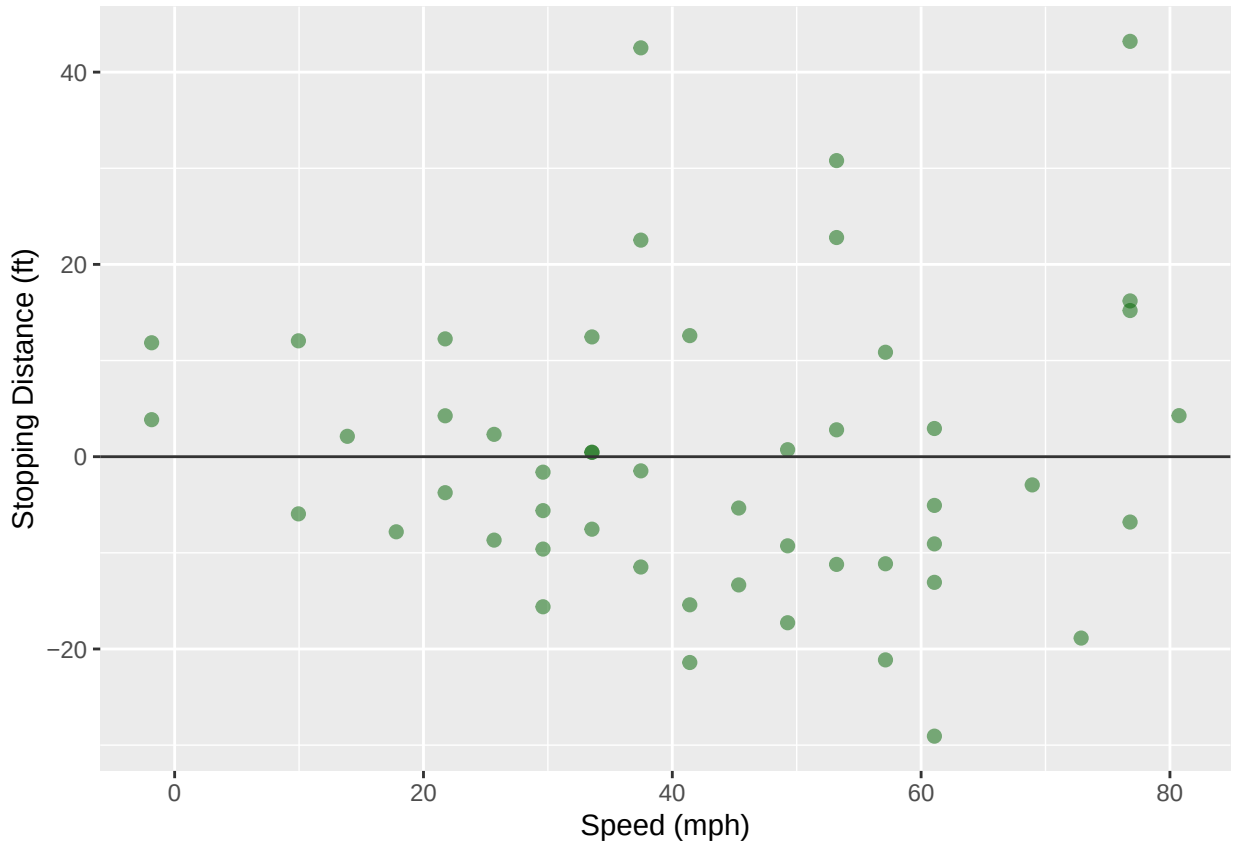
- The median residual value is around zero, which is a good sign (it's at -2, but not too far away from zero).
- Additionally, the minimum and maximum values of the residuals are somewhat around the same magnitude (roughly -30 and 43 respectively)
- The first and third quartile values are almost exactly the same magnitude, which is a good sign (-9 and 9 respectively)
- The standard error appears to be at least five times smaller than the corresponding coefficient (3.932 / 0.04155 = 94.632)
- Our Multiple R-squared value is 0.6511, which indicates that speed accounts for about 65% of the variability in the stopping distance.

Here's a plot of our model, with the line shown in black:



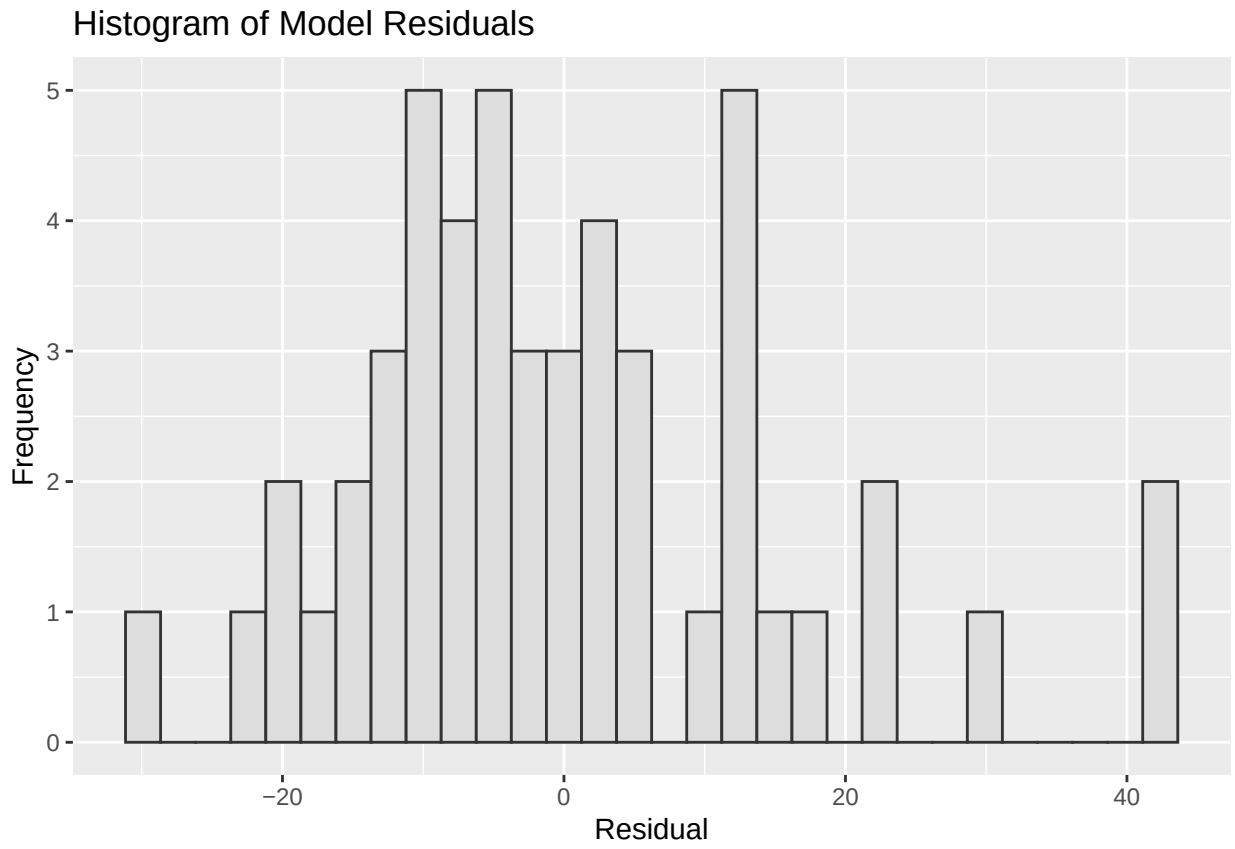
From above, we can see a pretty well-fit line, and when plotting residuals (below), we can confirm that most seem to be uniformly scattered above and below zero:

```
ggplot(cars, aes(x=fitted(cars_lm), y=resid(cars_lm))) +
  geom_point(size=2, color="darkgreen", alpha=0.5) +
  xlab("Speed (mph)") +
  ylab("Stopping Distance (ft)") +
  geom_hline(yintercept=0, color="#333333")
```

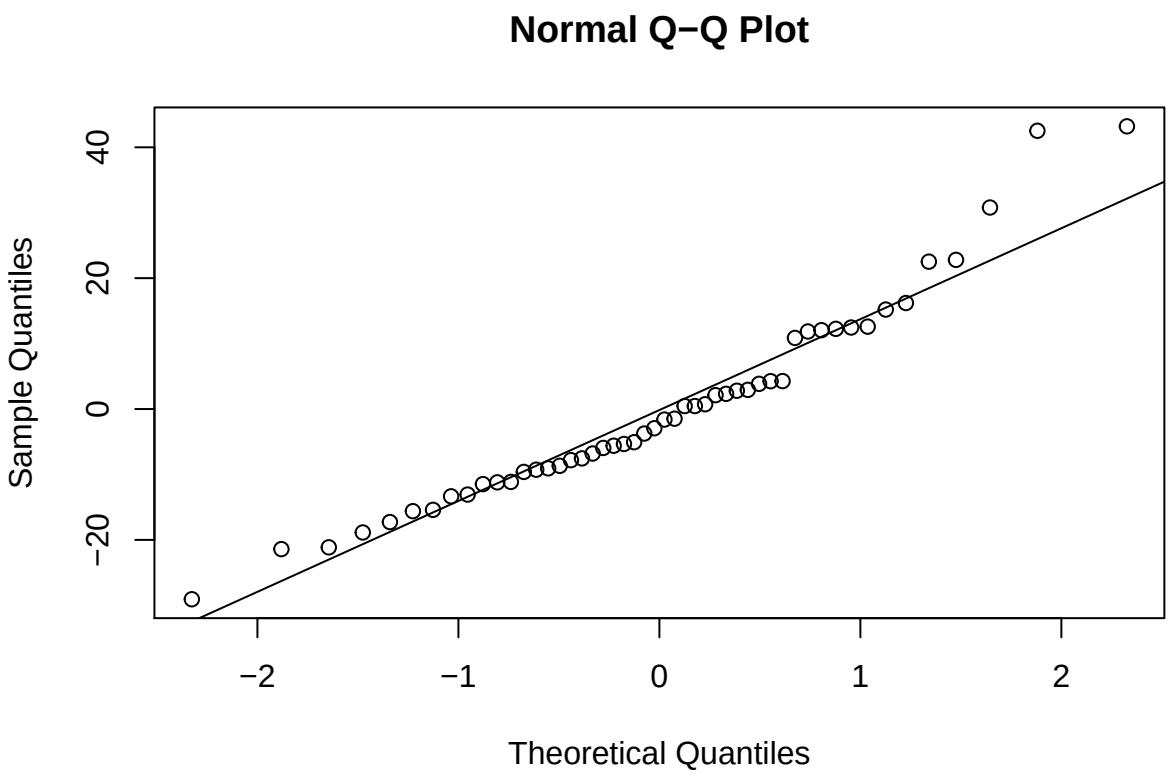


This can also be visualized in a basic histogram, or a quantile-versus-quantile (Q-Q) plot (see below):

```
p<-ggplot(cars, aes(x=resid(cars_lm))) +
  geom_histogram(color="#333333", fill="#dddddd", bins = 30)
p<- p + ggtitle("Histogram of Model Residuals") +
  xlab("Residual") + ylab("Frequency")
p
```




```
qqnorm(resid(cars_lm))
qqline(resid(cars_lm))
```



It does appear from these plots that the residuals are pretty normally distributed, since most points plotted on the Q-Q plot follow the straight line (with a few exceptions towards the two ends). The tails could indicate that they are slightly "heavier" than what we would expect form a normal distribution, but when plotting the residuals as a histogram, the residuals look only slightly right skewed.

Was the Linear Model Appropriate?

From this residual analysis, as well as assessing the quality of the model, we can conclude that using speed to predict the stopping distance of a car is a pretty good start, and would work as an okay predictor. However, there are many other factors that come into play as to why a car would stop after a certain distance (i.e. weather, road conditions, type of pavement/road, driver tendencies, etc.). Therefore, it would be worthwhile to add factors other than speed (or in addition to speed) to our model to see if we can create an even better prediction of stopping distance.