DATA 605 : Week 11 - Linear Regression Model

Ramnivas Singh

11/07/2021

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.)

```
head(cars)
```

Dataset : The "cars" dataset in R.

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

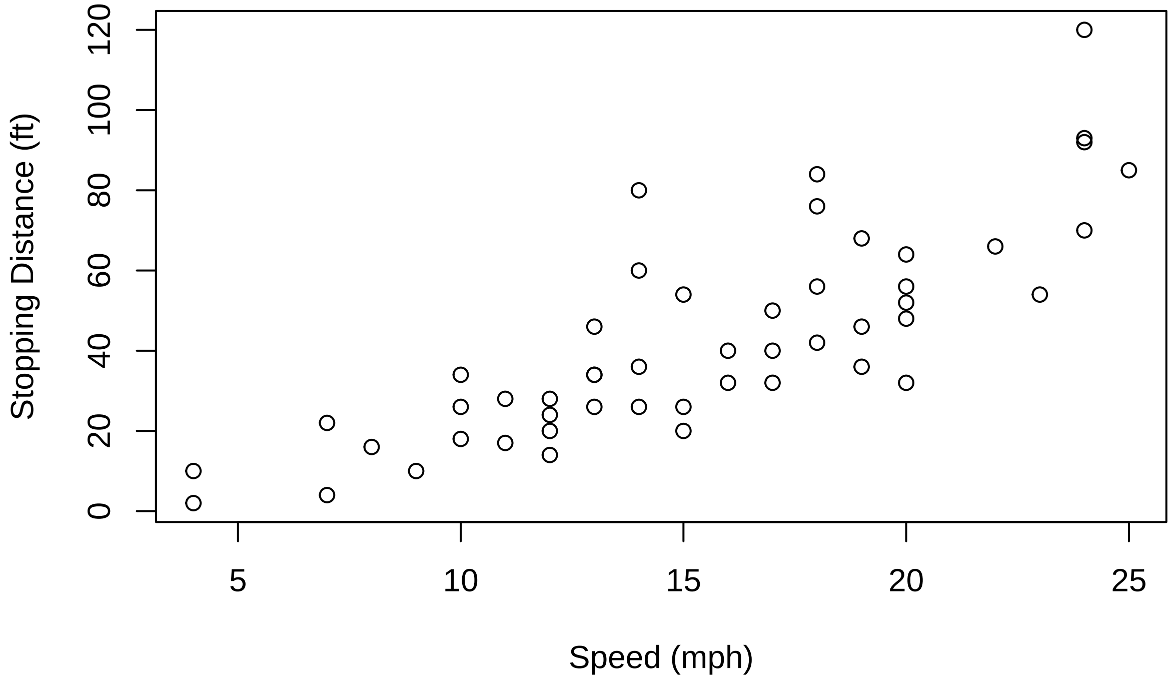
```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

This dataset has 50 rows and 2 columns. Each row is an observation that relates to a reading between car speed and the distance it takes for a car to stop. The columns in the dataset are "speed"" and "dist".

Plot speed and distance relationship, the plot shows that stopping distance increases as speed increases.

```
plot(cars$speed, cars$dist, xlab='Speed (mph)', ylab='Stopping Distance (ft)')
```



Lets create a regression model to show distance as a function of speed

```
lmod <- lm(dist ~ speed, data=cars)
summary(lmod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## 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 *
## 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
```

Lets generate Linear Model

```
cars.lm <- lm(dist ~ speed, data = cars)
summary(cars.lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## 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 *
## 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
```

Observations: The probability that the speed coefficient is not relevant in the model is 1.49e-12 (p-value), which means that speed is very relevant in modeling stopping distance.

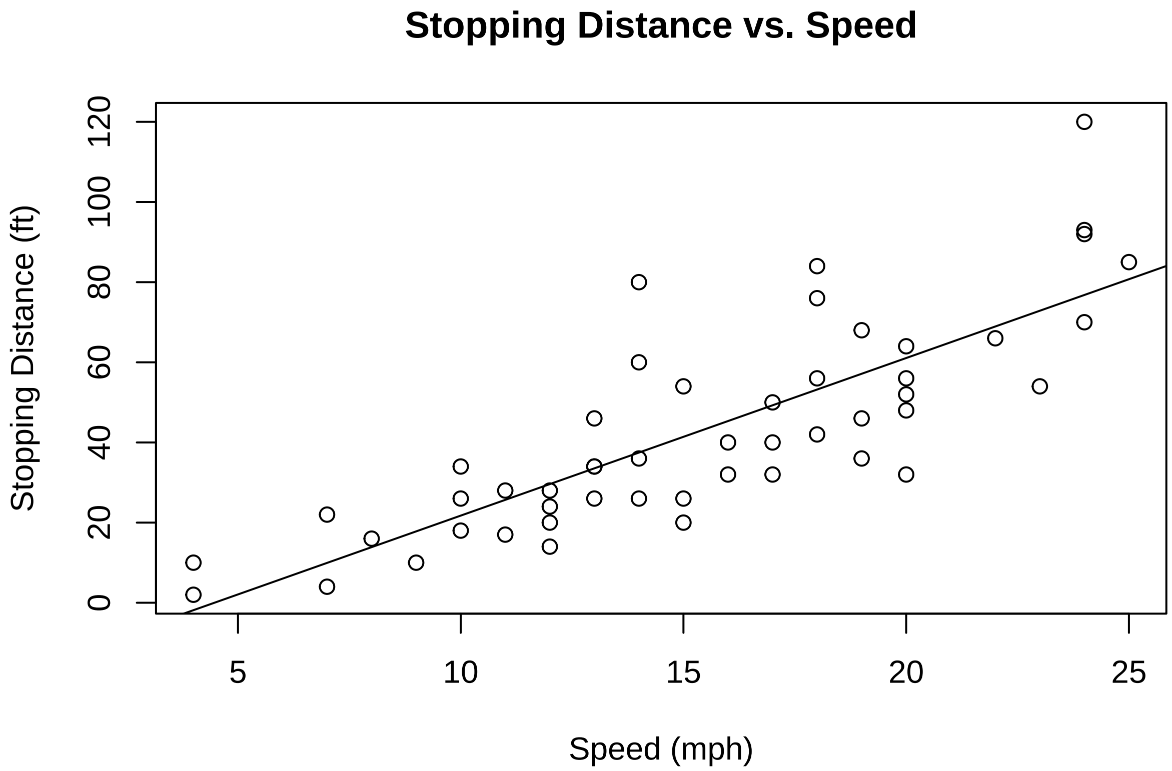
The p-value of the intercept is 0.0123, which means the intercept is pretty relevant in the model.

The multiple R-squared is 0.6511, which means that this model explains 65.11% of the data’s variation.

The standard error for the speed coefficient is ~ 9.4 (3.93/.42) times the coefficient value.

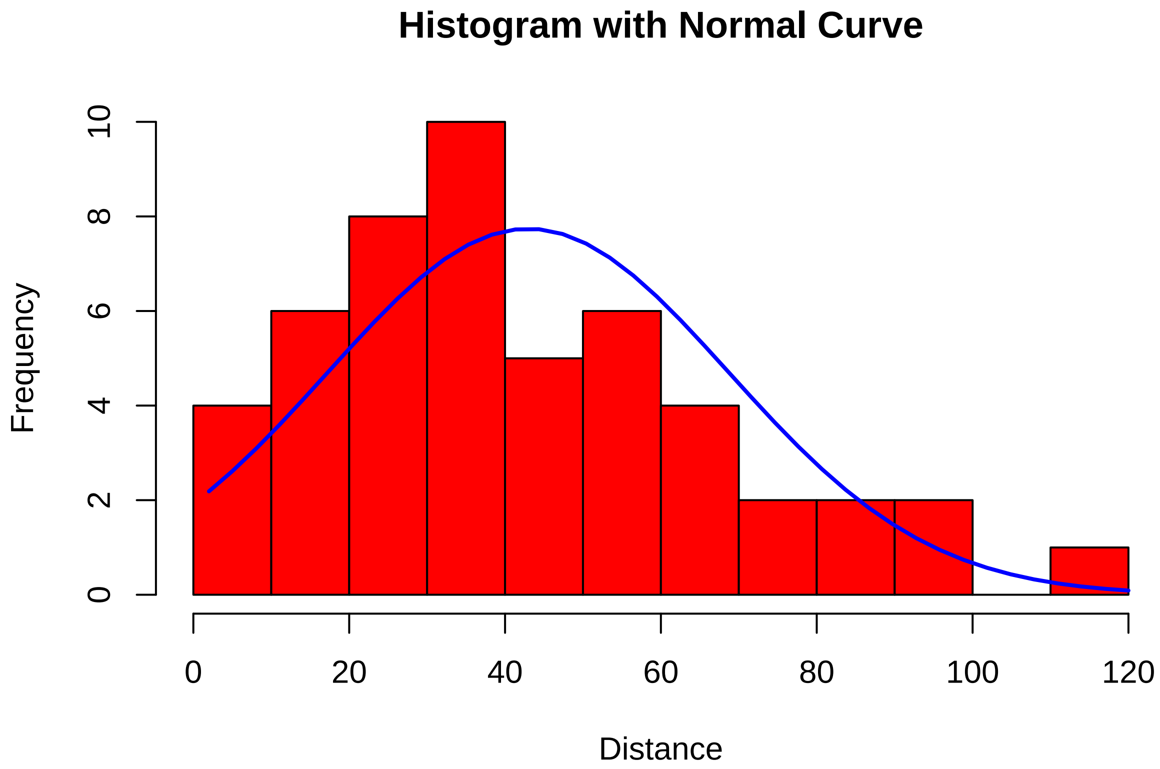
Now lets plot the Linear Model on the scatterplot to show distance as a function of speed

```
plot(cars$speed, cars$dist, xlab='Speed (mph)', ylab='Stopping Distance (ft)',
     main='Stopping Distance vs. Speed')
abline(lmod)
```



It seems that this model is a decent fit. It has an adjusted r squared of .6. Lets see how the distribution of our response variable holds against the normal distribution

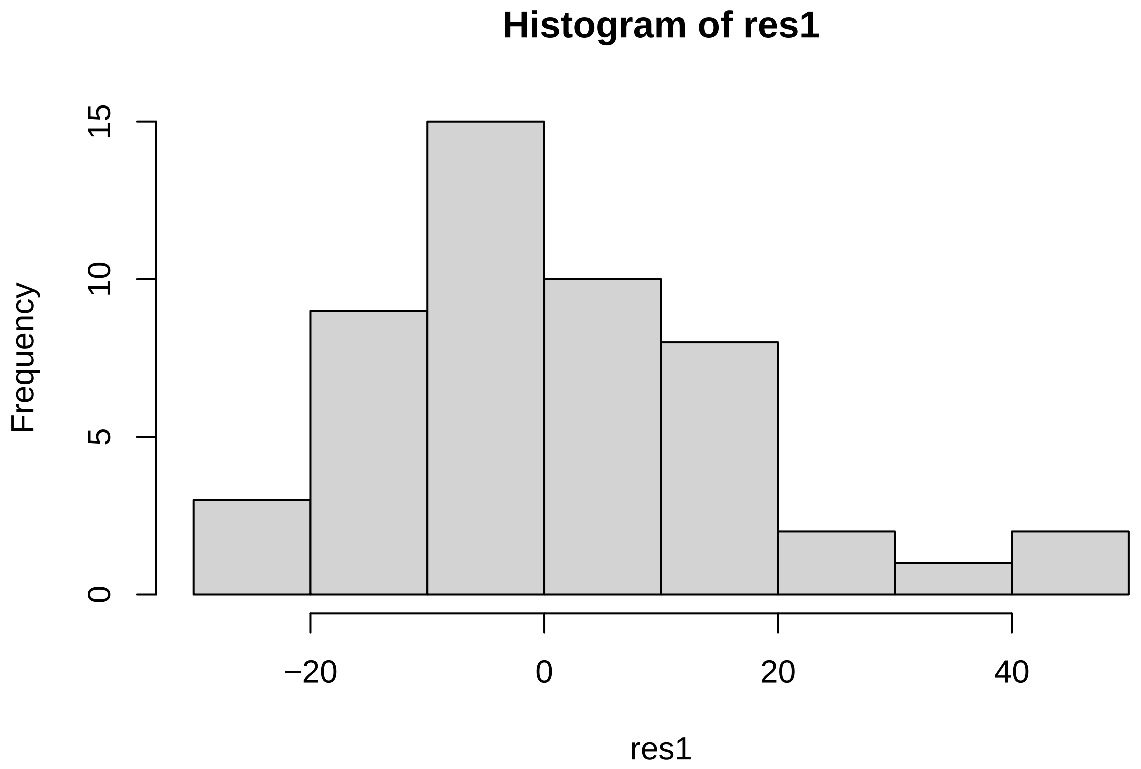
```
distance <- cars$dist
h<-hist(distance, breaks=10, col="red", xlab="Distance",
    main="Histogram with Normal Curve")
xfit<-seq(min(distance),max(distance),length=40)
yfit<-dnorm(xfit,mean=mean(distance),sd=sd(distance))
yfit <- yfit*diff(h$mids[1:2])*length(distance)
lines(xfit, yfit, col="blue", lwd=2)
```



Based on this chart, it seems as if though some sort of transform can help increase model accuracy.

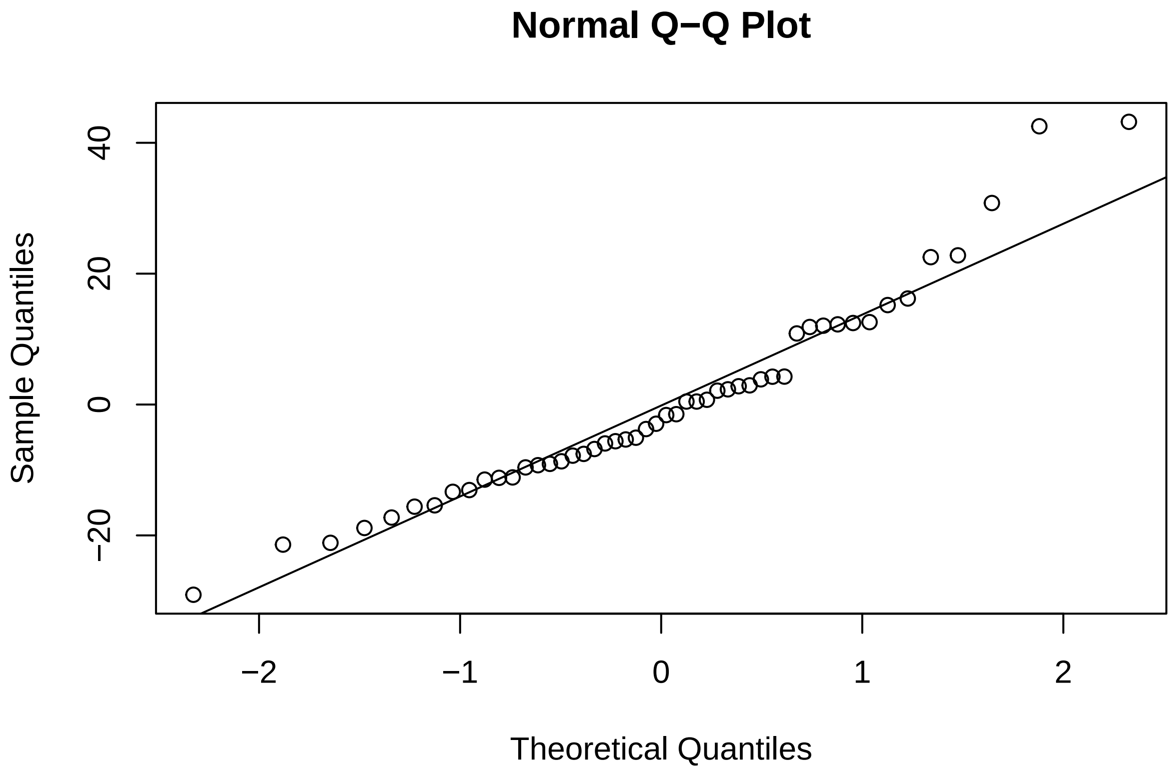
Lets run diagnostics on our model Residuals~There is a skew in our residuals as the spread does not exactly follow a normal distribution

```
res1<-lmod$residuals
hist(res1)
```



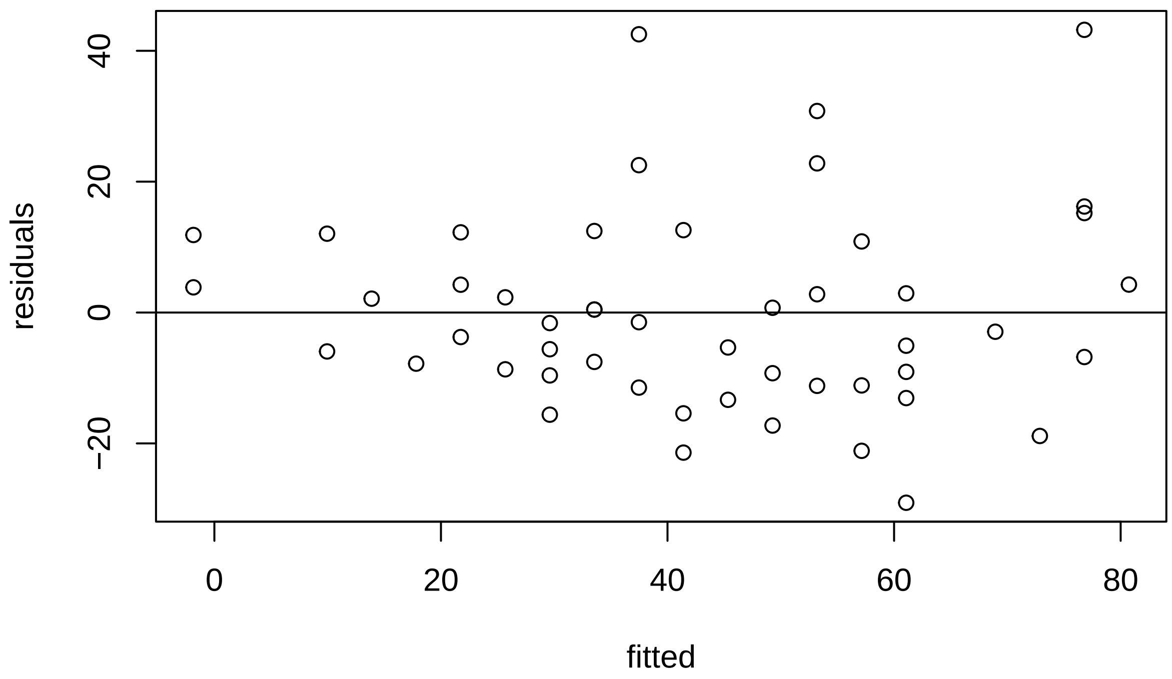
QQ-Plots

```
qqnorm(resid(cars.lm))
qqline(resid(cars.lm))
```

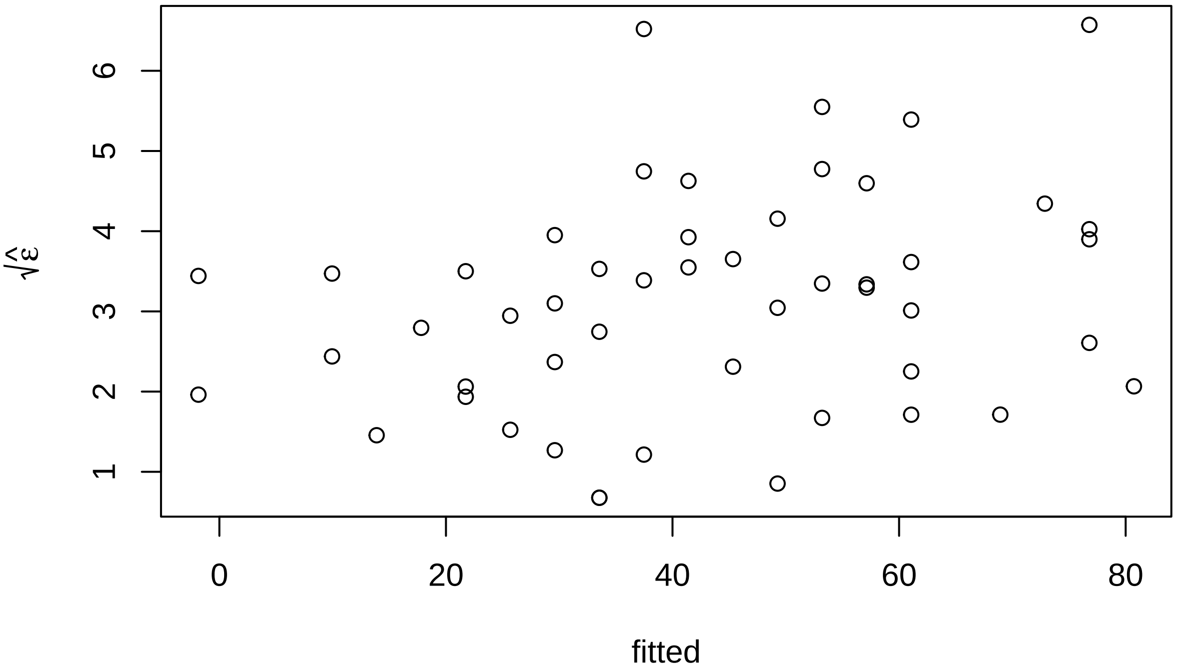


QQ-Plots seem to tell the same story. While the model we built is not bad, it can be better

```
plot(fitted(lmod), residuals(lmod), xlab="fitted", ylab="residuals")
abline(h=0);
```



```
plot(fitted(lmod), sqrt(abs(residuals(lmod))), xlab="fitted", ylab=expression(sqrt(hat(epsilon))))
```



It looks like variance is constant but one can also make the argument that there is some sort of curve. We need a more robust test. The square root of the residuals make a stronger visual argument that there is no constant variance.

Conclusion We found that the data has correlation 0.8069 with a Multiple R-squared value of 0.6511. Q-Q plot confirms that using only the speed as a predictor in the model is insufficient to explain the data. Therefore, we can say that there may be other factors like wind, tire type, fuel type also to be considered to accurately predict the stopping distance.