# 605-Wk11-Homework

*Jose Mawyin*

*11/9/2019*

**Linear regression model of stopping distance as a function of speed from the "Cars" dataset**

First, lets load the dataset cars that contains 50 observations of two variables. The independent variable is vehicle's speed and the response variable is the vehicle's stopping distance.

```
data("cars")
head(cars)
```

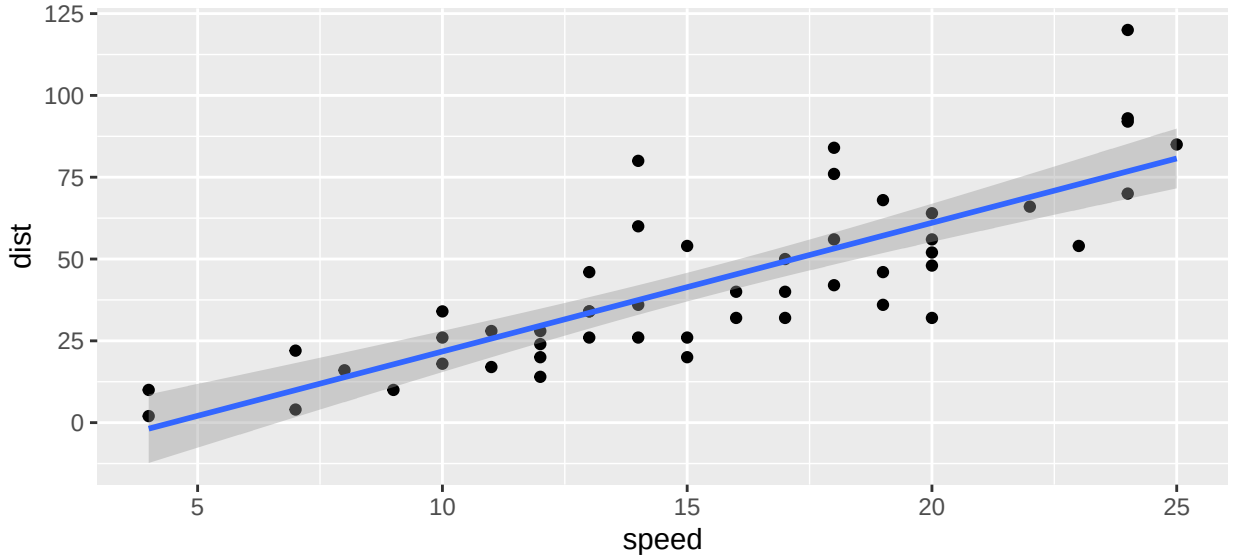
```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

We can easily create a linear fit regression model.

```
attach(cars)
Cars.lm <- lm(dist ~  speed)
```

Plotting a linear regression fit line we can see the speed vs. stopping distance response.

```
ggplot(cars, aes(x=speed, y=dist)) + geom_point() + geom_smooth(method='lm')
```



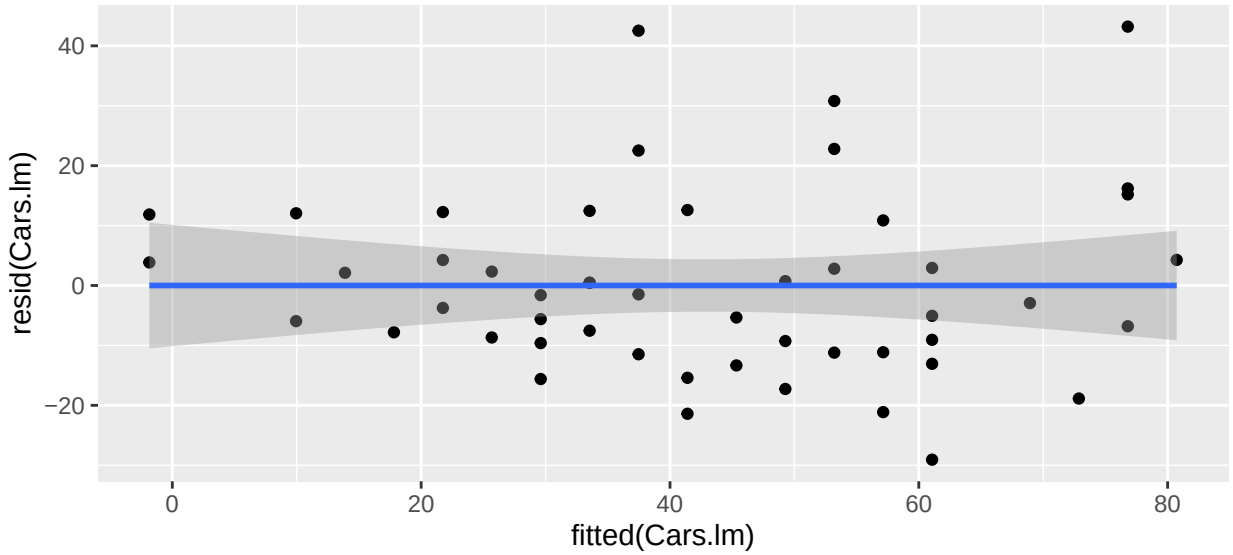

Looking into the statistics summary of the linear regression, we see that speed variable explains 65% of our data and that is has a P value (1.49e-12) well below the typical treshold of 0.05.

```r
summary(Cars.lm)
```

```
##
## Call:
## lm(formula = dist ~ 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 *
## 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
```

The residuals plot below show how the residuals are evenly distributed around the origin (residual = 0). We can visualize this using a linear fit line that in our case passes directly over the x-axis as shown below.

```r
ggplot(cars, aes(x=fitted(Cars.lm), y=resid(Cars.lm))) + geom_point() + geom_smooth(method='lm')
```

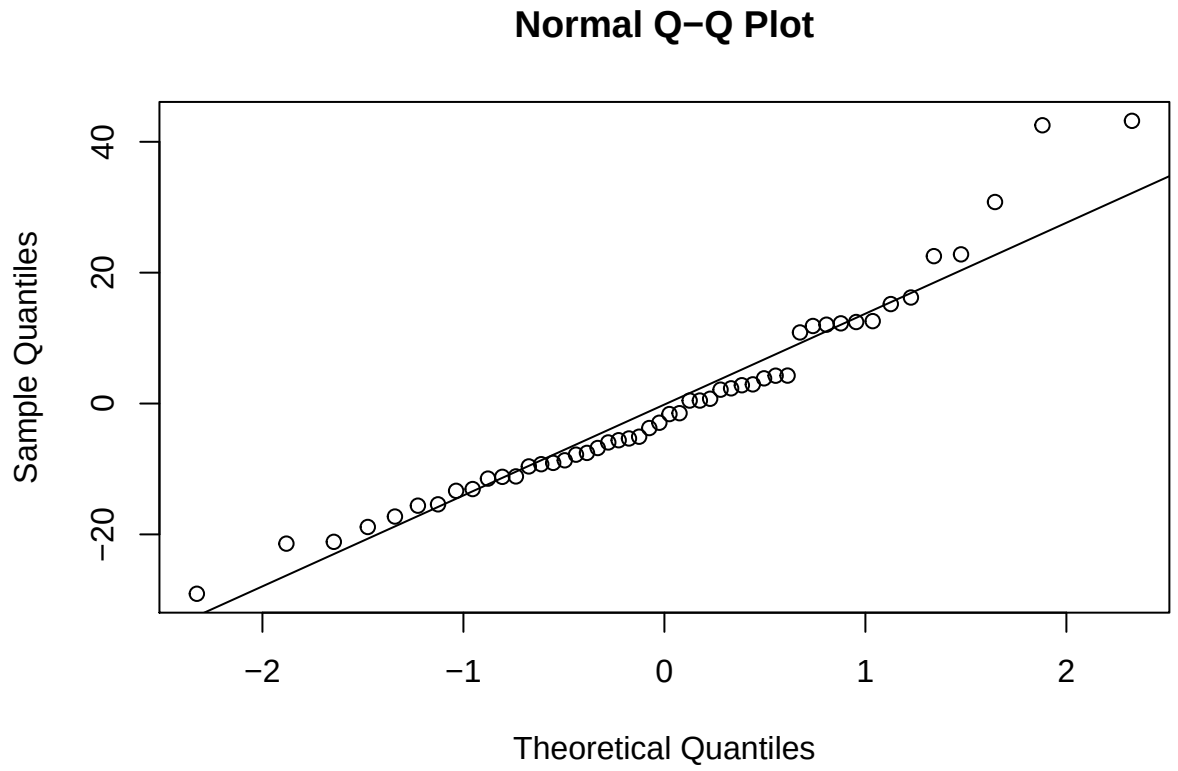



The residuals quantile-versus-quantile or Q-Q plot shows interesting results in that:

*. . . if the model fits the data well, we would expect the residuals to be normally (Gaussian) distributed around a mean of zero.*

We can see that in our Q-Q plot below that the residuals diverge around the tails indicating that the residuals are not normally distributed.

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




We can see this more clearly below in a density plot of the residuals. We can easily see how the distribution of the residuals is positively skewed.

```
ggplot(cars, aes(resid(Cars.lm))) + geom_density(kernel = "gaussian")
```