DATA605 - Assignment 11

Nick Oliver

Assignment 11

Question

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.)

Solution

Loaded the `tidyverse` library

View the data set to find the column names

```
glimpse(cars)
```

```
## Rows: 50
## Columns: 2
## $ speed <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 13~
## $ dist  <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, 26, 34~
```

Use the `lm` function to fit a linear regression model

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

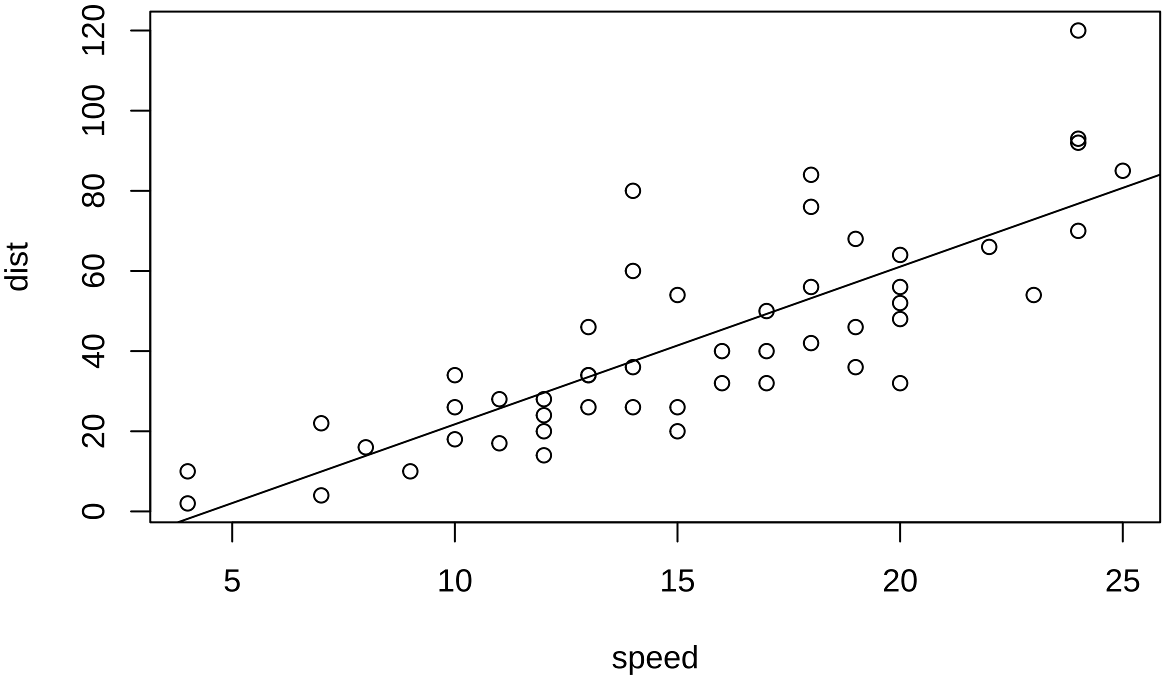
```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

The function to predict the stopping distance is $StoppingDistinace = -17.579 + 3.932 \times speed$

Visualizing the data

As the text does you can use the buil-in plot function to plot the scatter plot then add the regression line from generated by the `lm` function using `abline`

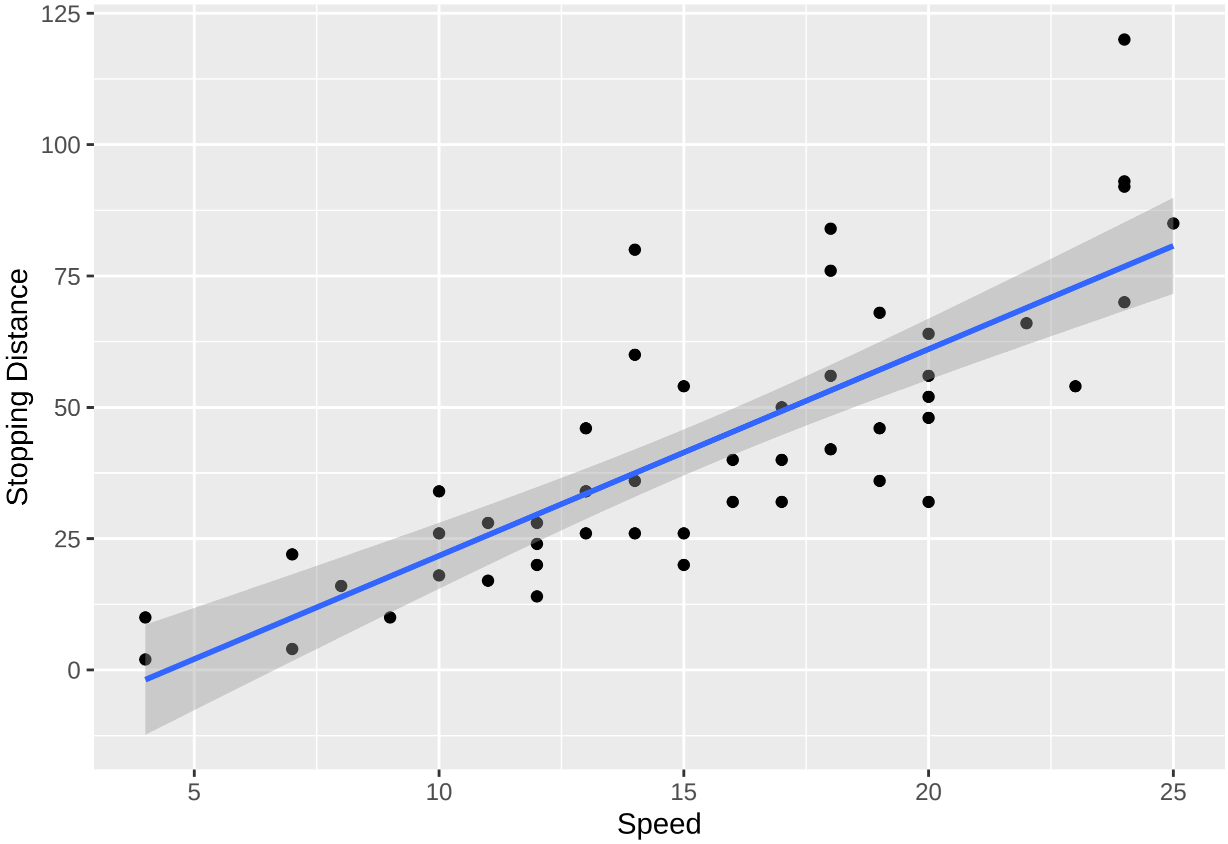
```
plot(cars)
abline(model)
```



You can also use ggplot to plot the scatter plot of speed vs stopping distance. Use stat_smooth to add a linear fitted line to the plot representing a linear regression model.

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() +
  stat_smooth(method = "lm", se = T) +
  xlab('Speed') +
  ylab('Stopping Distance')
```

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



Evaluating the quality of the model

```
summary(model)
```

```
##
## 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
```

Standard error at least five times to ten times smaller than corresponding coefficients? $3.9324/0.4155 \approx 9.464$ Yes it is about 9 times smaller

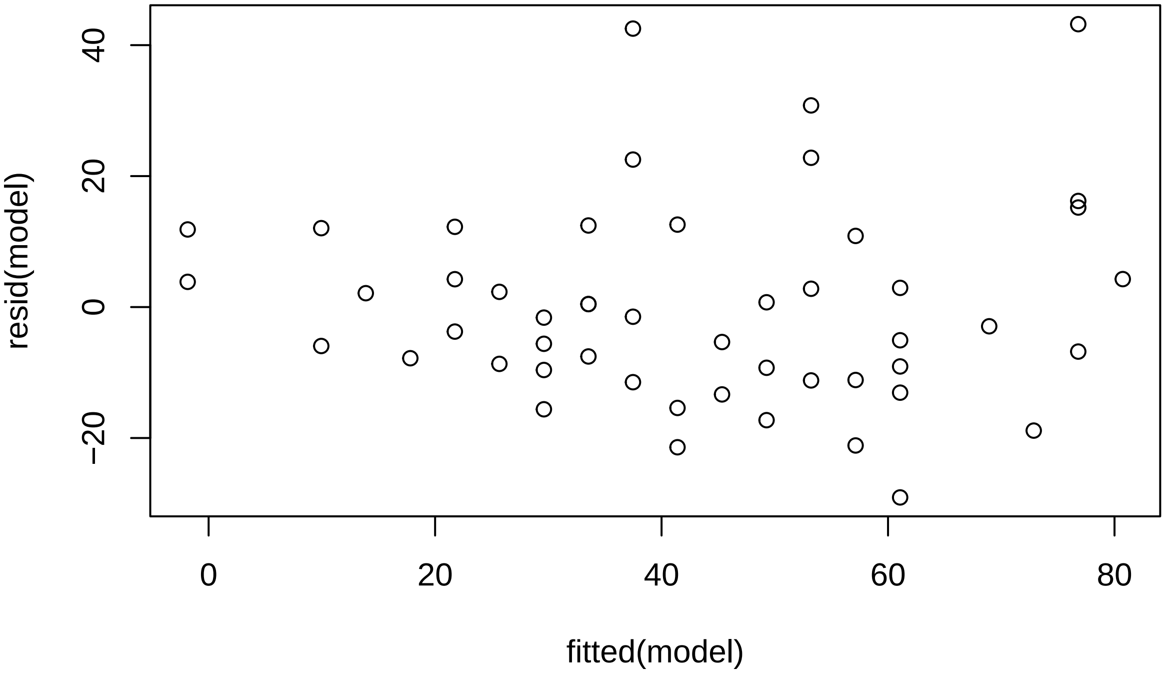
The value of 1.49e-12 and the `***` on the p-value of speed indicates high significance of the speed or to put it another way an extremely low probability that the speed is not a good predictor of stopping distance. A

0.0123 value for the intercept indicates a very low probability that the intercept is not relevant as well.

R^2 of 0.6511 indicates a good fitting model as it is relatively close to one.

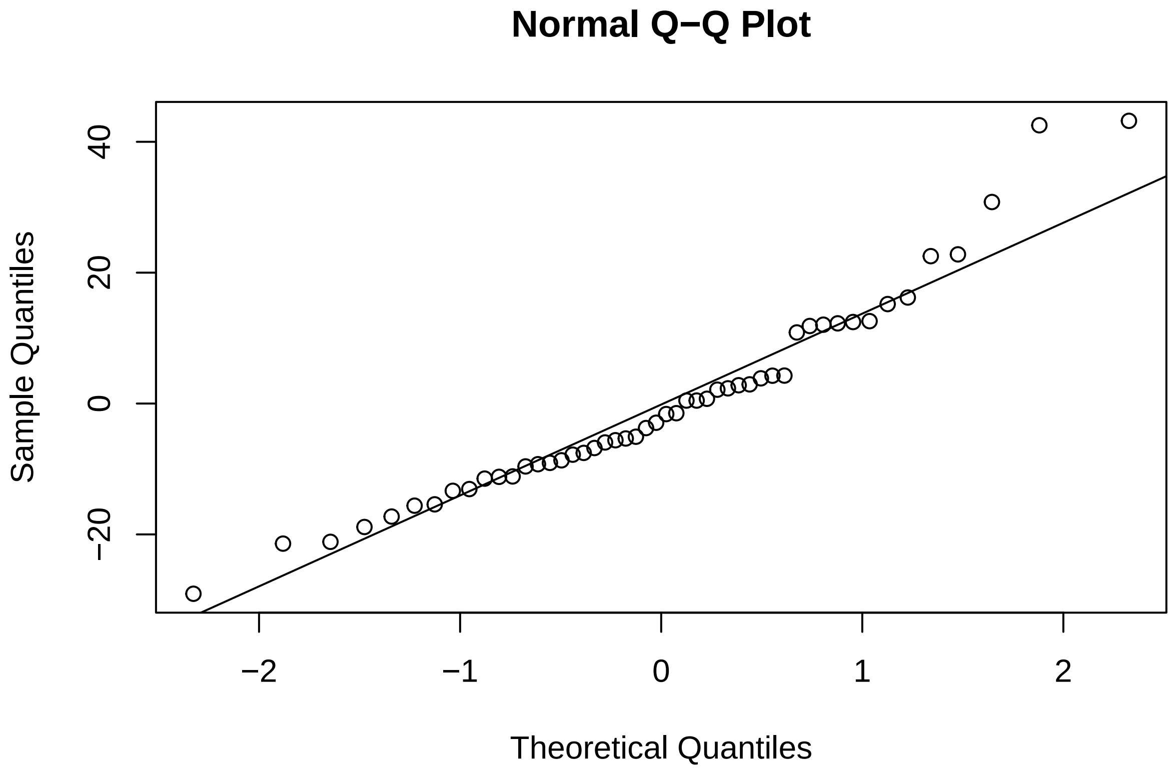
Residual Analysis

```
plot(fitted(model), resid(model))
```



Plotting the residuals vs. the input values does not show any obvious patterns which indicates that the variables used are sufficient to explain the relationship in the model.

```
qqnorm(resid(model))
qqline(resid(model))
```

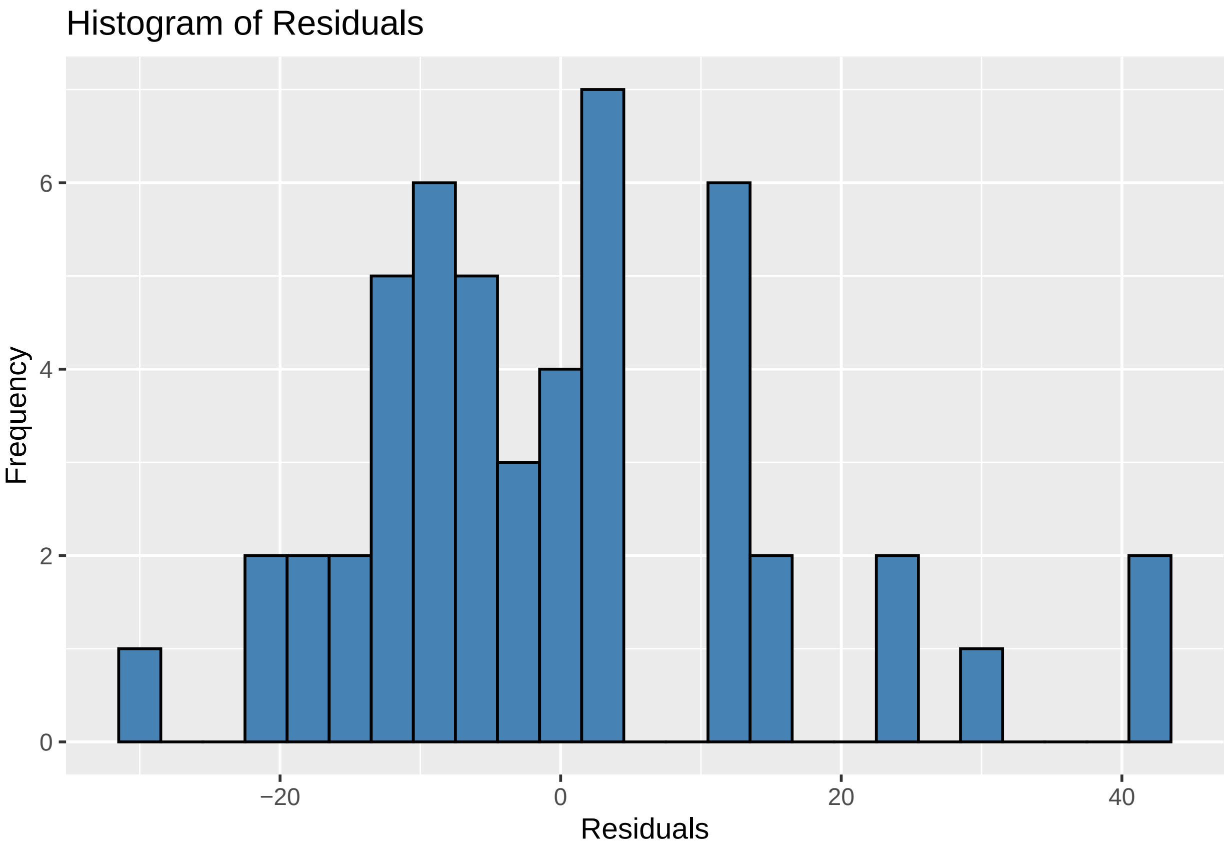


We should expect to see the residuals follow a straight line but at the beginning and most certainly near the end the do not follow a straight line. This shows that the residuals may not be normally distributed.

Plotting a histogram of the residuals shows similar results.

This indicates there may be better construct a model the produces better predictions with tighter residual values.

```
ggplot(data = cars, aes(x = model$residuals)) +
    geom_histogram(fill = 'steelblue', color = 'black', binwidth = 3) +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
```


Histogram of Residuals
Frequency
6
4
2
0
−20
0
20
40
Residuals