

Data 605 Assignment 11

Leticia Salazar

4/8/2022

Contents

Regression Analysis

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.)

```
# Load Libraries
library(ggplot2)
library(tidyverse)
```

```
# Loading dataset
head(cars)
```

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

Description of Data Set

```
# glimpse of data
glimpse(cars)
```

A data frame with 50 rows and 2 columns:

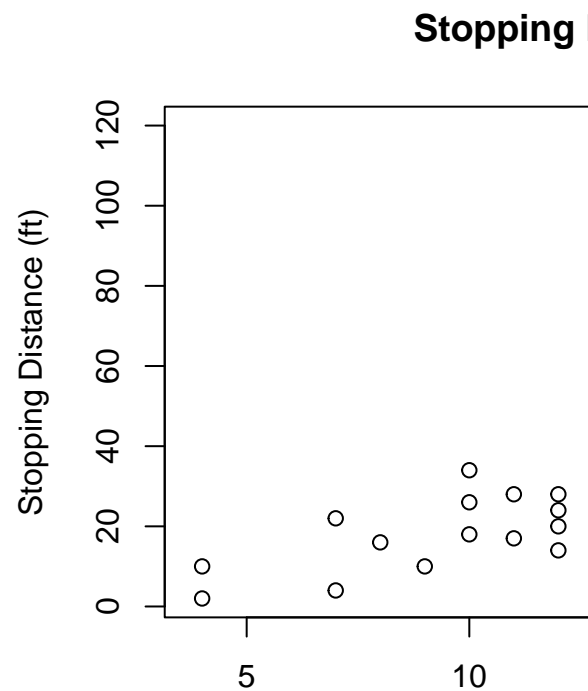
```
## Rows: 50
## Columns: 2
## $ speed <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 13~
## $ dist  <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, 26, 34~
```

```
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
```

Visualize the Data

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



From the plot we see that as speed increase so does the distance.

Linear Model Function

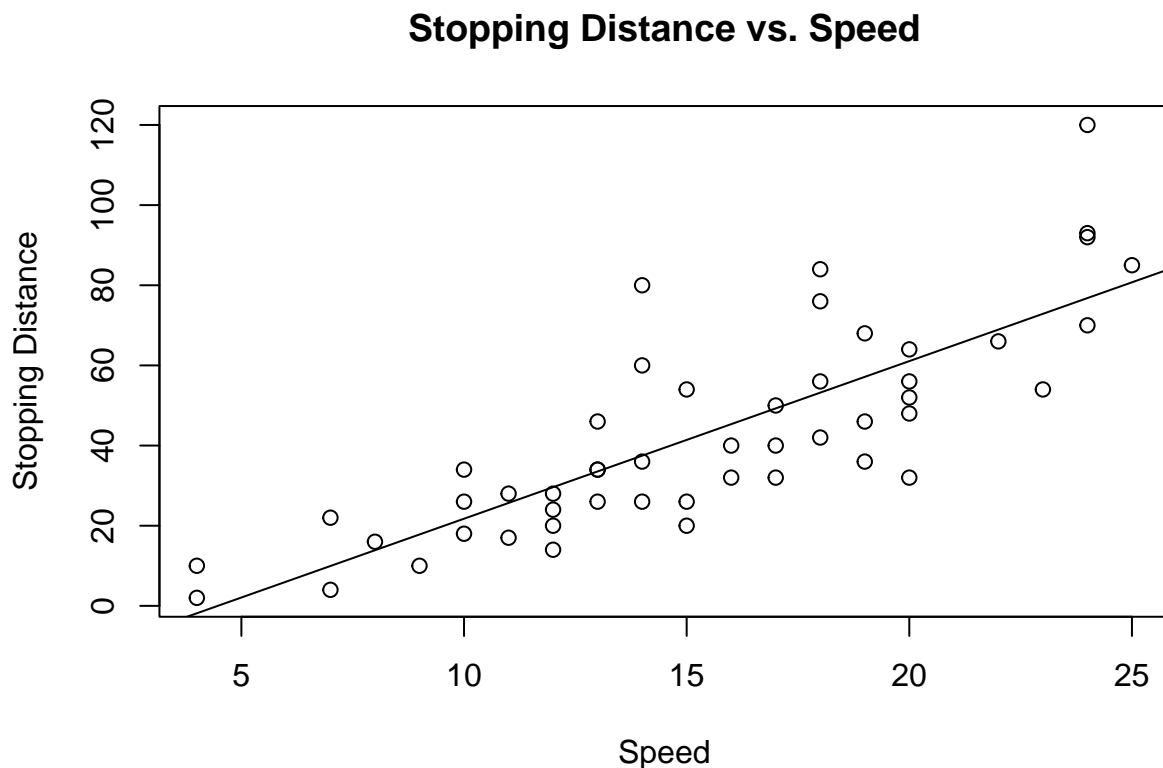
```
cars_lm <- lm(dist ~ speed, data = cars)
cars_lm
```

In this linear model we notice that there's a negative y-intercept, therefore we are assuming that if the speed is 0 then the distance is 0 since there can't be a negative distance.

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

Linear Model Plot

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



We have the following output for summary of cars_lm:

* According to the textbook: “The residuals are the differences between the actual measured values and the corresponding values on the fitted regression line. . . . we would expect residual values that are normally distributed around a mean of zero.” Based on this, the residual distribution seems to be normally distributed.

* For the standard error, the textbook mentions that you'd want to see the value to be at "at least five to ten times smaller than the corresponding coefficient" From the summary (T-value) we see that the speed's coefficient matches this criteria.

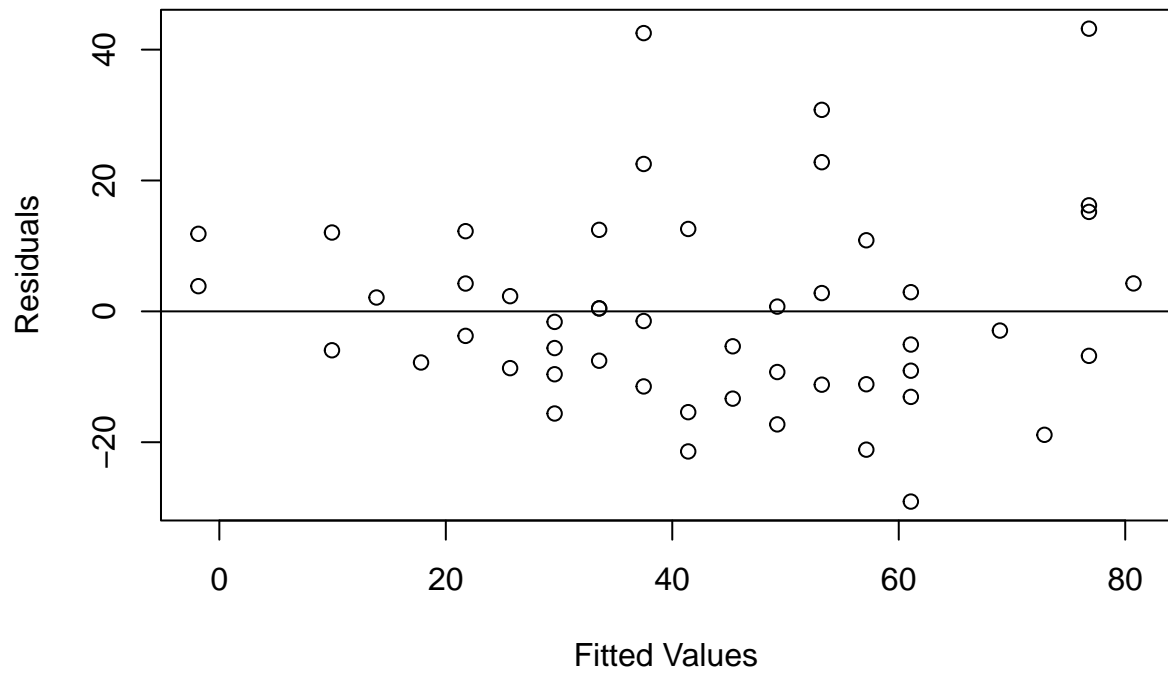
* The P-Value for both coefficients are also relevant to the model with a value close to 0.

```
summary(cars_lm)
```

* Finally, the multiple R^2 is 65.11%, explaining the variance.

```
##
## 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
```

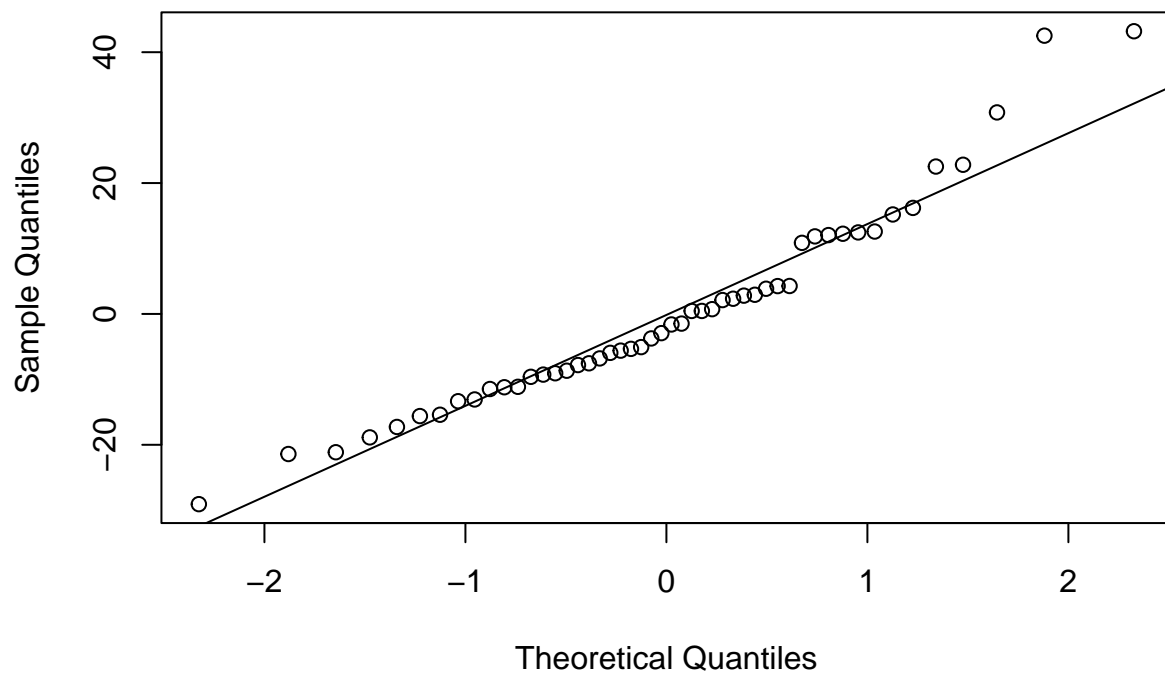
```
# Plot of the residuals
plot(cars_lm$fitted.values, cars_lm$residuals, xlab='Fitted Values', ylab='Residuals')
abline(0,0)
```



Residual Analysis

```
# Q-Q Plot
qqnorm(cars_lm$residuals)
qqline(cars_lm$residuals)
```

Normal Q-Q Plot



Conclusion:

From this analysis we can see that the residual model does a good job at exploring the data. For real world data you'd opt for other tests to be done to better analyze models. As per the analysis, this model shows a correlation between stopping distance and speed but it's only explaining 65.11% of the data, meaning there can be other factors that can influence the "normality" of this model.

References:

- Lilja, David J; Linse, Greta M. (2022). Linear Regression Using R: An Introduction to Data Modeling, 2nd Edition. University of Minnesota Libraries Publishing. Retrieved from the University of Minnesota Digital Conservancy, <https://hdl.handle.net/11299/189222>.