MThambeliyagodage\_Data605\_W11\_Assign11

Matheesha Thambeliyagodage

April 22, 2018

Table of Contents

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.4.4

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

# Data 605 Assignment 11

## Q1

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).

data(cars)  
head(cars)

## 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

colnames(cars)

## [1] "speed" "dist"

ncol(cars)

## [1] 2

nrow(cars)

## [1] 50

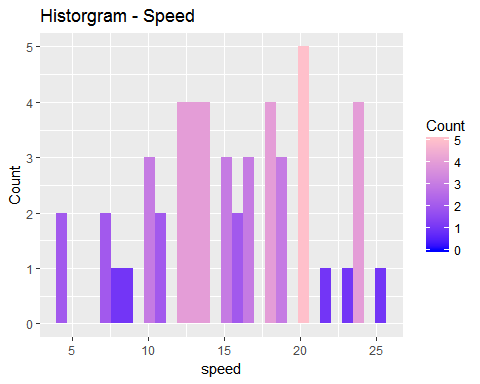
## Data Visualization:

cars\_speed\_df = arrange(cars, speed)

## Warning: package 'bindrcpp' was built under R version 3.4.4

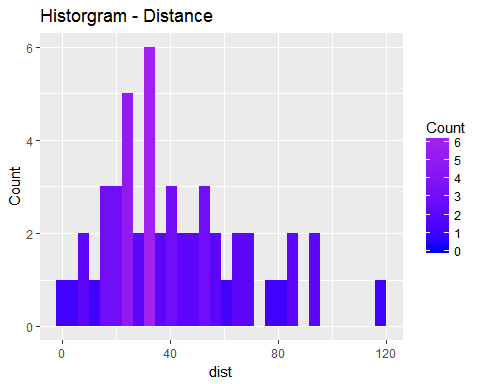
ggplot(data=cars\_speed\_df, aes(cars\_speed\_df$speed)) +   
 geom\_histogram(aes(fill = ..count..)) +  
 scale\_fill\_gradient("Count", low = "blue", high = "pink") +  
 labs(title = "Historgram - Speed") +  
 labs(x = "speed") +  
 labs(y = "Count")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



cars\_dist\_df = arrange(cars, dist)  
  
ggplot(data=cars\_dist\_df, aes(cars\_dist\_df$dist)) +   
 geom\_histogram(aes(fill = ..count..)) +  
 scale\_fill\_gradient("Count", low = "blue", high = "purple") +  
 labs(title = "Historgram - Distance") +  
 labs(x = "dist") +  
 labs(y = "Count")

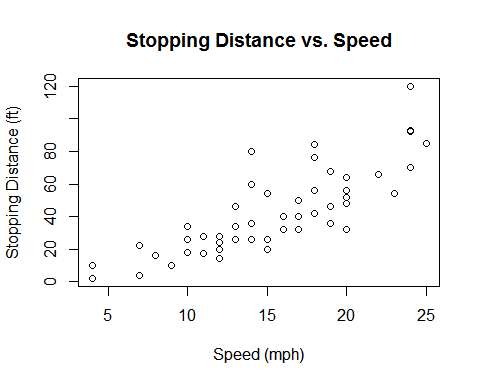
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 ## Statistical Analysis:

### Plot of two variables Speed & distance

The Speed is **explanatory** variable and stopping distance is the **response**.

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

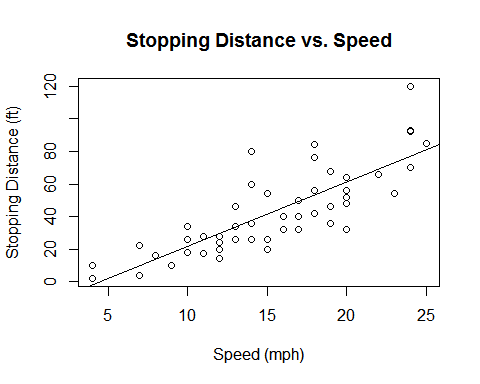


### Display the Linear Model

cars\_lm <- lm(cars$dist ~ cars$speed)  
cars\_lm

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

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



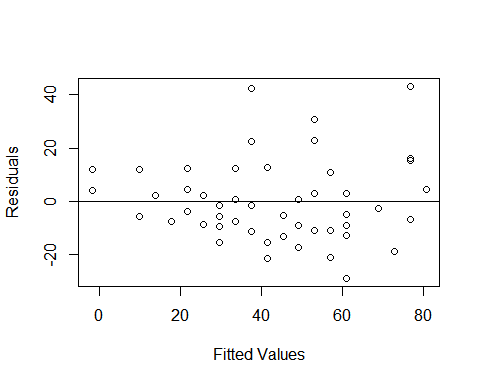
The above appears some correlation between two variables, but let us evaluate the linear model we have.

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

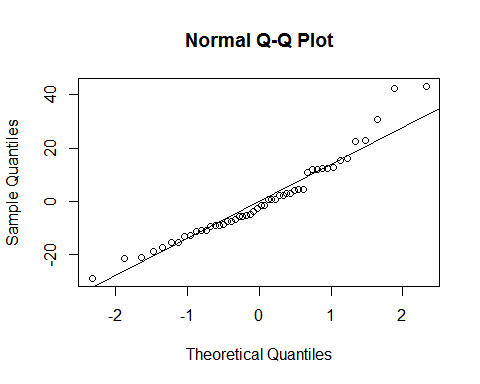
The median value of the residuals is somewhat close to zero and quartiles and min/max values are roughly the same magnitude. The standard error of the speed variable is more than 9 times smaller than the corresponding coefficient. There should not be a lot of variability in this coefficient. On the other hand, the difference between the intercept estimate and standard error is less significant, so there may be more variability. The speed coefficient is highly significant. The intercept coefficient is less significant, but it is still relevant depending on the confidence interval desired. Finally, explains about 65.11% of the data’s variation.

plot(cars\_lm$fitted.values, cars\_lm$residuals, xlab='Fitted Values', ylab='Residuals')  
abline(0,0)



It is possible to say that the outlier values do not show the same variance of the residuals; however, it is not very clear. I think it is reasonable to continue with the analysis and assume similar variance of residuals.

qqnorm(cars\_lm$residuals)  
qqline(cars\_lm$residuals)



Althought again there are some problems at the outlier levels, the normal Q-Q plot of the residuals appears to follow the theoretical line. Residuals are reasonably normally distributed.

### Conclusion

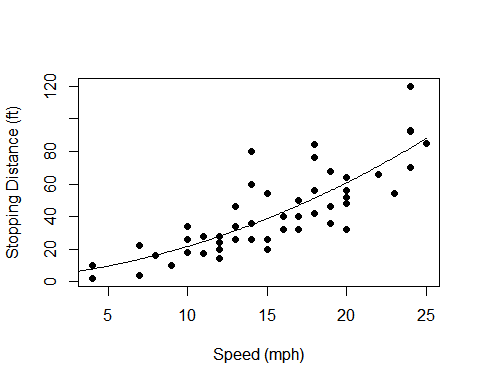
I believe the linear model does a good job at explaining the data. There appears to be some slight curvature in the main plot and in the residuals plot, so I decided to try a simple quadratic model (see below). It has it’s own problems - again varability of residuals is not constant enough, q-q plot has some deviations, coefficients are not very significant and is not increased by much. I don’t think it’s an improvement over a simplier linear model.

#### Quadtratic Model

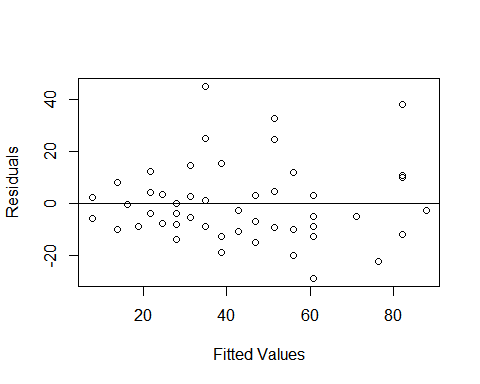
speed <- cars$speed  
speed2 <- speed^2  
dist <- cars$dist  
   
cars\_qm <- lm(dist ~ speed + speed2)  
summary(cars\_qm)

##   
## Call:  
## lm(formula = dist ~ speed + speed2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28.720 -9.184 -3.188 4.628 45.152   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.47014 14.81716 0.167 0.868  
## speed 0.91329 2.03422 0.449 0.656  
## speed2 0.09996 0.06597 1.515 0.136  
##   
## Residual standard error: 15.18 on 47 degrees of freedom  
## Multiple R-squared: 0.6673, Adjusted R-squared: 0.6532   
## F-statistic: 47.14 on 2 and 47 DF, p-value: 5.852e-12

speedvalues <- seq(0, 25, 0.1)  
predictedcounts <- predict(cars\_qm,list(speed=speedvalues, speed2=speedvalues^2))  
  
plot(speed, dist, pch=16, xlab='Speed (mph)', ylab='Stopping Distance (ft)')  
lines(speedvalues, predictedcounts)



plot(cars\_qm$fitted.values, cars\_qm$residuals, xlab='Fitted Values', ylab='Residuals')  
abline(0,0)



qqnorm(cars\_qm$residuals)  
qqline(cars\_qm$residuals)

