DATA 605 - Homework 11

Omer Ozeren

Table of Contents

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 the built-in R “cars” dataset

data("cars")  
head(cars, n = 10)  
## speed dist  
## 1 4 2  
## 2 4 10  
## 3 7 4  
## 4 7 22  
## 5 8 16  
## 6 9 10  
## 7 10 18  
## 8 10 26  
## 9 10 34  
## 10 11 17

### Summary of the dataset

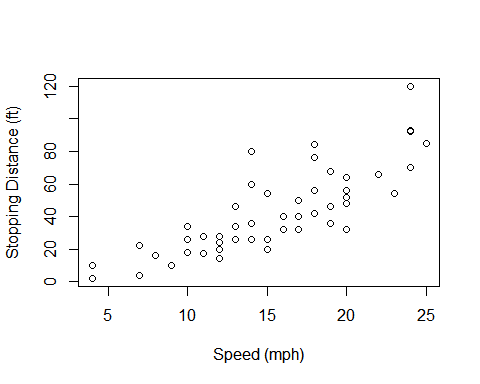
summary(cars)  
## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

### Dimensions of the dataset

dim(cars)  
## [1] 50 2

### Visualize the data

with(cars, plot(speed, dist,   
 xlab = "Speed (mph)",  
 ylab = "Stopping Distance (ft)"))

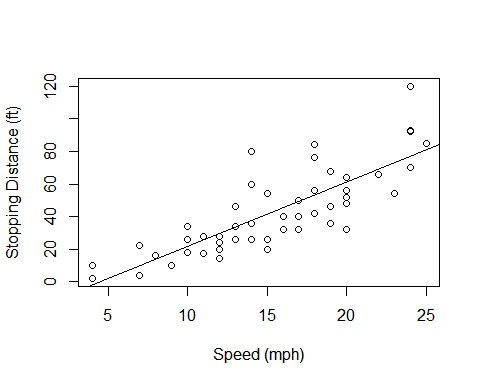


### Fit a Linear Model

# linear model  
lm\_cars <- lm(dist ~ speed, data = cars)  
lm\_cars  
##   
## Call:  
## lm(formula = dist ~ speed, data = cars)  
##   
## Coefficients:  
## (Intercept) speed   
## -17.579 3.932

* Thus y-intercept or and the slope and the linear model is

with(cars, plot(speed, dist,   
 xlab = "Speed (mph)",  
 ylab = "Stopping Distance (ft)"))  
abline(lm\_cars)



### Quality of the Model

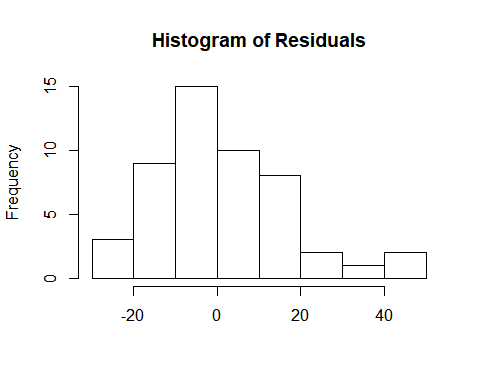
summary(lm\_cars)  
##   
## Call:  
## lm(formula = dist ~ speed, data = cars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.069 -9.525 -2.272 9.215 43.201   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.5791 6.7584 -2.601 0.0123 \*   
## speed 3.9324 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

* We can see the summary statistics of the residuals which are the differences between

the actual measured values and the values on the line. A good-fit model would have the residuals

to be nearly standard normal. The median should be near 0 which is the case here.

hist(lm\_cars$residuals, xlab = "", main = "Histogram of Residuals")



* A key statistic is value which shows that the model explains about 65% of the

data’s variation which for a linear model is not too bad.

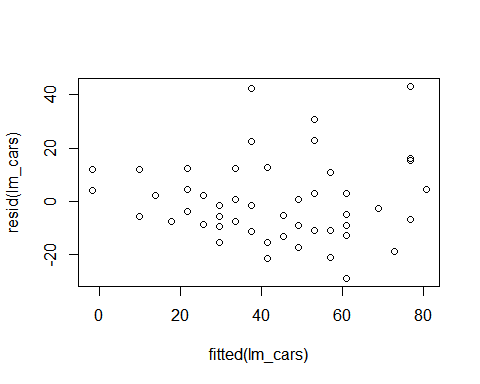
* Also looking at the p-values, we see that the probability that the speed variable is not

relevant is very small at about This means that speed plays a key predictor in

determining stopping distance and a strong dependency.

### Residual Analysis

plot(fitted(lm\_cars), resid(lm\_cars))



* Residuals are nearly uniformly scattered and approximately constant variance.
* A Quantile vs Quantile or Q-Q plot

qqnorm(resid(lm\_cars))  
qqline(resid(lm\_cars))

