STOR 455 Class 3 R Notebook

# message=FALSE, warning=FALSE supress warnings and messages from appearing in knitted html  
  
library(readr)  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
  
# If notebook and csv file are saved in the same folder  
# DistanceHome <- read\_csv("DistanceHome.csv")  
  
Domestic=subset(DistanceHome,Distance<250)

**Single Quantitative Predictor Model** - Notation:  
– Y = Response variable – X = Predictor variable

*Assume (for now) that both Y and X are quantitative variables.* Y = f(x) + Error

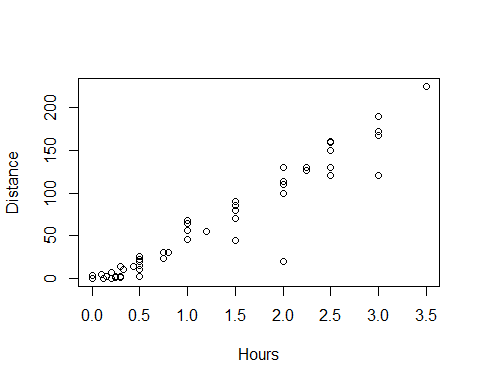
*Simple Linear Model* - X = Single quantitative predictor - Y = Quantitative response

*Goal*: Find a line that best summarizes the trend in the data.

Y = Bo + B1x + Error Response = Intercept + Slope*Predictor + Random Error* Assumptions:\* - Assume: Error ~ Follows a normal distribution and independent - There are 3 parameters to estimate: Bo, B1, and std error

**Scatterplot in R** *See below*

plot(Distance~Hours, data=Domestic)



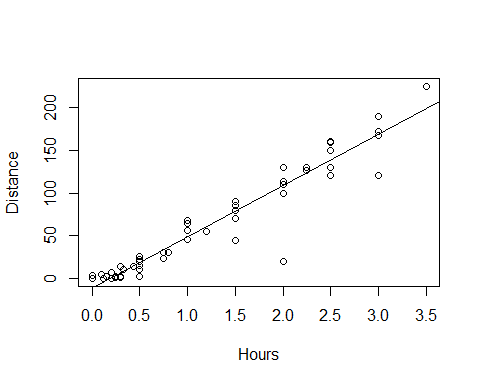
**Least Square Regression in R** *Syntax:* Syntax: lm(Response~Predictor,data= )

lm(Distance~Hours, data=Domestic)

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Coefficients:  
## (Intercept) Hours   
## -11.06 59.98

*R – Plot with Fitted Line*

mod1=lm(Distance~Hours, data=Domestic)  
plot(Distance~Hours, data=Domestic)  
abline(mod1)



*Simple Linear Model- Conditions* **Model:** 1. Linearity: The means for Y vary as a linear function of X. **Error:** 2.Zero Mean: The distribution of the errors is centered at zero. 3.Constant variance: The variance for Y is the same at each X. (Homoscedasticity) 4.Independence: No relationships among errors. 5.Normality: - Residuals are normally distributed - (sometimes) At each X, the Y’s follow a normal distribution.

*Linear* Look for consistent curvature or non-linear patterns

*Constant Variance* Look for “fan-shaped” pattern - Fan-shaped is **bad**

summary(mod1)

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

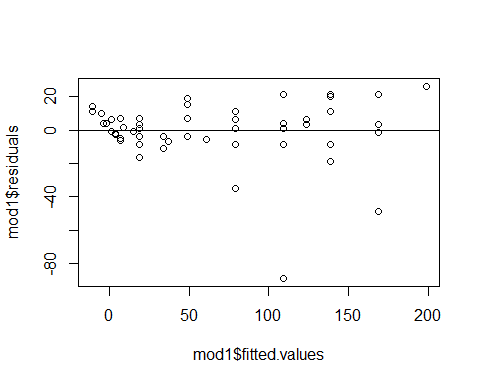
mod1$residuals

## 1 2 3 4 5 6   
## 26.1414099 -48.8698461 21.1301539 3.1301539 -1.7698461 21.1301539   
## 7 8 9 10 11 12   
## -8.8811021 20.1188979 -18.8811021 11.1188979 21.1188979 6.1132699   
## 13 14 15 16 17 18   
## 3.1132699 -8.8923581 21.1076419 -88.8923581 4.1076419 1.1076419   
## 19 20 21 22 23 24   
## 6.0963859 6.0963859 -34.9036141 -8.9036141 6.0963859 11.0963859   
## 25 26 27 28 29 30   
## 1.0963859 -5.9103678 -3.9148702 19.0851298 15.0851298 7.0851298   
## 31 32 33 34 35 36   
## -6.9193726 -3.9204982 -10.9204982 -16.4261262 3.0738738 7.0738738   
## 37 38 39 40 41 42   
## -3.9261262 -8.9261262 1.0738738 -0.7277020 1.2700468 -4.9306286   
## 43 44 45 46 47 48   
## 7.0693714 -6.4306286 -2.9317542 -1.9317542 -1.9317542 -0.7328798   
## 49 50 51 52 53 54   
## 6.4671202 4.0659946 4.0153193 10.0648690 14.0626178 11.0626178

mod1$fitted.values

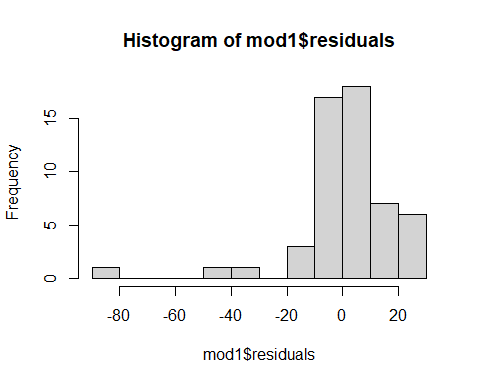
## 1 2 3 4 5 6   
## 198.8585901 168.8698461 168.8698461 168.8698461 168.8698461 168.8698461   
## 7 8 9 10 11 12   
## 138.8811021 138.8811021 138.8811021 138.8811021 138.8811021 123.8867301   
## 13 14 15 16 17 18   
## 123.8867301 108.8923581 108.8923581 108.8923581 108.8923581 108.8923581   
## 19 20 21 22 23 24   
## 78.9036141 78.9036141 78.9036141 78.9036141 78.9036141 78.9036141   
## 25 26 27 28 29 30   
## 78.9036141 60.9103678 48.9148702 48.9148702 48.9148702 48.9148702   
## 31 32 33 34 35 36   
## 36.9193726 33.9204982 33.9204982 18.9261262 18.9261262 18.9261262   
## 37 38 39 40 41 42   
## 18.9261262 18.9261262 18.9261262 14.7277020 8.7299532 6.9306286   
## 43 44 45 46 47 48   
## 6.9306286 6.9306286 3.9317542 3.9317542 3.9317542 0.9328798   
## 49 50 51 52 53 54   
## 0.9328798 -2.0659946 -3.8653193 -5.0648690 -11.0626178 -11.0626178

plot(mod1$residuals~mod1$fitted.values)  
abline(0,0)



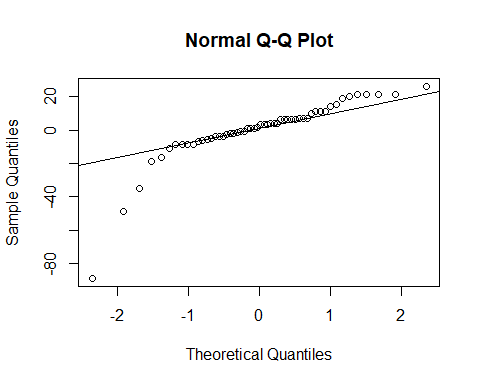
*Residuals* Look at a histogram of the residuals Look for clear skewness and outliers - skew and outliers are **bad**

hist(mod1$residuals, breaks=10)

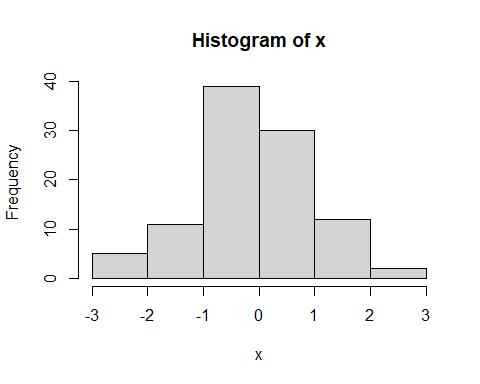


*How much Variability is Expected?*

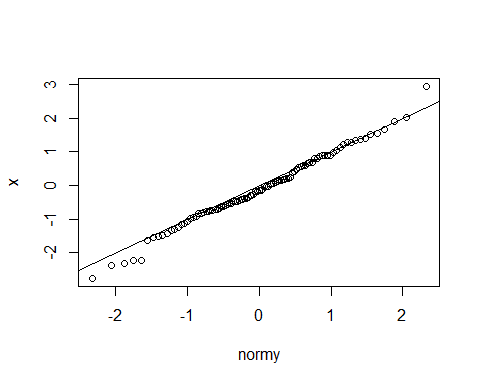
qqnorm(mod1$residuals)  
qqline(mod1$residuals)



#Sets randomization  
set.seed(455001)  
  
#Sample of 99 values from a Normal distribution; mean=0; sd=1; sorted ascending  
x = sort(rnorm(99,0,1))  
hist(x)



#list of integers 1 through 99.  
y = c(1:99)  
  
#z-scores of dataset of 99 values if perfectly normally distributed  
normy = qnorm(y/100)  
  
plot(x~normy)  
abline(0,1)



x <- rnorm(54, 0, 18.26)  
qqnorm(x)  
qqline(x)

