linear regression

harsha

April 23, 2020

library(moments)  
library(readr)  
library(readr)  
Salary\_Data <- read\_csv("D:/Modules/Module 6 - SLR/Salary\_Data.csv")

## Parsed with column specification:  
## cols(  
## YearsExperience = col\_double(),  
## Salary = col\_double()  
## )

View(Salary\_Data)  
attach(Salary\_Data)  
  
  
#EDA  
#1st business moments  
mean(YearsExperience) #5.31

## [1] 5.313333

mean(Salary\_Data$Salary) #76003

## [1] 76003

median(Salary\_Data$YearsExperience) #4.7

## [1] 4.7

median(Salary\_Data$Salary) #65237

## [1] 65237

# 2nd business moments  
var(Salary\_Data$YearsExperience) #8.053

## [1] 8.053609

var(Salary\_Data$Salary)

## [1] 751550960

sd(Salary\_Data$YearsExperience) #2.83

## [1] 2.837888

sd(Salary\_Data$Salary) #27414.43

## [1] 27414.43

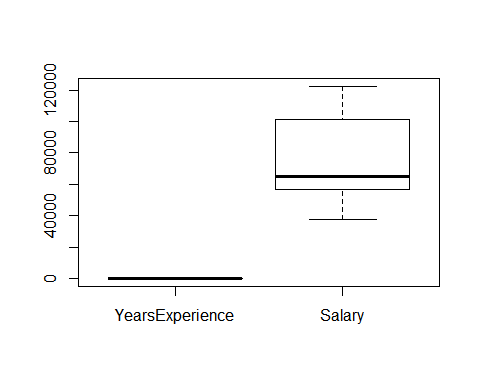
range(Salary\_Data$YearsExperience) #1.1 10.5

## [1] 1.1 10.5

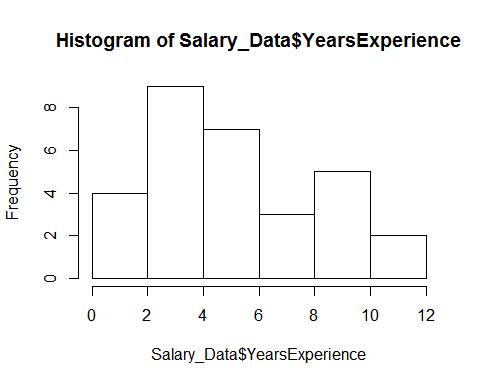
range(Salary\_Data$Salary) #37731 122391

## [1] 37731 122391

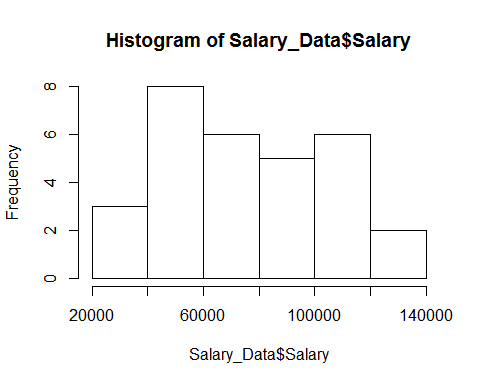
#different visulaizations  
boxplot(Salary\_Data) #no outliers



hist(Salary\_Data$YearsExperience)



hist(Salary\_Data$Salary)



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

#3rd business moment  
library(moments)  
skewness(Salary\_Data$YearsExperience) #positive skewness

## [1] 0.3603123

skewness(Salary\_Data$Salary) #positive skewness

## [1] 0.3361619

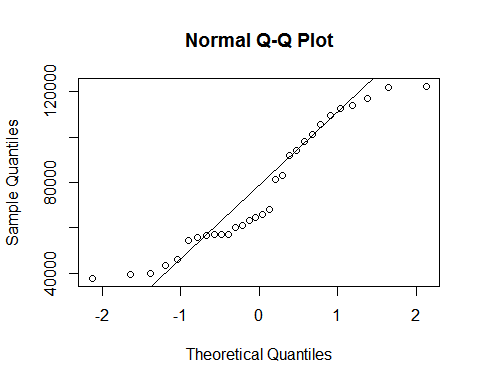
#4th business moment  
kurtosis(Salary\_Data$YearsExperience) #positive kurtosis

## [1] 1.955248

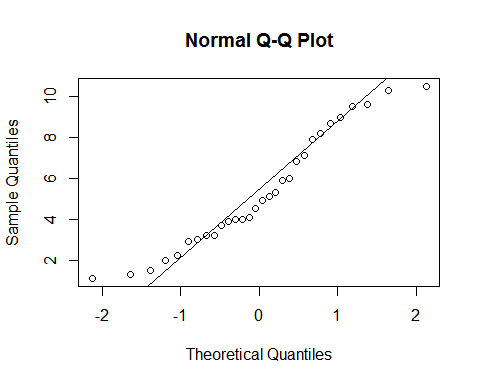
kurtosis(Salary\_Data$Salary)#positive kurtosis

## [1] 1.717087

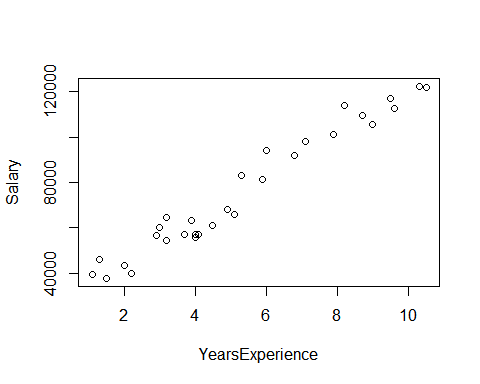
#Normal Quantile-Quantile Plot  
qqnorm(Salary)  
qqline(Salary)



qqnorm(YearsExperience)##Normally distributed  
qqline(YearsExperience)



plot(YearsExperience,Salary) #positive relation



cor(YearsExperience,Salary)

## [1] 0.9782416

attach(Salary\_Data)

## The following objects are masked from Salary\_Data (pos = 3):  
##   
## Salary, YearsExperience

#model building  
model<- lm(YearsExperience~Salary)  
summary (model)

##   
## Call:  
## lm(formula = YearsExperience ~ Salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.12974 -0.46457 0.04105 0.54311 0.79669   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.383e+00 3.273e-01 -7.281 6.3e-08 \*\*\*  
## Salary 1.013e-04 4.059e-06 24.950 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5992 on 28 degrees of freedom  
## Multiple R-squared: 0.957, Adjusted R-squared: 0.9554   
## F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

#p is significant and r=0.95

confint(model,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -3.053603e+00 -1.7127178614  
## Salary 9.295173e-05 0.0001095796

predict(model,interval="predict")

## Warning in predict.lm(model, interval = "predict"): predictions on current data refer to \_future\_ responses

## fit lwr upr  
## 1 1.600934 0.3165619 2.885307  
## 2 2.295819 1.0237773 3.567861  
## 3 1.437694 0.1500755 2.725313  
## 4 2.024427 0.7478589 3.300996  
## 5 1.656428 0.3731291 2.939727  
## 6 3.352729 2.0947042 4.610754  
## 7 3.707969 2.4533424 4.962595  
## 8 3.130248 1.8697562 4.390740  
## 9 4.142905 2.8915256 5.394284  
## 10 3.408121 2.1506703 4.665572  
## 11 4.018652 2.7664480 5.270856  
## 12 3.266856 2.0079094 4.525802  
## 13 3.384628 2.1269353 4.642320  
## 14 3.397185 2.1396216 4.654747  
## 15 3.805285 2.5514728 5.059097  
## 16 4.496626 3.2471410 5.746111  
## 17 4.303310 3.0528728 5.553747  
## 18 6.030801 4.7817265 7.279875  
## 19 5.856117 4.6076374 7.104597  
## 20 7.129735 5.8731707 8.386300  
## 21 6.906748 5.6522247 8.161272  
## 22 7.568520 6.3071722 8.829867  
## 23 7.875253 6.6099641 9.140542  
## 24 9.142087 7.8554139 10.428759  
## 25 8.698442 7.4201795 9.976704  
## 26 8.308670 7.0369817 9.580359  
## 27 9.461782 8.1684469 10.755118  
## 28 9.022897 7.7385799 10.307214  
## 29 10.010845 8.7049141 11.316775  
## 30 9.958288 8.6536250 11.262951

model$residuals

## 1 2 3 4 5 6   
## -0.50093427 -0.99581922 0.06230598 -0.02442725 0.54357215 -0.45272891   
## 7 8 9 10 11 12   
## -0.70796884 0.06975175 -0.94290484 0.29187878 -0.11865188 0.73314437   
## 13 14 15 16 17 18   
## 0.61537241 0.70281547 0.69471486 0.40337421 0.79669035 -0.73080053   
## 19 20 21 22 23 24   
## 0.04388273 -1.12973546 -0.10674848 -0.46851956 0.02474675 -0.94208664   
## 25 26 27 28 29 30   
## 0.00155821 0.69132973 0.03821767 0.57710304 0.28915527 0.54171214

rmse<-sqrt(mean(model$residuals^2))  
rmse #0.578

## [1] 0.5788774

#least RMSE value model is evualted