R Notebook

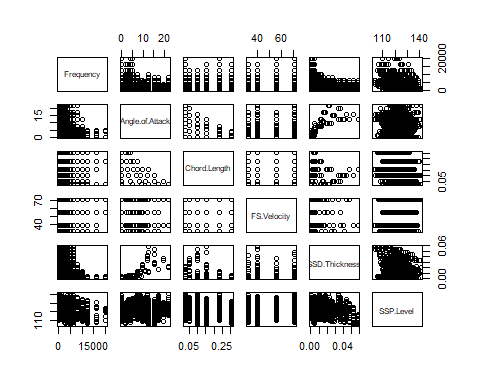
This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

# UCI Data sets: Airfoil Self Noise Data Set  
# Importing the data Set into R  
ASN <- read.csv(file = "C:/Venu/UCI DataSets/Airfoil Self Noise.csv", header = TRUE,stringsAsFactors = FALSE)  
# Exploring the data using summary statistics  
summary(ASN)

## Frequency Angle.of.Attack Chord.Length FS.Velocity   
## Min. : 200 Min. : 0.000 Min. :0.0254 Min. :31.70   
## 1st Qu.: 800 1st Qu.: 2.000 1st Qu.:0.0508 1st Qu.:39.60   
## Median : 1600 Median : 5.400 Median :0.1016 Median :39.60   
## Mean : 2886 Mean : 6.782 Mean :0.1365 Mean :50.86   
## 3rd Qu.: 4000 3rd Qu.: 9.900 3rd Qu.:0.2286 3rd Qu.:71.30   
## Max. :20000 Max. :22.200 Max. :0.3048 Max. :71.30   
## SSD.Thickness SSP.Level   
## Min. :0.0004007 Min. :103.4   
## 1st Qu.:0.0025351 1st Qu.:120.2   
## Median :0.0049574 Median :125.7   
## Mean :0.0111399 Mean :124.8   
## 3rd Qu.:0.0155759 3rd Qu.:130.0   
## Max. :0.0584113 Max. :141.0

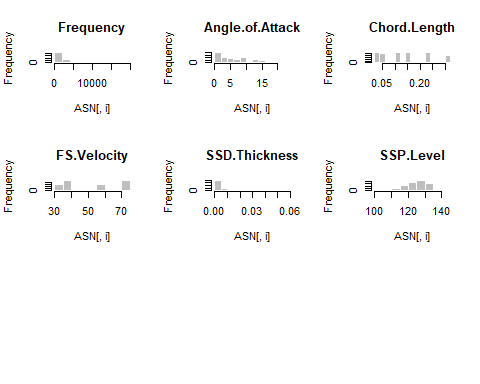
# Scatter Plot: Exploring relationships among the variables  
plot(ASN)



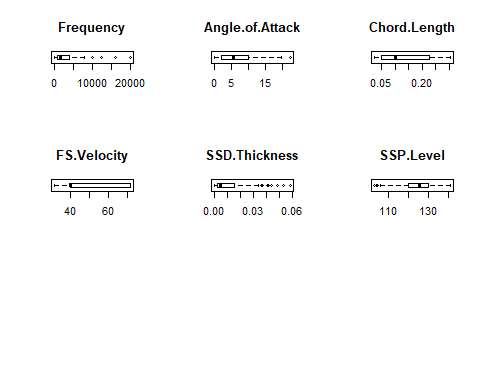
# Correlation Analysis: Exploring linear relationships among variables using Pearson's correlation coerfficient  
cor(ASN)

## Frequency Angle.of.Attack Chord.Length FS.Velocity  
## Frequency 1.000000000 -0.27276454 -0.003660639 0.133663831  
## Angle.of.Attack -0.272764536 1.00000000 -0.504868150 0.058759565  
## Chord.Length -0.003660639 -0.50486815 1.000000000 0.003786629  
## FS.Velocity 0.133663831 0.05875957 0.003786629 1.000000000  
## SSD.Thickness -0.230107353 0.75339378 -0.220842431 -0.003974013  
## SSP.Level -0.390711412 -0.15610753 -0.236161512 0.125102801  
## SSD.Thickness SSP.Level  
## Frequency -0.230107353 -0.3907114  
## Angle.of.Attack 0.753393785 -0.1561075  
## Chord.Length -0.220842431 -0.2361615  
## FS.Velocity -0.003974013 0.1251028  
## SSD.Thickness 1.000000000 -0.3126695  
## SSP.Level -0.312669506 1.0000000

attach(ASN)  
# Distribution of the variables using histogram  
  
colnames <- dimnames(ASN)[[2]]  
par(mfrow = c(3,3))  
for (i in 1:6)  
 {  
   
 hist(ASN[,i],main = colnames[i],col = "gray",border = "white")  
 }  
# Box Plots  
par(mfrow = c(3,3))



for (i in 1:6) {  
 boxplot(ASN[,i], horizontal = TRUE, main = colnames[i])  
}

  
# Multiple Linear Regression: SSP noise level as dependent and all others as independent variables  
ASN.Reg <- lm(SSP.Level~.,data = ASN)  
  
# Results of the regression analysis  
summary(ASN.Reg)

##   
## Call:  
## lm(formula = SSP.Level ~ ., data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.480 -2.882 -0.209 3.152 16.064   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.328e+02 5.447e-01 243.87 <2e-16 \*\*\*  
## Frequency -1.282e-03 4.211e-05 -30.45 <2e-16 \*\*\*  
## Angle.of.Attack -4.219e-01 3.890e-02 -10.85 <2e-16 \*\*\*  
## Chord.Length -3.569e+01 1.630e+00 -21.89 <2e-16 \*\*\*  
## FS.Velocity 9.985e-02 8.132e-03 12.28 <2e-16 \*\*\*  
## SSD.Thickness -1.473e+02 1.501e+01 -9.81 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.809 on 1497 degrees of freedom  
## Multiple R-squared: 0.5157, Adjusted R-squared: 0.5141   
## F-statistic: 318.8 on 5 and 1497 DF, p-value: < 2.2e-16

It may observed that all the coefficients (t-tests) are significant as well as the overall regression (F-test).

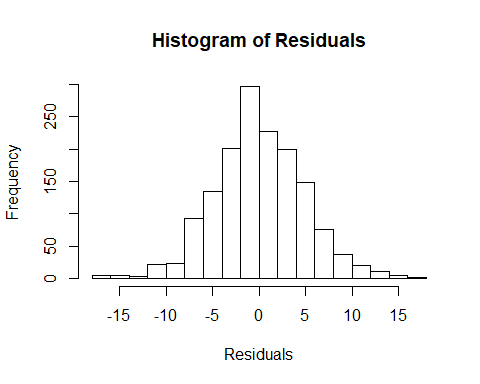
Note that the Multiple R-squared is 51.5% and Adjusted R-squared is about 51.4%.

ASN.Reg$coefficients

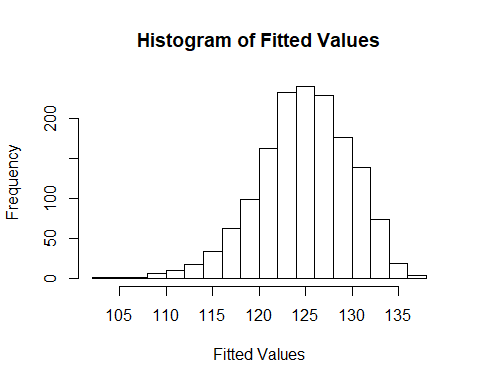
## (Intercept) Frequency Angle.of.Attack Chord.Length FS.Velocity   
## 1.328338e+02 -1.282207e-03 -4.219117e-01 -3.568800e+01 9.985404e-02   
## SSD.Thickness   
## -1.473005e+02

par(mfrow=c(1,1))

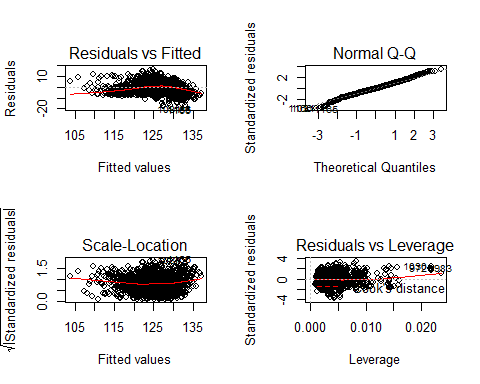
# Histogram of residuals  
hist(ASN.Reg$residuals,xlab = "Residuals", main = "Histogram of Residuals")



# Distribution of fitted value s  
hist(ASN.Reg$fitted.values, xlab = "Fitted Values", main = "Histogram of Fitted Values")



# Validating Assumptions of Regression Analysis; Normality, Heteroscadasticity, Multicollinearity  
par(mfrow=c(2,2))  
plot(ASN.Reg)



library(car)

## Loading required package: carData

vif(ASN.Reg)

## Frequency Angle.of.Attack Chord.Length FS.Velocity SSD.Thickness   
## 1.144444 3.441658 1.510754 1.041698 2.532127

It may be noted that above diagrams and table of Variance Inflation Factors:

1. absence of heteroskedasticity,
2. Normality of Residuals
3. Absence of multicollinearity

# Contructing the confidence intervals for the Regression parameters  
confint(ASN.Reg)

## 2.5 % 97.5 %  
## (Intercept) 1.317653e+02 1.339023e+02  
## Frequency -1.364799e-03 -1.199615e-03  
## Angle.of.Attack -4.982083e-01 -3.456151e-01  
## Chord.Length -3.888617e+01 -3.248983e+01  
## FS.Velocity 8.390221e-02 1.158059e-01  
## SSD.Thickness -1.767525e+02 -1.178485e+02

# Construction of confidence and prediction intervals for the mean value of dependent variable  
predict (ASN.Reg ,data.frame(Frequency=800,Angle.of.Attack=5.4,Chord.Length=0.1016,FS.Velocity=39.6,SSD.Thickness=0.0049),interval = "confidence")

## fit lwr upr  
## 1 129.1363 128.7473 129.5252

# Construction of prediction interval for a randomly chosen value of the dependent variable  
predict (ASN.Reg ,data.frame(Frequency=800,Angle.of.Attack=5.4,Chord.Length=0.1016,FS.Velocity=39.6,SSD.Thickness=0.0049),interval = "prediction")

## fit lwr upr  
## 1 129.1363 119.6954 138.5771

# Regression analysis using interaction of variables  
ASN.R1 <- lm(SSP.Level~Frequency+Chord.Length+SSD.Thickness+Frequency:SSD.Thickness, data = ASN)  
summary(ASN.R1)

##   
## Call:  
## lm(formula = SSP.Level ~ Frequency + Chord.Length + SSD.Thickness +   
## Frequency:SSD.Thickness, data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.3311 -3.0231 0.1292 3.2401 15.0356   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.333e+02 3.155e-01 422.493 <2e-16 \*\*\*  
## Frequency -7.485e-04 4.664e-05 -16.047 <2e-16 \*\*\*  
## Chord.Length -2.439e+01 1.361e+00 -17.918 <2e-16 \*\*\*  
## SSD.Thickness -1.221e+02 1.354e+01 -9.016 <2e-16 \*\*\*  
## Frequency:SSD.Thickness -7.187e-02 4.601e-03 -15.621 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.794 on 1498 degrees of freedom  
## Multiple R-squared: 0.5184, Adjusted R-squared: 0.5171   
## F-statistic: 403.1 on 4 and 1498 DF, p-value: < 2.2e-16

It may be noted that using interaction of variables yield a significant regression results. However,

does not improve the R-squared or Adjusted R-squared.

# Regression analysis using quadratic functions  
ASN.R2 <- lm(SSP.Level~Frequency+I(Frequency^2),data = ASN)  
summary(ASN.R2)

##   
## Call:  
## lm(formula = SSP.Level ~ Frequency + I(Frequency^2), data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.470 -4.267 0.086 4.147 17.820   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.283e+02 2.742e-01 467.994 < 2e-16 \*\*\*  
## Frequency -1.561e-03 1.262e-04 -12.375 < 2e-16 \*\*\*  
## I(Frequency^2) 5.649e-08 9.217e-09 6.129 1.13e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.276 on 1500 degrees of freedom  
## Multiple R-squared: 0.1734, Adjusted R-squared: 0.1723   
## F-statistic: 157.3 on 2 and 1500 DF, p-value: < 2.2e-16

ASN.R3 <- lm(SSP.Level~Chord.Length+I(Chord.Length^2),data = ASN)  
summary(ASN.R3)

##   
## Call:  
## lm(formula = SSP.Level ~ Chord.Length + I(Chord.Length^2), data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.0417 -4.5202 0.7471 5.2396 16.8241   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 127.7335 0.4715 270.916 < 2e-16 \*\*\*  
## Chord.Length -27.8393 7.4324 -3.746 0.000187 \*\*\*  
## I(Chord.Length^2) 32.9997 22.7930 1.448 0.147882   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.703 on 1500 degrees of freedom  
## Multiple R-squared: 0.05709, Adjusted R-squared: 0.05583   
## F-statistic: 45.41 on 2 and 1500 DF, p-value: < 2.2e-16

ASN.R4 <- lm(SSP.Level~SSD.Thickness+I(SSD.Thickness^2),data = ASN)  
summary(ASN.R4)

##   
## Call:  
## lm(formula = SSP.Level ~ SSD.Thickness + I(SSD.Thickness^2),   
## data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.8203 -4.5623 0.7384 4.7406 18.6726   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 126.2263 0.2793 451.925 <2e-16 \*\*\*  
## SSD.Thickness -66.6654 40.1074 -1.662 0.0967 .   
## I(SSD.Thickness^2) -2181.3511 851.3100 -2.562 0.0105 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.543 on 1500 degrees of freedom  
## Multiple R-squared: 0.1017, Adjusted R-squared: 0.1005   
## F-statistic: 84.9 on 2 and 1500 DF, p-value: < 2.2e-16

# Regression analysis using higher order polynomials functions (Non-linear functions)  
ASN.R5 <- lm(SSP.Level~poly(Frequency,5),data = ASN)  
summary(ASN.R5)

##   
## Call:  
## lm(formula = SSP.Level ~ poly(Frequency, 5), data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.7596 -4.0277 0.0768 3.8827 19.2372   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 124.8359 0.1604 778.512 < 2e-16 \*\*\*  
## poly(Frequency, 5)1 -104.4613 6.2166 -16.804 < 2e-16 \*\*\*  
## poly(Frequency, 5)2 38.4702 6.2166 6.188 7.84e-10 \*\*\*  
## poly(Frequency, 5)3 1.2528 6.2166 0.202 0.840313   
## poly(Frequency, 5)4 -21.0053 6.2166 -3.379 0.000746 \*\*\*  
## poly(Frequency, 5)5 28.1765 6.2166 4.532 6.29e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.217 on 1497 degrees of freedom  
## Multiple R-squared: 0.1907, Adjusted R-squared: 0.188   
## F-statistic: 70.53 on 5 and 1497 DF, p-value: < 2.2e-16

ASN.R6 <- lm(SSP.Level~poly(Chord.Length,5),data = ASN)  
summary(ASN.R6)

##   
## Call:  
## lm(formula = SSP.Level ~ poly(Chord.Length, 5), data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.7224 -4.5580 0.7475 5.2123 15.8530   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 124.836 0.172 725.882 < 2e-16 \*\*\*  
## poly(Chord.Length, 5)1 -63.141 6.667 -9.470 < 2e-16 \*\*\*  
## poly(Chord.Length, 5)2 9.705 6.667 1.456 0.14571   
## poly(Chord.Length, 5)3 -7.850 6.667 -1.177 0.23923   
## poly(Chord.Length, 5)4 -21.541 6.667 -3.231 0.00126 \*\*   
## poly(Chord.Length, 5)5 -18.136 6.667 -2.720 0.00660 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.667 on 1497 degrees of freedom  
## Multiple R-squared: 0.06904, Adjusted R-squared: 0.06594   
## F-statistic: 22.21 on 5 and 1497 DF, p-value: < 2.2e-16

ASN.R7 <- lm(SSP.Level~log(Frequency),data = ASN)  
summary(ASN.R7)

##   
## Call:  
## lm(formula = SSP.Level ~ log(Frequency), data = ASN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.6827 -4.3373 0.2065 4.4321 16.6844   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 141.5836 1.1517 122.9 <2e-16 \*\*\*  
## log(Frequency) -2.2554 0.1535 -14.7 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.452 on 1501 degrees of freedom  
## Multiple R-squared: 0.1258, Adjusted R-squared: 0.1252   
## F-statistic: 216 on 1 and 1501 DF, p-value: < 2.2e-16

It is observed that higher order polynomials in different variables fail to improve the value of R-squared or

Adjusted R-Squared. Hence, it may be concluded that the multiple linear regression provides the best model among the

alternatives tried with Prediction accuracy of 51.4%.