Examining Distributions

Seema Singh Saharan

December 22, 2016

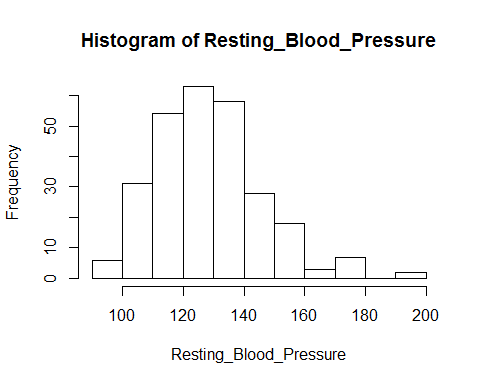
# Included Packages

# File Reading and Creation of Histogram

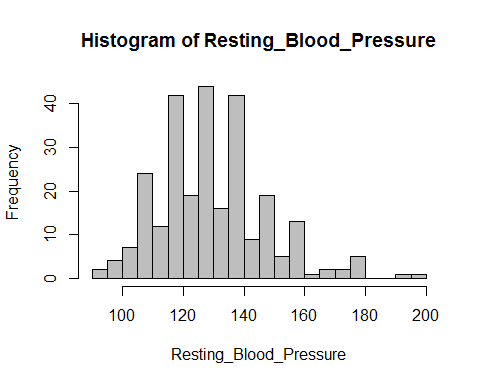
HeartAttack<-read\_excel("HeartAttack.xlsx")  
head(HeartAttack)

## Age Sex Chest\_Pain\_Type Resting\_Blood\_Pressure Serum\_Cholesterol  
## 1 70 1 4 130 322  
## 2 67 0 3 115 564  
## 3 57 1 2 124 261  
## 4 64 1 4 128 263  
## 5 74 0 2 120 269  
## 6 65 1 4 120 177  
## Fasting\_Blood\_MoreThan\_120 Resting\_Electrocardiographic\_Reading  
## 1 0 2  
## 2 0 2  
## 3 0 0  
## 4 0 0  
## 5 0 2  
## 6 0 0  
## Maximum\_Heart\_Rate Exercise\_Induced\_Angina Old\_Peak Slope  
## 1 109 0 2.4 2  
## 2 160 0 1.6 2  
## 3 141 0 0.3 1  
## 4 105 1 0.2 2  
## 5 121 1 0.2 1  
## 6 140 0 0.4 1  
## Number\_Blood\_Vessels\_Calcified thal Heart\_Attack\_Diagnosis Residual  
## 1 3 3 2 52.9060  
## 2 0 7 1 298.6518  
## 3 0 7 2 8.1378  
## 4 1 7 1 1.3976  
## 5 1 3 1 -5.0884  
## 6 0 7 1 -85.8510

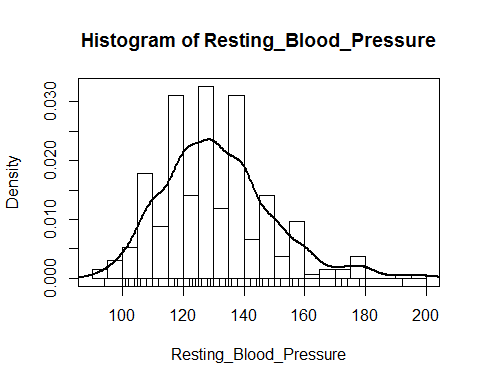
with(HeartAttack,hist(Resting\_Blood\_Pressure))



with(HeartAttack,hist(Resting\_Blood\_Pressure,breaks="FD",col="gray"))

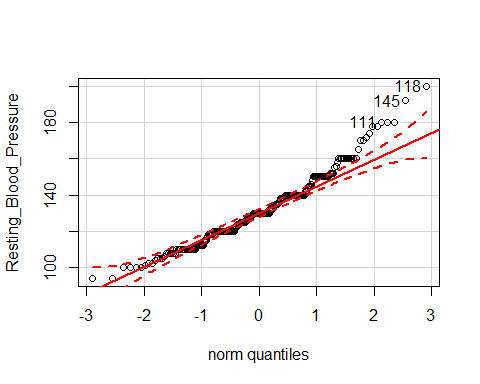


with(HeartAttack, {  
 hist(Resting\_Blood\_Pressure,breaks="FD",freq=FALSE,ylab="Density")  
 #ylab command defines the y axis. Freq=FALSE defines Density graph instead of Frequency  
 lines(density(Resting\_Blood\_Pressure),lwd=2)  
 #lines command draws the density. lwd=2 draws double thick line.  
 rug(Resting\_Blood\_Pressure)  
 box()  
 #box creates a box around the image.  
})



# Quantile Quantile Plots

with(HeartAttack,qqPlot(Resting\_Blood\_Pressure,labels = row.names(HeartAttack),id.n=3))

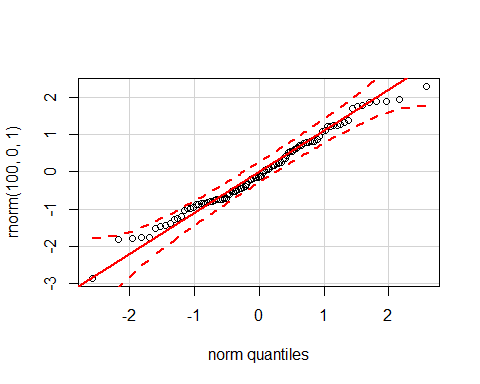


## 118 145 111   
## 270 269 266

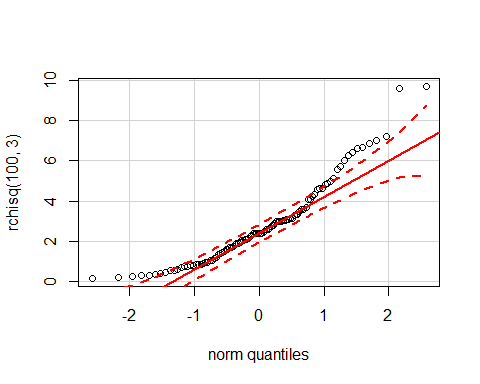
#The qqplot() function in the car package provides the 95% CI around the line fit to the plot.The #row name provides the Resting\_Blood\_Pressure being evaluated and the id.n=3 provides marking the #three most extreme points.

# Demonstrating normality,right skew,thick tails using QQ PLots

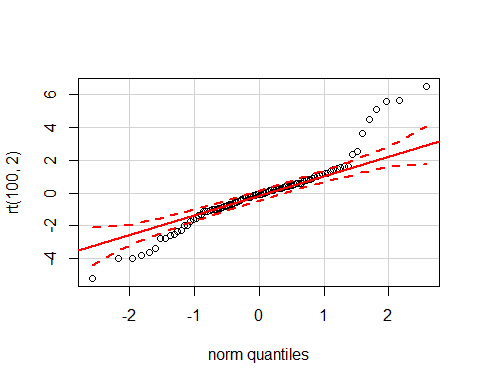
qqPlot(rnorm(100,0,1))



qqPlot(rchisq(100,3))

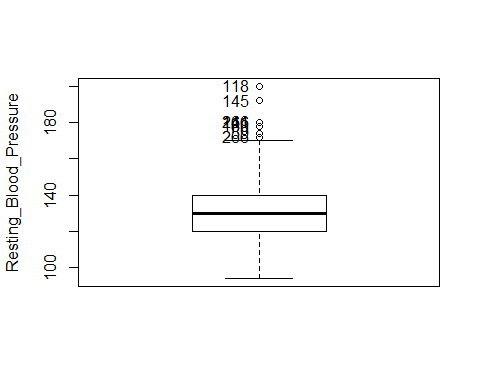


qqPlot(rt(100,2))



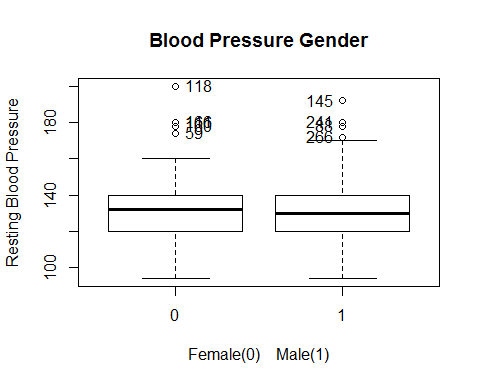
# Boxplots

Boxplot(~ Resting\_Blood\_Pressure,data=HeartAttack)



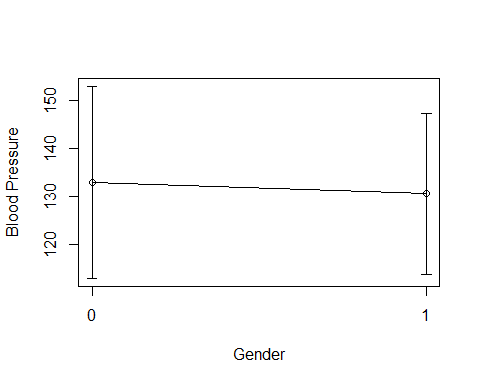
## [1] "59" "88" "111" "118" "145" "160" "166" "241" "266"

#Parallel Box Plots  
Boxplot(Resting\_Blood\_Pressure~Sex,data=HeartAttack, main="Blood Pressure Gender", xlab="Female(0) Male(1)", ylab="Resting Blood Pressure")



## [1] "59" "111" "118" "160" "166" "88" "145" "241" "266"

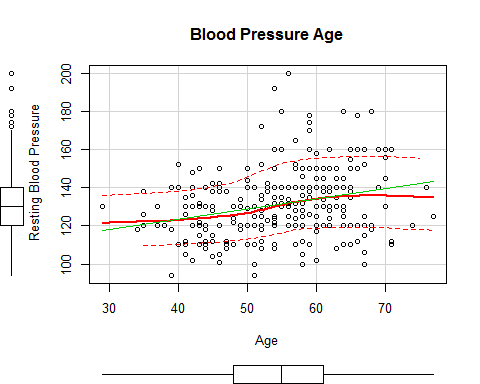
#Boxplots of Means  
mean\_values<-with(HeartAttack,tapply(Resting\_Blood\_Pressure,Sex,mean))  
standard\_deviation<-with(HeartAttack,tapply(Resting\_Blood\_Pressure,Sex,sd))  
# tapply() computes means ans sds for each gender  
plotCI(1:2,mean\_values,standard\_deviation,xaxt="n",xlab="Gender",ylab="Blood Pressure")  
# plotCI() function helps draw the graph.The parameters are coordinates on horizontal axis ,means   
# on vertical axis,standard deviations,xaxt="n" suppresses the x ticks.  
lines(1:2,mean\_values)   
#lines joins the means with the lines  
axis(1,at=1:2,labels=names(mean\_values))



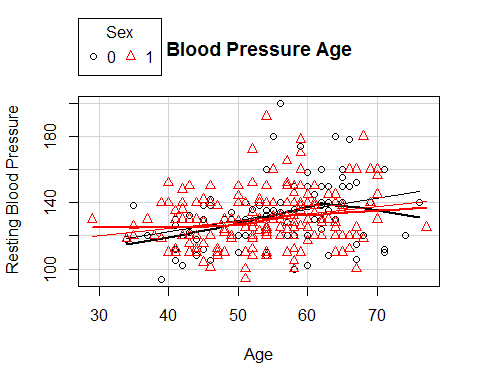
# axis specifies the groups on the x axis.

# Scatterplots

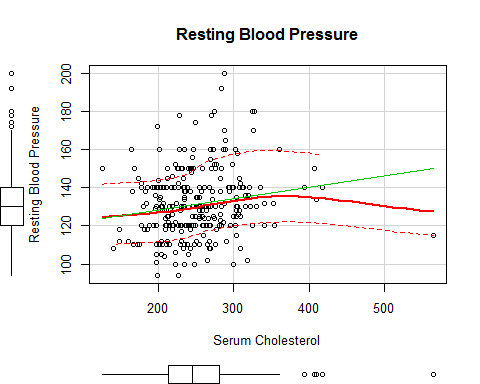
scatterplot(Resting\_Blood\_Pressure~Age,data=HeartAttack, main="Blood Pressure Age", xlab="Age", ylab=" Resting Blood Pressure")



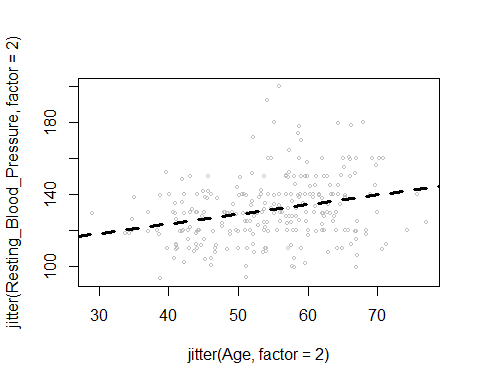
# scatterplot is a function in the car package that provides two smoothers on the scatterplot.  
# The first smoother is the Ordinary Least Squares(OLS) which can be suppressed by reg.line=FALSE  
# The second is a solid curved line which is the non parameteric regression smoother produced by # # the lowess smmoother(locally weighted scatterplot smoother) and this can be supressed by lowess # function in R.The variance is seen by the dotted line around the lowess curve.The conditional # # distribution of Resting Blood Pressure with Age can be evaluated by the vertical line between   
# dotted variance curve line.This also gives us a boxplot for both variables.  
  
# Coded Scatter plot  
scatterplot(Resting\_Blood\_Pressure~Age|Sex,data=HeartAttack, main="Blood Pressure Age", xlab="Age", ylab=" Resting Blood Pressure")



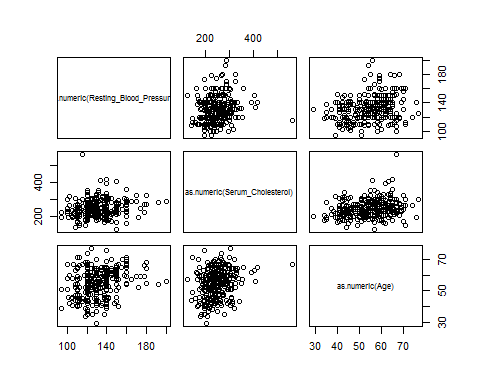
scatterplot(Resting\_Blood\_Pressure~Serum\_Cholesterol,data=HeartAttack, main="Resting Blood Pressure", xlab="Serum Cholesterol", ylab="Resting Blood Pressure")



#Jittered Scatterplot  
plot(jitter(Resting\_Blood\_Pressure,factor=2)~jitter(Age,factor=2),col="gray",cex=.5,data=HeartAttack)  
with(HeartAttack, {  
 abline(lm(Resting\_Blood\_Pressure~Age),lwd=3,lty="dashed")  
 lines(lowess(Resting\_Blood\_Pressure,Age,f=0.2),lwd=3)  
 # least square line is created by abline using the lm function  
})



# Three dimensional scatterplot  
scatter3d(Resting\_Blood\_Pressure~Age+Serum\_Cholesterol,id.n=3,data=HeartAttack)  
# The 3d graph shows Resting blood pressure on y axis and Age and Serum Chaolesterol on the   
# x axis and z axis.The three furthest points were identified by the Mahalanobis distances  
# (Point of means)   
  
# Scatterplot Matrices  
pairs(~ as.numeric(Resting\_Blood\_Pressure) + as.numeric(Serum\_Cholesterol) + as.numeric(Age) ,data=HeartAttack)



# Hexbin

# High Density Scatterplot with Binning

bin<-hexbin(HeartAttackResting\_Blood\_Pressure,xbins=50) plot(bin,main="Hexagonal Binning")

boxplot(Serum\_Cholesterol~Sex, data = HeartAttack, lwd = 2, ylab = 'Cholesterol') stripchart(Serum\_Cholesterol~Sex, vertical = TRUE, data = HeartAttack, method = "jitter", add = TRUE, pch = 20, col = 'blue')

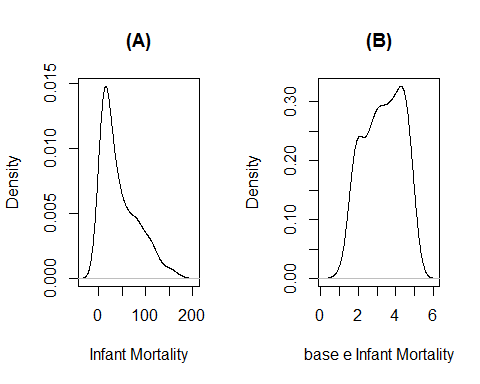
with(HeartAttack,  
{plot(Serum\_Cholesterol ~ Age)  
abline(lm(Serum\_Cholesterol ~ Age), lcol="red")  
lines(lowess(Serum\_Cholesterol ~ Age), lcol="blue") } )

# TRANSFORMATIONS using infant.mortality data set in the UN dataframe of the car package

# Logarithm Transformations  
UN <- na.omit(UN)  
head(UN)

## infant.mortality gdp  
## Afghanistan 154 2848  
## Albania 32 863  
## Algeria 44 1531  
## Angola 124 355  
## Antigua 24 6966  
## Argentina 22 8055

par(mfrow=c(1,2))  
with(UN,plot(density(infant.mortality),xlab="Infant Mortality",main="(A)"))  
with(UN,plot(density(log(infant.mortality)),  
 xlab="base e Infant Mortality",main="(B)"))



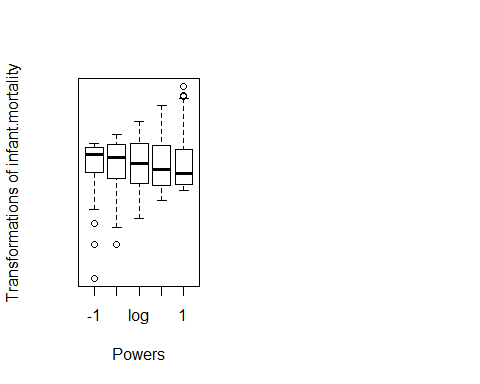
Median\_Value=median(UN$infant.mortality)  
Lower\_Hinge=quantile(UN$infant.mortality,prob=.25)  
Higher\_Hinge=quantile(UN$infant.mortality,.75)  
Tukey\_Criterion=(Higher\_Hinge-Median\_Value)/(Median\_Value-Lower\_Hinge)  
Tukey\_Criterion

## 75%   
## 2.235294

Infant\_logfunction = log(UN$infant.mortality)  
Median\_Value\_log=median(Infant\_logfunction)  
Lower\_Hinge\_log=quantile(Infant\_logfunction,prob=.25)  
Higher\_Hinge\_log=quantile(Infant\_logfunction,prob=.75)  
Tukey\_Criterion\_log=(Higher\_Hinge\_log-Median\_Value\_log)/(Median\_Value\_log-Lower\_Hinge\_log)  
Tukey\_Criterion\_log

## 75%   
## 0.9785498

symbox(~infant.mortality,data=UN)   
  
#symbox uses the bc function and shows the transformation down and up the   
 # ladder.The log function gives the best transformation

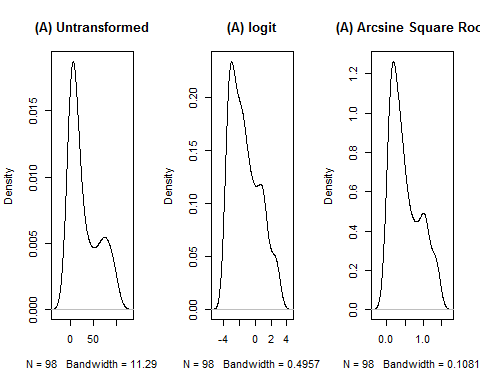
 #Proportion Transformations using data set in the Prestige dataframe of the car package

Prestige <- na.omit(Prestige)  
head(Prestige)

## education income women prestige census type  
## gov.administrators 13.11 12351 11.16 68.8 1113 prof  
## general.managers 12.26 25879 4.02 69.1 1130 prof  
## accountants 12.77 9271 15.70 63.4 1171 prof  
## purchasing.officers 11.42 8865 9.11 56.8 1175 prof  
## chemists 14.62 8403 11.68 73.5 2111 prof  
## physicists 15.64 11030 5.13 77.6 2113 prof

par(mfrow=c(1,3))  
  
with(Prestige,{  
 plot(density(women),main="(A) Untransformed")  
 plot(density(logit(women),adjust=.75),main="(A) logit")  
 plot(density(asin(sqrt(women/100)),adjust=.75),main="(A) Arcsine Square Root")  
  
})# adjust factor smoothes the multiple modes by the decreasing the bandwidth fraction

## Warning in logit(women): proportions remapped to (0.025, 0.975)

 # Transformations for equalizing variance

Ornstein <- na.omit(Ornstein)  
head(Ornstein)

## assets sector nation interlocks  
## 1 147670 BNK CAN 87  
## 2 133000 BNK CAN 107  
## 3 113230 BNK CAN 94  
## 4 85418 BNK CAN 48  
## 5 75477 BNK CAN 66  
## 6 40742 FIN CAN 69

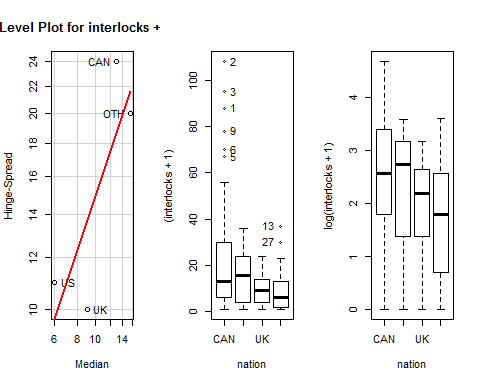
par(mfrow=c(1,3))  
  
spreadLevelPlot(interlocks+1~nation,Ornstein)

## LowerHinge Median UpperHinge Hinge-Spread  
## US 2 6.0 13 11  
## UK 4 9.0 14 10  
## CAN 6 13.0 30 24  
## OTH 4 15.5 24 20  
##   
## Suggested power transformation: 0.1534487

Boxplot((interlocks+1)~nation,data=Ornstein)

## [1] "1" "2" "3" "5" "6" "9" "13" "27"

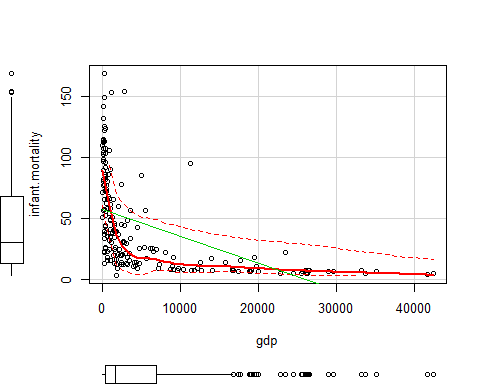
Boxplot(log(interlocks+1)~nation,data=Ornstein)



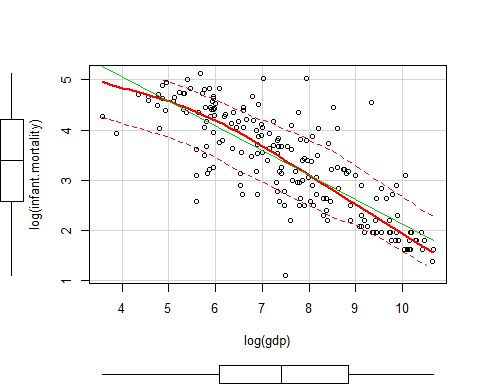
# The spread is less variable across groups   
   
# There are no outliers and within group spread  
   
 # is less skewed.

# Transforming Non Linearity using Tukey's Bulging Rule:

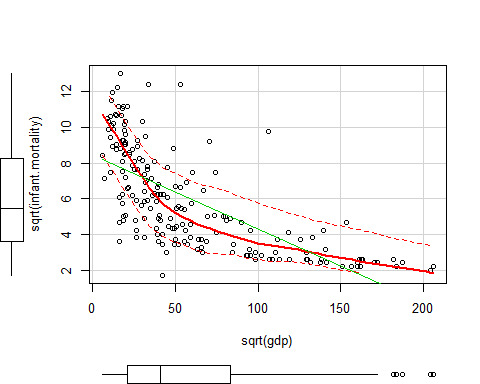
scatterplot(infant.mortality~gdp,data=UN)



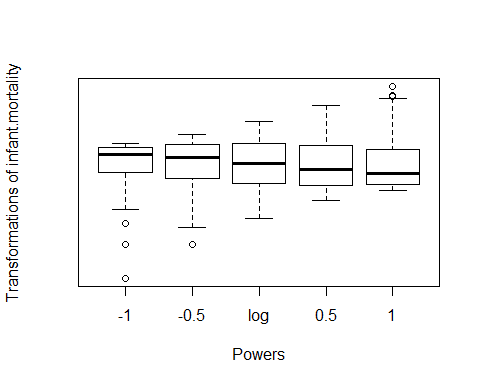
scatterplot(log(infant.mortality)~log(gdp),data=UN)



# Notice the relationship is now linear though it might be overcorrecting this relationship.  
scatterplot(sqrt(infant.mortality)~sqrt(gdp),data=UN)



# Cube root can also be used  
symbox(~infant.mortality,data=UN)



symbox(~gdp,data=UN)

