LECTURE 4 Linear Modelling

Objectives:

- To understand how transformations like log sqrt can help make data symmetrical, equalize variance and linearlize relationships.
- Box Cox family of transformation and transformation of proportions.
- Application of Tukey's bulging rule

Transforming Data

- Data transformation is useful when the data does not conform to restrictions and assumptions of the classical statistical models. Example of a classical model is the linear least square regression
- An alternative approach can be used by applying the models that unlike the conventional classical models are not constrained by assumptions. For example non parametric regression.
- The alternative approaches have their advantages and complexities.
- Transformation helps in envisioning data by making it symmetrical or by linearlizing a relationship or by equalizing variances amongst groups.
- Box Cox Family of transformation are used for their convenience. These are called scaled transformations. This can be denoted as:

$$X \rightarrow X^{(p)} = T(X, p) = \begin{cases} \dfrac{X^p - 1}{p} & p \neq 0 \\ \log_e X & p = 0 \end{cases}$$

Since $\lim\limits_{p \to 0} \dfrac{X^p - 1}{p} = \log_e X$

Using L'Hospital rule Differentiating the numerator and denominator

$\lim\limits_{p \to 0} \dfrac{X^p \log_e X}{1} = X^0 \log_e X = \log_e X$ substituting zero for p

Therefore $X^{(0)} \approx \log_e X$

Therefore $X^{(1)} = \dfrac{X^1 - 1}{1}$

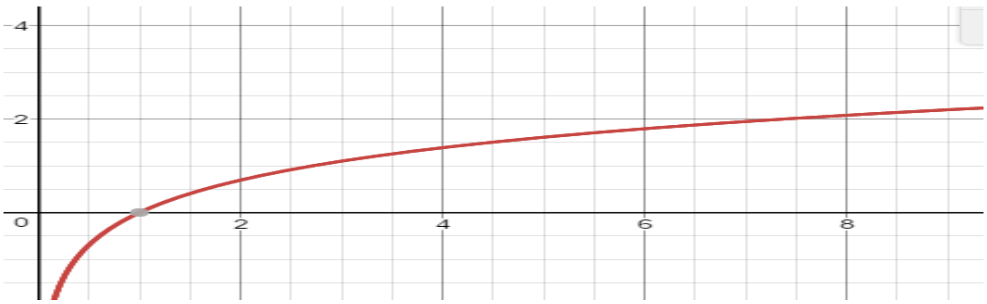
Therefore $X^{(2)} = \dfrac{X^2 - 1}{2}$

Ascending up the ladder of powers

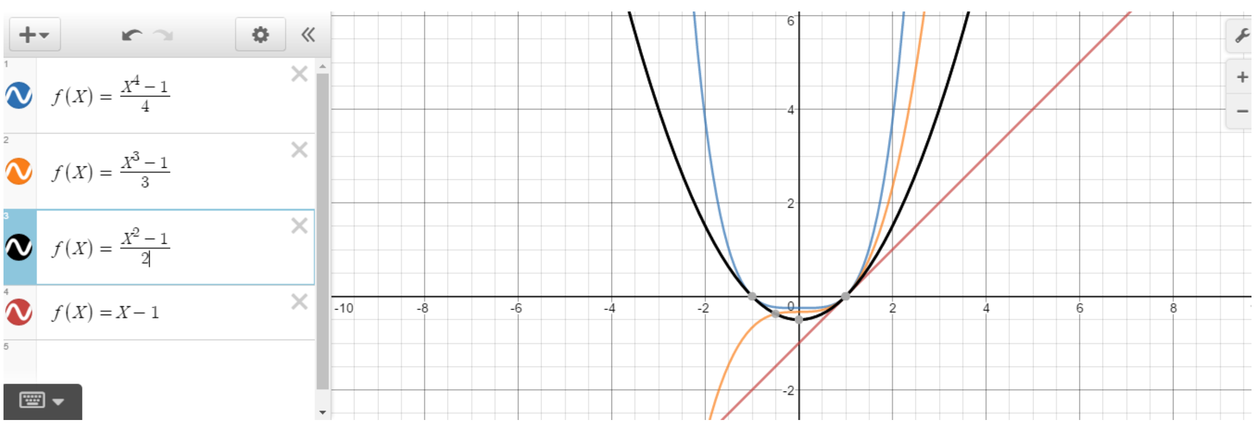
- By dividing the Box Cox transformation by p the direction of X are preserved
- The subtraction of 1 equalizes the differences of adjacent transformation (Page 56).

X	X^{-1}	$X^{-1}/-1$	$X^{-1}-1/-1$
1	1	-1	0
2	1/2	-1/2	1/2
3	1/3	-1/3	2/3
4	1/4	-1/4	3/4

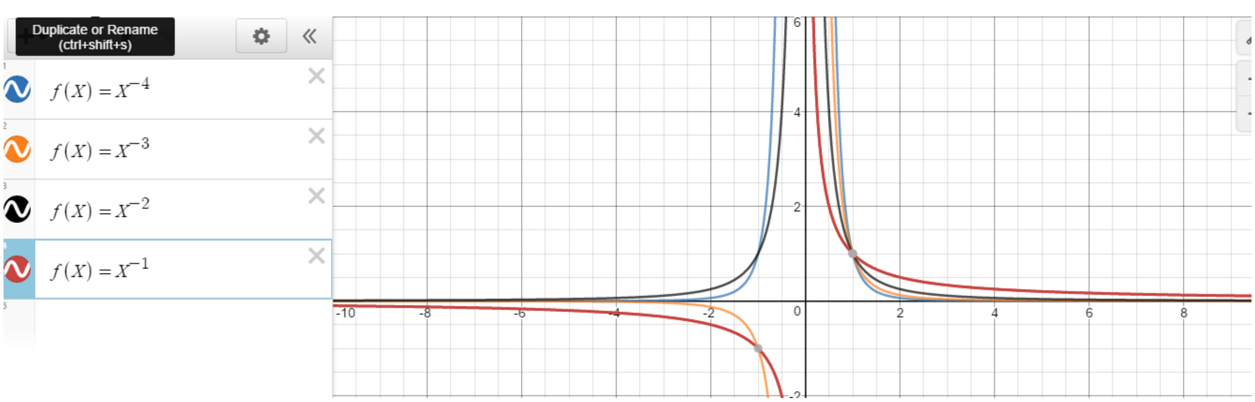
- Logarithm function spreads the smaller values and compresses the larger value therefore making the distribution symmetrical. Visually log looks like:

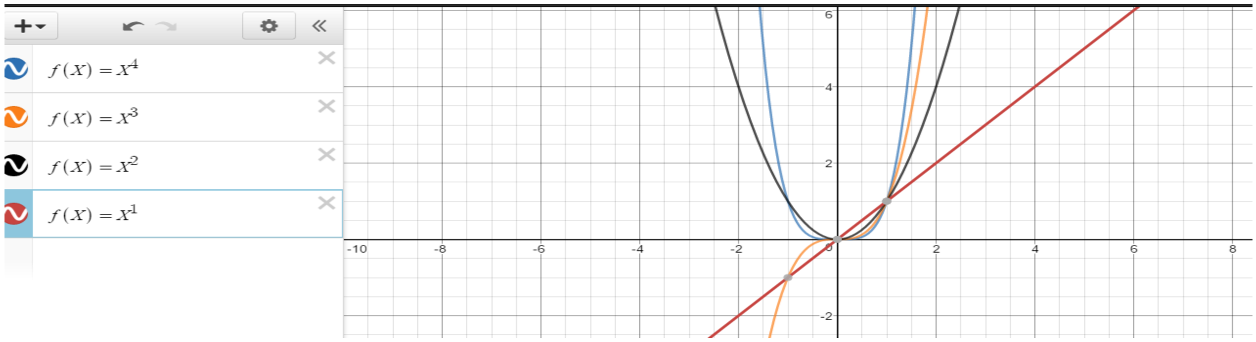


By Desmos grapher: https://www.desmos.com/calculator



- The values of negative powered X descends towards X^{-1} shows a pattern of compressing for larger values of X and spreading out for smaller values of X. For the positive powers the effect is reversed.





Constraints of Power Transformations:

- Power transformations are often used only for positive values of X since the log function for negative values or zero values are undefined, for square root function the negative values are imaginary, for inverse function zero values is undefined.
- If there are some negative and some positive values the monotone function gets transformed to non monotone functions. This is resolved by adding a constant before effecting the transforming. $X \to (X+s)^p$
- The power transformations are effective only if ratio of largest data value to smallest data value is large. If the ratio is close to 1 then power transformation is linear and therefore cannot bend the data.

Skewed Distributions:

- Distributions that are negative or positive are not interpretable and the mean cannot be considered as the appropriate measure of central tendency.
- Typically for a right skewed distribution a log transformation is used to compress the right tail and the then the distribution is more symmetrical.
- Typically for a left skewed distribution a power transformation(X^2 or X^3) is used to equalize the left tail and the right tail and then the distribution is more symmetrical.
- To decide which transformation is the best we can use the property that after the transformation which new data set has the value $\dfrac{UpperHinge - Median}{Median - LowerHinge}$ closest to 1 .

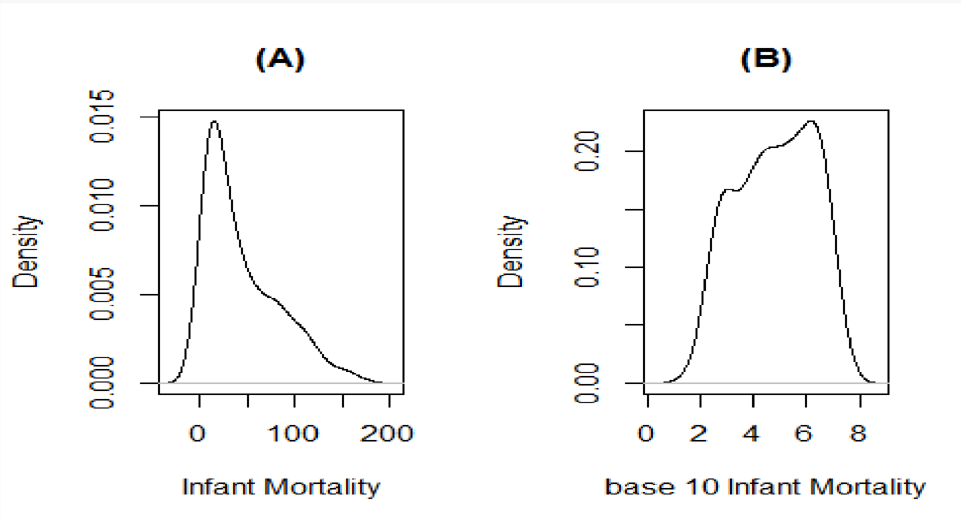
  Upper hinge is the Q3 and Lower Hinge is Q1
  For a left skewed data this ratio will be less than 1 and for a right skewed data the ratio is more than 1.
- The decision to determine the best transformation depends on the interpretability. The inverse time function might be used for the speed to traverse a given distance.

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(log10(infant.mortality)),
                      xlab="base 10 Infant Mortality",main="(B)"))
```



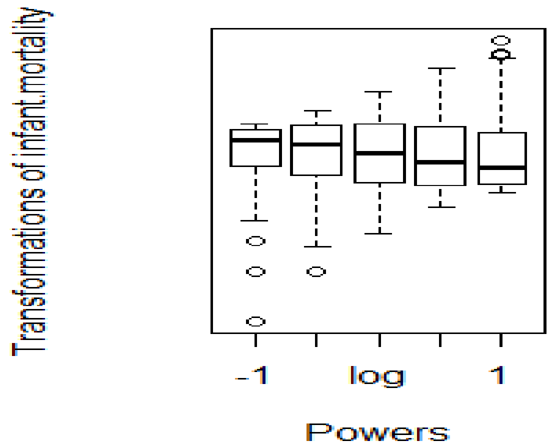

```
Median_Value=median(UN$infant.mortality)
Lower_Hinge=quantile(UN$infant.mortality,prob=.25)
Higher_Hinge=quantile(UN$infant.mortality,.75)
Tukey_Criterion=(Higher_Hinge-Median_Value)/(Median_Value-Lower_Hinge)
Tukey_Criterion
```

```
##      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-Low
er_Hinge_log)
Tukey_Criterion_log
```

```
##       75%
## 0.9785498
```

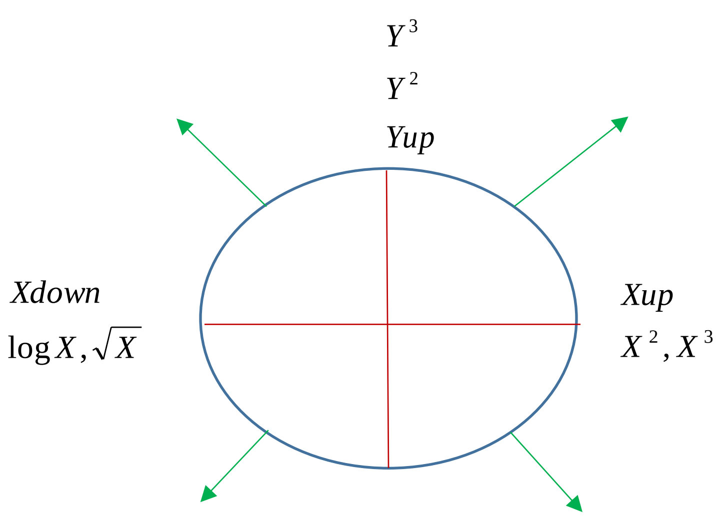
```
symbox(~infant.mortality,data=UN)
```



You can clearly see the log function best normalizes the infant.mortality. If the variable is overfitted then we use the square root function

Non Linear Transformation

- Transforming data to linear data is that the Linear Statistical model can be used to interpret data.
- For a data which has more than one explanatory variables the Linear models are easier to interpret and visualize as compared to the non parametric regression techniques.
- For a simple monotone function the transformations of a power or square root can be used to linearize the function. For a function that is not simple monotone or non monotone these transformations cannot be used.
- The Mosteller and Tukey's bulging rule helps in determining the transformation. For instance if the bulge is down and to the right then to linearlize the data we can transform Y down the ladder of powers or transform the X up the ladder.



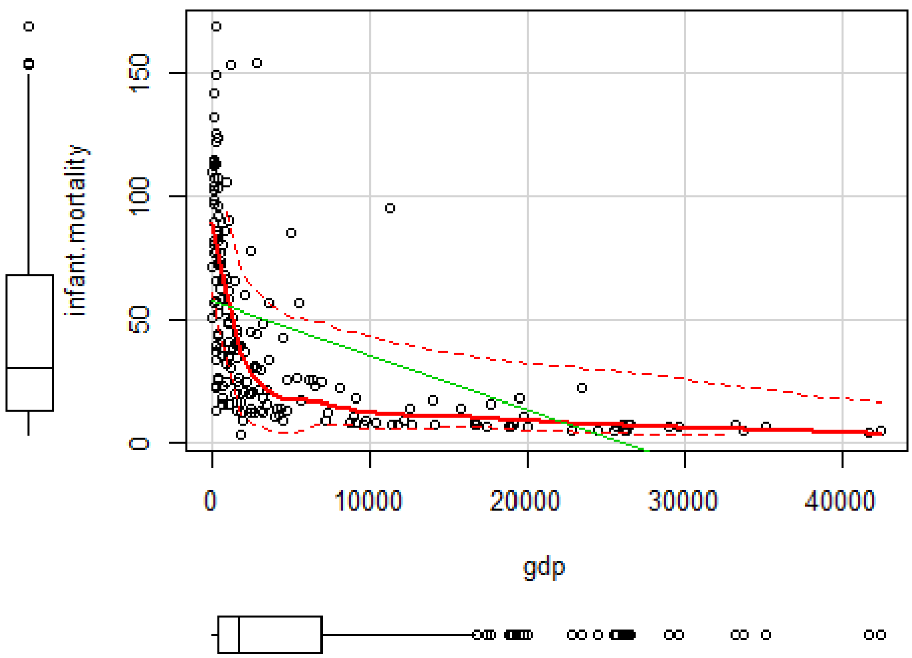
$$Ydown$$

$$\sqrt{Y}$$

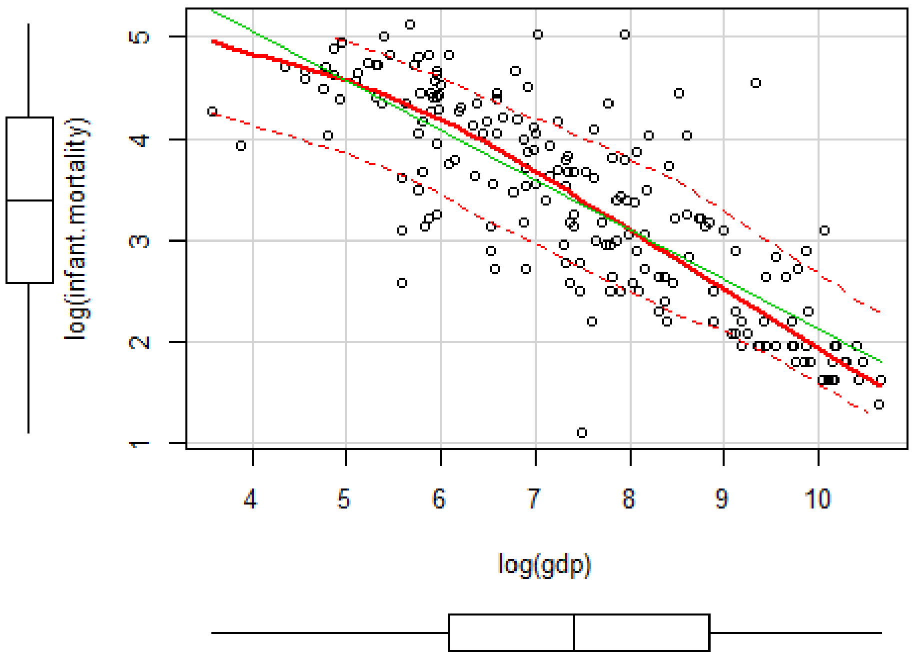
$$\log(Y)$$

Non Linearity using Tukey's Bulging Rule:

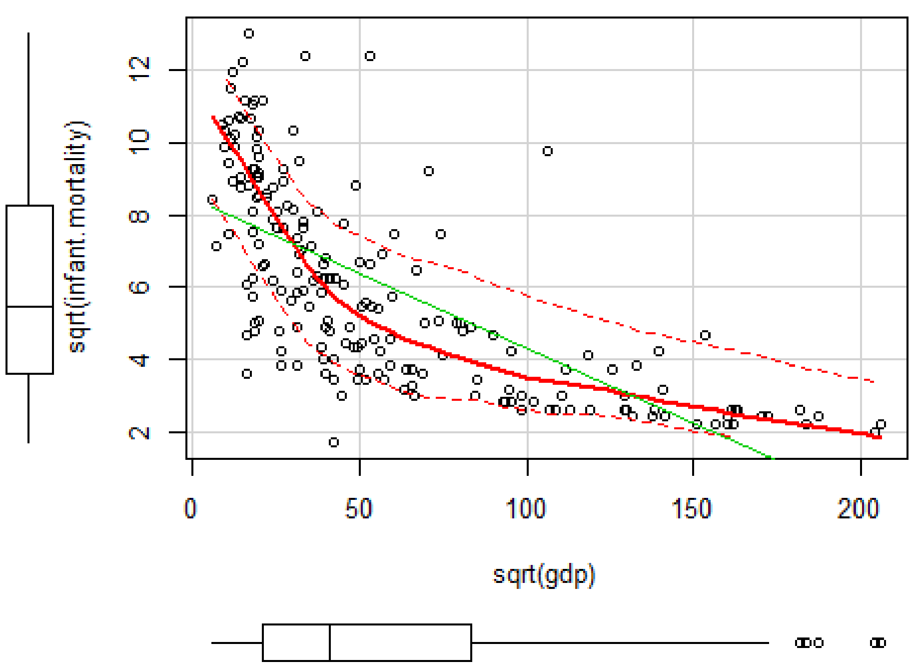
```
scatterplot(infant.mortality~gdp,data=UN)
```



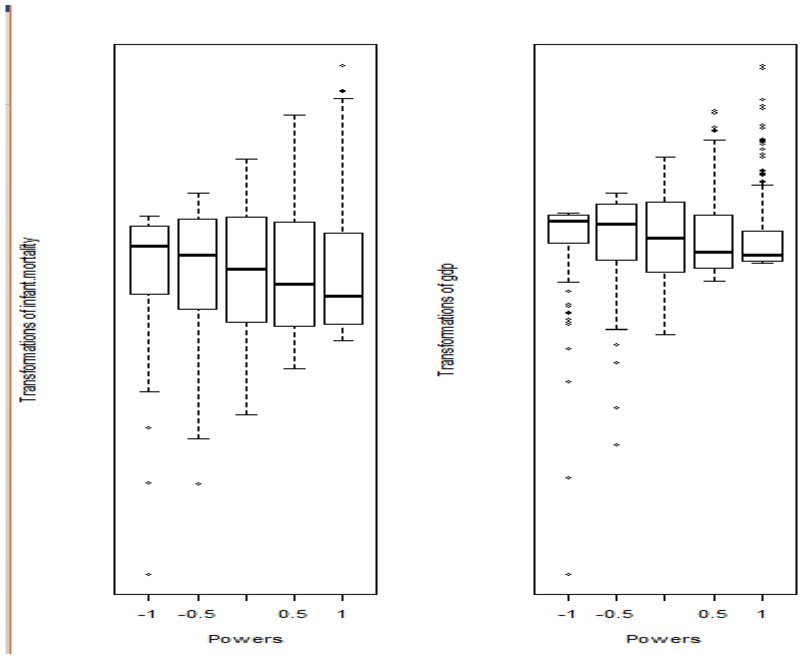

```
scatterplot(log(infant.mortality)~log(gdp),data=UN)
```




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



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



- It is clear that the log transformation is the best but if it overfits the data then the next best ie square root can be used for both explanatory and response variable.
- If the distribution is nonlinear and monotone the Mosteller and Tukeys rule can be used. The only additional rule to follow is to transform out of X or the variable that is individually skewed.
- To decide the choice between log(X) and $\sqrt{X}$ and other square roots of X we need to choose a transformation that makes the relationship approximates linear but does not over correct the relationship.
-

Transforming non constant variance:

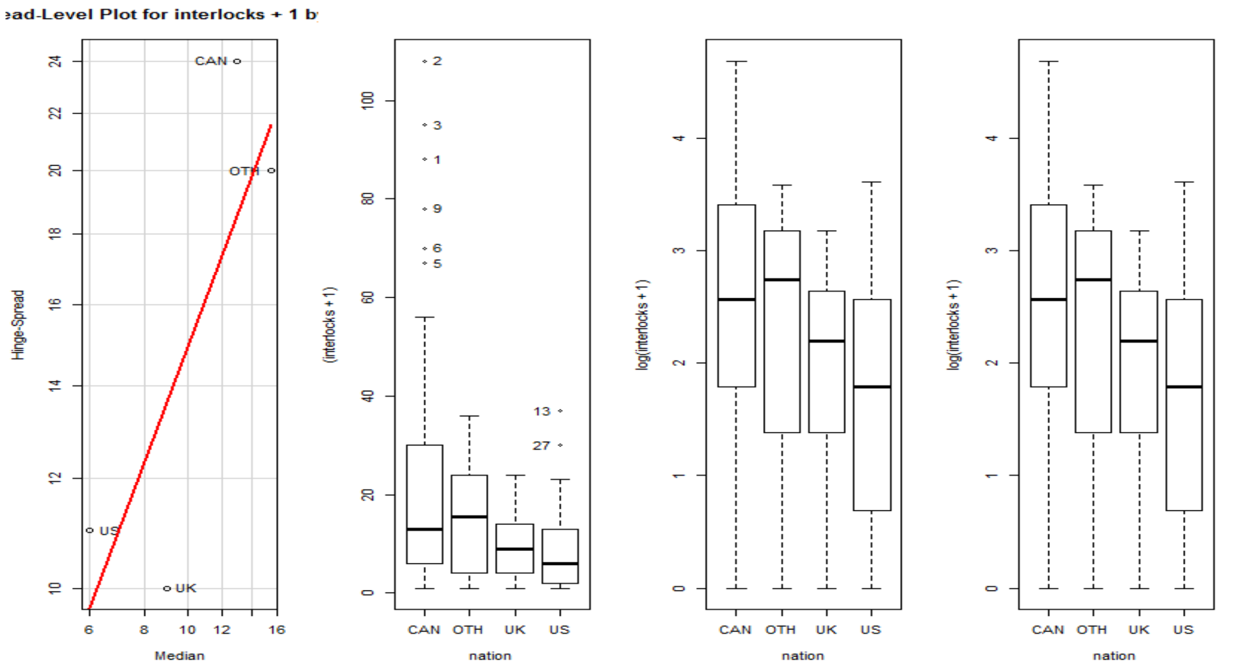
- If a variable depicted by various groups has different variations across groups, the data comparisons across groups is not interpretative. To ensure substantive analysis, transformations are used.
- The pattern for differences of variations often systematically increases or decreases with higher levels (the indices of levels can be median and hinge spread).
- According to Tukey if the aforementioned constraint is satisfied then the median and hinge spread can be plotted on the X and Y axis. If the slope is a positive (with increase levels of spread increases) one then the transformation that should be used are the descending powers of powers and roots..
  Linear model: Log spread = a + b Log level

The corresponding spread stabilizing transformation uses power is p=1-b. If 0<p<1 then the log transformation is used and the linear fit can be expressed as

- Data which has unequal variances across groups often shows skewness. Typically for different levels the variance increases/decreases so does the skewness.
- If the data is both negatively skewed and the slope is negative that is with the increase in levels the spread decreases then the transformation used is ascending up like the X^2. This is not seen often.

Ornstein dataset https://cran.r-project.org/web/packages/car/car.pdf

```
Interlocks are the number of work units that a country owes to another

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)

# we added one to remove the chance of a zero as then we cannot apply a log

function

##     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 Remember the closest value to

# this is p=0 which is the log transformation in the Box CoxTransformation ru
le # therefore we can use log.

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

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

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

Boxplot(log(interlocks+1)~(nation^2),data=Ornstein)

# Here I used the bulging rule
```




```
# The spread is less variable across groups

# There are no outliers and within group spread

 # is less skewed.
```

Transforming Proportions

- If the data is in a proportions format like the data that has a dichotomous answers the Power transformations are not as effective especially if the data approaches the bounds of 0 and 1.
- The common proportion transformation is logit. It can be represented mathematically by the formula: $P \to \log it(P) = \log_e \dfrac{P}{1-P}$

  The P/1-P is the odds of an event. The logit transformation eliminates the lower and higher boundaries of the scale, spreads the tails of the distributions and therefore making the distribution more symmetric about 0.
  Images to be uploaded from page 73 and 74

- Probit transformation $P \to probit(P) = \phi^{-1}(P)$ Is the inverse distribution function of a normal distribution or quantile function. If the scales are equated then the probit and logit are approximately the same. $\log it = (\pi / \sqrt{3})\,probit$
- Arcsine square root transformation has a similar shape. $P \to \sin^{-1}\sqrt{P}$
- Tukey set the rules for these proportions transformations. These are folded powers and roots their power q takes on values between 0 and 1.

Tukey's Propotional transformation can be represented as

$P \to P^q - (1-P)^q$

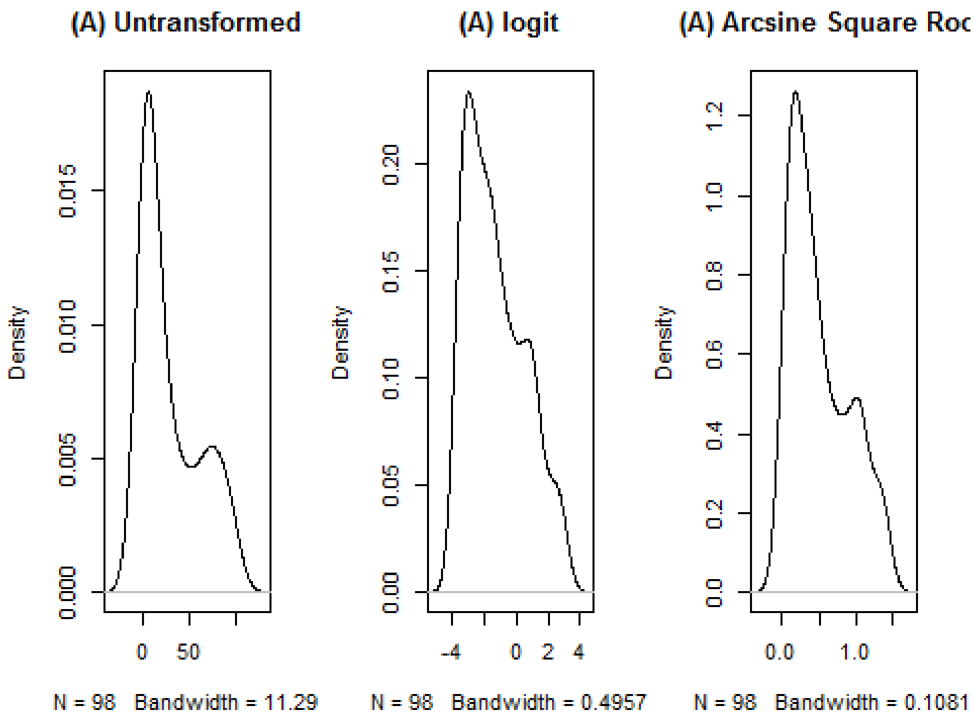
- If q=0 the logit transformation is used. If q=.14 (or close approximation) then multiple of probit transformation is used and if q=.41 produces a multiple of arcsine square root transformation.
- Power transformations are not applied if the proportions are exactly 0 or 1 or are always close to this value .If we have the original counts then we can transform the data by the applying using P' instead of P where P' is as follows:

  $\boxed{P' = \frac{F+1}{N+1}}$ F is the frequency count in focal category and N is the total count.
- If the original data is not available then the following proportion is used:

  $P' = .005 + .99P$ This maps the proportions in the interval [.005,.995]

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 Roo
t")

})# adjust factor smoothes the multiple modes by the decreasing the bandwidth
fraction

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


(A) Untransformed
(A) logit
(A) Arcsine Square Roc
Density
0.000
0.005
0.010
0.015
0
50
N = 98 Bandwidth = 11.29
0.00
0.05
0.10
0.15
0.20
-4
0
2
4
N = 98 Bandwidth = 0.4957
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
1.0
N = 98 Bandwidth = 0.1081