Chapter 04.
Transforming data

Likan Zhan
Beijing Language and Culture University

2020-11-13

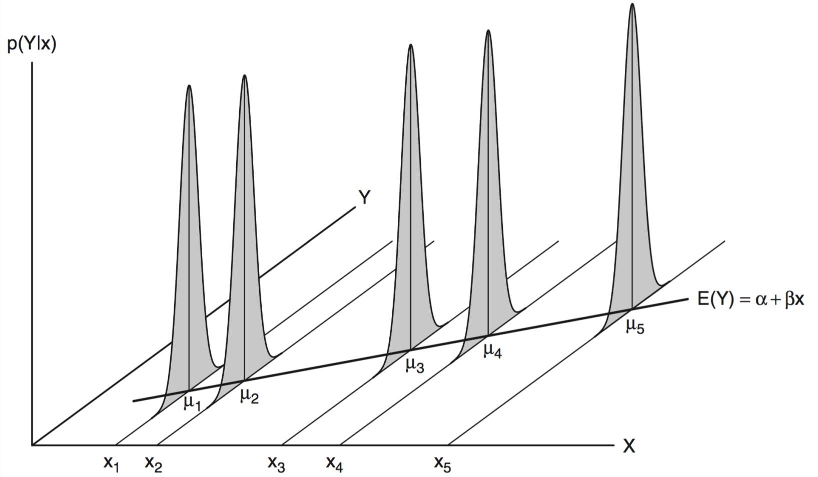
https://likan.info
zhanlikan@blcu.edu.cn

Table of Contents

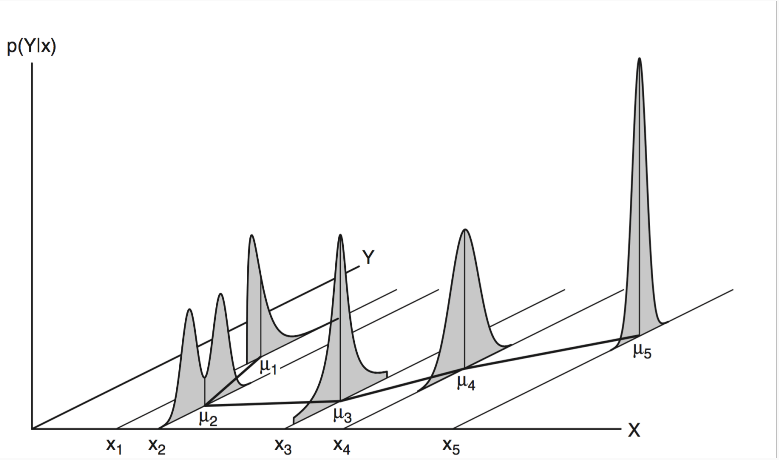
Table of Contents

Why transformations are important

Why transformations are important



Why transformations are important



Family of powers and roots

Family of powers and roots

- Powers and roots transformations

$$X \longrightarrow X^p \qquad (1)$$

Family of powers and roots

- Powers and roots transformations

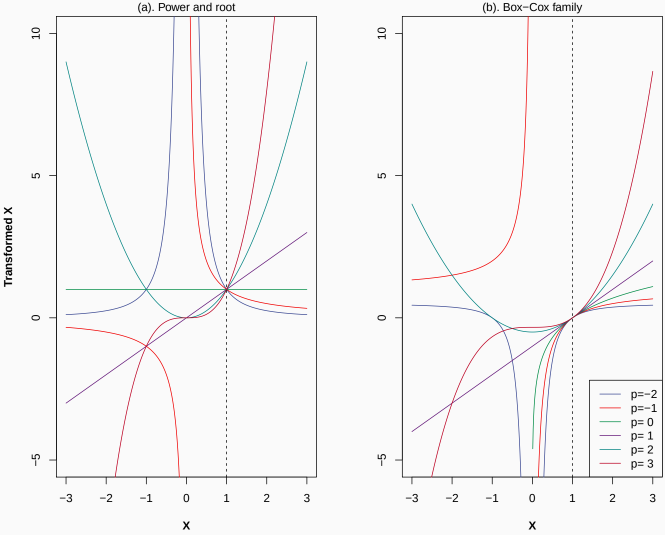
$$X \longrightarrow X^p \tag{1}$$

- The Box-Cox family of transformations

$$X \longrightarrow X^{(p)} \equiv \begin{cases} \frac{X^p - 1}{p} & p \neq 0 \\ \log_e(X) & p = 0 \end{cases} \tag{2}$$

$$\lim_{p \to 0} \frac{X^p - 1}{p} = \log_e X$$

Power and root transformations



Effects of the various power transformations

X	X^(-2)	X^(-1)	log2(X)	X^2	X^3
1	1.00	1.00	0.00	1	1
2	0.25	0.50	1.00	4	8
3	0.11	0.33	1.58	9	27
4	0.06	0.25	2.00	16	64

Effects of the various power transformations

Effects of the various power transformations

- Add a positive start

X	X^2	(X+3)^2
-2	4	1
-1	1	4
0	0	9
1	1	16
2	4	25

Effects of the various power transformations

Effects of the various power transformations

- Add a negative start

X	log10(X)	X-2010	log10(X - 2010)^2
2011	3.3034	1	0.000
2012	3.3036	2	0.301
2013	3.3038	3	0.477
2014	3.3041	4	0.602
2015	3.3043	5	0.699

Applications of power transformations

Applications of power transformations

- Power transformations can make a skewed distribution more symmetric.

Applications of power transformations

- Power transformations can make a skewed distribution more symmetric.
- Power transformations can also be used to make many nonlinear relationships more nearly linear.

Table of Contents

Why bother

Why bother

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data.

Why bother

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data.
- The outlying values in the direction of the skew are brought in toward the main body of the data when the distribution is made more symmetric.

Why bother

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data.
- The outlying values in the direction of the skew are brought in toward the main body of the data when the distribution is made more symmetric.
- Unusual values in the direction opposite to the skew can be hidden prior to transforming the data.

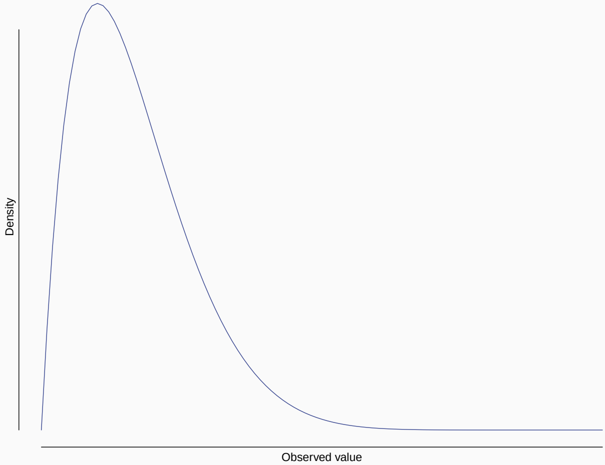
Why bother

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data.
- The outlying values in the direction of the skew are brought in toward the main body of the data when the distribution is made more symmetric.
- Unusual values in the direction opposite to the skew can be hidden prior to transforming the data.
- Some of the most common statistical methods summarize distributions using means.

Why bother

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data.
- The outlying values in the direction of the skew are brought in toward the main body of the data when the distribution is made more symmetric.
- Unusual values in the direction opposite to the skew can be hidden prior to transforming the data.
- Some of the most common statistical methods summarize distributions using means.
- Least-squares regression, for example, traces the mean of Y conditional on X. The mean of a skewed distribution is not, however, a good summary of its center.

Why bother



How a power transformation can eliminate a skew

How a power transformation can eliminate a skew

- Descending the ladder of powers to log X makes the a positive skew more symmetric by pulling in the right tail.

X	log10(X)
1	0
10	1
100	2
1000	3

How a power transformation can eliminate a skew

How a power transformation can eliminate a skew

- Ascending the ladder of powers (toward X^2 and X^3) can correct a negative skew.

X	X^2
0.00000	0
1.00000	1
1.41421	2
1.73205	3

Use order statistics to find an effective transformation

Use order statistics to find an effective transformation

- Order statistics, such as the median and hinges, are preserved under nonlinear monotone transformations of the data, such as powers and roots, that is

Use order statistics to find an effective transformation

- Order statistics, such as the median and hinges, are preserved under nonlinear monotone transformations of the data, such as powers and roots, that is

$$\text{If } X' = X^{(p)}, \text{ then}$$

Use order statistics to find an effective transformation

- Order statistics, such as the median and hinges, are preserved under nonlinear monotone transformations of the data, such as powers and roots, that is

$$\text{If } X' = X^{(p)}, \text{ then}$$
$$X'_{(i)} = \left[X_{(i)}\right]^{(p)}$$

Use order statistics to find an effective transformation

- Order statistics, such as the median and hinges, are preserved under nonlinear monotone transformations of the data, such as powers and roots, that is

$$\text{If } X' = X^{(p)}, \text{ then}$$
$$X'_{(i)} = \left[X_{(i)}\right]^{(p)}$$
$$\text{median}(X') = [\text{median}(X)]^{(p)}$$

Use order statistics to find an effective transformation

- Order statistics, such as the median and hinges, are preserved under nonlinear monotone transformations of the data, such as powers and roots, that is

$$\text{If } X' = X^{(p)}, \text{ then}$$
$$X'_{(i)} = \left[X_{(i)}\right]^{(p)}$$
$$\text{median}(X') = [\text{median}(X)]^{(p)}$$

- This is not the case for the mean and standard deviation.

Use order statistics to find an effective transformation

$$\text{Ratio} = \frac{\text{Upper hinge (3rd Qu)} - \text{Median (2nd Qu)}}{\text{Median (2nd Qu)} - \text{Lower hinge (1st Qu)}}$$

Use order statistics to find an effective transformation

$$\text{Ratio} = \frac{\text{Upper hinge (3rd Qu)} - \text{Median (2nd Qu)}}{\text{Median (2nd Qu)} - \text{Lower hinge (1st Qu)}}$$

$$\Rightarrow \begin{cases} \approx 1 & \text{Symmetric distribution} \\ > 1 & \text{Positive skew} \\ < 1 & \text{Negative skew} \end{cases}$$

Use order statistics to find an effective transformation

```
library(carData)
infM <- UN[complete.cases(UN), "infantMortality"]
Trns <- rbind(
        "X^(-1)"= fivenum((infM) ^ (-1)),
     "X^(-0.5)" = fivenum((infM) ^ (-1 / 2)),
     "log10(X)" = fivenum(log10(infM)),
      "X^(0.5)" = fivenum(sqrt(infM)),
            "X" = fivenum(infM))
colnames(Trns) <-
  c("Min", "HL", "Median", "HU", "Max")
Ratio <- (Trns[, "HU"] - Trns[, "Median"]) /
         (Trns[, "Median"] - Trns[, "HL"])
Trns <- cbind(Trns, Ratio)
```

Use order statistics to find an effective transformation

	Min	HL	Median	HU	Max	Ratio
X^(-1)	0.01	0.02	0.05	0.14	0.52	2.99
X^(-0.5)	0.09	0.15	0.23	0.37	0.72	1.87
log10(X)	0.28	0.86	1.29	1.66	2.10	0.85
X^(0.5)	1.38	2.69	4.43	6.77	11.16	1.35
X	1.92	7.24	19.64	45.89	124.53	2.12

Asymmetry and power transformation

Asymmetry and power transformation

- The symbox function from package *car* first transforms x to each of a series of selected powers, with each transformation standardized to mean 0 and standard deviation 1.

Asymmetry and power transformation

- The symbox function from package *car* first transforms x to each of a series of selected powers, with each transformation standardized to mean 0 and standard deviation 1.
- The results are then displayed side-by-side in boxplots, permitting a visual assessment of which power makes the distribution reasonably symmetric.

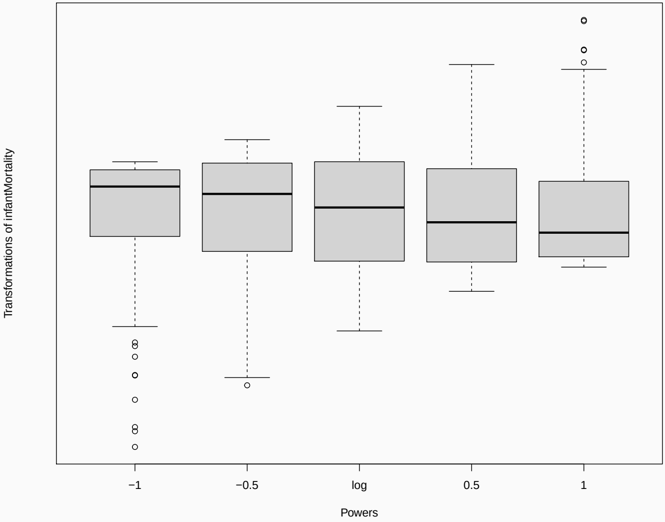
Asymmetry and power transformation

- The symbox function from package *car* first transforms x to each of a series of selected powers, with each transformation standardized to mean 0 and standard deviation 1.
- The results are then displayed side-by-side in boxplots, permitting a visual assessment of which power makes the distribution reasonably symmetric.
- The boxplots for various power transformations of *infant mortality*:

Asymmetry and power transformation

```
car::symbox( ~ infantMortality,
    trans = car::bcPower,
    powers = c(-1, -0.5, 0, 0.5, 1), data = UN)
```

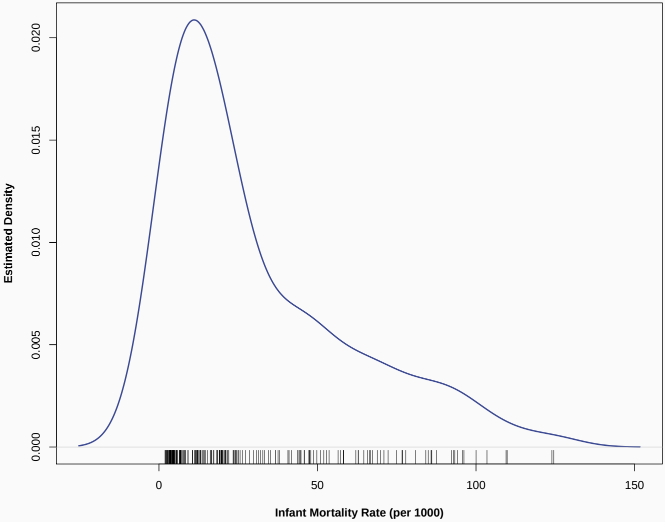
Asymmetry and power transformation



Infant mortality rates: Original data

```
UN <- UN[complete.cases(UN), ]
dnst <- density(UN[, "infantMortality"])
plot(dnst, font.lab = 2, main = "",
  xlab = "Infant Mortality Rate (per 1000)",
  ylab = "Estimated Density",
  lty = 1, lwd = 2, col = lkz[1])
rug(UN[, "infantMortality"])
```

Infant mortality rates: Original data



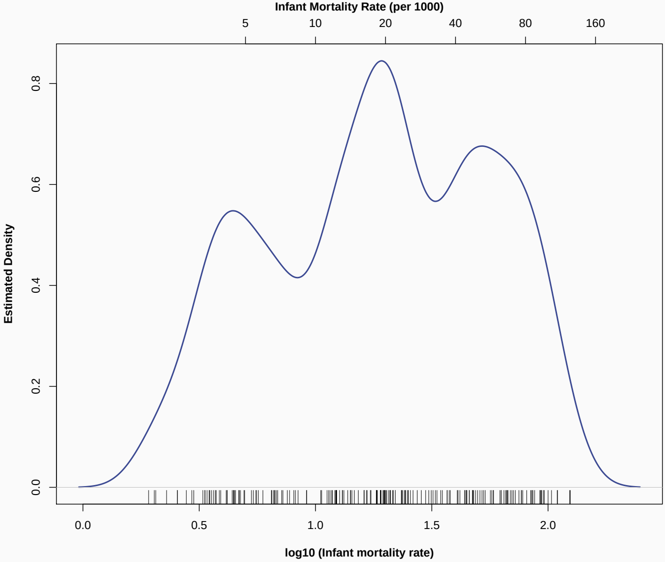
Infant mortality rates: Transformed results

```r
UN[, "log.infantMortality"] <-
  log10(UN[, "infantMortality"])
dnstl <-
  density(UN[, "log.infantMortality"], bw = 0.1)
```

Infant mortality rates: Transformed results

```
plot(dnstl, font.lab = 2, main = "",
  xlab = "log10 (Infant mortality rate)",
  ylab = "Estimated Density",
  lty = 1, lwd = 2, col = lkz[1])
rug(UN[, "log.infantMortality"])
LO <- c(5, 10, 20, 40, 80, 160)
axis(side = 3, at = log10(LO), labels = LO)
mtext("Infant Mortality Rate (per 1000)",
  font = 2, side = 3, padj = -3.7)
```

Infant mortality rates: Transformed results



Choose a meaningful transformation

Original scale	Transformation	Transformed scale
Time duration	Inverse	Speed
Response latency	Inverse	Frequency
Area	Square root	Length
Length	Cube	Volume

Table of Contents

Why bother

Why bother

- Linear relationships are particularly simple.

Why bother

- Linear relationships are particularly simple.
- Especially when there are several explanatory variables, the alternative of nonparametric regression may not be feasible because of the sparseness of the data.

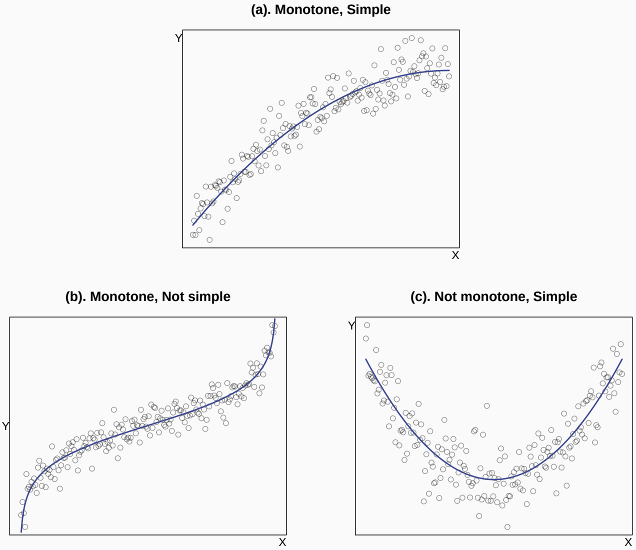
Why bother

- Linear relationships are particularly simple.
- Especially when there are several explanatory variables, the alternative of nonparametric regression may not be feasible because of the sparseness of the data.
- There are certain technical advantages to having linear relationships among the explanatory variables in a regression analysis.

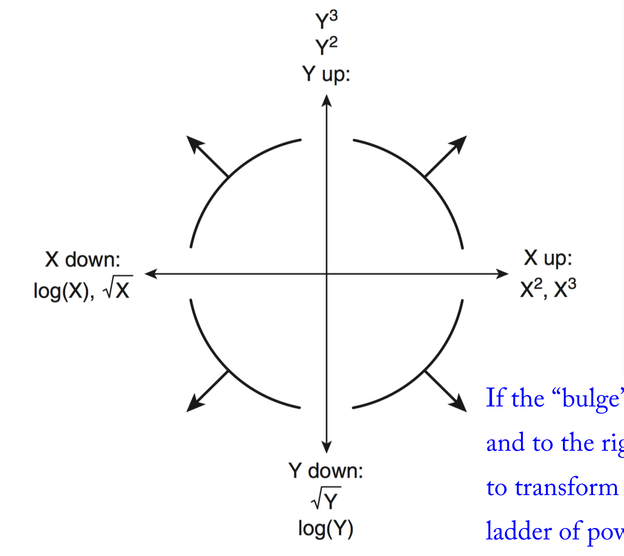
Why bother

- Linear relationships are particularly simple.
- Especially when there are several explanatory variables, the alternative of nonparametric regression may not be feasible because of the sparseness of the data.
- There are certain technical advantages to having linear relationships among the explanatory variables in a regression analysis.
- There is a simple and elegant statistical theory for linear models, which we explore in subsequent chapters.

Simple and monotone relationship

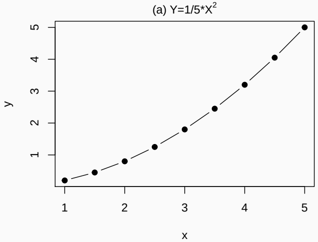


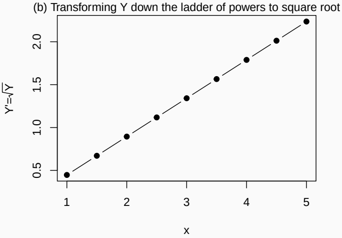
Bulging rule for selecting a transformation

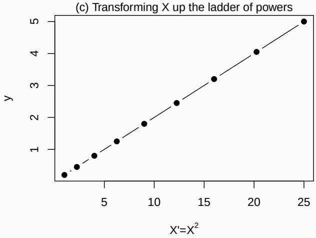


If the "bulge" points down and to the right, we need to transform Y down the ladder of powers or X up (or both)

Use power transformation to linearize relations







Prestige of Canadian Occupations

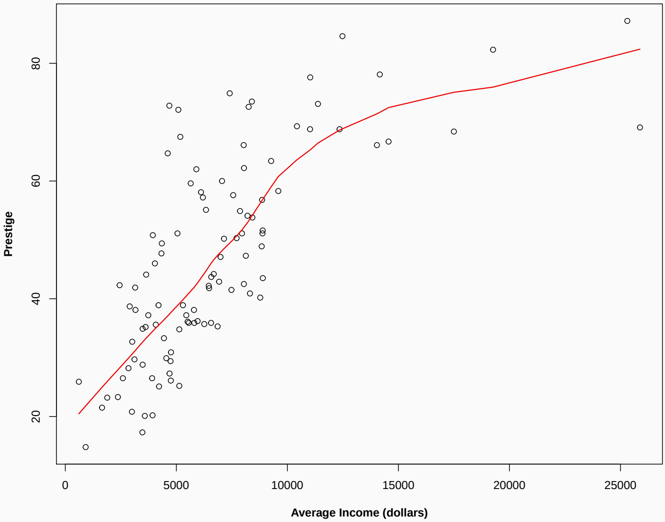
```
str(Prestige)

'data.frame':^^I102 obs. of  6 variables:
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ income   : int  12351 25879 9271 8865 8403 11030 825
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 7
 $ census   : int  1113 1130 1171 1175 2111 2113 2133 2
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2
```

Prestige of Canadian Occupations

```r
lowl <- lowess(
  Prestige[,   "income"],
  Prestige[, "prestige"], f = 0.6)
plot(prestige ~ income, data = Prestige,
  type = "p", font.lab = 2,
  xlab = "Average Income (dollars)",
  ylab = "Prestige")
lines(lowl, lwd = 1.5, col = lkz[2])
```

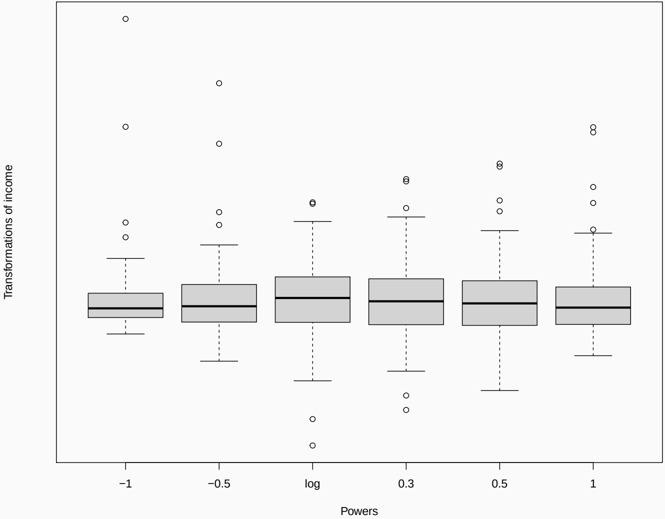
Prestige of Canadian Occupations



Choose a transformation

```
car::symbox( ~ income,
    trans = car::basicPower, data = Prestige,
    powers = c(-1, -0.5, 0, 0.3, 0.5, 1))
```

Choose a transformation



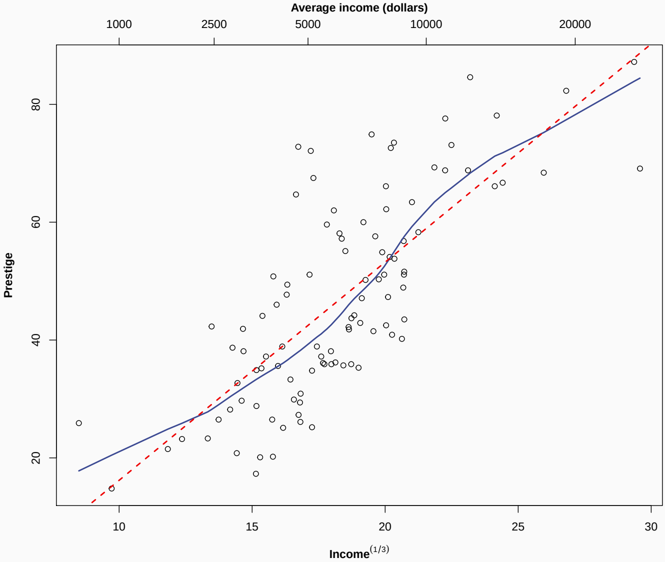
Transforming income down the ladder of powers

```
Prestige[, "incomecr"] <-
  (Prestige[, "income"]) ^ (1 / 3)
Prestige[, "incomelg"] <-
  log10(Prestige[, "income"])
locr <- lowess(Prestige[, "incomecr"],
               Prestige[, "prestige"], f = 0.6)
lolg <- lowess(Prestige[, "incomelg"],
               Prestige[, "prestige"], f = 0.6)
lmcr <- lm(prestige ~ incomecr, data = Prestige)
lmlg <- update(lmcr, . ~ incomelg)
```

Transforming down the ladder of powers: cube root

```r
plot(prestige ~ incomecr, data = Prestige,
  xlab = expression(bold(paste(Income^(1/3)))),
  ylab = "Prestige", font.lab = 2, type = "p")
lines(locr, lty = 1, lwd = 2, col = lkz[1])
abline(lmcr, lty = 2, lwd = 2, col = lkz[2])
LO <- c(1000, 2500, 5000, 10000, 20000)
axis(3, at = (LO) ^ (1 / 3), labels = LO)
mtext("Average income (dollars)",
  font = 2, side = 3, padj = -3.7)
```

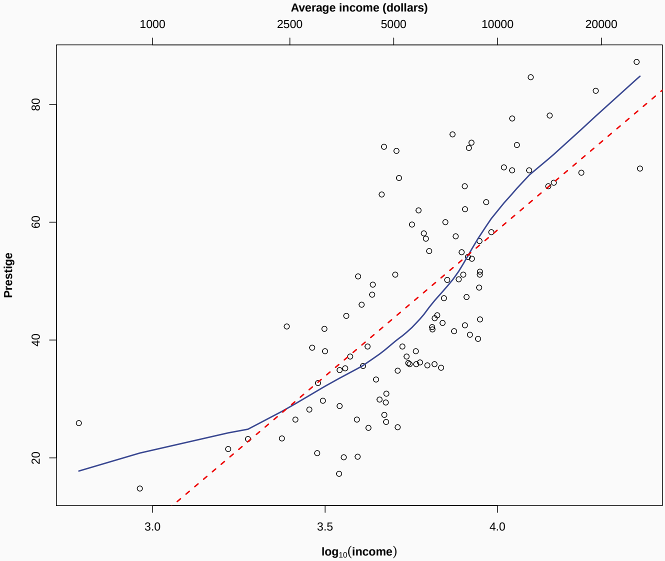
Transforming down the ladder of powers: cube root



Transforming income down the ladder of powers: log

```
plot(prestige ~ incomelg, data = Prestige,
  xlab = expression(bold(paste(log[10](income)))),
  ylab = "Prestige", font.lab = 2, type = "p")
lines(lolg, lty = 1, lwd = 2, col = lkz[1])
abline(lmlg, lty = 2, lwd = 2, col = lkz[2])
LO <- c(1000, 2500, 5000, 10000, 20000)
axis(3, at = log10(LO), labels = LO)
mtext("Average income (dollars)",
  font = 2, side = 3, padj = -3.7)
```

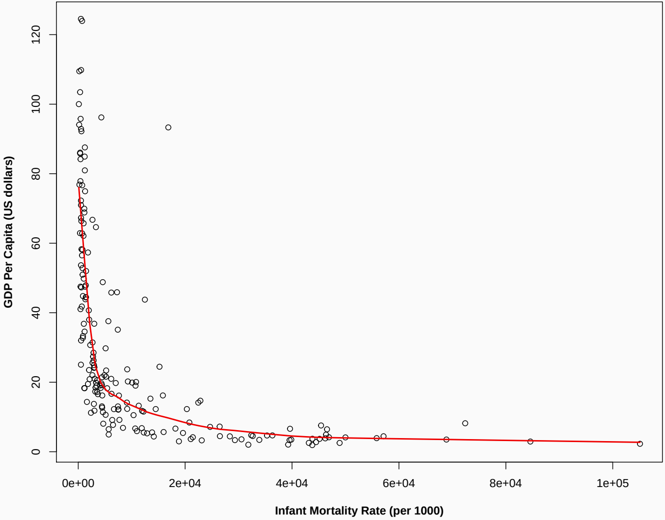
Transforming income down the ladder of powers: log



Infant mortality and GDP

```
low <- lowess(UN[, "ppgdp"], UN[, "infantMortality"], 
plot(infantMortality ~ ppgdp, data = UN, type = "p",
     font.lab = 2,
     xlab = "Infant Mortality Rate (per 1000)",
     ylab = "GDP Per Capita (US dollars)")
lines(low, lwd = 2, col = lkz[2])
```

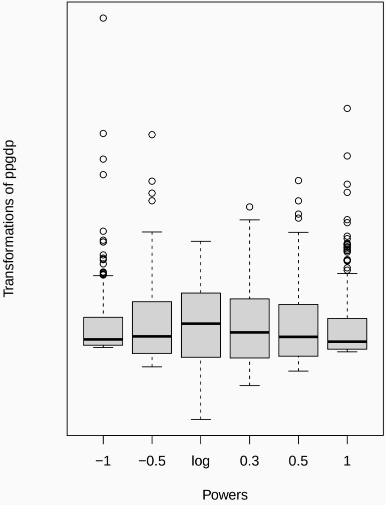
Infant mortality and GDP

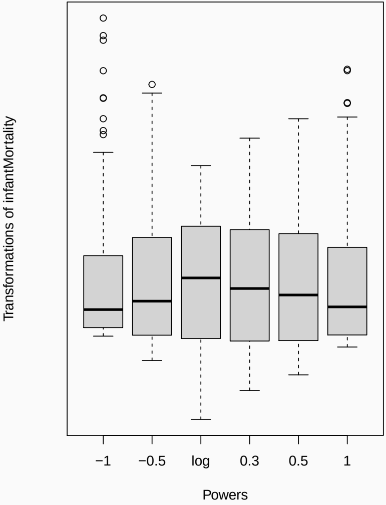


Choose a transformation

```
par(mfrow = c(1, 2))
car::symbox( ~ ppgdp,
  trans = car::basicPower,
  data = UN, powers = c(-1, -0.5, 0, 0.3, 0.5, 1))
car::symbox( ~ infantMortality,
  trans = car::basicPower,
  data = UN, powers = c(-1, -0.5, 0, 0.3, 0.5, 1))
```

Choose a transformation





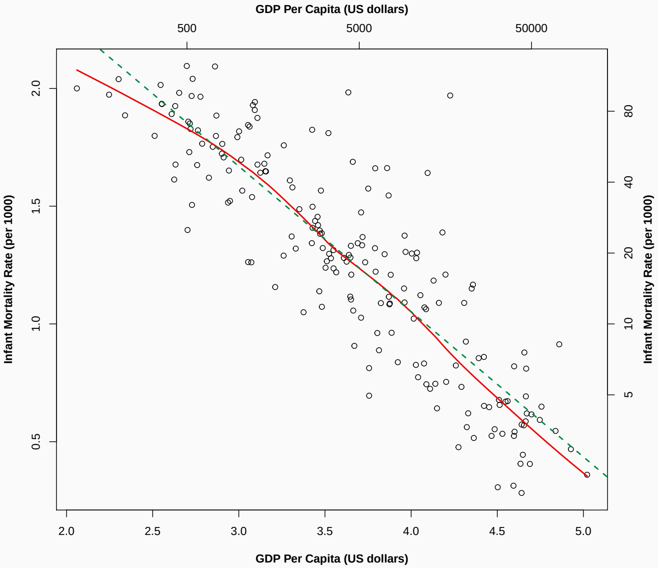
Transforming both mortality and GDP down

```
UN <- UN[complete.cases(UN), ]
UN[, "log.infantMortality"] <-
                      log10(UN[, "infantMortality"])
UN[, "log.ppgdp"] <- log10(UN[, "ppgdp"])
lws <- lowess(UN[, "log.ppgdp"],
    UN[, "log.infantMortality"], f = 0.5)
lm <- lm(log.infantMortality ~ log.ppgdp, data = UN)
```

Transforming both mortality and GDP down

```
par(oma = c(0, 0, 0, 4)) #set margins
plot(log.infantMortality ~ log.ppgdp, data = UN,
  type = "p", font.lab = 2,
  xlab = "GDP Per Capita (US dollars)",
  ylab = "Infant Mortality Rate (per 1000)")
lines(lws, lwd = 2, col = lkz[2])
abline(lm, lty = 2, lwd = 2, col = lkz[3])
Original.X <- c(50, 500, 5000, 50000)
axis(3, at = log10(Original.X), labels = Original.X)
mtext("GDP Per Capita (US dollars)",
  font = 2, side = 3, padj = -4)
Original.Y <- c(5, 10, 20, 40, 80, 160)
axis(4, at = log10(Original.Y), labels = Original.Y)
mtext("Infant Mortality Rate (per 1000)",
  font = 2, side = 4, padj = 4)
```

Transforming both mortality and GDP down



The least-squares regression line

```
fm <- lm(log(infantMortality, 10) ~
         log(ppgdp, 10), data = UN)
```

The least-squares regression line

```
fm <- lm(log(infantMortality, 10) ~
         log(ppgdp, 10), data = UN)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5194	0.0916	38.43	0.0000
log(ppgdp, 10)	-0.6168	0.0247	-25.02	0.0000

The least-squares regression line

```r
fm <- lm(log(infantMortality, 10) ~
         log(ppgdp, 10), data = UN)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5194	0.0916	38.43	0.0000
log(ppgdp, 10)	-0.6168	0.0247	-25.02	0.0000

$$\log_{10} \widehat{\text{Infant mortality}} = 3.52 - 0.617 \times \log_{10}\text{GDP}$$

Table of Contents

Why bother?

- When a variable has very different degrees of variation in different groups, it becomes difficult to examine the data and to compare differences in level across the groups.
- Differences in spread are often systematically related to differences in level: Groups with higher levels tend to have higher spreads.

Transforming nonconstant spread

Transforming nonconstant spread

- To explore the nonconstant spread, Tukey (1977) suggests graphing the **log hinge-spread** against the **log median**.

Transforming nonconstant spread

- To explore the nonconstant spread, Tukey (1977) suggests graphing the **log hinge-spread** against the **log median**.
- The **slope** of the linear trend, if any, in the spread-level plot can be used to suggest a spread-stabilizing power transformation of the data.

Transforming nonconstant spread

- To explore the nonconstant spread, Tukey (1977) suggests graphing the **log hinge-spread** against the **log median**.
- The **slope** of the linear trend, if any, in the spread-level plot can be used to suggest a spread-stabilizing power transformation of the data.
- Express the linear fit as

$$\log \text{spread} \approx \alpha + b \log \text{level}$$

Then the corresponding spread-stabilizing transformation uses the power

$$p = 1 - b$$

Transforming nonconstant spread

Transforming nonconstant spread

- When spread is positively related to level (i.e., $b > 0$), therefore, we select a transformation down the ladder of powers and roots.

Transforming nonconstant spread

- When spread is positively related to level (i.e., $b > 0$), therefore, we select a transformation down the ladder of powers and roots.
- A **negative** association between level and spread is less common but can be corrected by ascending the ladder of powers.

Interlocking Directorates Among Major Canadian Firms

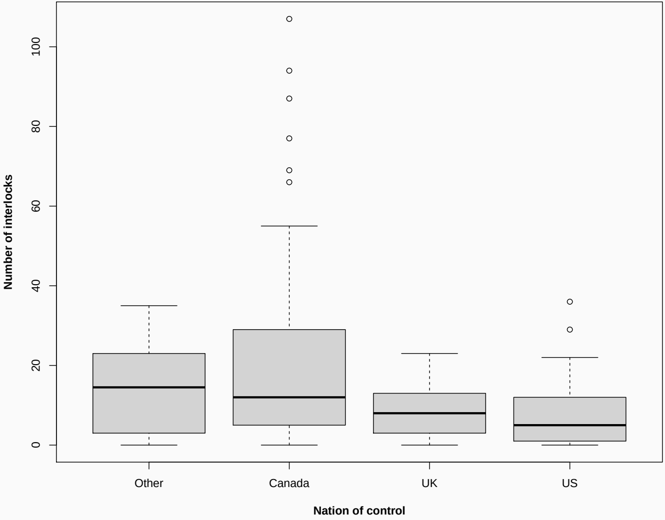
```
str(Ornstein)

'data.frame':^^I248 obs. of  4 variables:
 $ assets    : int  147670 133000 113230 85418 75477 40
 $ sector    : Factor w/ 10 levels "AGR","BNK","CON",..
 $ nation    : Factor w/ 4 levels "CAN","OTH","UK",..:
 $ interlocks: int  87 107 94 48 66 69 46 16 77 6 ...
```

Interlocking Directorates Among Major Canadian Firms

```
Ornstein[, "nation"] <- factor(Ornstein[, "nation"],
              levels = c("OTH", "CAN", "UK", "US"))
levels(Ornstein[, "nation"]) <-
                  c("Other", "Canada", "UK", "US")
boxplot(interlocks ~ nation,
    data = Ornstein, main = "",
    xlab = expression(bold("Nation of control")),
    ylab = expression(bold("Number of interlocks")))
```

Interlocking Directorates Among Major Canadian Firms



The hinge-spread of the Ornstein data

```
Ornstein[, "interlocks"] <-
    Ornstein[, "interlocks"] + 1
hs <- aggregate(interlocks ~ nation,
    data = Ornstein, FUN = fivenum)
hss <- data.frame(hs[, 1], hs[, 2])
colnames(hss) <- c("Nation", "Min", "Lower Hinge",
                   "Median", "Upper Hinge", "Max")
hss[, "Hinge Spread"] <- hss[, "Upper Hinge"] -
                         hss[, "Lower Hinge"]
hss <- hss[, c(1, 3, 4, 5, 7)]
```

The hinge-spread of the Ornstein data

Nation	Lower Hinge	Median	Upper Hinge	Hinge Spread
Other	4	15.5	24	20
Canada	6	13.0	30	24
UK	4	9.0	14	10
US	2	6.0	13	11

Transforming nonconstant spread

Transforming nonconstant spread

- Log transform the hinge-spread and the median

```
hss[, "log.Median"] <-
    log10(hss[, "Median"])
hss[, "log.Hinge_Spread"] <-
    log10(hss[, "Hinge Spread"])
```

Transforming nonconstant spread

- Log transform the hinge-spread and the median

```
hss[, "log.Median"] <-
    log10(hss[, "Median"])
hss[, "log.Hinge_Spread"] <-
    log10(hss[, "Hinge Spread"])
```

- Find the least-squares fit of the transformed data

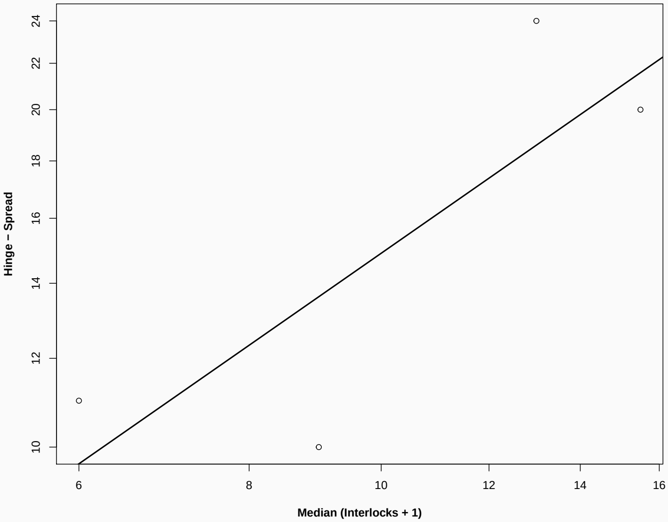
```
fmhs <- lm(log.Hinge_Spread ~ log.Median, data = hs
```

Transforming nonconstant spread

- Plot the log the hinge-spread against the log median

```r
plot(log.Hinge_Spread ~ log.Median, data = hss,
     xlab = list("Median (Interlocks + 1)",
                 font = 2),
     ylab = list("Hinge - Spread", font = 2),
     axes = FALSE, frame = TRUE)
abline(fmhs, lwd = 2)
axis(1, at = log10(seq(6, 16, by = 2)),
    labels = seq(6, 16, by = 2))
axis(2, at = log10(seq(10, 24, by = 2)),
    labels = seq(10, 24, by = 2))
```

Transforming nonconstant spread



Transforming nonconstant spread

Transforming nonconstant spread

- The summary of the fitted line

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3263	0.4242	0.77	0.5221
log.Median	0.8466	0.4152	2.04	0.1783

Transforming nonconstant spread

- The summary of the fitted line

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3263	0.4242	0.77	0.5221
log.Median	0.8466	0.4152	2.04	0.1783

- The fitted equation is

$$\log \text{Hinge Spread} = 0.33 + 0.85 \times \log \text{Median}$$

Transforming nonconstant spread

- The summary of the fitted line

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3263	0.4242	0.77	0.5221
log.Median	0.8466	0.4152	2.04	0.1783

- The fitted equation is

$$\log \text{Hinge Spread} = 0.33 + 0.85 \times \log \text{Median}$$

- The p of the power transformation

$$p = 1 - 0.85 = 0.15 \approx 0$$

Transforming nonconstant spread

- The summary of the fitted line

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3263	0.4242	0.77	0.5221
log.Median	0.8466	0.4152	2.04	0.1783

- The fitted equation is

$$\log \text{Hinge Spread} = 0.33 + 0.85 \times \log \text{Median}$$

- The p of the power transformation

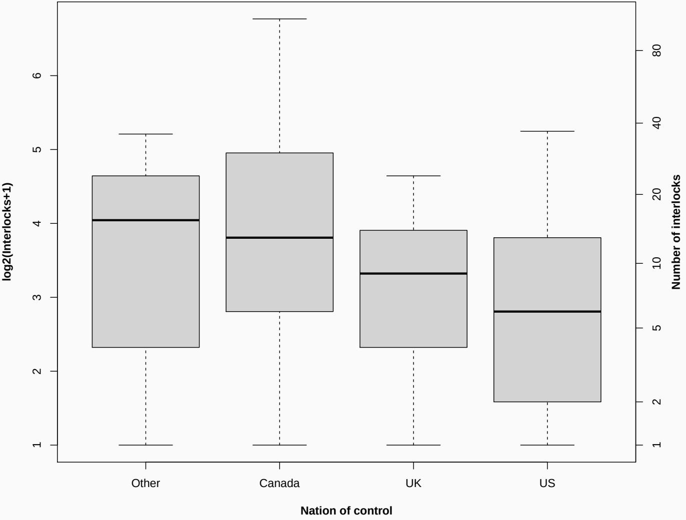
$$p = 1 - 0.85 = 0.15 \approx 0$$

- Henceforth, a log transformation should be used.

Transforming nonconstant spread: Ornstein

```r
par(oma = c(0, 0, 0, 2)) #set margins
boxplot(log2(interlocks + 1) ~ nation,
    data = Ornstein,
    ylab = expression(bold("log2(Interlocks+1)")),
    xlab = expression(bold("Nation of control")))
Original.Scale <- c(0, 1, 2, 5, 10, 20, 40, 80)
axis(side = 4,
    at = log2(Original.Scale + 1),
    labels = Original.Scale)
mtext("Number of interlocks",
      font = 2, side = 4, padj = 4)
```

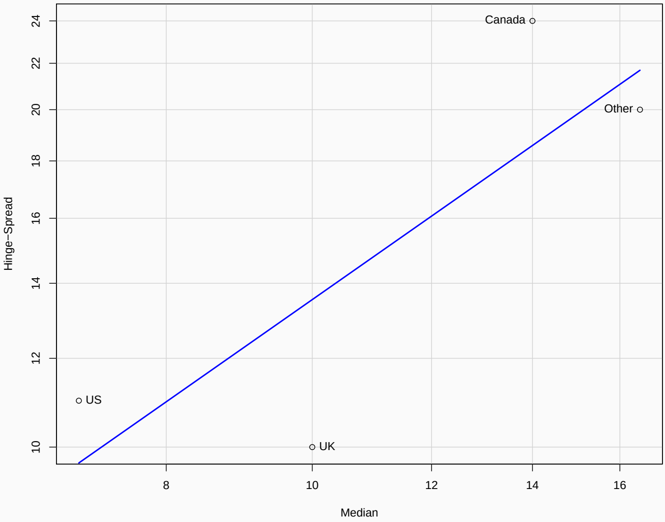
Transforming nonconstant spread: Ornstein



Transforming nonconstant spread: Ornstein

```
car::spreadLevelPlot(interlocks + 1 ~ nation,
                     data = Ornstein, main = "")
```

Transforming nonconstant spread: Ornstein



LowerHinge Median UpperHinge Hinge−Spread

The common origin of unequal spread and skewness

Bound	Skewness	Unequal spread	Transform	Example
Below	Positive	↑Level - ↑Spread	Ladder down	Frequency counts
Up	Negative	↑Level - ↓Spread	Ladder up	Simple exam
Below + Up	??	??	Not helpful	Percentages, Rates

Table of Contents

Transforming proportions

Transforming proportions

- Power transformations are often unhelpful for proportions because these quantities are bounded below by 0 and above by 1.

Transforming proportions

- Power transformations are often unhelpful for proportions because these quantities are bounded below by 0 and above by 1.
- If the data values do not approach the two boundaries, then proportions can be handled much like other sorts of data.

Transforming proportions

- Power transformations are often unhelpful for proportions because these quantities are bounded below by 0 and above by 1.
- If the data values do not approach the two boundaries, then proportions can be handled much like other sorts of data.
- Percentages and many sorts of rates are simply rescaled proportions and are similarly affected.

Common transformations for proportions

Name	Formula

Common transformations for proportions

Name	Formula
probit	$P \Rightarrow \text{probit}(P) = \Phi^{-1}(P) = \text{qnorm}(P)$

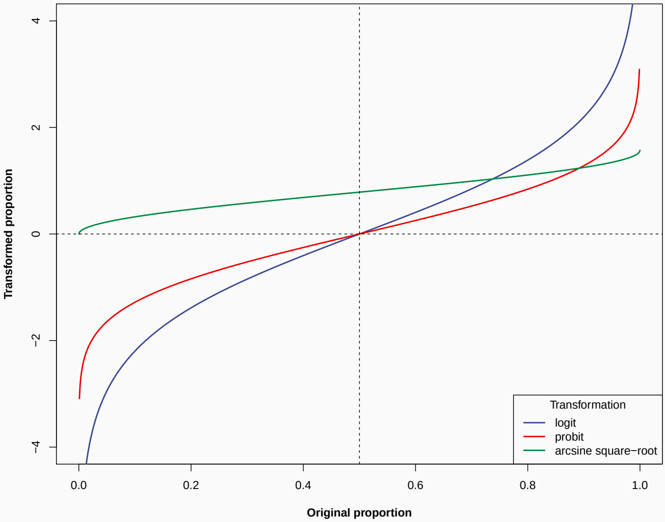
Common transformations for proportions

Name	Formula
probit	$P \Rightarrow \text{probit}(P) = \Phi^{-1}(P) = \text{qnorm}(P)$
logit	$P \Rightarrow \text{logit}(P) = \log_e \dfrac{P}{1-P} \approx \dfrac{\pi}{\sqrt{3}} \times \text{probit}$

Common transformations for proportions

Name	Formula
probit	$P \Rightarrow \text{probit}(P) = \Phi^{-1}(P) = \text{qnorm}(P)$
logit	$P \Rightarrow \text{logit}(P) = \log_e \frac{P}{1-P} \approx \frac{\pi}{\sqrt{3}} \times \text{probit}$
Arcsine-square-root	$P \Rightarrow \sin^{-1}\sqrt{P}$

Common transformations for proportions



The family of folded powers and roots by Tukey (1977)

The family of folded powers and roots by Tukey (1977)

- Tukey (1977) has embedded these common transformations for proportions into the family of folded powers and roots indexed by the power q, which takes on values between 0 and 1:

The family of folded powers and roots by Tukey (1977)

$$P \Rightarrow P^q - (1-P)^q \,, \quad 0 \leq p \leq 1$$

The family of folded powers and roots by Tukey (1977)

$$P \Rightarrow P^q - (1-P)^q \,, \quad 0 \leq p \leq 1$$

$$= \begin{cases} q = 0 & \text{logit transformation} \\ q = 0.14 & \text{probit transformation} \\ q = 0.41 & \text{arcsine-square-root transformation} * \text{a} \\ q = 1 & \text{difference between P and 1/2} \end{cases}$$

Adjust the proportions of 0 and 1

Adjust the proportions of 0 and 1

- The logit and probit transformations cannot be applied to proportions of exactly 0 or 1.

Adjust the proportions of 0 and 1

- The logit and probit transformations cannot be applied to proportions of exactly 0 or 1.
- To avoid this embarrassment,

Adjust the proportions of 0 and 1

- The logit and probit transformations cannot be applied to proportions of exactly 0 or 1.
- To avoid this embarrassment,
- If the original counts are avaliable,

$$P \Rightarrow P' = \frac{F + \frac{1}{2}}{N + 1}$$

F is the frequency count in the focal category, and N is the total count.

Adjust the proportions of 0 and 1

- The logit and probit transformations cannot be applied to proportions of exactly 0 or 1.
- To avoid this embarrassment,
- If the original counts are avaliable,

$$P \Rightarrow P' = \frac{F + \frac{1}{2}}{N + 1}$$

  F is the frequency count in the focal category, and N is the total count.

- If the original counts are not available,

$$P \Rightarrow P' = \text{adjust} + (1 - 2 * \text{adjust}) \times P$$

Adjust the proportions of 0 and 1

Adjust the proportions of 0 and 1

- For example,

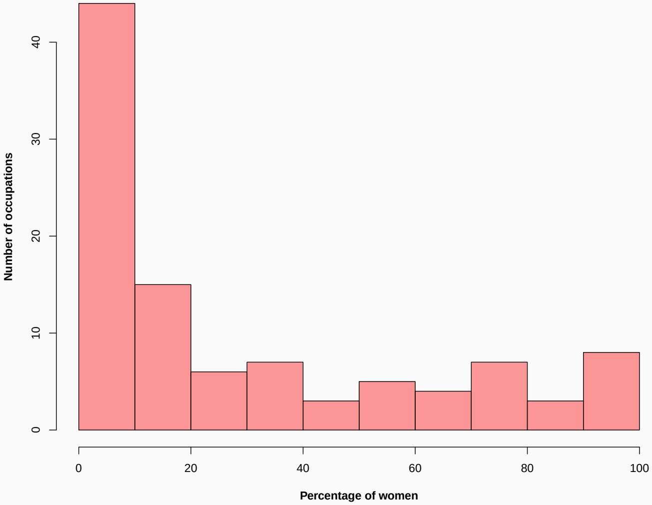
$$\begin{aligned} \text{adjust} &= 0.005 \\ P' &= 0.005 + 0.99 \times P \\ &\in [0.005, 0.995] \end{aligned}$$

The percentage of women among occupations

```
str(Prestige)

'data.frame':^^I102 obs. of  8 variables:
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ income   : int  12351 25879 9271 8865 8403 11030 825
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 7
 $ census   : int  1113 1130 1171 1175 2111 2113 2133 2
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2
 $ incomecr : num  23.1 29.6 21 20.7 20.3 ...
 $ incomelg : num  4.09 4.41 3.97 3.95 3.92 ...
```

The percentage of women among occupations



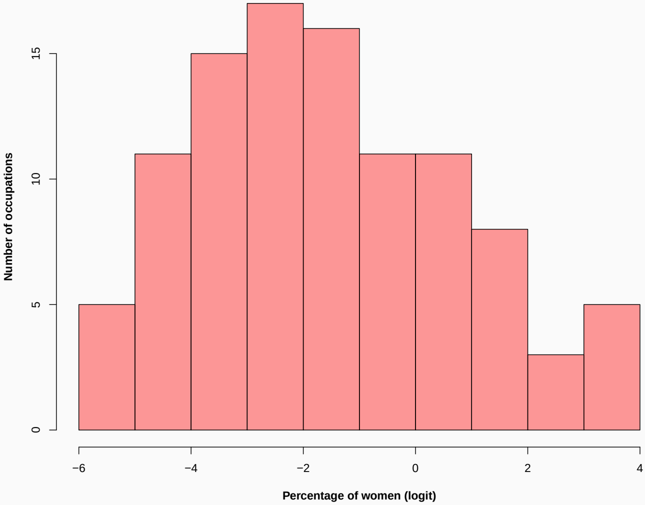
The percentage of women among occupations

```
1 | 2: represents 12
 leaf unit: 1
           n: 102
  32     0* | 00000000000000111111222233334444
  44     0. | 555566777899
  (8)    1* | 01111333
  50     1. | 5557779
  43     2* | 1344
  39     2. | 57
  37     3* | 01334
  32     3. | 99
         4* |
  30     4. | 678
  27     5* | 224
  24     5. | 67
  22     6* | 3
  21     6. | 789
  18     7* | 024
  15     7. | 5667
  11     8* | 233
         8. |
   8     9* | 012
   5     9. | 56667
```

Back to the Prestige data

```
logitwm <- car::logit(
    Prestige[, "women"],
    adjust = 0.005)
```

The percentage of women among occupations



Back to the Prestige data

```
aplpack::stem.leaf(logitwm)


1 | 2: represents 1.2
 leaf unit: 0.1
            n: 102
    5    -5* | 22222
    8    -4. | 555
   16    -4* | 44332111
   21    -3. | 98875
   31    -3* | 4432111000
   39    -2. | 98887655
   48    -2* | 443220000
  (10)   -1. | 9888666555
   44    -1* | 331110
   38    -0. | 987666
   32    -0* | 44110
   27     0* | 00122
   22     0. | 577889
   16     1* | 01111
   11     1. | 556
    8     2* | 23
    6     2. | 5
    5     3* | 00014
```

Table of Contents

Multivariate-normal distribution

Multivariate-normal distribution

- If the vector random variable

  $$\mathbf{x} = (X_1, X_2, \cdots, X_p)'$$

  with population mean vector

  $$\underset{(p \times 1)}{\boldsymbol{\mu}}$$

  and covariance matrix

  $$\underset{(p \times p)}{\Sigma}$$

  is **multivariate-normal distributed**,

Multivariate-normal distribution

Multivariate-normal distribution

- then its probability density function is:

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

Multivariate-normal distribution

- then its probability density function is:

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

- In shorthand, $\mathbf{x} \sim \mathbf{N}_p(\boldsymbol{\mu}, \Sigma)$.

Multivariate-normal distribution

Multivariate-normal distribution

- For a sample of n observations

$$\underset{(n \times p)}{\mathrm{X}}$$

Multivariate-normal distribution

- For a sample of n observations

$$\underset{(n \times p)}{\mathbf{X}}$$

- we have

$$p(\mathbf{X}|\boldsymbol{\mu}, \Sigma) = \left[\frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}}\right]^n \exp\left\{\sum_{i=1}^{n}\left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]\right\}$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$.

Multivariate-normal distribution

Multivariate-normal distribution

- The log-likelihood for the parameters is:

$$\log_e L(\boldsymbol{\mu}, \Sigma|\mathbf{X}) = -\frac{np}{2}\log_e(2\pi) - \frac{n}{2}\log_e \det \Sigma - \frac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]$$

Multivariate-normal distribution

- The log-likelihood for the parameters is:

$$\log_e L(\boldsymbol{\mu}, \Sigma|\mathbf{X}) = -\frac{np}{2}\log_e(2\pi) - \frac{n}{2}\log_e \det \Sigma$$
$$- \frac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]$$

- The maximum-likelihood estimators (MLEs) of the mean and covariance matrix are

$$\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = (\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p)'$$
$$\widehat{\Sigma} = \{\widehat{\sigma}_{jj'}\} = \left\{\frac{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ij'} - \overline{X}_{j'})}{n}\right\}$$

Choose a transformation as parameter testing

Choose a transformation as parameter testing

- If $\mathbf{x}$ is not multivariate-normal, but it can be made so by a power transformation, then we need to find a parameter vector $\boldsymbol{\lambda}$ that makes $\mathbf{x}^{(\lambda)}$ multivariate-normal distributed,

$$\boldsymbol{\lambda} \equiv (\lambda_1, \lambda_2, \cdots, \lambda_p)'$$
$$\mathbf{x}^{(\lambda)} \equiv \left[x_1^{(\lambda_1)}, x_2^{(\lambda_2)}, \cdots, x_p^{(\lambda_p)}\right]'$$

Choose a transformation as parameter testing

Choose a transformation as parameter testing

- Then the probability density becomes

$$p(\mathrm{x}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\lambda}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\mathrm{x}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\Sigma^{-1}(\mathrm{x}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})\right] \prod_{j=1}^{p} X_j^{\lambda_j - 1} \quad (3)$$

where $\boldsymbol{\mu} = E\left[\mathbf{x}^{(\lambda)}\right]$ and $\Sigma = V\left[\mathbf{x}^{(\lambda)}\right]$ are the mean vector and covariance matrix of the transformed variables, and $\prod_{j=1}^{p} X_j^{\lambda_j - 1}$ is the Jacobian of the transformation from $\mathbf{x}^{(\lambda)}$ to $\mathbf{x}$.

Choose a transformation as parameter testing

Choose a transformation as parameter testing

- The log-likelihood for the model becomes

$$\begin{aligned}\log_e L(\boldsymbol{\lambda}, \boldsymbol{\mu}, \Sigma | X) = &-\frac{np}{2}\log_e(2\pi) - \frac{n}{2}\log_e \det \Sigma \\ &- \frac{1}{2}\sum_{i=1}^{n}\left[(\mathrm{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\Sigma^{-1}(\mathrm{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})\right] + \sum_{j=1}^{p}(\lambda_j - 1)\sum_{i=1}^{n}\log_e X_{ij}\end{aligned}$$

Choose a transformation as parameter testing

- The log-likelihood for the model becomes

$$\begin{aligned}\log_e L(\boldsymbol{\lambda}, \boldsymbol{\mu}, \Sigma | X) = &-\frac{np}{2}\log_e(2\pi) - \frac{n}{2}\log_e \det \Sigma \\ &- \frac{1}{2}\sum_{i=1}^{n}\left[(\mathrm{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\Sigma^{-1}(\mathrm{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})\right] + \sum_{j=1}^{p}(\lambda_j - 1)\sum_{i=1}^{n}\log_e X_{ij}\end{aligned}$$

- There is no closed-form solution for the MLEs of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and Σ.

Choose a transformation as parameter testing

- The log-likelihood for the model becomes

$$\log_e L(\boldsymbol{\lambda}, \boldsymbol{\mu}, \Sigma | \mathrm{X}) = -\frac{np}{2}\log_e(2\pi) - \frac{n}{2}\log_e \det \Sigma$$
$$- \frac{1}{2}\sum_{i=1}^{n}\left[(\mathrm{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\Sigma^{-1}(\mathrm{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})\right] + \sum_{j=1}^{p}(\lambda_j - 1)\sum_{i=1}^{n}\log_e X_{ij}$$

- There is no closed-form solution for the MLEs of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and Σ.
- Standard errors for the estimated transformations are available in the usual manner from the inverse of the information matrix, and both Wald and likelihood-ratio tests can be formulated for the transformation parameters.

Choose a transformation as parameter testing

Choose a transformation as parameter testing

- Given $\widehat{\boldsymbol{\lambda}}$, the MLEs of $\boldsymbol{\mu}$ and Σ are just the sample mean vector and covariance matrix of $\mathbf{x}^{(\lambda)}$.

Choose a transformation as parameter testing

- Given $\widehat{\boldsymbol{\lambda}}$, the MLEs of $\boldsymbol{\mu}$ and Σ are just the sample mean vector and covariance matrix of $\mathbf{x}^{(\lambda)}$.
- Let us define the modified Box-Cox family of transformations as follows:

$$X^{[\lambda]} = \begin{cases} \widetilde{X}^{1-\lambda} \frac{X^{\lambda}-1}{\lambda} & \text{for } \lambda \neq 0 \\ \widetilde{X} \log eX & \text{for } \lambda = 0 \end{cases}$$

where

$$\widetilde{X} \equiv \left(\prod_{i=1}^{n} X_i \right)^{1/n}$$

is the geometric mean of X.

Choose a transformation as parameter testing

- Given $\widehat{\boldsymbol{\lambda}}$, the MLEs of $\boldsymbol{\mu}$ and Σ are just the sample mean vector and covariance matrix of $\mathbf{x}^{(\lambda)}$.
- Let us define the modified Box-Cox family of transformations as follows:

$$X^{[\lambda]} = \begin{cases} \widetilde{X}^{1-\lambda} \frac{X^{\lambda}-1}{\lambda} & \text{for } \lambda \neq 0 \\ \widetilde{X} \log eX & \text{for } \lambda = 0 \end{cases}$$

  where

$$\widetilde{X} \equiv \left(\prod_{i=1}^{n} X_i \right)^{1/n}$$

  is the geometric mean of X.
- Multiplication by $\widetilde{X}^{1-\lambda}$ is a kind of standardization, equating the scales of different power transformations of X.

Choose a transformation as parameter testing

Choose a transformation as parameter testing

- Let $\mathbf{V}^{[\lambda]}$ represent the sample covariance matrix of

$$\mathbf{x}^{[\lambda]} \equiv \left[x_1^{[\lambda_1]}, x_2^{[\lambda_2]}, \cdots, x_p^{[\lambda_p]}\right]'$$

Choose a transformation as parameter testing

- Let $\mathbf{V}^{[\lambda]}$ represent the sample covariance matrix of

$$\mathbf{x}^{[\lambda]} \equiv \left[x_1^{[\lambda_1]}, x_2^{[\lambda_2]}, \cdots, x_p^{[\lambda_p]}\right]'$$

- Velilla (1993) shows that the MLEs of $\boldsymbol{\lambda}$, in Equation 3 are the values that minimize the determinant of $\mathbf{V}^{[\lambda]}$.

Transformation as parameter testing

```
library(car)
pt1 <- car::powerTransform(
    infantMortality ~ 1,
    data = UN,
    family = "bcPower")
```

Transformation as parameter testing

```
summary(pt1)

bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd
Y1    0.0945           0      -0.0435
   Wald Upr Bnd
Y1       0.2326

Likelihood ratio test that transformation parameter is equal to
 (log transformation)
                           LRT df    pval
LR test, lambda = (0) 1.809774  1 0.17854

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 146.7235  1 < 2.22e-16
```

Transformation as parameter testing

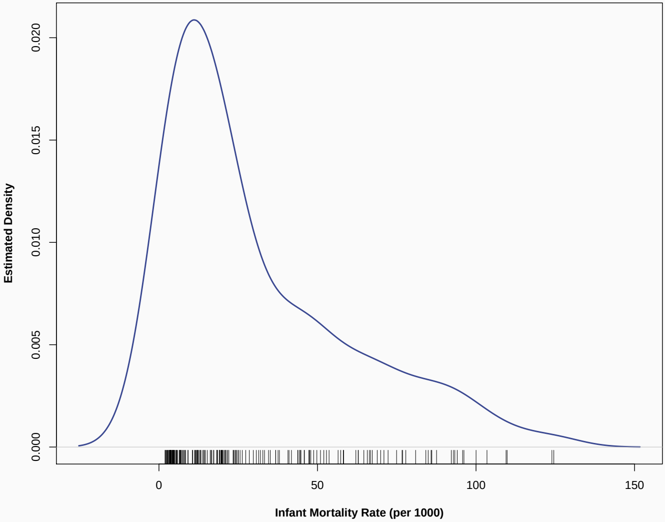
```
car::testTransform(pt1, lambda = 0.5)

                                LRT df         pval
LR test, lambda = (0.5) 31.85218  1 1.6636e-08

car::testTransform(pt1, lambda = 0.1)

                                   LRT df    pval
LR test, lambda = (0.1) 0.006021638  1 0.93815
```

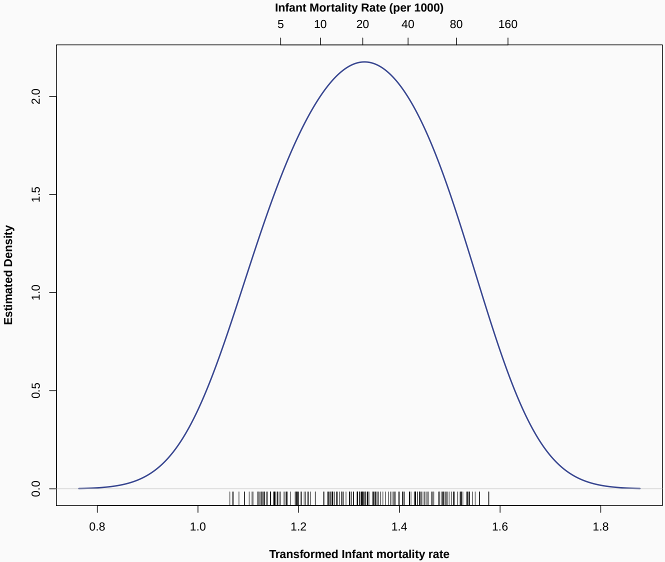
Transformation as parameter testing



Transformation as parameter testing

```
UN <- transform (UN,
   infantMortality.tran =
   infantMortality ^ pt1[["lambda"]])
```

Transformation as parameter testing



Transformation as parameter testing

```
pt2 <- car::powerTransform(
    cbind(income, prestige) ~ 1,
    data = Prestige,
    family = "bcPower")
```

Transformation as parameter testing

```
summary(pt2)


bcPower Transformations to Multinormality
         Est Power Rounded Pwr Wald Lwr Bnd
income      0.1982        0.33       0.0235
prestige    0.4596        0.50       0.0307
         Wald Upr Bnd
income         0.3730
prestige       0.8885

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                             LRT df     pval
LR test, lambda = (0 0) 8.073059  2 0.017659

Likelihood ratio test that no transformations are needed
                             LRT df       pval
LR test, lambda = (1 1) 66.30043  2 3.9968e-15
```

Transformation as parameter testing

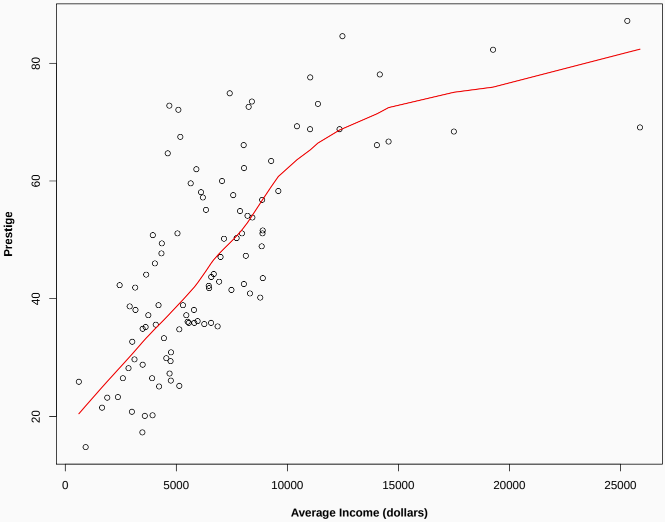
```
testTransform(pt2, lambda = c(1/3, 0.5))

                                  LRT df    pval
LR test, lambda = (0.33 0.5) 2.242532  2 0.32587

testTransform(pt2, lambda = c(0, 1))

                             LRT df      pval
LR test, lambda = (0 1) 13.48008  2 0.0011826
```

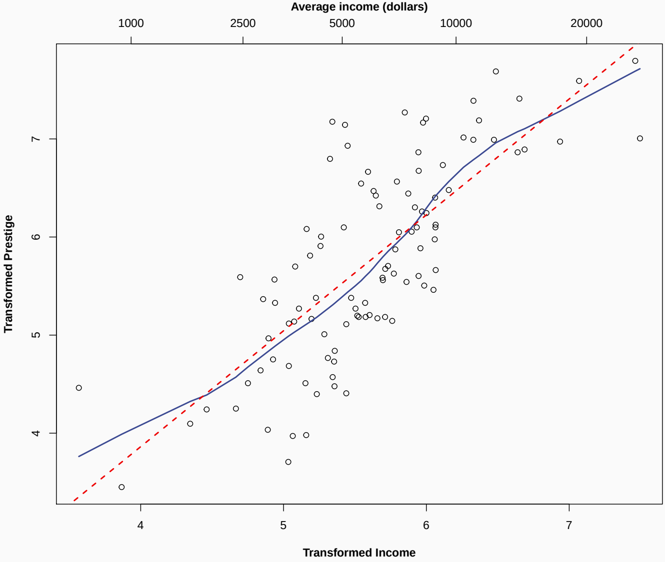
Transformation as parameter testing



Transformation as parameter testing

```r
pnc <- pt2[["lambda"]][["income"]]
ppr <- pt2[["lambda"]][["prestige"]]
Prestige <- transform (Prestige,
    income.tran = income ^ pnc,
    prestige.tran = prestige ^ ppr)
lws <- lowess(Prestige[, "income.tran"],
              Prestige[, "prestige.tran"], f = 0.6)
lm <- lm(prestige.tran ~ income.tran,
    data = Prestige)
```

Transformation as parameter testing



Questions?