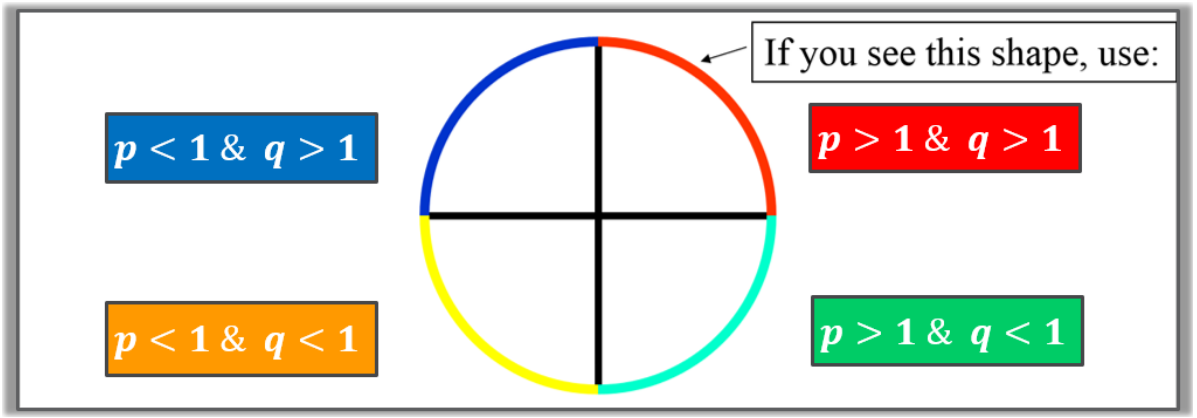
Non-Linear Transformations: Data Pre-processing

- Ladder of Roots of Power
- Log transformation
- Box-Cox transformation
- Rank transformation
- Change the shape of distribution
- Transforming for linearity, constant spread, and normality
- Logit & probit for skewed proportions



Prof. Nagiza F. Samatova
samatova@csc.ncsu.edu
Department of Computer Science
North Carolina State University

Motivating Example #1:

- Linear least squares regression makes strong assumptions about the data:
  - Linear relation between predictors and response
  - Equal variance (i.e., constant spread)
  - Normal distribution (of the response variable)
- Transforming the data can help satisfy these assumptions.
- It can also assist in examining the data.

Motivational Example #2

- A common task is to model the joint distribution of many random variables
  - E.g., for imputing missing values
- BUT modeling moderate-sized joint distribution is difficult
  - Exception: multivariate normal model
  - Idea: First transform each variable to have marginal normal distribution, model, then transform back
  - Extremely useful approach for many modeling tasks
  - Mathematical details can be complex, but implementation is straightforward

Why to Transform Data?

- Better examine a distribution
- Meet modeling assumptions
- Allow simpler modeling options

- Many statistical models are based on the mean
  - thus require that the mean is an appropriate measure of central tendency (*i.e.*, the distribution is approximately normal)
- Linear least squares regression assumes that the relationship between two variables is linear:
  - Often we can "straighten" a nonlinear relationship by transforming one or both of the variables
  - Often transformations will 'fix' problem distributions so that we can use least-squares regression
  - When transformations fail to remedy these problems, another option is to use nonparametric regression, which makes fewer assumptions about the data
- Disadvantage of transformations:
  - interpretation may become more difficult

Transformations

- **Quantitative Variables**
  - **Powers, Inverse Powers, and Roots based transformations**
  - **Logarithmic transformations**
  - **Box-Cox transformation**
  - **Rank transformation**

- **Proportions, Percentages, and Rates**
  - **logit**
  - **probit**

Transformations

QUANTITATIVE VARIABLES

Quantitative Variable Transformations: Powers and Roots

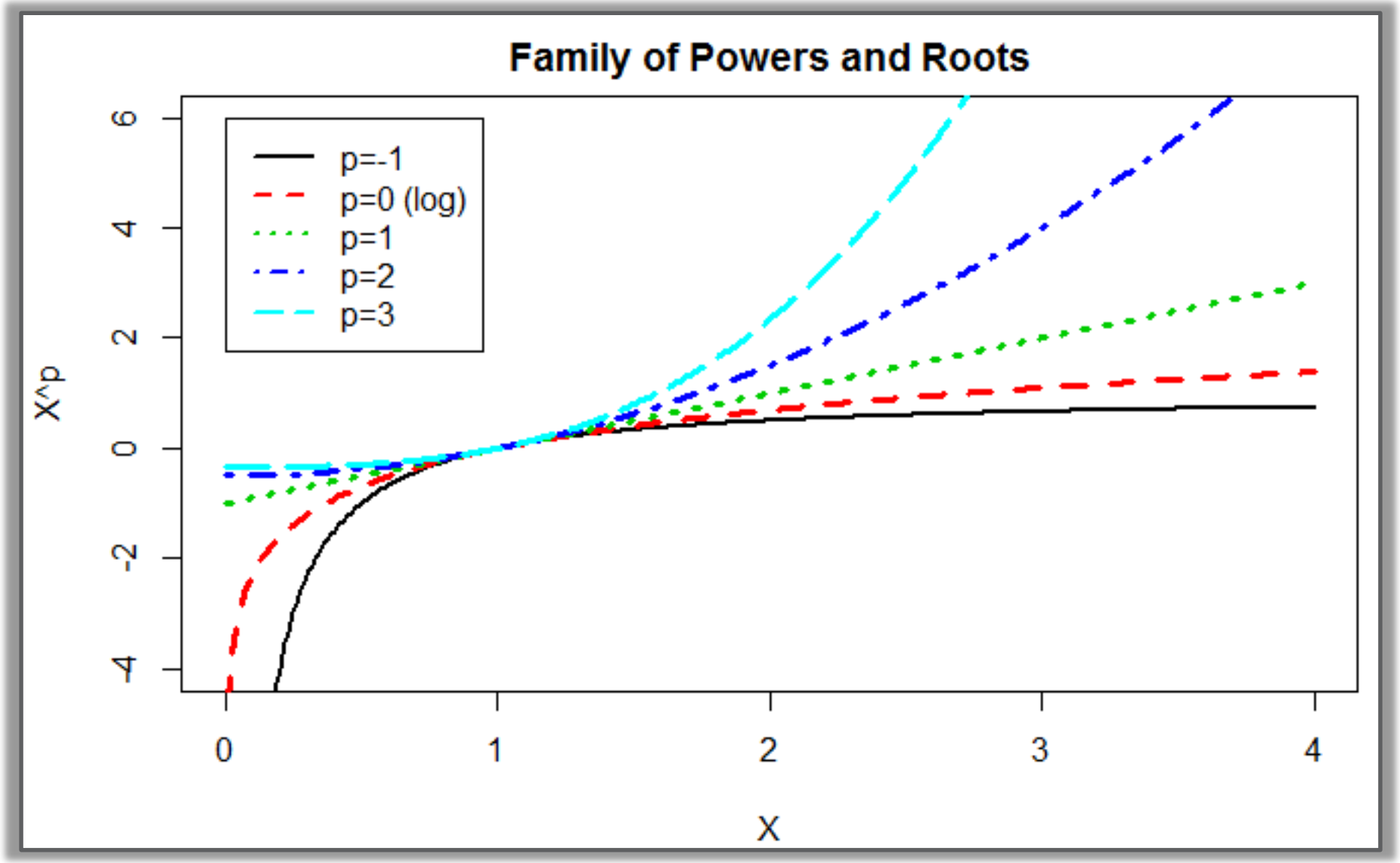
$X > 0$

$X \rightarrow X^p$

$p = 0$ should not be used because it changes all values to 1 (i.e., it makes the variable a constant)

$p > 1$	$p \leq -1$	$-1 < p < 1$
power	inverse power	root
	$X^{-1} = \frac{1}{X}$ $X^{-2} = \frac{1}{X^2}$ $X^{-3} = \frac{1}{X^3}$	$X^{1/2} = \sqrt{X}$ $X^{-1/2} = 1/\sqrt{X}$

Family of Powers and Roots



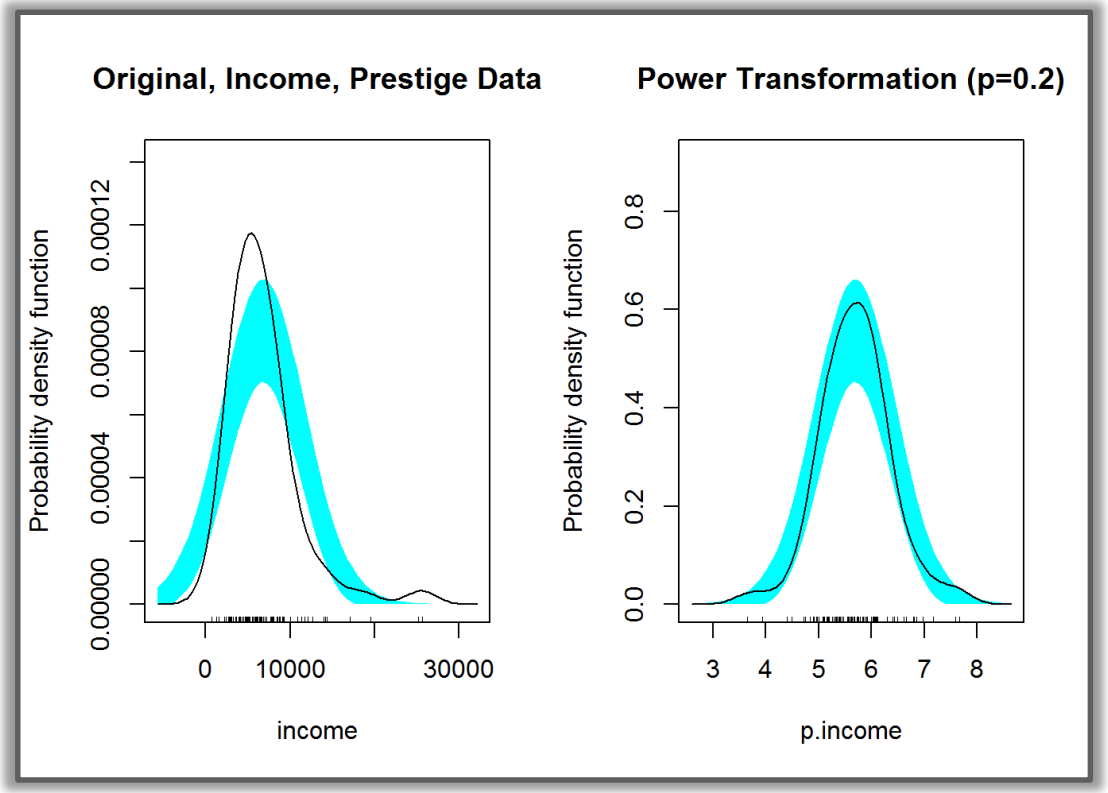
Transforming Skewness

- Problems with skewed distribution
  - Data difficult to examine because most observations are in a small part of the data range.
    - Outlying values in the direction opposite to the skew may be invisible.
  - Least squares regression traces the conditional mean of Y given the X's.
    - The mean is not a good summary of the center of a skewed distribution.

Right skew (positive skew) → compress large values → descend the ladder of powers → p < 1
Left skew (negative skew) → compress small values → ascend the ladder of powers → p > 1

The Ladder of Powers and Roots

- Descending the ladder ($p < 1$):
  - compresses large values of X and
  - spreads out small values of X
- Ascending the ladder ($p > 1$):
  - compresses small values of X and
  - spreads out large values of X
- As p moves away from 1 in either direction:
  - the transformation becomes more powerful



Right skew (positive skew) → compress large values → descend the ladder of powers → p < 1
Left skew (negative skew) → compress small values → ascend the ladder of powers → p > 1

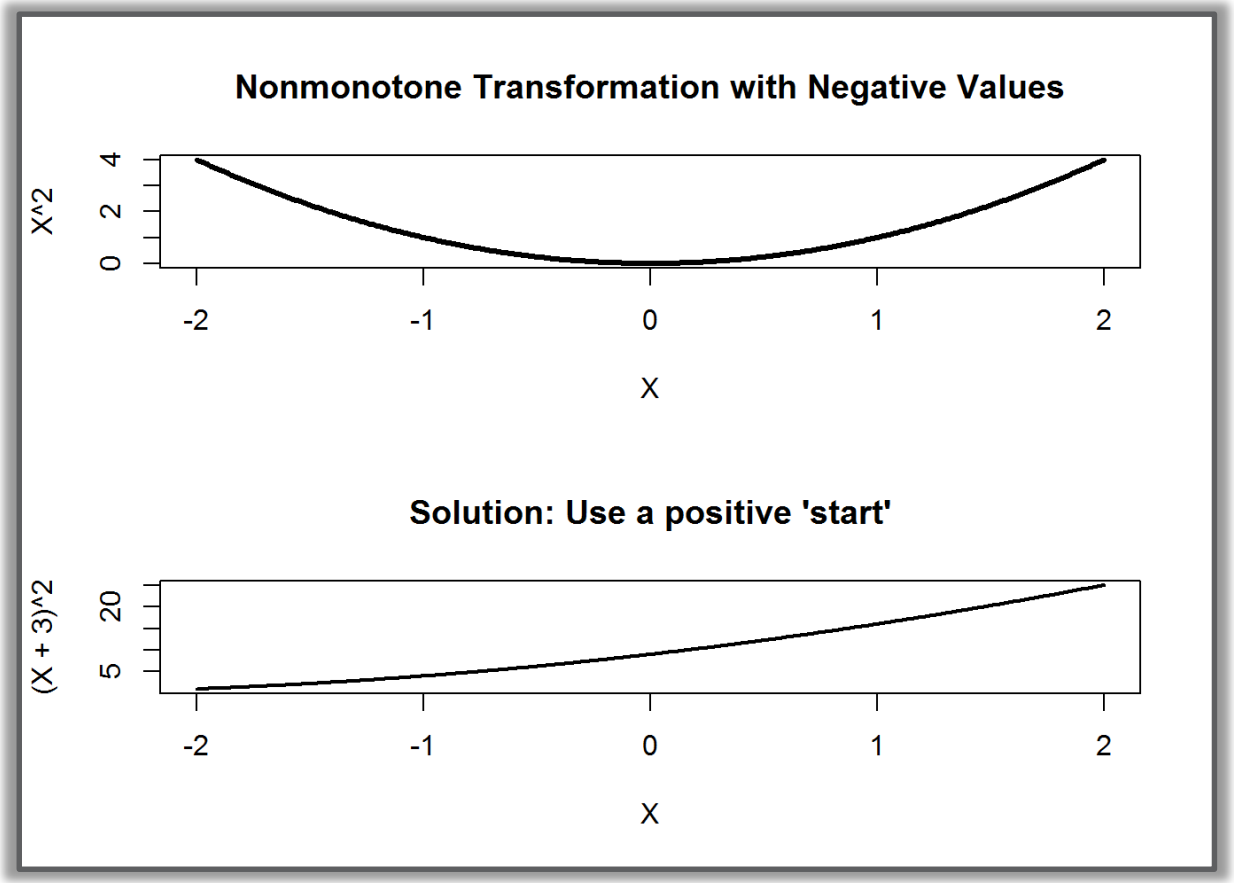
The Negative Values: Use a Positive Start

- All the X values must be POSITIVE
  - If not, this can be solved by adding a start value
  - Otherwise, power transformations will not be monotone:
    - thus changing the order of the data

X	X^2
-2	4
0	0
1	1
2	4

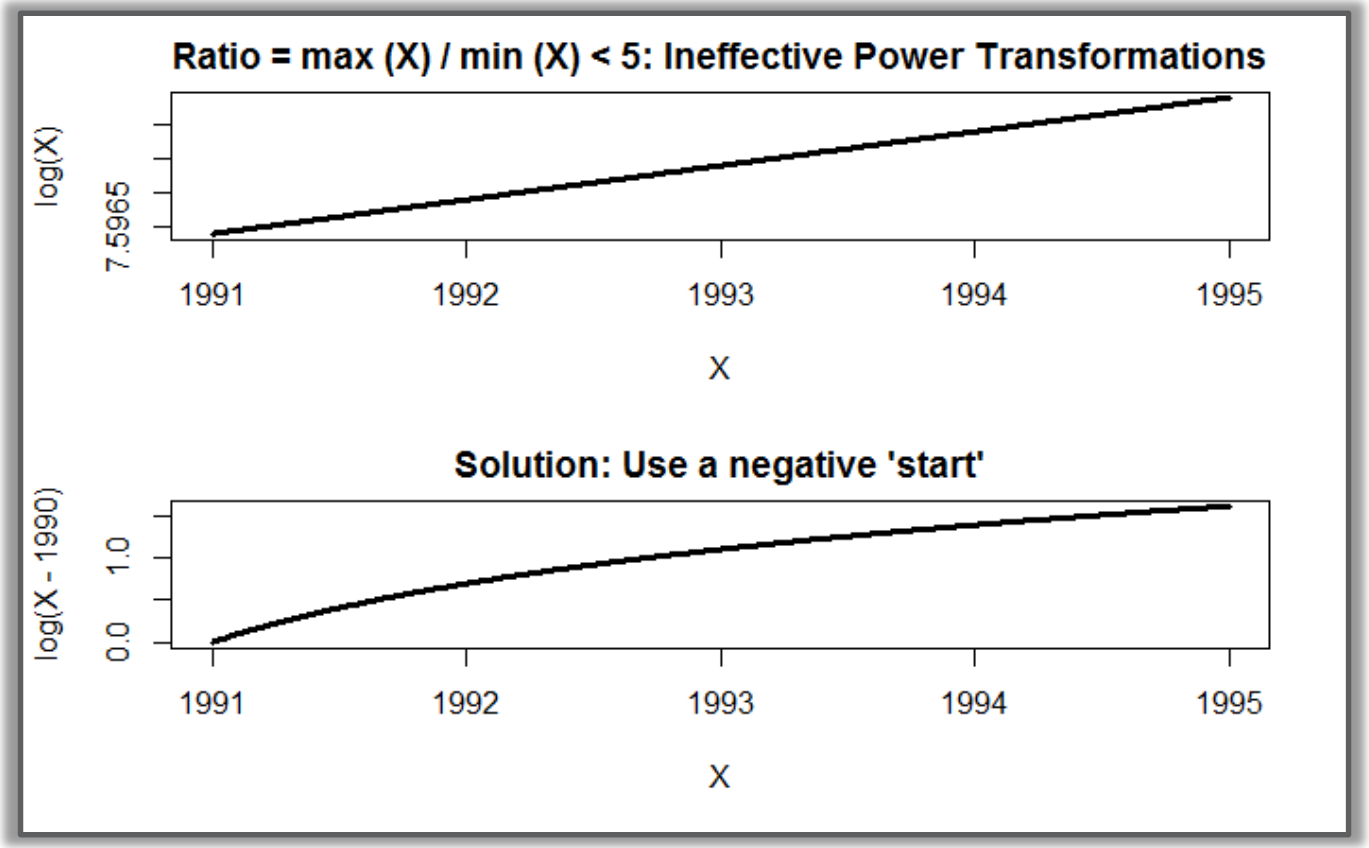
X	$(X+3)^2$
-2	1
0	9
1	16
2	25

add a start value to make X values positive



Effectiveness of Power Transformations

- Power transformations are ONLY effective if the ratio of the largest data value to the smallest data value is large (> 5)
  - If the ratio is very close to 1, the power transformation is almost linear and ineffective



X	$\log_{10} X$	X	$\log_{10}(X - 1990)$
1991	3.2990	1991	0
1992	3.2992	1992	.301
1993	3.2995	1993	.477
1994	3.2997	1994	.602

Quantitative Variable: Log Transformations

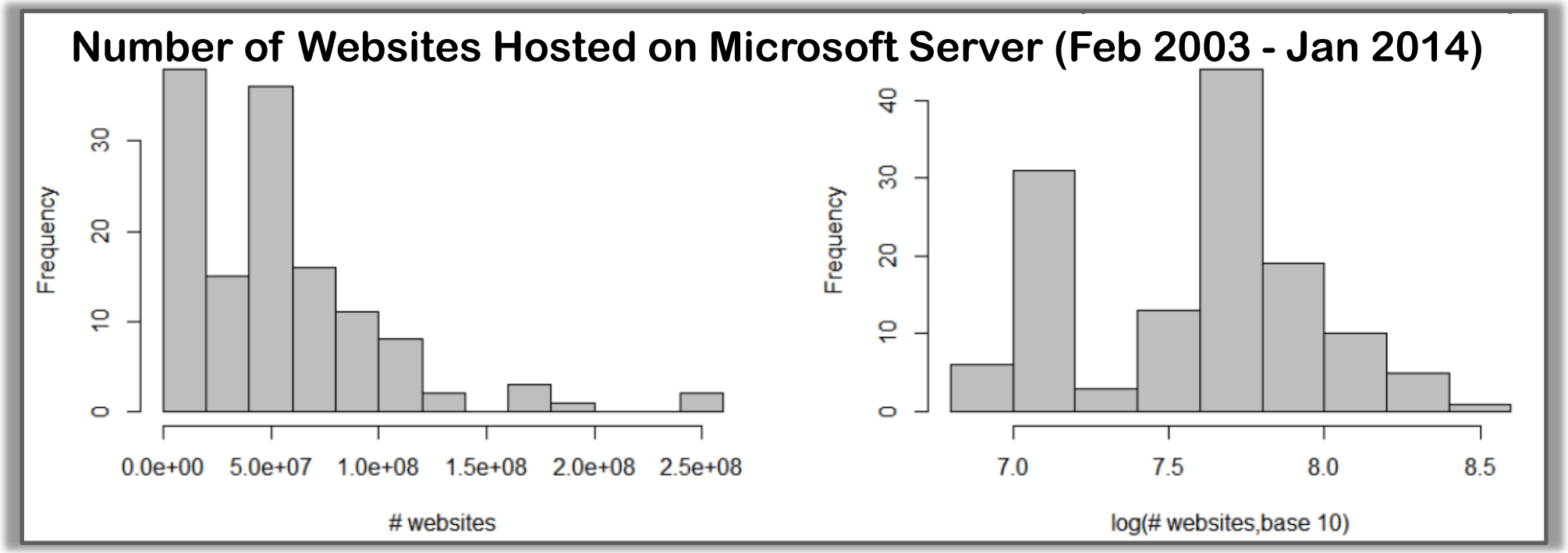
- $\mathbf{X_{new}} = \log_{10} X$
  - preferred in practice because it is easier to interpret
  - increasing X_{new} by 1 is the same as multiplying X by 10
  - in terms of results, it matters little which base is used
    - changing base is equivalent to multiplying X by a constant
- The base of the natural logarithm (e=2.718)

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

- Log transformations are undefined for $X \leq 0$

Log Transformations

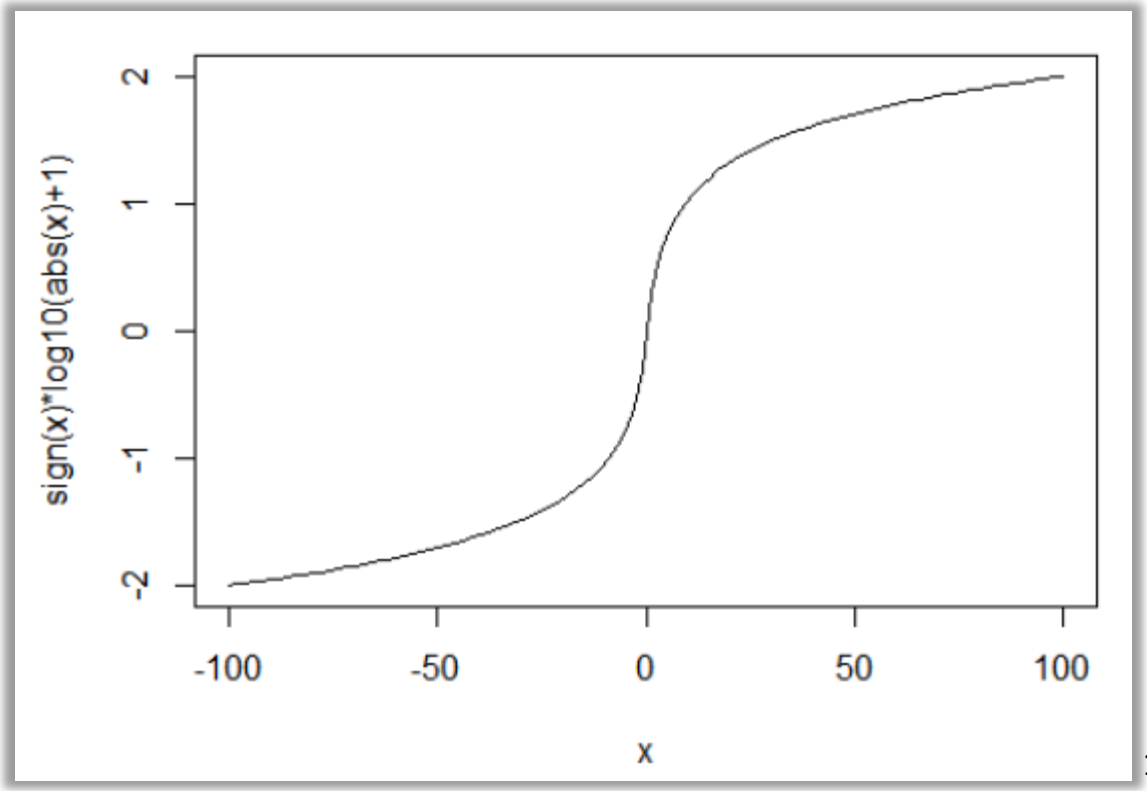
- Data values varying by several orders of magnitude may lead to unstable model fits or make patterns difficult to visualize
  - E.g., income, sales, account sizes, etc.
  - Log transformation can reduce the large value effects, and make trends more apparent



Source: http://www.internetlivestats.com/total-number-of-websites/

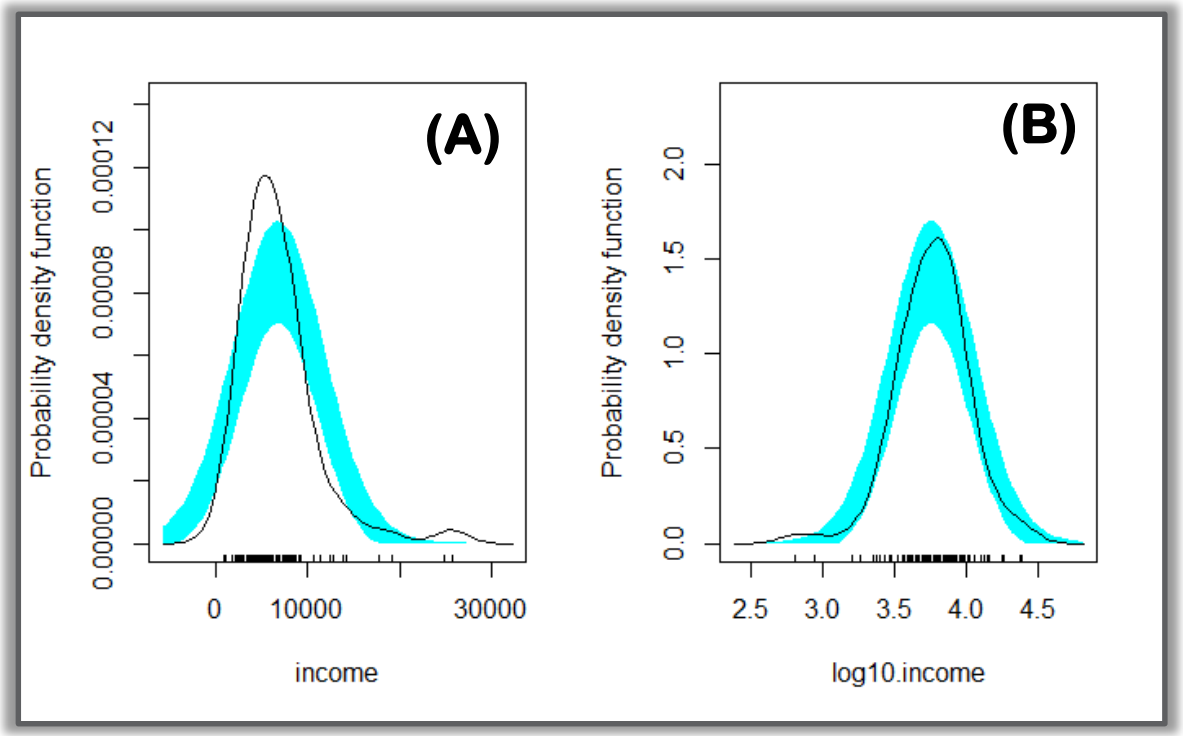
Log Transformations with Non-Positive Data

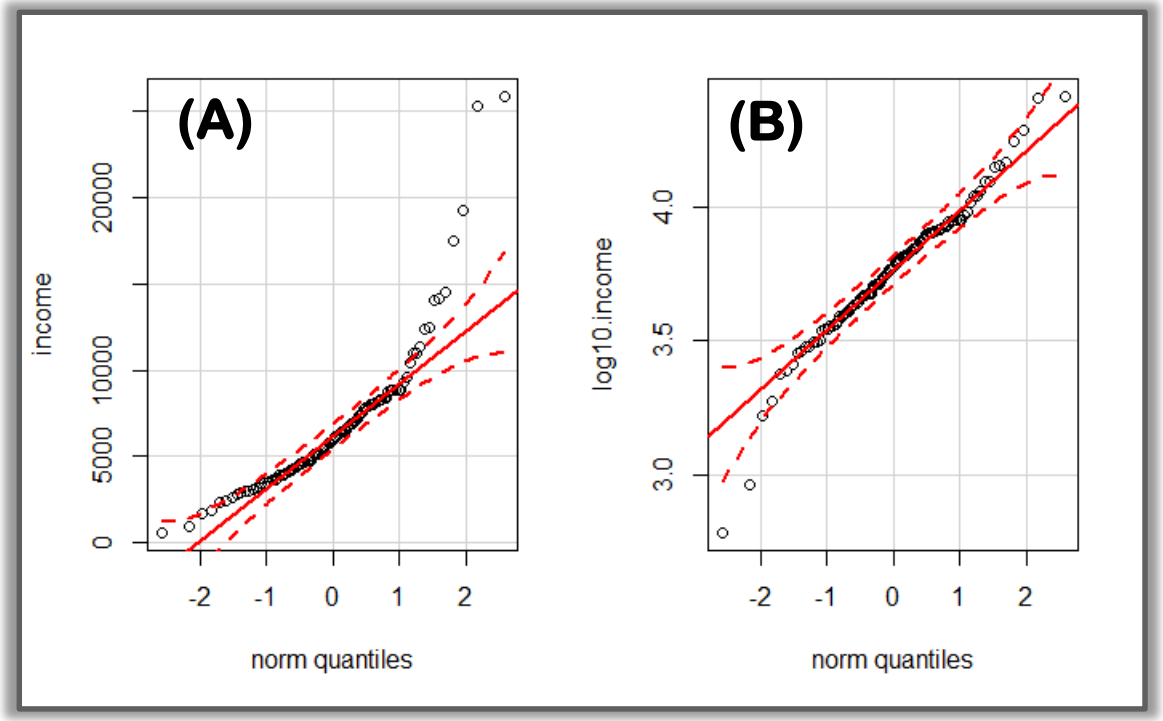
- **Log transformation only applies to positive data**
  - **If you have few non-positive values, use $log\ (x - c)$, where $c < \min_i x_i$**
  - **If you have all negative values, use $log(-x)$**
  - **If have many positive and negative values, use signed log transform:**
  - **$sign(x) \times log(|x| + 1)$, where 0 is preserved**



Transforming Positively Skewed Distributions

- The example below shows how a log10 transformation can fix a positive skew
  - (A): The density estimate for average income for occupations from the Canadian Prestige data and a Q–Q (quantile-quantile) plot
  - (B): the density estimate for the transformed income and Q–Q (quantile-quantile) plot:



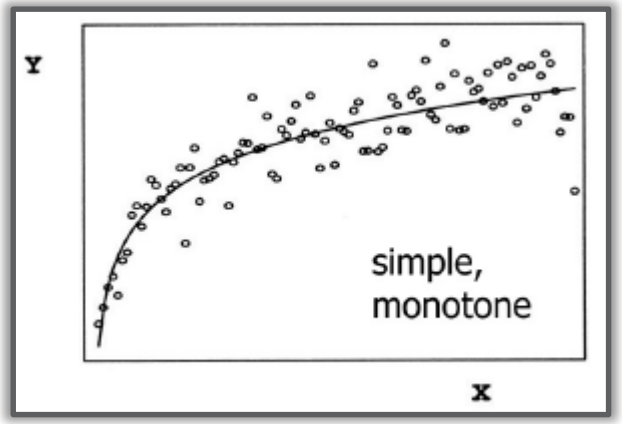


Take-aways: Family of Powers and Roots

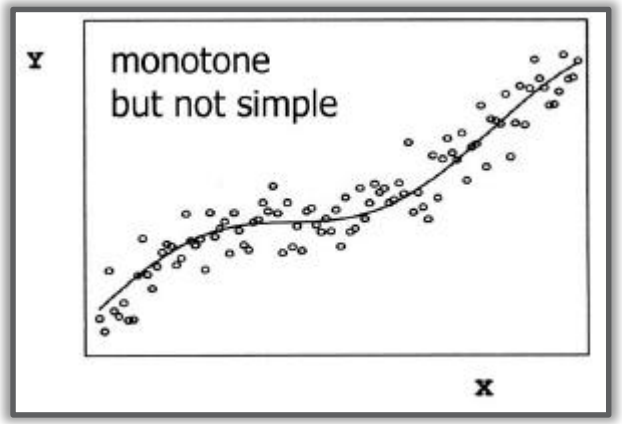
- The family of powers and roots (X^p or $(X^p - 1)/p$):
- Ascending the ladder of powers (p > 1) spreads out large values and compresses small values.
- Descending the ladder of powers (p < 1) does the opposite.

Transforming Non-Linearity between X and Y

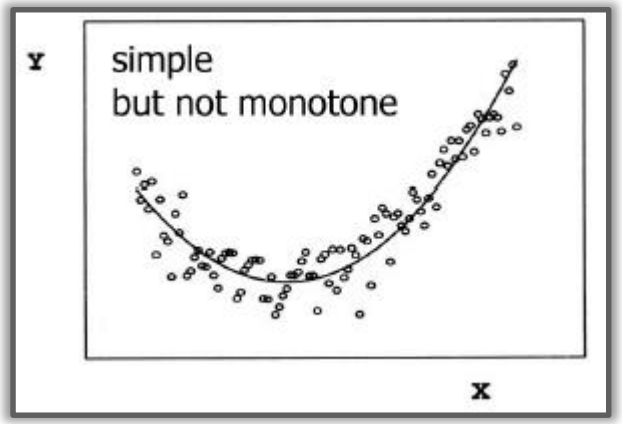
- An important use of transformations is to "straighten" the relationship between two variables
  - This is possible only when the nonlinear relationship is simple and monotone
  - Simple implies that the curvature does not change
    - there is one curve
  - Monotone implies that the curve is always positive or always negative



simple and monotone



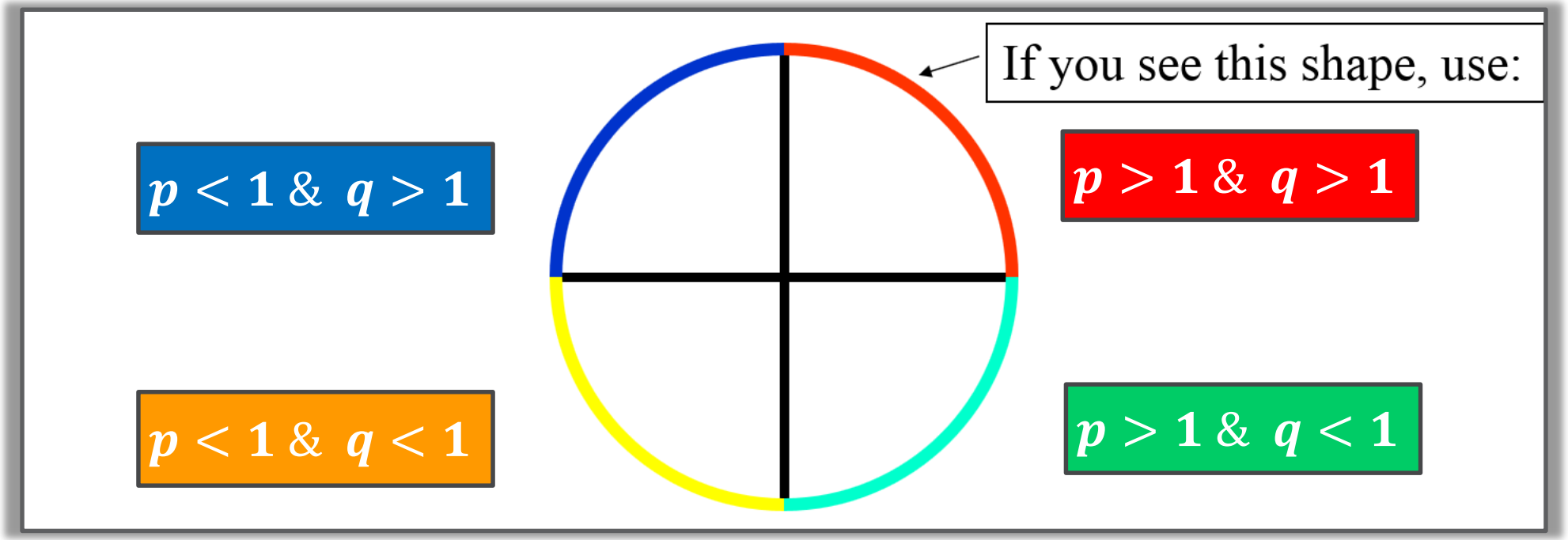
monotone but not simple



simple but not monotone

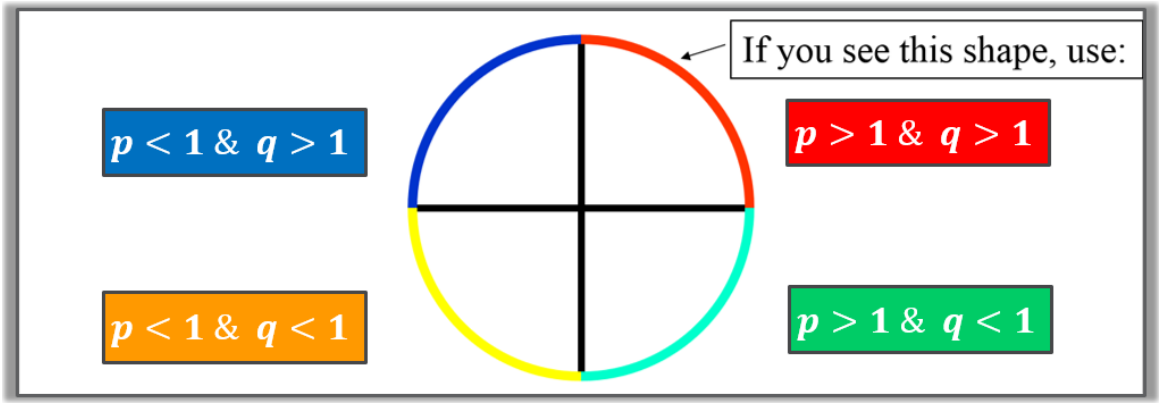
The Bulging Rule: Correct Non-Linearity between X and Y

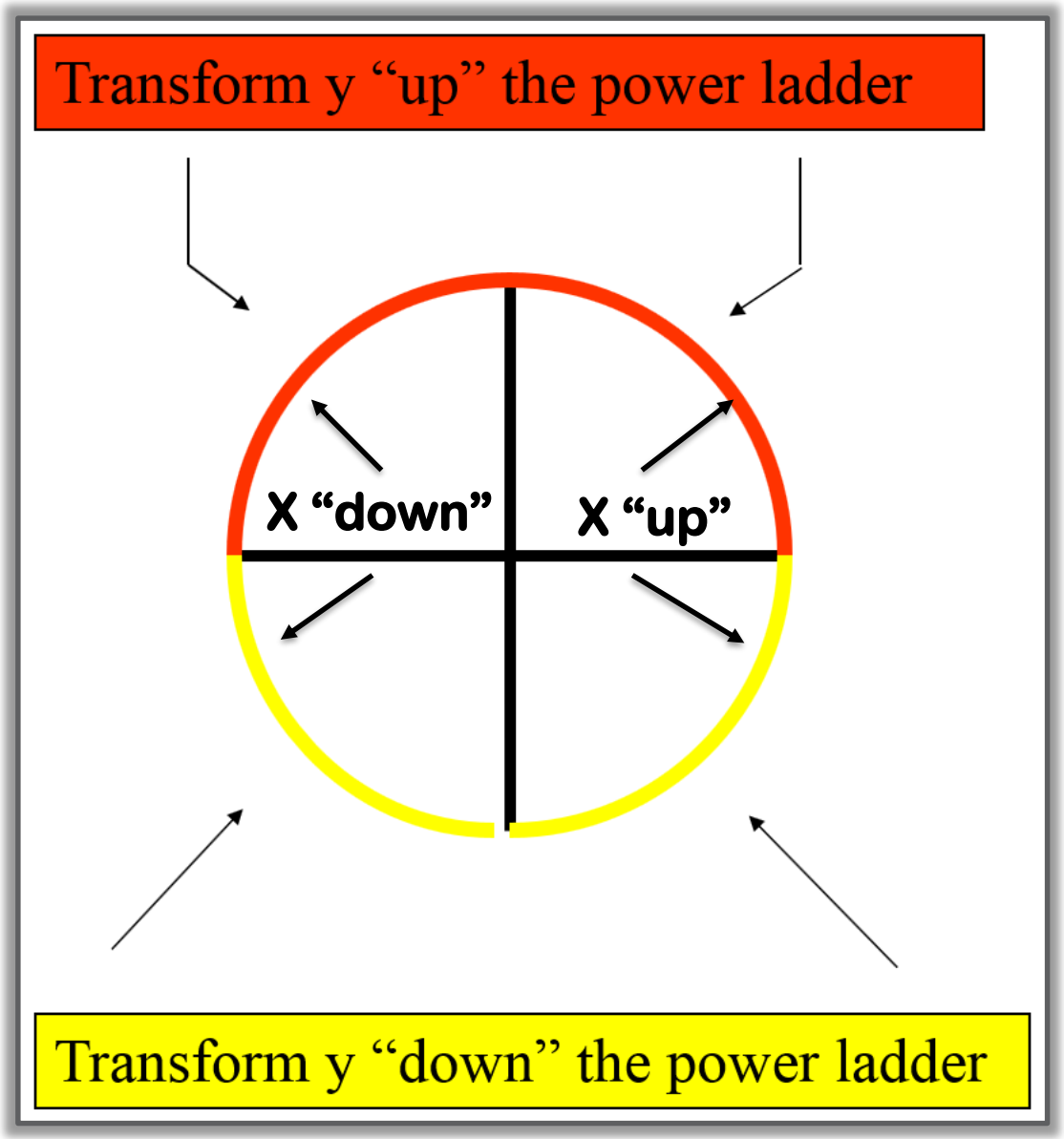
- Tukey and Mosteller's Rule, 1977
- Add proper constants to the original X and Y so that they are both positive
  - If it is multiple regression, transform X only
- Transform X to X^p and/or Y to Y^q
  - p and q are determined by the shape in each quadrant of the plot



The Bulging Rule

- 1st quadrant: Data shape
  - go up-ladder for X, Y or both
- 2nd quadrant: Data shape
  - go down-ladder for X, and/or
  - go up-ladder for Y
- 3rd quadrant: Data shape
  - go down-ladder for X, Y or both
- 4th quadrant: Data shape
  - go up-ladder for X, and/or
  - go down-ladder for Y

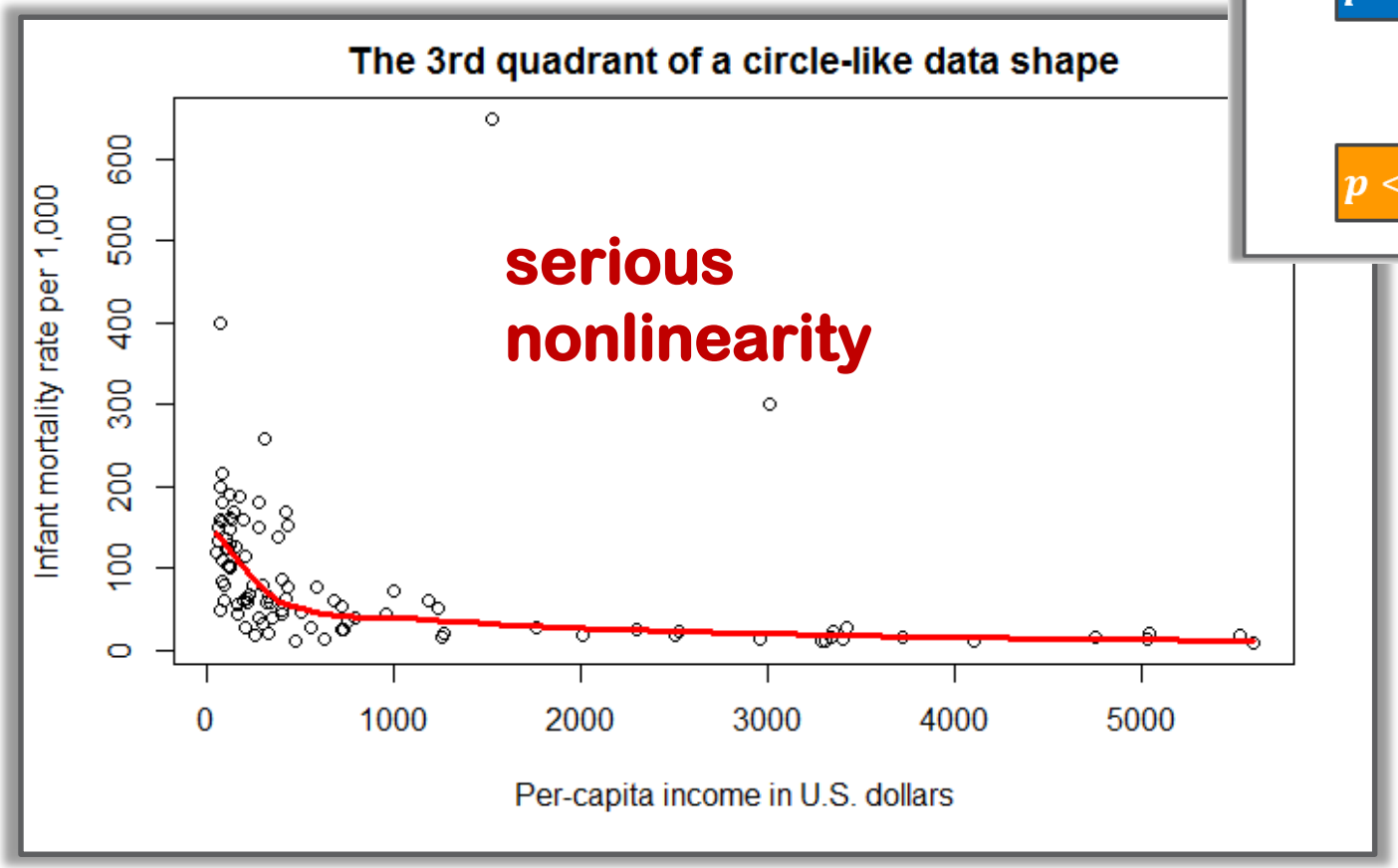


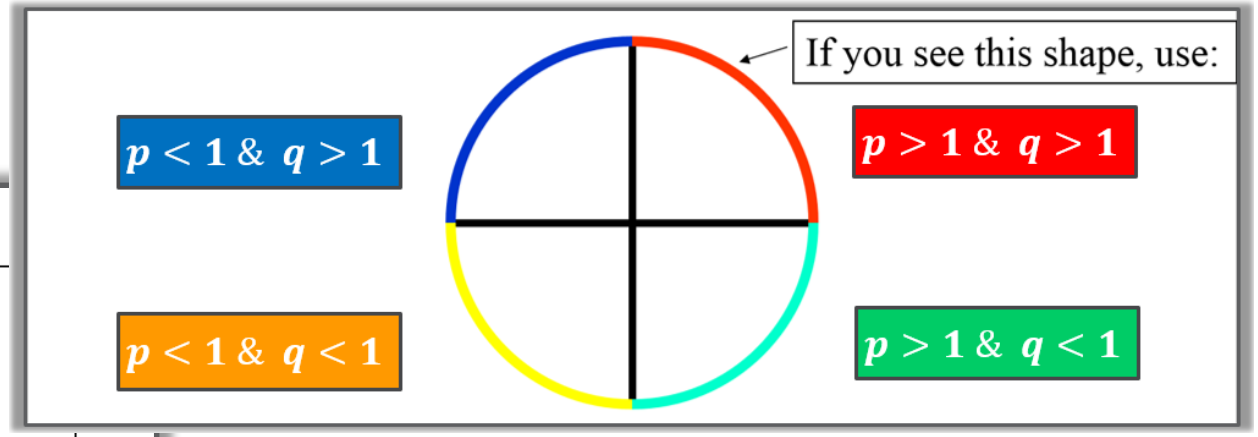


Up the Ladder: Power Greater than 1
Down the Ladder: Power Less Than 1

Example: Non-Linearity: Income & Infant Mortality

Both Y and X need to be transformed down the ladder of powers

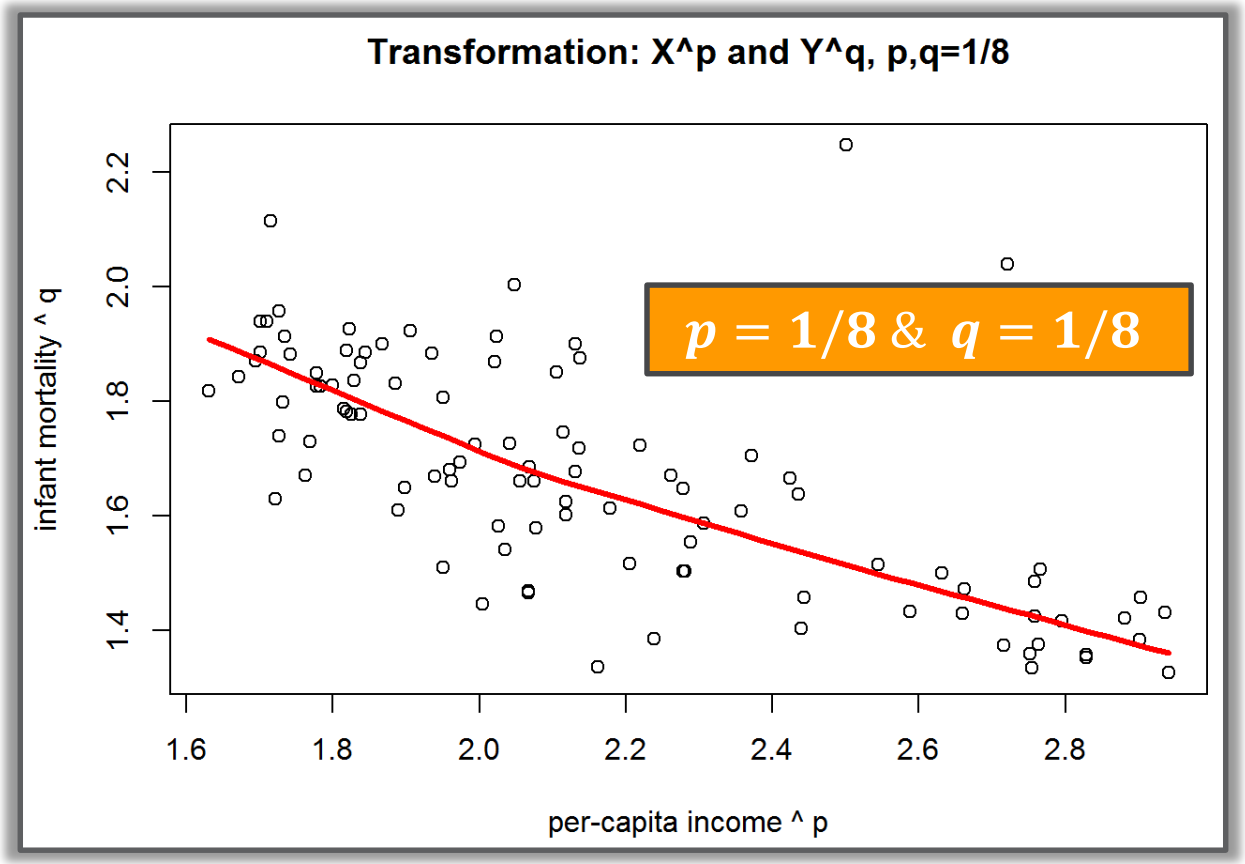


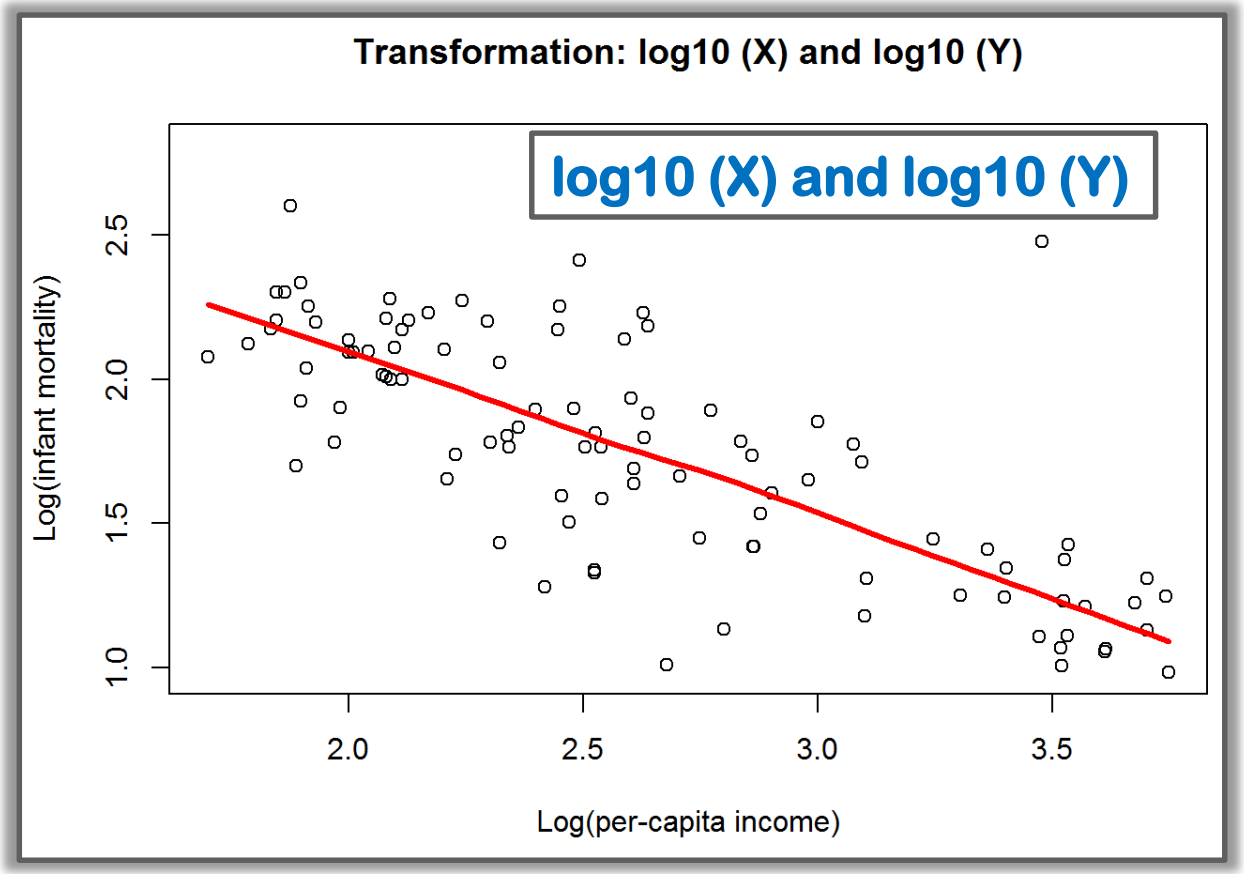


3rd quadrant shape: transform as X^p and Y^q

$$p < 1 \;\&\; q < 1$$

Transforming Non-Linearity: Income & Infant Mortality





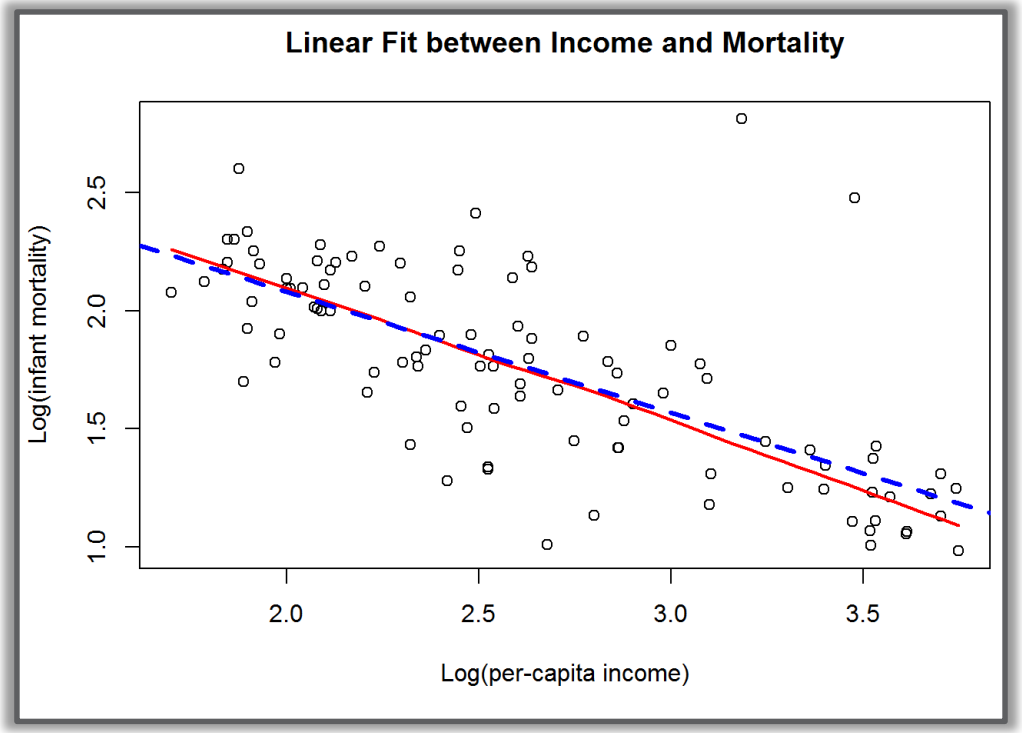
Keep interpretability in mind. If p = .1 seems best for the data, it is often better to use the log transformation (p = 0), because this is easier to interpret.

Linear Model: Income & Infant Mortality

Linear model after transformations

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1034	0.1375	22.57	0.0000
log.income	−0.5118	0.0512	−9.99	0.0000

- A linear model fits well after transformation
  - blue dashed regression line
- Since both variables are transformed by the log10 the coefficients are easy to interpret:
  - An increase in income by 1% is associated, on average, with a .51% decrease in infant mortality



Caution: Non-Linear Transformation

- When a nonlinear transformation is performed strictly on an empirical basis, extrapolation beyond the range of the data is extremely dangerous.
- On the other hand, if guided by strong theory, such extrapolation might be reasonable.

Take-aways: Transforming Nonlinearity

- Why do we want things to be linear?
  - Linear relationships are simple, and there is nice statistical theory for these models.
  - If there are several independent variables, nonparametric regression may be infeasible
- Simple monotone nonlinearity can be corrected
  - direction of curvature does not change
  - correction using
    - a transformation in the family of powers and roots
- Example: quadratic function - two possible transformations
- Mosteller and Tukey's Bulging rule
- Consider how transformation affects symmetry.
  - If the dependent variable already was symmetric,
  - then try to leave this one untouched.
- And again, keep in mind interpretability.

Quantitative Variable

BOX-COX TRANSFORMATION

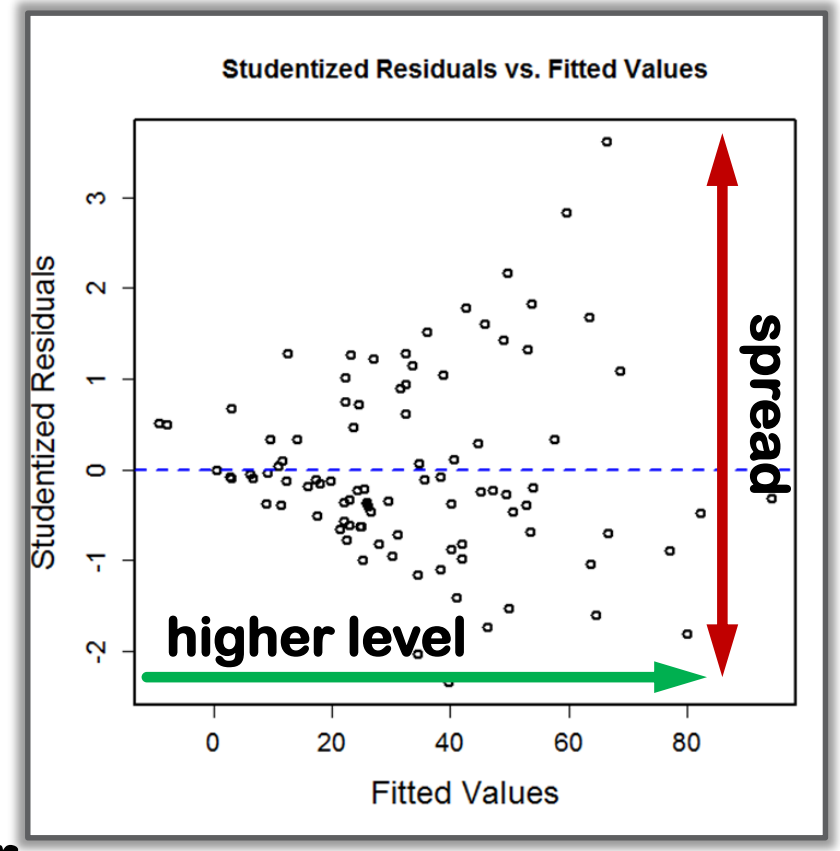
Box-Cox Transformations

- Box and Cox (1964) present a class of transformations of positive values of y designed to stabilize residual variances and linearize fit
- Such a transformation of y can work well when the raw data show heterogeneity of error variance
- The Box-Cox family is defined in terms of a parameter λ:

$$y_\lambda = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$$

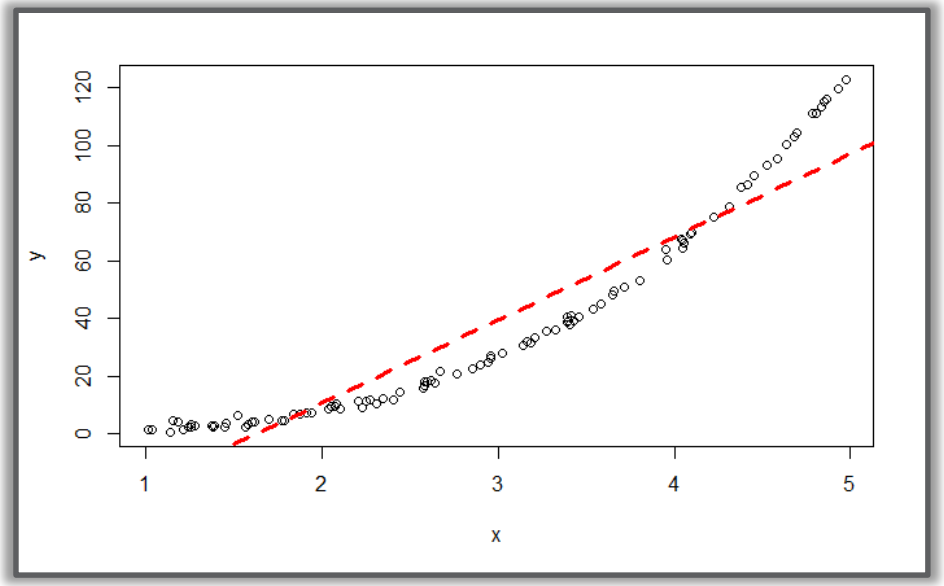
Transforming Non-constant Spread (Heteroscedasticity)

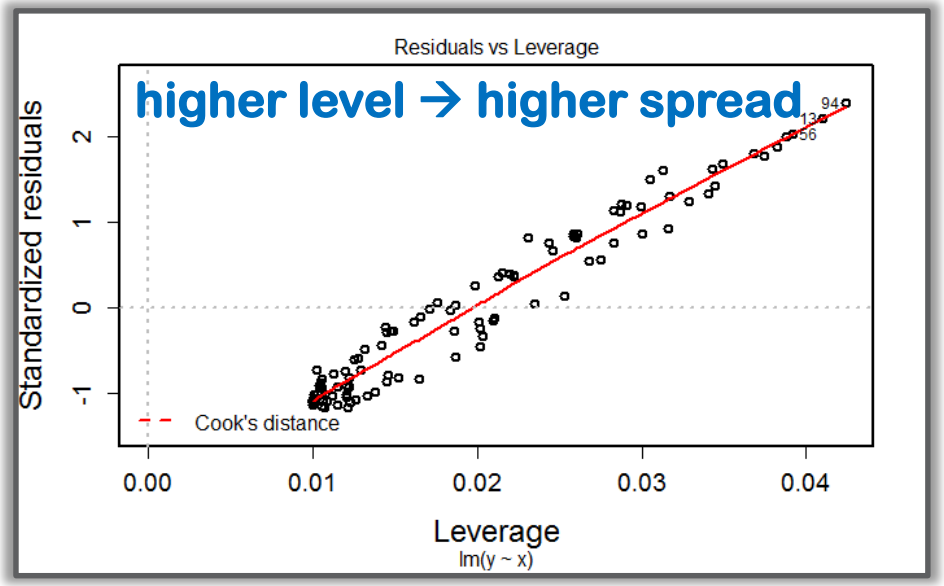
- Differences in spread are often related to differences in level
- When spread is positively related to level:
  - higher level → higher spread (common)
  - compress large values
  - transform down the ladder of powers and roots: p < 1
- When spread is negatively related to level:
  - lower level → higher spread (rare)
  - spread out large values
  - transform up the ladder of powers and roots: p > 1
- Unequal spread and skewness often occur together
  - can be corrected together
  - E.g.: transform Y down the ladder to correct both issues



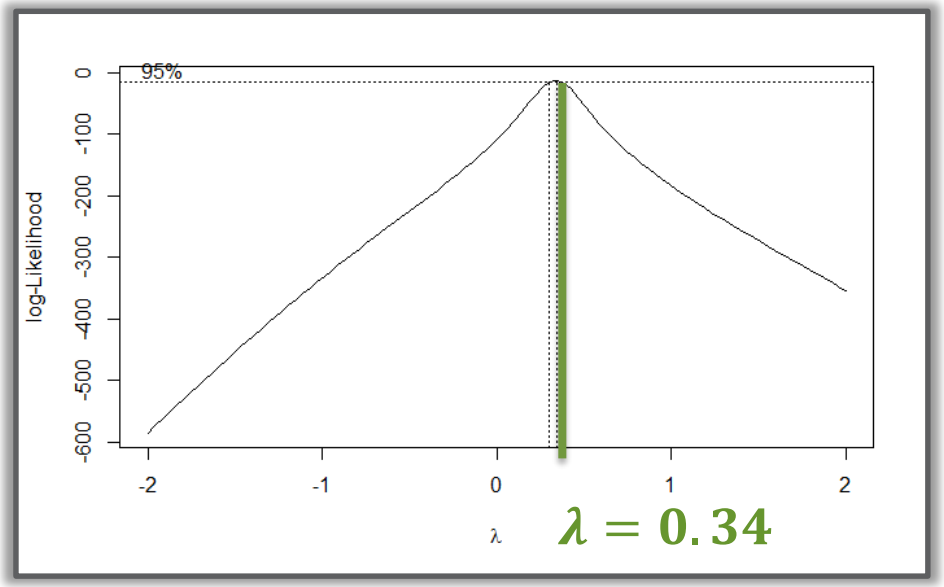
higher level → higher spread

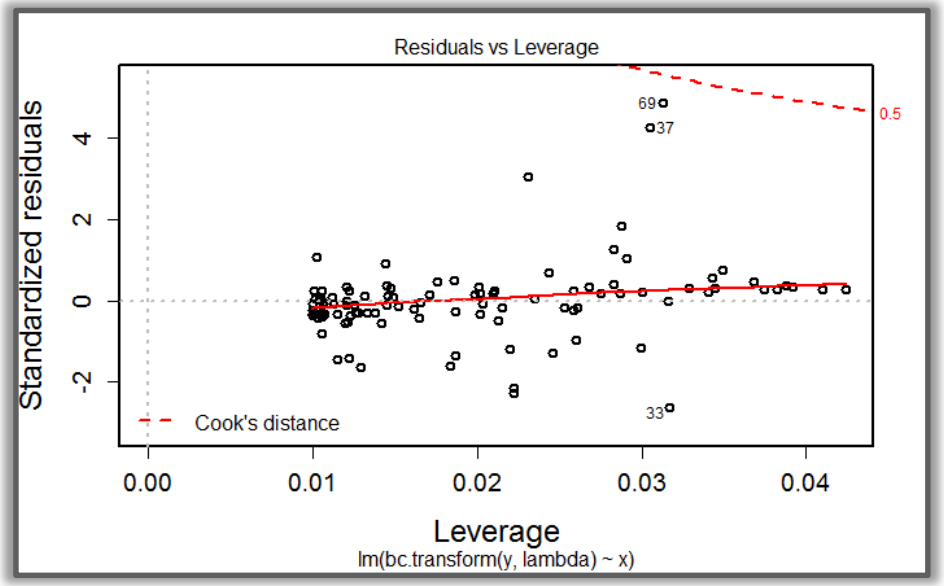
Box-Cox Transformations: Dealing with Heteroscedasticity



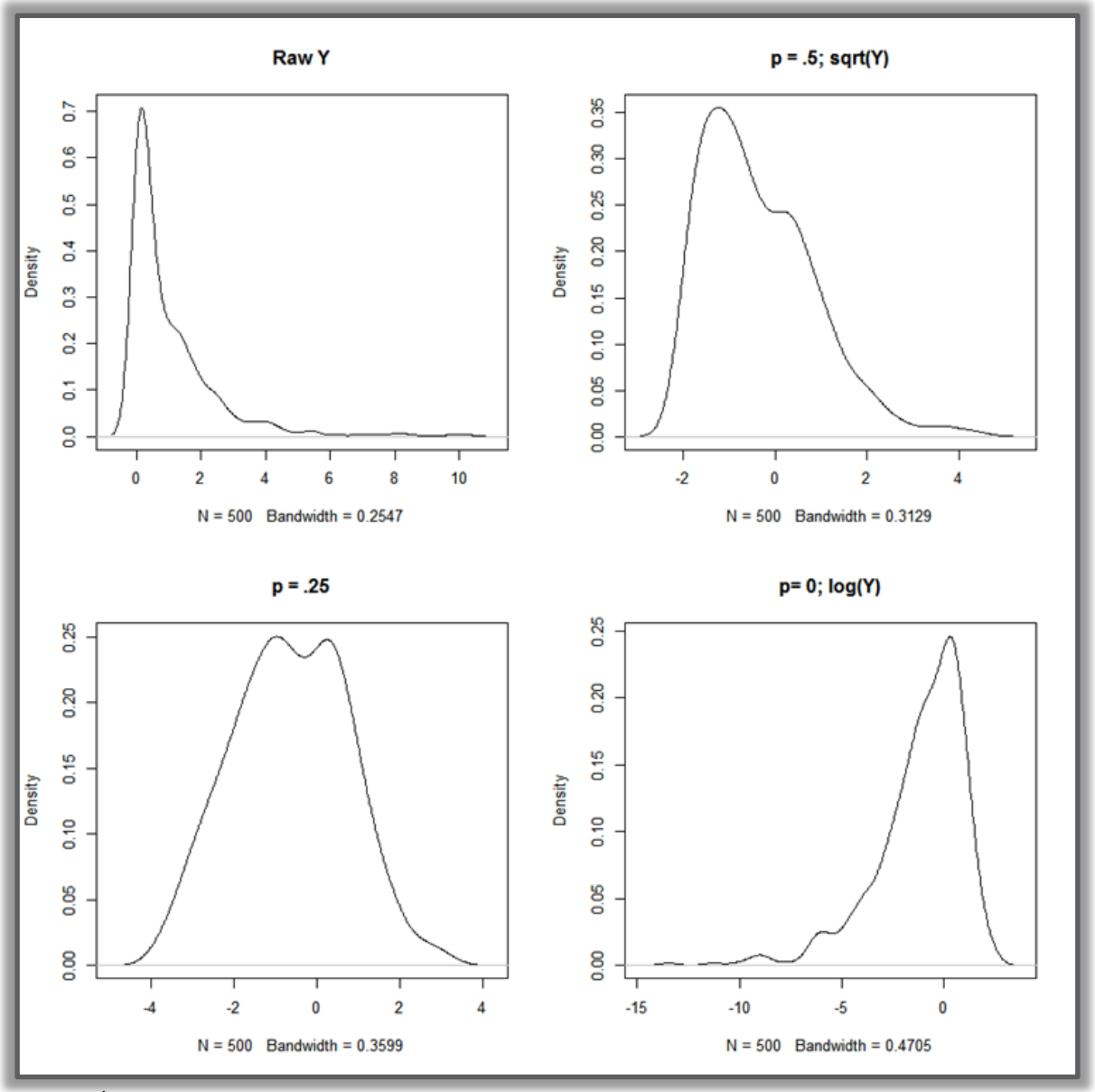


After Box-Cox Transformation with $\lambda = 0.34$





Box-Cox Transformation: Dealing with Skew



Take-Aways: Powers & Roots

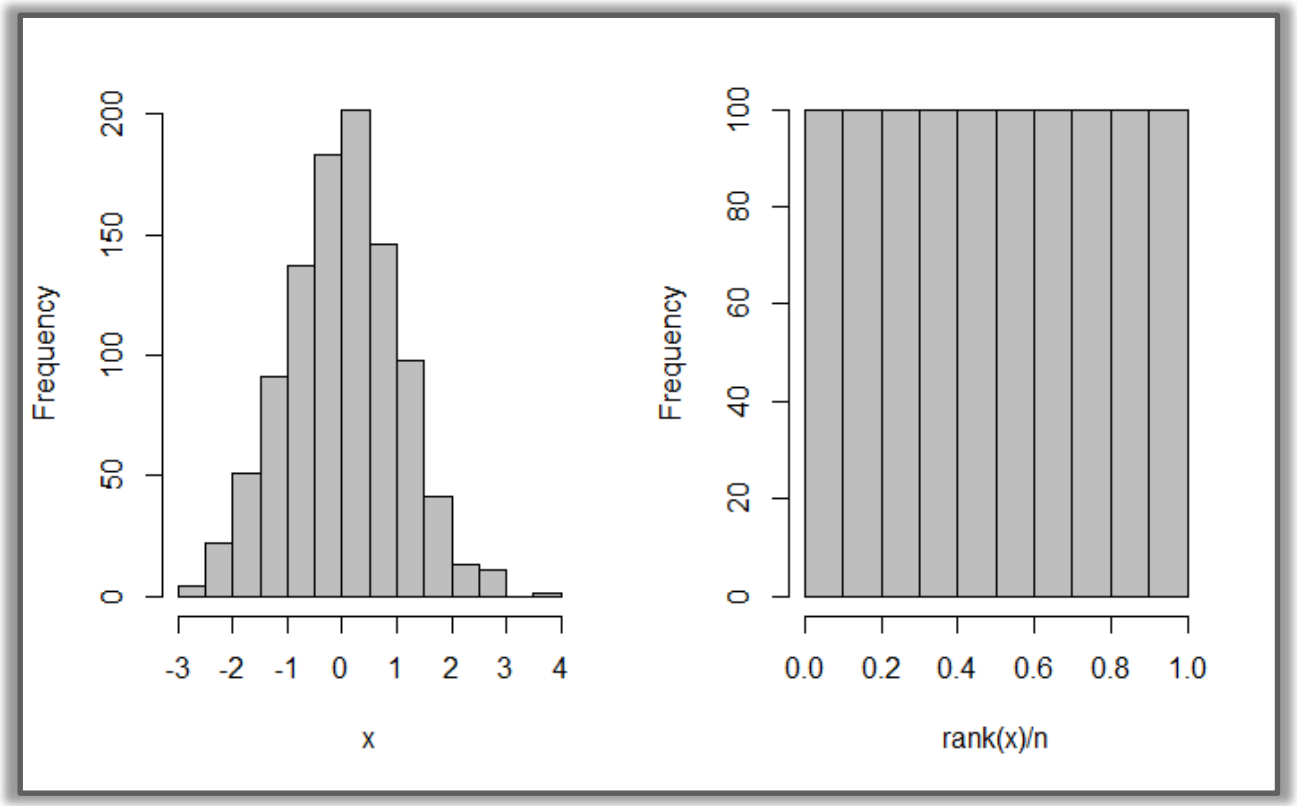
- Power transformations of Y (or sometimes X) can correct problems with normality of errors, or residuals:
  - Transforming Y down the ladder can correct positive skew in the errors (common problem)
  - Transforming Y up the ladder corrects negative skew in the errors
- Power transformations of Y (or X) can stabilize the variance of the errors
- Power transformations of X (or sometimes Y) can make many nonlinear relationships nearly linear
  - Use Box-Cox transformations of X or Y in R using:
    - xp = car::bcPower(x, p)

Quantitative Variable

RANK TRANSFORMATION

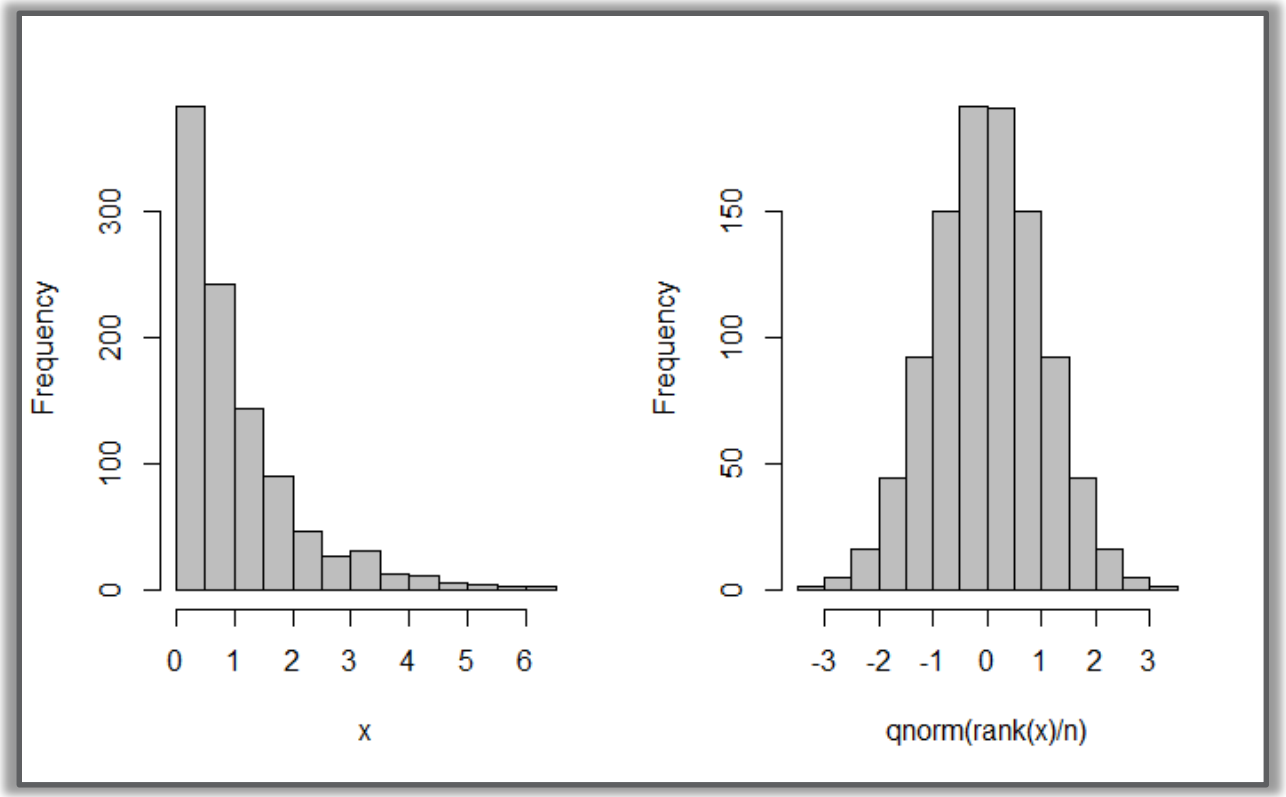
Rank Transformation

- **Rank transformation is central to methods in nonparametric statistics**
  - **Assume sample values $x_1, x_2, \ldots, x_n$**
  - **Define r_i as the rank of x_i, with rank 1 for the smallest and rank n for the largest**
- **Rank-transformed data:**
  - **$r_1, r_2, \ldots, r_n$ or**
  - **$\frac{r_1}{n}, \frac{r_2}{n}, \ldots, \frac{r_n}{n}$**



Rank-based Transformation for Normality

- Let Φ^{-1} denote the inverse cumulative distribution function of a standard normal random variable:
  - E.g.: $\Phi^{-1}(0.975) = 1.96$
  - qnorm() function in R
- $\Phi^{-1}\left(\frac{r_1}{n}\right), \Phi^{-1}\left(\frac{r_2}{n}\right), \ldots, \Phi^{-1}\left(\frac{r_n}{n}\right)$
  - follow normal distribution
  - $\Phi^{-1}(p)$ represents $100 \times p^{th}$ percentile
  - of the standard normal distribution



Take-Aways: Rank Transformations

- **Data transformation can simplify interpretation and analysis**
  - **Many methods critically assume normality**
  - **“Nonparametric” or “distribution-free methods tend to work from rank transformations**

Transformations

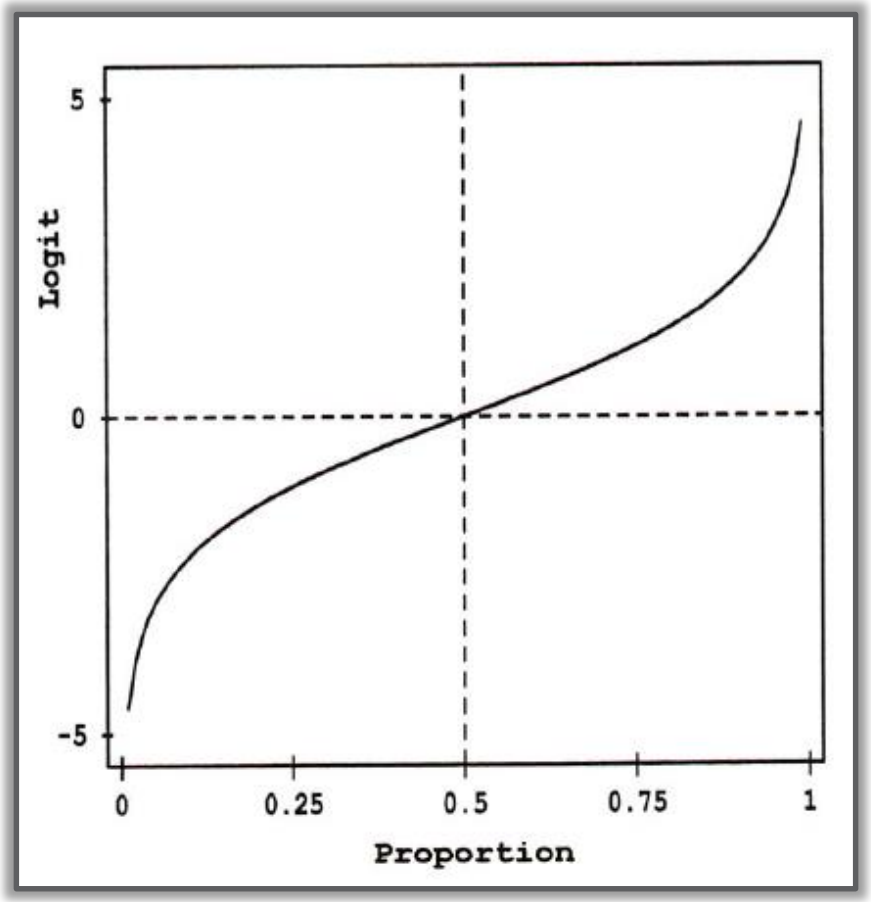
PROPORTIONS, PERCENTAGES AND RATE

Proportions: Transformations: logit and probit

- Power transformations will not work for proportions (including percentages and rates) if the data values approach the boundaries of 0 and 1
- Skewed Proportion Distributions
  - logit transformation
  - probit transformation
- Logit Transformation
  - removes the boundaries of the scale
  - spreads out the tails of the distribution and
  - makes the distribution symmetric about 0
- Probit Transformation
  - if their scales are equated

$$\text{logit} \simeq (\pi/\sqrt{3}) \times \text{probit}$$

$$P \rightarrow \text{logit}(P) = \log_e \frac{P}{1-P}$$



Proportions: logit transformations

- Transformation is nearly linear for proportions between .20 and .80
- Values close to 0 and 1 are spread out at an increasing rate, however
- Transformed variable is now centered at 0 rather than .5

$$P \rightarrow \text{logit}(P) = \log_e \frac{P}{1-P}$$