

Neuendorf



Transforming Data

General guidelines:

1. Data transformations are used primarily for: (a) correcting the distribution of a variable that is not normal, (b) correcting for heteroscedasticity, or (c) linearizing a nonlinear relationship.
2. Transformations should be applied to the IV except in the case of heteroscedasticity. This is particularly true for multivariate situations, such as multiple regression (Fox, 1997).
3. As Fox (1997) says, an effective transformation can be selected analytically or by trial and error. ("It's a creative process" some have said.)
4. As noted by Hair et al. (2006), for a noticeable effect from transformations, the ratio of a variable's mean to its standard deviation should be less than 4.0. When the transformation may be performed on either of two variables, select the variable with the smallest such ratio.
5. Always remember that a transformed variable is now a *new* variable, and needs to be interpreted as such. (See Figure D below.)
6. Always check the effect of the transform before proceeding with further analyses.

Types of transformation:

- I. Transforming data to correct for deviations from normality in a univariate distribution. As noted by Fox, descending the ladder of powers (e.g., to the square root of X or log X) tends to correct a positive skew; ascending the ladder of powers (e.g., to X^2 or X^3) tends to correct a negative skew.

Tukeys ladder of powers	5		Up  -----  Down
	4	X^4	
	3	X^3	
	2	X^2	
	1	X^1	
	$\frac{1}{2}$	\sqrt{X}	
	$\frac{1}{3}$	$\sqrt[3]{X}$	
	(0)	$\log_{10} X$	
	$-\frac{1}{2}$	$-\frac{1}{\sqrt{X}}$	
	-1	$-\frac{1}{X}$	
	-2	$-\frac{1}{X^2}$	
	-3		

<u>Problem</u>	<u>Remedy</u>
Positive skew	take a root (e.g., square root) or logarithm (e.g., \ln , \log_{10}): from less extreme to more extreme, they are: square root, \ln , \log_{10}
Negative skew	use an exponent (e.g., square, cube): from less extreme to more extreme, they are: square, cube, $\exp(e^x)$, 10^x
Positive kurtosis (lepto, “pointy”)	??—in search of a good transform
Negative kurtosis (platy, “flat”)	take the inverse ($1/X$ or $1/Y$)

II. Transforming data to correct for heteroscedasticity (DV exhibits unequal variance across the range of an IV). As Hair et al. note, heteroscedasticity is also due to the distribution(s) of the variable(s); hence, the first step is to check for non-normality of each of the variables, and transform accordingly.

III. Transforming data to correct for heteroscedasticity of residuals in multiple regression (residuals exhibit unequal variance of residuals (errors of prediction) across the range of the predicted DV). Hair et al. offer the following:

<u>Problem</u>	<u>Remedy</u>
Cone opens to the right	take the inverse ($1/X$)
Cone opens to the left	take a root (e.g., square root of X)

IV. Transforming data to linearize a nonlinear relationship. (See Figures A through D below.)

*The introduction of nonlinear components in a multiple regression model—*polynomials*. For example, X^2 and X^3 are both polynomial versions of X that may be included in a regression model.

*Monotonic vs. non-monotonic relationships

References:

Fink, E. L. (2009). The FAQs on data transformation. *Communication Monographs*, 76, 379-397.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage Publications.

Hair, J. F. Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Figure A: An example of linearizing a monotonic, nonlinear relationship

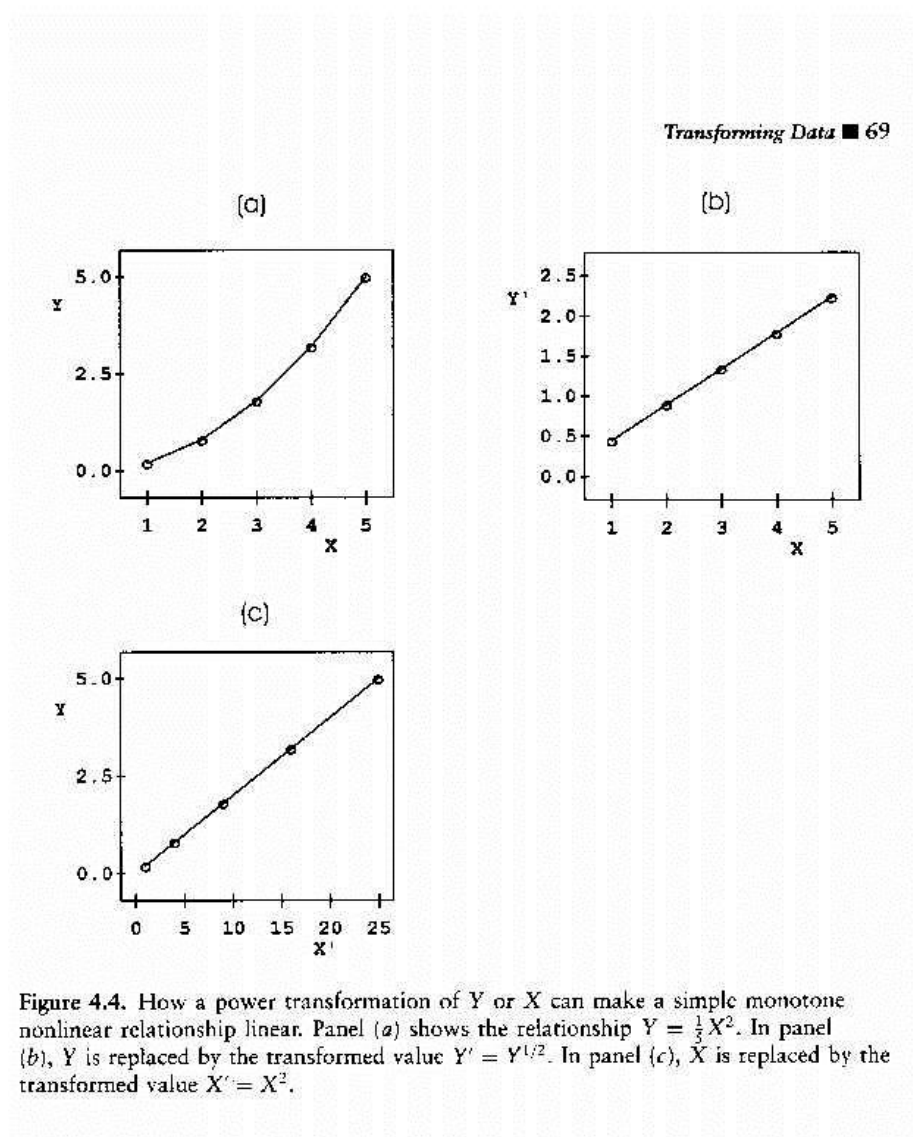


Figure 4.4. How a power transformation of Y or X can make a simple monotone nonlinear relationship linear. Panel (a) shows the relationship $Y = \frac{1}{3}X^2$. In panel (b), Y is replaced by the transformed value $Y' = Y^{1/2}$. In panel (c), X is replaced by the transformed value $X' = X^2$.

Figure B: Examples of monotonic and non-monotonic nonlinear relationships

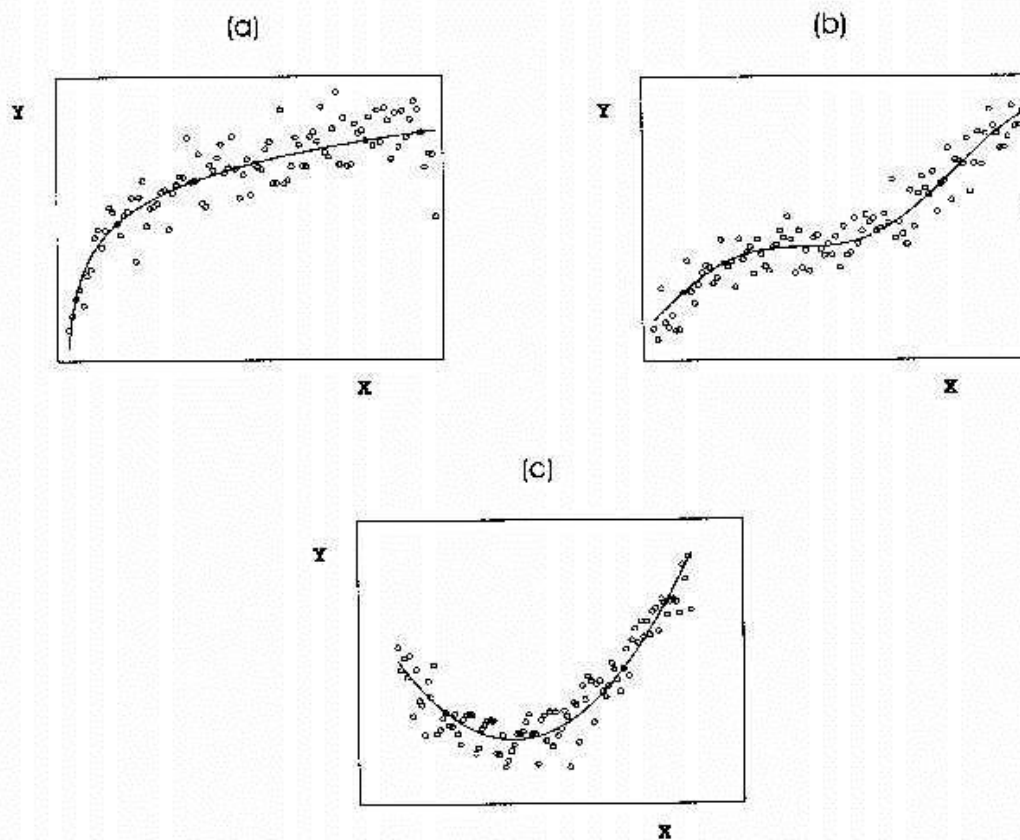


Figure 4.5. (a) A simple monotone relationship between Y and X ; (b) a monotone relationship that is not simple; (c) a relationship that is simple but not monotone. A power transformation of Y or X can straighten (a), but not (b) or (c).

Figure C: The “bulging rule” for transforming nonlinear, monotonic relationships

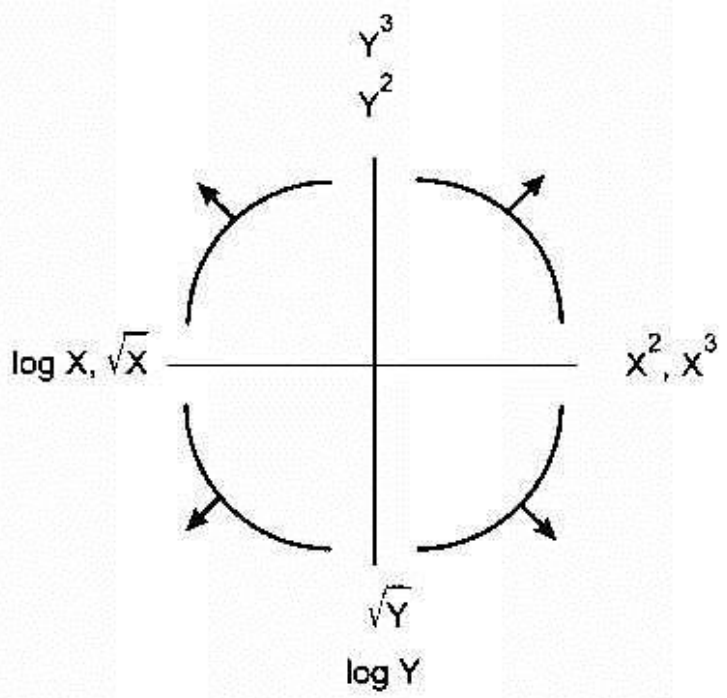


Figure 4.6. Tukey and Mosteller’s “bulging rule”: The direction of the “bulge” indicates the direction of the power transformation of Y and/or X to straighten the relationship between them.

Simple monotone nonlinearity can often be corrected by a power transformation of X , of Y , or of both variables. Mosteller and Tukey’s “bulging rule” assists in the selection of a transformation.

Figure D: An application of the bulging rule. . . that also addressed positive skewness in both variables

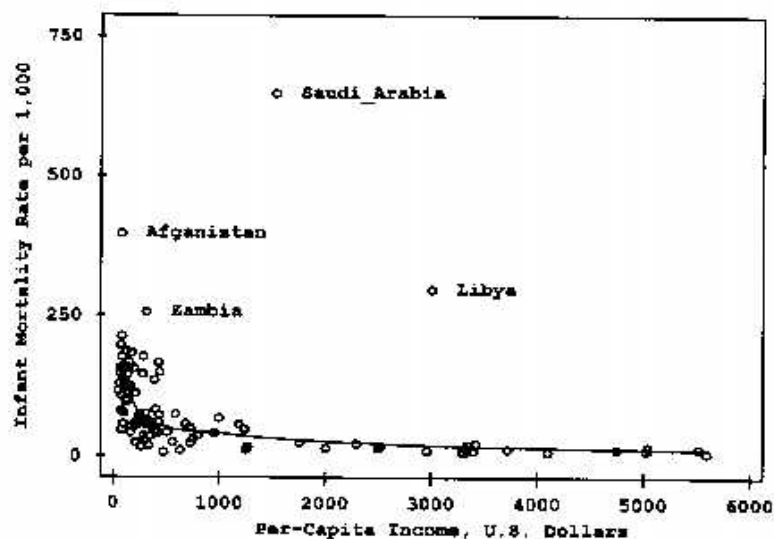


Figure 4.9. Scatterplot of infant mortality rate versus income in U.S. dollars, for 101 nations *circa* 1970. The nonparametric regression shown on the plot was calculated by robust local regression. Several outlying observations are flagged.

Source of Data: Leinhardt and Wasserman (1978).

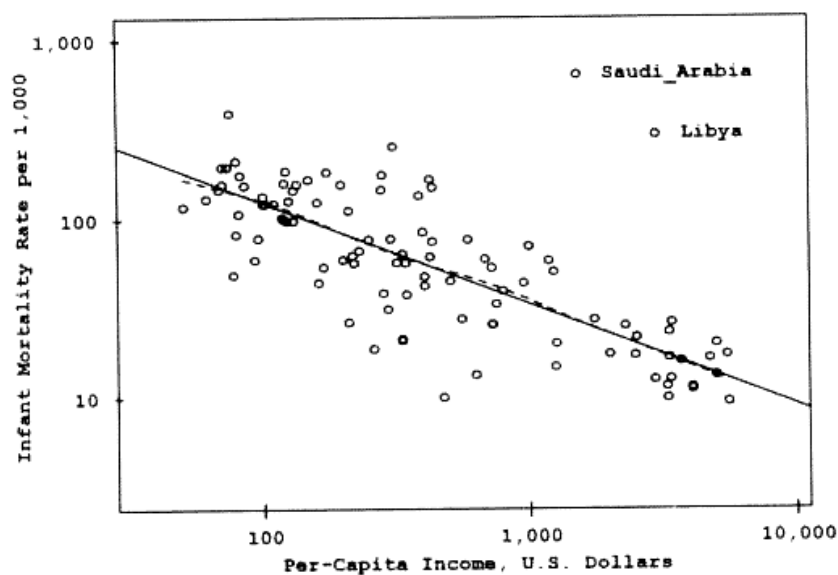


Figure 4.10. Scatterplot of \log_{10} infant mortality rate versus \log_{10} per-capita income for 101 nations. The solid line was calculated by least-squares linear regression, omitting Saudi Arabia and Libya; the broken line was calculated by robust local regression.