STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 7: Transformations

Transformation

- data are messy
- they seldom fit our model assumptions
- why transformation? we transform the data so that the usual linear regression assumptions apply
- we either transform (i) the predictor, (ii) the response or (iii) both, so that in the transformed domain we have

$$E(Y|X = x) \approx \beta_0 + \beta_1 x$$

- note: we used "$\approx$" not "$=$"
- transformation also works for multiple predictors

BodyWt and BrainWt

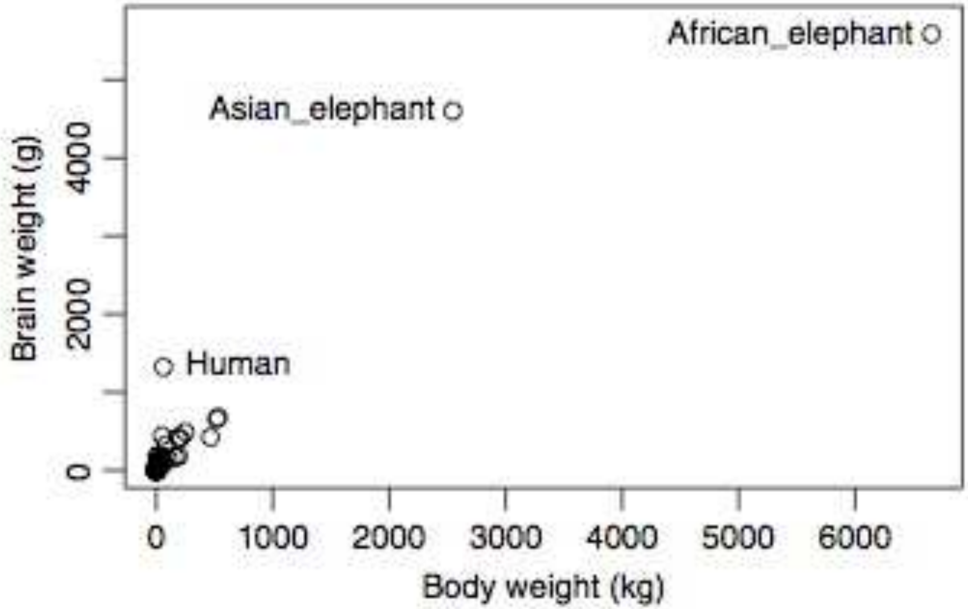
- 

FIG. 7.1 Plot of *BrainWt* versus *BodyWt* for 62 mammal species.

- due to the elephants, it is hard to observe any patterns

Power Transformation

- can be applied to the response, or the predictor, or both
- U: original variable, strictly positive

$$\psi(U, \lambda) = U^{\lambda}$$

- usual range for λ: -2 to 2
- $\lambda = 1 \rightarrow$ no transformation,
  $\lambda = \frac{1}{2} \rightarrow$ square root transformation,
  $\lambda = -1 \rightarrow$ inverse,
  $\lambda = 0 \rightarrow$ taken as the log transformation (not 1)

Power Transformation - con't

- transform both predictor and response

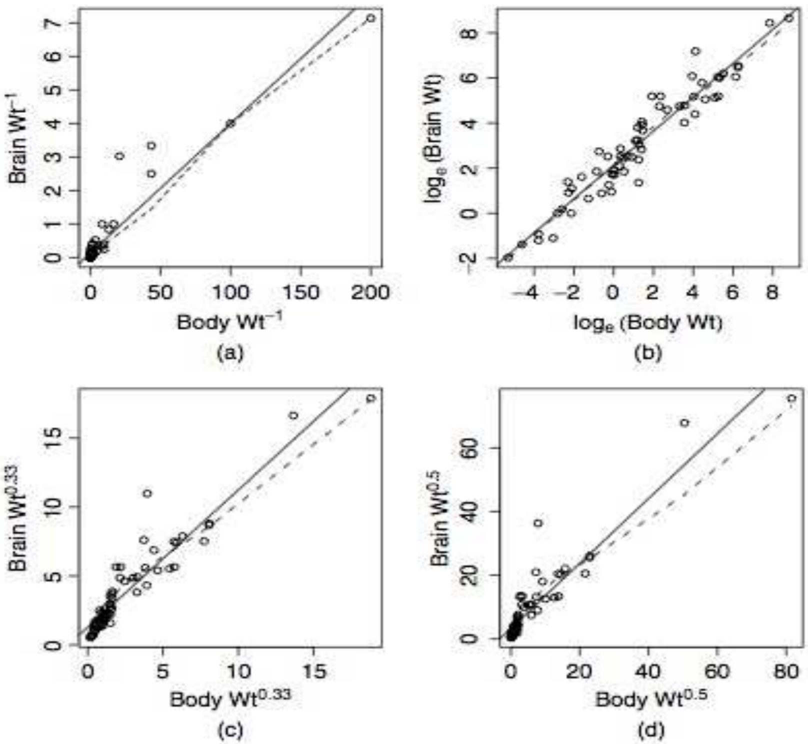


FIG. 7.2 Scatterplots for the brain weight data with four possible transformations. The solid line on each plot is the OLS line; the dashed line is a *loess* smooth.

Power Transformation - con't

- applying log transformation to both the response and predictor, the linear model is given by

$$\log(BrainWt) = \beta_0 + \beta_1 \log(BodyWt) + e$$

- this means we are actually fitting a multiplicative model

$$BrainWt = \beta_0 \times BodyWt^{\beta_1} \times e,$$

- in this example, we choose λ by visual inspection

Transforming only the Predictor

- scaled power transformation
- $\psi_s(X,\lambda) = \begin{cases} (X^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log_e(X) & \text{if } \lambda = 0 \end{cases}$
- $\psi_s(X,\lambda)$ is a continuous function of λ
- $\lim_{\lambda \to 0} \psi_s(X,\lambda) = \log_e(X)$
- How to choose λ?
- fit $(\psi_s(X,\lambda), Y)$ for different values of λ
- note Y is not transformed, thus one can choose λ by minimizing $RSS(\lambda)$, e.g., $\lambda \in \{-1, -\frac{1}{2}, 0, \frac{1}{3}, \frac{1}{2}, 1\}$

Transforming only the Predictor - con’t

- tree height v.s. diameter at 137cm above ground (Dbh)
- scaled power transform only for predictor, plot $(Dbh, \hat{y}_\lambda)$, where $\hat{y}_\lambda = \hat{\beta}_0 + \hat{\beta}_1 \psi_s(Dbh, \lambda),\ \lambda = 1, 0, -1$

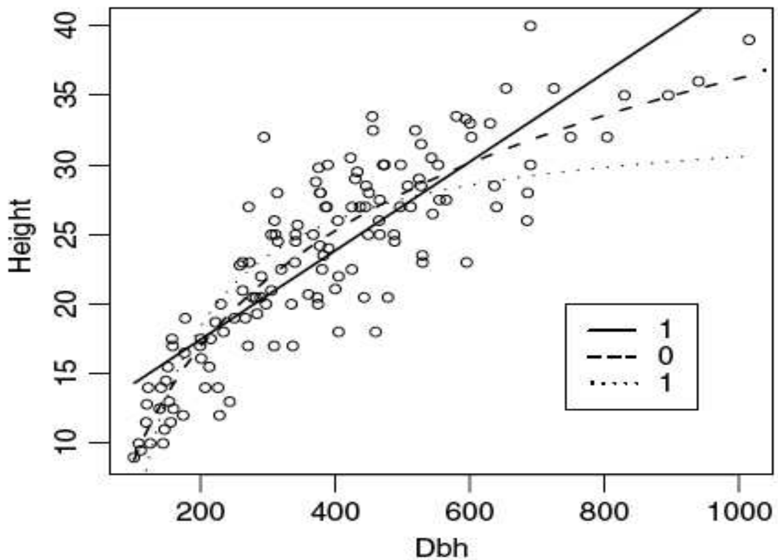


FIG. 7.3 *Height* versus *Dbh* for the red cedar data from Upper Flat Creek.

Transforming only the Predictor - con't

- $\mathrm{E}(Y|X) = \beta_0 + \beta_1 \psi_s(X, \lambda)|_{\lambda=0} = \beta_0 + \beta_1 \log X$
- plot the fitted model with log-transformed predictor
- transform predictor is to improve linearity assumption

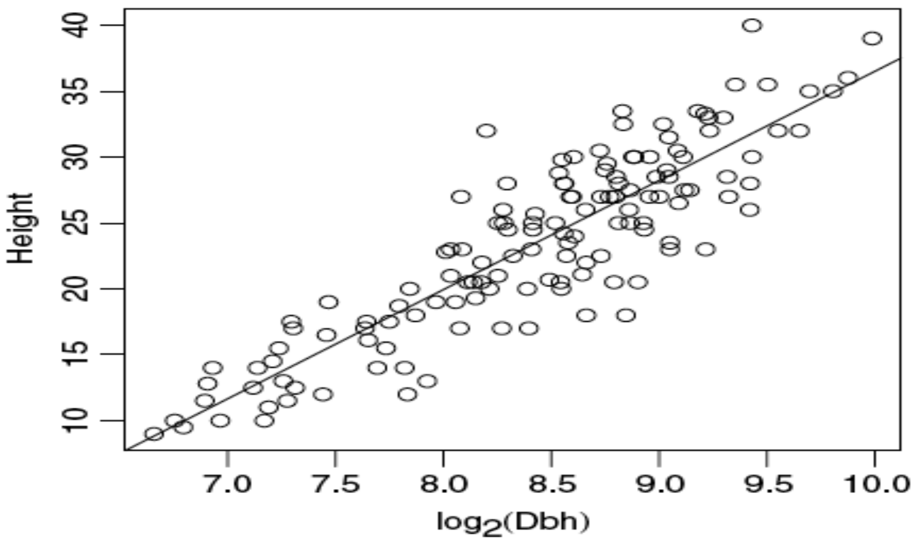


FIG. 7.4 The red cedar data from Upper Flat Creek transformed.

Box-Cox Transformation for Response

- modified power transformation: for response $Y > 0$
- $$\begin{aligned}\psi_M(Y, \lambda_y) &= \psi_S(Y, \lambda_y) \times \mathrm{gm}(Y)^{1-\lambda_y} \\ &= \begin{cases} \mathrm{gm}(Y)^{1-\lambda_y} \times (Y^{\lambda_y} - 1)/\lambda_y & \text{if } \lambda_y \neq 0 \\ \mathrm{gm}(Y) \times \log(Y) & \text{if } \lambda_y = 0 \end{cases}\end{aligned}$$
- $\mathrm{gm}(Y)$: geometric mean of Y, i.e.,

$$\mathrm{gm}(Y) = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} \log_{\mathrm{e}}(y_i)\right\}$$

Box-Cox Transformation for Response - con't

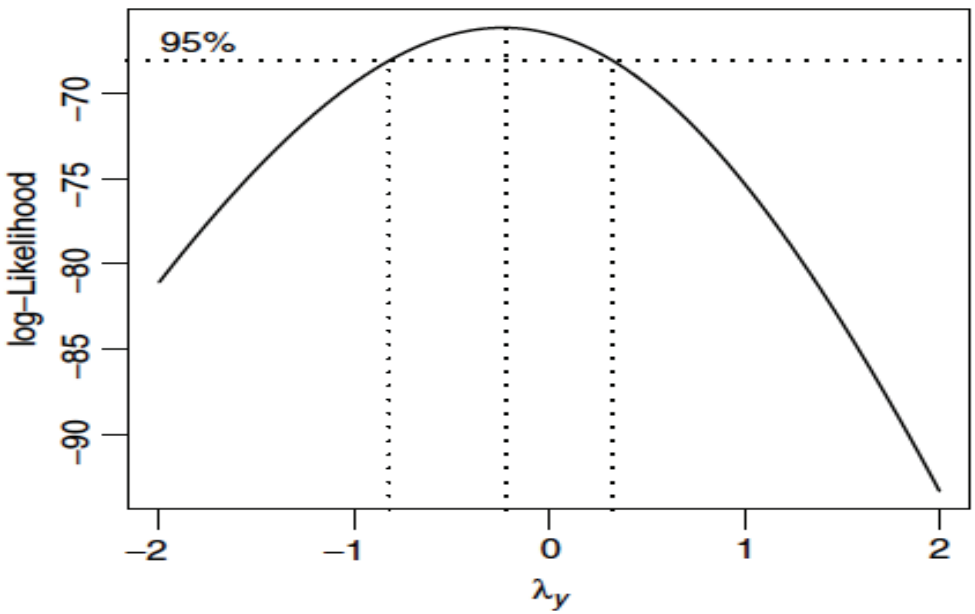
- Box-Cox method assumes

$$E(\psi_M(Y, \lambda_y)|X = \boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$$

- $\mathrm{gm}(Y)^{1-\lambda_y}$: guarantees that the unit of $\psi_M(Y, \lambda_y)$ are the same for all values of λ_y
- so λ_y can be chosen as the one that minimizes $RSS(\lambda_y)$
- goal of Box-Cox: not for linearity, but for normality
- i.e., try to make $\hat{e}_i$ as normal as possible
- R function: $\mathrm{boxcox(object, lambda = ...)}$

Box-Cox Transformation for Response - con't

- Box-Cox graph for highway data: $\hat{\lambda} \approx -0.2$ with the approximate 95% confidence interval $(-0.8, 0.3)$



Moreover...

- what happens if we have negative variables?
- how about multiple regression?
- what you have seen are simple methods: might not work all the times
- that is, it may not be possible for “simultaneous corrections”