Logarithmic transformation of response

Often, support $\mathcal{S}$ of Y is $\mathcal{S} = (0, \infty)$.

Logarithm is then one of transformations to consider when trying to obtain a correct (wrong but useful) model.

Suppose that the following model is correct:

$$\log(Y) = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, \sigma^2).$$

Then

$$Y = \exp\big(\mathbf{x}^\top \boldsymbol{\beta}\big)\, \eta, \quad \eta = \exp(\varepsilon) \sim \mathcal{LN}(0,\, \sigma^2),$$

i.e., errors and a regression function are combined multiplicatively.

Properties of the log-normal distribution:

$$\mathbb{E}(\eta) = M = \exp\Big(\frac{\sigma^2}{2}\Big) > 1 \;\text{ for } \sigma^2 > 0,$$

$$\text{var}(\eta) = V = \big\{\exp(\sigma^2) - 1\big\} \exp(\sigma^2).$$

$\mathrm{var}(Y;\,x)$ increases with $\mathbb{E}(Y;\,x)$

Captured by a normal linear model for $\log(Y)$ as then

$$\mathbb{E}(Y;\,x) = M\,\exp\bigl(\mathbf{x}^\top\boldsymbol{\beta}\bigr),$$

$$\mathrm{var}(Y;\,x) = V\,\exp\bigl(2\,\mathbf{x}^\top\boldsymbol{\beta}\bigr) = V\left(\frac{\mathbb{E}(Y;\,x)}{M}\right)^2,$$

which is increasing function of $\mathbb{E}(Y;\,x)$ for Y with a support $\mathcal{S} = (0,\,\infty)$.

It is then said that the logarithmic transformation stabilizes the variance.

$\mathcal{D}(Y;\,x)$ skewed

Often sufficiently captured (leads to a model which is wrong but useful) by a normal linear model for $\log(Y)$ as then

$$\mathcal{D}(Y;\,x) = \mathcal{LN}\bigl(\mathbf{x}^\top\boldsymbol{\beta},\,\sigma^2\bigr),$$

and a log-normal distribution is one of the “benchmark” skewed distributions.

Interpretation of regression coefficients

Let
$$\mathbf{x}_1 = \left(x_{1,0}, \ldots, x_{1,j}, \ldots, x_{1,k-1}\right)^\top,$$
$$\mathbf{x}_2 = \left(x_{2,0}, \ldots, x_{1,j} + 1, \ldots, x_{2,k-1}\right)^\top,$$
$$\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_j, \ldots, \beta_{k-1}\right)^\top.$$

Then

$$\frac{\mathbb{E}(Y;\, \mathbf{x}_2)}{\mathbb{E}(Y;\, \mathbf{x}_1)} = \frac{M\, \exp\left(\mathbf{x}_2^\top \boldsymbol{\beta}\right)}{M\, \exp\left(\mathbf{x}_1^\top \boldsymbol{\beta}\right)} = \exp(\beta_j).$$

When $\left(\beta_j^L,\ \beta_j^U\right)$ is the confidence interval for β_j with a coverage of $1-\alpha$ then

$$\left(\exp\left(\beta_j^L\right),\ \exp\left(\beta_j^U\right)\right)$$

is the confidence interval for $\frac{\mathbb{E}(Y;\, \mathbf{x}_2)}{\mathbb{E}(Y;\, \mathbf{x}_1)}$ with a coverage of $1-\alpha$.

Logarithmic transformation of response

Interpretation of regression coefficients

When ANOVA linear model with log-transformed response is fitted, estimated differences between the group means of log-response are equal to estimated ratios between the group means of the original response.

When a model with logarithmically transformed response if fitted, estimated regression coefficients, estimates of estimable parameters etc. and their confidence intervals are often reported back-transformed (exponentiated) due to above interpretation.