

NOTES ON THE BOX-COX TRANSFORMATIONS

Stanley L. Sclove

*Professor, Information & Decision Sciences Department
University of Illinois at Chicago*

Outline

1. Background
 2. Some Related Mathematics
 3. Confidence Intervals
 4. Multivariate Data
-

1 Background

Often, one has to consider the issue of whether or not to pre-process the data by transforming them. Often the variables are positive, such as lengths, weights, prices, or blood pressures. Then usually the transformations under consideration are the log, square root, or reciprocal.

The reciprocal transform might mean, for example, analyzing gallons per miles instead of miles per gallon.

The log transform can be done with natural logs or common (base ten) logs. This is just a change of scale, for $\ln y = (\ln 10)(\log y)$. To see this, suppose $y = e^l = 10^m$. Then $\ln e^l = \ln 10^m$, and $l = m \ln 10$; that is, $\ln y = (\log y)(\ln 10)$.

2 Some Related Mathematics

Box and Cox defined a family of power transformations which includes any positive or negative power, as well as the log. The transformation is

$$Y^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda}$$

for $\lambda \neq 0$, and

$$Y^{(0)} = \ln Y.$$

The transform $Y^{(0)}$, corresponding to $\lambda = 0$, is defined by continuity as

$$\lim_{\lambda \rightarrow 0} Y^{(\lambda)}.$$

This limit gives an indeterminate form of the type zero-over-zero; applying *l'Hôpital's rule* shows the limit to be $\ln Y$.

Note that t and F statistics are invariant under linear transformation $y \rightarrow a + by$, so in this sense it makes no difference whether $Y^{(\lambda)}$ or Y^λ is analyzed. But, as will be seen, using $Y^{(\lambda)}$ puts things on a common scale, enabling the choice of power transformations.

For $\lambda = -1$ the transform is $(Y^{-1} - 1)/(-1)$, or $-1/Y + 1$, essentially the negative reciprocal.

To see how the Box-Cox transformations work, write the likelihood of the N observations, consider it as a function of λ and maximize (its log) with respect to λ . To derive the likelihood, let $F(y)$ be the cumulative distribution function (c.d.f.) of the random variable (r.v.) Y , evaluated at y ; that is $F(y) = \Pr\{Y \leq y\}$, and let $f(y)$ denote the probability density function (p.d.f.). It is assumed that there is a value of λ for which the r.v. $Y^{(\lambda)}$ is Normally distributed. Let $F_\lambda(y^{(\lambda)})$ denote the c.d.f. of the random variable Y^λ , evaluated at $y^{(\lambda)}$. Let $\phi(y; \mu, \sigma^2)$ denote the p.d.f. of the Normal distribution with parameters μ and σ^2 . Then $\phi(y; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp[-(y - \mu)^2/2\sigma^2]$. First obtain the p.d.f. of any single observation Y . Write

$$\begin{aligned} f(y) &= \frac{dF(y)}{dy} = \frac{d\Pr\{Y \leq y\}}{dy} \\ &= \frac{d\Pr\{Y^{(\lambda)} \leq y^{(\lambda)}\}}{dy} \\ &= \frac{dF_\lambda(y^{(\lambda)})}{dy} \\ &= f_\lambda(y^{(\lambda)}) \frac{dy^{(\lambda)}}{dy} \\ &= f_\lambda(y^{(\lambda)}) y^{\lambda-1} \\ &= \phi(y^\lambda; \mu_\lambda, \sigma_\lambda^2) y^{\lambda-1} \end{aligned}$$

Let $L(\lambda)$ denote the likelihood, that is, the joint p.d.f. of Y_1, Y_2, \dots, Y_N at y_1, y_2, \dots, y_N considered as a function of the parameters λ, μ_λ , and σ_λ^2 . This is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^N f(y_i) = \prod_{i=1}^N \phi(Y_i^{(\lambda)}; \mu_\lambda, \sigma_\lambda^2) y_i^{\lambda-1} \\ &= (2\pi\sigma_\lambda^2)^{-N/2} \exp\left[-\sum_{i=1}^N (y_i^{(\lambda)} - \mu_\lambda)^2 / (2\sigma_\lambda^2)\right] \prod_{i=1}^N y_i^{\lambda-1}. \end{aligned}$$

The log likelihood is

$$\ln L(\lambda) = (-N/2) \ln 2\pi - (N/2) \ln \sigma_\lambda^2 - \sum_{i=1}^N (y_i^{(\lambda)} - \mu_\lambda)^2 / (2\sigma_\lambda^2) + (\lambda - 1) \sum_{i=1}^N \ln y_i.$$

The maximum likelihood estimate (MLE) of μ_λ is $\overline{y^{(\lambda)}}$ and that of σ_λ^2 is $SSD(\lambda)/N$, where

$$SSD(\lambda) = \sum_{i=1}^n (y_i^{(\lambda)} - \overline{y^{(\lambda)}})^2.$$

This gives

$$\begin{aligned} \max_{\mu_\lambda, \sigma_\lambda^2} \ln L(\lambda) &= (-N/2) \ln 2\pi - (N/2) \ln SSD(\lambda) + (N/2) \ln N - N/2 + (\lambda - 1) \sum_{i=1}^N \ln y_i \\ &= (-N/2) \ln SSD(\lambda) + (\lambda - 1) \sum_{i=1}^N \ln y_i + \text{Constant}. \end{aligned}$$

The MLE of λ is the value for which

$$(-N/2) \ln SSD(\lambda) + (\lambda - 1) \sum_{i=1}^N \ln y_i$$

is maximized, i.e., for which the criterion

$$\ln SSD(\lambda) - (2/N)(\lambda - 1) \sum_{i=1}^N \ln y_i$$

is minimized.

Further analysis shows that if the transform is modified to

$$Y^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda(g.m.)^{\lambda-1}},$$

where

$$g.m. = \left(\prod_{i=1}^N y_i\right)^{1/N} = \exp\left(\sum_{i=1}^N \ln y_i / N\right),$$

then the criterion can be taken to be simply $SSD(\lambda)$.

Note that equivalently the criterion can be s_λ^2 or s_λ . In practice one can evaluate the criterion for various values of λ , e.g. $\lambda = -2.0(0.1)+2.0$, that is, from -2.0 to +2.0 in steps of size 0.1, and see where it is minimized.

3 Confidence Interval for λ

Let β denote any one of the regression parameters in a Normal multiple linear regression model. The F test statistic for $H_0 : \beta = \beta_0$ is

$$F(\beta_0) = \frac{RSS(\beta_0) - RSS(\hat{\beta})}{RSS(\hat{\beta})/RDF},$$

where $RSS(\hat{\beta})$ is the Residual Sum of Squares corresponding to the MLE $\hat{\beta}$, $RSS(\beta_0)$ is the Residual Sum of Squares corresponding to the hypothesized β_0 , RDF is the Residual Degrees of Freedom, $N - (p + 1)$, and p is the number of explanatory variables. The statistic $F(\beta_0)$ is distributed according to the F distribution with 1 and $N - (p + 1)$ degrees of freedom (d.f.); this is the same as the distribution of t_{N-p-1}^2 . A $100(1-\alpha)\%$ confidence interval (CI) for β is the set of acceptable β_0 , that is

$$\{\beta_0 : F(\beta_0) \leq t^{*2}, \},$$

where $t^* = t_{N-p-1; \alpha/2}$, the upper $\alpha/2$ percentage point of the t distribution with $N - p - 1$ d.f. This CI can be written

$$\{\beta_0 : RSS(\beta_0) \leq RSS(\hat{\beta})(1 + t^{*2}/RDF)\}.$$

By analogy with this, a $100(1-\alpha)\%$ CI for λ is

$$\{\lambda : RSS(\lambda) \leq RSS(\hat{\lambda})(1 + t^{*2}/RDF)\},$$

where $RDF = N - p - 1$. This applies to multiple regression models. When the model involves only a mean, the CI is

$$\{\lambda : SSD(\lambda) \leq SSD(\hat{\lambda})(1 + t^{*2}/RDF)\},$$

where $RDF = N - 1$. In terms of $s_{\hat{\lambda}}^2$, this is

$$\{\lambda : s_{\lambda}^2 \leq s_{\hat{\lambda}}^2(1 + t^{*2}/RDF)\},$$

or in terms of s_{λ} , it is

$$\{\lambda : s_{\lambda} \leq s_{\hat{\lambda}}\sqrt{(1 + t^{*2}/RDF)}\}.$$

4 Multivariate Data

Usually it is preferred to make the same transformation for all n variables. This would be the case if all the variables are lengths, or all are weights, or all are prices, or all are blood pressures, etc. Here is a suggestion for such situations.

To choose λ , apply the Box-Cox method to each variable separately. Typically the CIs will overlap, i.e., will have a non-null intersection S . Here is one way to proceed: If $\lambda = 1$ is in S , don't transform any of the variables. If $\lambda = 1$ is not in S but $\lambda = 0$ is, make the log transform on all variables. If $\lambda = 1$ and $\lambda = 0$ are not in S but $\lambda = 1/2$ is, make the square root transform on all variables. If $\lambda = 1, 0$, and $1/2$ are not in S but $\lambda = -1$ is, make the reciprocal transform on all variables.

If the CIs don't overlap, consider increasing the confidence level (e.g., from 95% to 99%) so that the CIs become wider and overlap.