Transformacja Boxa-Coxa

To keep the mathematics uncluttered, the theory will be described in the context of the simple regression model, where we are choosing between:

$$Y = \beta_1 + \beta_2 X + u$$

and:

$$\log Y = \beta_1 + \beta_2 X + u.$$

It generalises with no substantive changes to the multiple regression model.

The two models are actually special cases of the more general model:

$$Y_\lambda = \frac{Y^\lambda - 1}{\lambda} = \beta_1 + \beta_2 X + u$$

with $\lambda = 1$ yielding the linear model (with an unimportant adjustment to the intercept) and $\lambda = 0$ yielding the logarithmic specification at the limit as λ tends to zero. Assuming that u is iid (independently and identically distributed) $N(0, \sigma^2)$, the density function for u_i is:

$$f(u_i) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-u_i^2/2\sigma^2}$$

and hence the density function for $Y_{\lambda i}$ is:

$$f(Y_{\lambda i}) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2/2\sigma^2}.$$

From this we obtain the density function for Y_i:

$$f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2/2\sigma^2} \left| \frac{\partial Y_{\lambda i}}{\partial Y_i} \right| = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2/2\sigma^2} Y_i^{\lambda - 1}.$$

The factor $\left| \frac{\partial Y_{\lambda i}}{\partial Y_i} \right|$ is the Jacobian for relating the density function of $Y_{\lambda i}$ to that of Y_i. Hence the likelihood function for the parameters is:

$$L(\beta_1, \beta_2, \sigma, \lambda) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \prod_{i=1}^{n} \mathrm{e}^{-(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2/2\sigma^2} \prod_{i=1}^{n} Y_i^{\lambda - 1}$$

and the log-likelihood is:

$$\begin{aligned}\log L(\beta_1, \beta_2, \sigma, \lambda) &= -\frac{n}{2}\log 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 + \sum_{i=1}^{n} \log Y_i^{\lambda-1} \\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 + (\lambda - 1)\sum_{i=1}^{n}\log Y_i.\end{aligned}$$

From the first-order condition $\partial \log L/\partial\sigma = 0$, we have:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 = 0$$

giving:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2.$$

Substituting into the log-likelihood function, we obtain the concentrated log-likelihood:

$$\log L(\beta_1, \beta_2, \lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\frac{1}{n}\sum_{i=1}^{n}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n}\log Y_i.$$

The expression can be simplified (Zarembka, 1968) by working with Y_i^* rather than Y_i, where Y_i^* is Y_i divided by Y_{GM}, the geometric mean of the Y_i in the sample, for:

$$\begin{aligned}\sum_{i=1}^{n}\log Y_i^* &= \sum_{i=1}^{n}\log(Y_i/Y_{\mathrm{GM}}) = \sum_{i=1}^{n}(\log Y_i - \log Y_{\mathrm{GM}}) \\ &= \sum_{i=1}^{n}\log Y_i - n\log Y_{\mathrm{GM}} = \sum_{i=1}^{n}\log Y_i - n\log\left(\prod_{i=1}^{n} Y_i\right)^{1/n} \\ &= \sum_{i=1}^{n}\log Y_i - \log\left(\prod_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n}\log Y_i - \sum_{i=1}^{n}\log Y_i = 0.\end{aligned}$$

With this simplification, the log-likelihood is:

$$\log L(\beta_1, \beta_2, \lambda) = -\frac{n}{2}\left(\log 2\pi + \log\frac{1}{n} + 1\right) - \frac{n}{2}\log\sum_{i=1}^{n}(Y_{\lambda i}^* - \beta_1 - \beta_2 X_i)^2$$

and it will be maximised when β_1, β_2 and λ are chosen so as to minimise $\sum_{i=1}^{n}(Y_{\lambda i}^* - \beta_1 - \beta_2 X_i)^2$, the residual sum of squares from a least squares regression of the scaled, transformed Y on X. One simple procedure is to perform a grid search, scaling and transforming the data on Y for a range of values of λ and choosing the value that leads to the smallest residual sum of squares (Spitzer, 1982).

A null hypothesis $\lambda = \lambda_0$ can be tested using a likelihood ratio test in the usual way. Under the null hypothesis, the test statistic $2(\log L_\lambda - \log L_0)$ will have a chi-squared distribution with one degree of freedom, where $\log L_\lambda$ is the unconstrained log-likelihood and L_0 is the constrained one. Note that, in view of the preceding equation:

$$2(\log L_\lambda - \log L_0) = n(\log RSS_0 - \log RSS_\lambda)$$

where RSS_0 and RSS_λ are the residual sums of squares from the constrained and unconstrained regressions with Y^*.

The most obvious tests are $\lambda = 0$ for the logarithmic specification and $\lambda = 1$ for the linear one. Note that it is not possible to test the two hypotheses directly against each other. As with all tests, one can only test whether a hypothesis is incompatible with the sample result. In this case we are testing whether the log-likelihood under the restriction is significantly smaller than the unrestricted log-likelihood. Thus, while it is possible that we may reject the linear but not the logarithmic, or vice versa, it is also possible that we may reject both or fail to reject both.