Why is the Box-Cox Transformation so effective?

Haim Shore

November 2013

This question has intrigued my curiosity for many years. I have had the opportunity to talk both to Box and to Cox about their transformation (Box and Cox, 1964). I conversed with the late George Box (deceased last March at age 94) when I was a visitor in Madison, Wisconsin, back in 1993-4. A few years later I talked to David Cox at a conference on reliability in Bordeaux (MMR'2000).

I asked both the same question, I received the same response.

The question was: What was the theory that led to the derivation of the Box-Cox transformation?

The answer was: "No theory. This was a purely empirical observation".

The question therefore remains: Why is the Box-Cox transformation so effective, in particular when applied to a response variable in the framework of linear regression analysis?

A common answer, which appears often in the literature, is that the BC transformation "puts" the response in its "correct" scale, thus eliminating unnecessary (artificial) interaction effects while invoking the basic set of assumptions (denoted the normal scenario), needed to implement linear regression analysis. Applying the latter to the transformed data, so the explanation goes, is expected to lead to the most "natural" (read "simplest"!) relationship between the response and its covariates. As a result, the normal scenario is preserved, thus validating the use of linear regression analysis and allied statistical analyses.

While this explanation is valid, I doubt that it provides a full and satisfactory explanation for the cumulative evidence of the exceptional modeling power of the BC transformation.

In this post I would like to suggest an alternative explanation.

Let Y be the response and let η denote a linear combination of effects (or covariates), often denoted the linear predictor (LP). The one-parameter BC transformation is:

$$\frac{Y^{\lambda}-1}{\lambda}=\eta+\varepsilon \quad , \qquad (1)$$

where ε is a zero-mean normal error and we have ignored in (1) a geometric mean that is commonly added to the BC transformation in order to standardize the mean-squared error. Expressing ε in terms of the standard normal variate, Z, we obtain:

$$\frac{Y^{\lambda}-1}{\lambda}=\eta+\sigma_{\varepsilon}Z\ . \qquad (2)$$

This model may also serve to express the 100p-th percentile of Y, denoted y_p, in terms of the corresponding percentile of Z, z_p:

$$\frac{y_p^{\lambda}-1}{\lambda}=\eta+\beta z_p+\varepsilon_m\ , \qquad (3)$$

where β is a parameter and ε_m is a the model's random error (different from ε used in (1)).

The inverse BC transformation (the inverse of (1)), which expresses Y explicitly in terms of a nonlinear function of η and ε, is:

$$Y=[\lambda(\eta+\varepsilon)+1]^{(1/\lambda)}, \qquad (4)$$

and the inverse of (2) is:

$$Y=[\lambda(\eta+\sigma_{\varepsilon}\,\mathrm{Z})+1]^{(1/\lambda)}. \qquad (5)$$

Note that both (4) and (5) express the relationships between two random variables: Y with ε and Y with Z. Similarly, the inverse of (3) (ignoring the model's error) is:

$$y_p=[\lambda(\eta+\beta z_p)+1]^{(1/\lambda)}\ . \qquad (6)$$

We denote (6) "the normal-based quantile function" since it expresses the p-th quantile of Y, y_p, in terms of the corresponding quantile of a normal variate, z_p. For the median of Y (Med) we obtain from (6) the following model (z_p=0):

$$Med=\{(1+\lambda\eta)^{[1/(\lambda\eta)]}\}^{\eta} \qquad (7)$$

We realize that this median model is a nonlinear function of LP (η), and it comprises three distinct models: linear (λ=1), power (λ≠1, λ≠0) and exponential (λ→0). These are not some arbitrary models. These are fundamental functions that keep re-appearing in various scientific and engineering models (Shore, 2004, 2005, 2011, 2012, 2013a). Furthermore, they represent different levels of monotone convexity: As we move from linear to power to exponential models, the intensity of monotone convexity increases. In other words, the trio of models, "linear, power, exponential," represents three modes of modeling monotone convexity, where each represents a fundamentally different degree of monotone convexity. These models may therefore be arranged in a hierarchy of models, according to their level of monotone convexity. We denote this hierarchy: "The ladder of monotone convex functions" (Shore, 2005, 2012).

Let us see what effect this hierarchy of models has on the effectiveness of the BC transformation. Using (7) as a model for the median of the data on hand, one does not have to decide in advance which of the three levels of monotone convexity, as conveyed by the three functions, best fit the data. Estimating the parameter λ would determine that. In other words: the median model (7), derived from the BC transformation, renders three separate models, which convey different levels of convexity intensity, into mere points on a "Continuous spectrum of monotone convexity", as the latter is spanned by the parameter λ.

We have arrived at an answer to our initial question: The BC transformation is so powerful in modeling relationships between a response and its covariates (included in the LP) because it represents three fundamental models, having widely different levels of monotone convexity. However, the data alone determine which point on the continuous spectrum of monotone convexity, spanned by λ, best represents them. There is no need to specify a model.

A good analogy that may help understand how separate entities (like models, in our case) may become mere points on a continuous spectrum is provided by the colors. Formerly treated as separate entities, we now know that each color is just a point on the continuous spectrum of electromagnetic radiation. Likewise, the "Linear, power, exponential" trio of fundamental models are in fact not separate entities but mere points belonging to the continuous spectrum of monotone convexity, partially spanned by the inverse BC transformation (we will elaborate later on the use of "partial" here).

We say that the inverse BC transformation owns the "Continuous Monotone Convexity (CMC)" property (Shore, 2011, 2013ab). We believe that this property alone may explain the extreme capability of the BC transformation to provide high-quality models (as judged by various goodness-of-fit criteria, as MSE, and various stability criteria, like AIC and BIC).

An immediate consequence of this new understanding should be a modification of how the parameter λ and the LP are estimated. Both are presently estimated within the framework of linear regression analysis (with λ identified as the value that minimizes the associated mean-squared-error). We suggest that once the parameters of the LP and λ have been estimated (within linear regression), the parameter λ should be re-estimated by minimizing the sum of absolute values obtained on fitting the median model given by (7). Using software that performs quantile regression analysis may be helpful in that regard. If the obtained estimate of λ is close to the value obtained in the linear regression stage – the model is well estimated. Otherwise, iteratively estimating the BC transformation and the derived median model (7) may be good practice.

All of the above, regarding the median model, may naturally be extended to estimating any percentile, using the normal-based quantile function given by (6) (see details in Shore, 2012).

While the inverse BC transformation provides the first three "steps" of the "Ladder of monotone convex functions", the "story" does not end there. The BC transformation provides only a partial representation to the "Ladder". As it turns out, the "Ladder" has more steps to climb in order to obtain ever more convex models. This is done by cyclically repeating the basic trio of functions in order to obtain increasing level of monotone convexity. Examples for models, where the "linear, power, exponential" trio is repeated in order to obtain models with stronger monotone convexity, may be found abundantly in various published scientific and engineering models (find some examples in Shore, 2004, 2011).

Recently, a new modeling methodology, denoted "Response Modeling Methodology (RMM)", has been introduced that captures all the models belonging to the "Ladder" in a single model (of which the inverse BC transformation is a special case). RMM naturally owns the CMC property so that it extends that property to all models of the "Ladder". Furthermore, RMM demonstrates that for the "cost" of two additional parameters, the "cycle" of "linear, power, exponential" functions may be repeated indefinitely to obtain more convex models. In other words: the "Ladder" is practically unbounded from above and models with any desirable degree of monotone convexity may be provided, using RMM.

An introductory exposition of the basic principles of RMM will be given in a future post.

References

[1] Box, George E. P.; Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B 26 (2): 211–252

[2] Shore, H. (2004). Response Modeling Methodology (RMM) – Validating evidence from engineering and the sciences. *Quality & Reliability Engineering International*, 20: 61-79.

[3] Shore, H. (2005). *Response Modeling Methodology – Empirical Modeling for Engineering and Science*. World Scientific Publishing Co. Ltd., Singapore.

[4] Shore, H. (2011). Response Modeling Methodology – Advanced review. *WIREs Computational Statistics*, 3 (4): 357-372.

[5] Shore, H. (2012). Estimating RMM Models – Focus article. *WIREs Computational Statistics*, 4(3): 323-333.

[6] Shore, H. (2013a). Modeling and monitoring ecological systems – A statistical process control approach. *Quality and Reliability Engineering International*. Published on line, July. DOI: 10.1002/qre.1544.

[7] Shore, H. (2013b). A general model of random variation. *Communication in Statistics (Theory & Methods)*. In press.