Box-Cox transformations for generalized linear models

Patrick R Johnston
246 Brattle Street, Cambridge, MA 02138

Abstract
The original Box-Cox method attempts to transform the response variable of a regression model such that it follows a Normal linear model with constant variance. We extend the method to generalized linear models (GLMs) based on natural exponential dispersion (NED) distributions. Thus a power, λ, is sought such that y^λ follows a GLM based on a NED with constant dispersion. Extensions to NEDs with non-constant dispersions are also considered. We apply the method to a two-group blink rate study.

Key Words: Box-Cox transformation, generalized linear models, exponential dispersion models, blink rate, interblink interval.

1. Introduction

The original Box-Cox method is based on an attempt to transform the response variable of a regression model such that it follows a Normal linear model with constant variance (Box and Cox (1964)). We extend target models from linear models based on the Normal distribution with constant variance to generalized linear models (GLMs) based on natural exponential dispersion (NED) distributions with constant dispersion. Thus a power, λ, is sought such that y^λ follows a GLM based on a NED with constant dispersion. Extensions to NEDs with non-constant dispersions are also considered.

We apply the method to a two-group blink rate study (Johnston et al (2013) provide clinical details as well as an alternative analysis based on time series methods). The study comprised 10 Healthy and 11 Dryeye subjects whose blinks, N(i), were counted over recorded time intervals, T(i) (Figures 1 and 4). Time intervals ranged from 5.7 to 14.5 minutes, while blink counts ranged from 64 to 651. We consider two responses: blink rate, BR = N(i)/T(i), and interblink interval, IBI = T(i)/N(i). For convenience, an identity link is used, and thus models such as $y^{1/4} \sim \text{Normal}(\alpha+\beta x,\varphi)$ and $y^{1/2} \sim \text{Gamma}(\alpha+\beta x,\varphi)$ are compared. Likelihood computations were carried out using the nlmixed procedure of SAS 9.3 (2011).

2. Theory

GLMs were originally based on natural exponential distributions with a single mean parameter depending on a vector of predictors, x_i. Thus $y_i \sim NE(\mu_i)$, where $\mu_i = g(x_i\beta)$ with inverse link g (Nelder and Wedderburn (1972), McCullagh and Nelder (1983, 1989)). These include the one parameter Poisson as well as the Normal, Gamma, and Wald with known dispersions. Natural exponential distributions were investigated by Morris (1982), and subsequently generalized by Jorgensen (1987, 1997) to natural exponential dispersion (NED) distributions containing an additional dispersion parameter, φ. Thus for a NED-based GLM with constant dispersion, y_i ~ NED(μ_i, φ), where $\mu_i = g(x_i\beta)$. Important examples include the Normal(μ_i, φ), Gamma(μ_i, φ), and Wald(μ_i, φ) with unknown dispersions, as well as an extension of the Poisson containing an unknown dispersion, PoisED(μ_i, φ).

In what follows we suppress the subscript *i* with the understanding that y_i ~ NED(μ_i, φ) for a NED with constant dispersion, while y_i ~ NED(μ_i, φ_i) for a NED with non-constant dispersion. Equation (1) gives the log-likelihood for a single observation y ~ NED(μ,φ). Here d(y,μ) is the deviance and a(y,φ) is a normalizing term free from μ.

$$l(\mu,\varphi)=a(y,\varphi)-\frac{d(y,\mu)}{2\varphi} \tag{1}$$

When a(y,φ) is not in closed form, equation (1) can be approximated by equation (2). Here v(y) is the variance function evaluated at y, such as $v(y) = y^2$ for the Gamma. In this paper we use exact likelihoods (1) for the Normal, Gamma, and Wald, and an approximate likelihood (2) for the PoisED. The approximation allows us to treat the PoisED as a continuous NED with variance proportional to the mean.

$$l(\mu,\varphi)=-\frac{1}{2}\log\left(2\pi\varphi v(y)\right)-\frac{d(y,\mu)}{2\varphi} \tag{2}$$

NEDs maintain two central properties of the Normal, namely separability and robustness. As used here, separable likelihoods enable means to be estimated separately from dispersions, and robust models provide consistent maximum likelihood estimates for means even under misspecified distributions.

In addition to the canonical NEDs noted above, we encounter two types of transformed NEDs which are neither separable nor robust in the above senses. The first type are power transformations of NEDs (eg y ~ $(\text{Gamma})^2$) corresponding to inverse power transformations of the response (eg $y^{1/2}$ ~ Gamma). The second type are NEDs for which the dispersion depends on the mean, or equivalently, for which the variance function, v(μ), is noncanonical. For example, the canonical setup for the Gamma is {y ~ Gamma(μ,φ) with $v(\mu) = \mu^2$}, while a noncanonical setup is {y ~ Gamma(μ, $\varphi = \psi/\mu^2$) with v(μ) = 1}.

3. Method

Our generalization of the Box-Cox method allows for transformations to models based on nonlinear mean functions and non-Normal distributions. We consider transformations of Normal, PoisED, Gamma, and Wald distributions, with canonical variances proportional to 1, μ, μ^2, and μ^3, respectively. Transformations of y are compared by AIC (Akaike (1974)).

Box and Cox used an extended power family incorporating log(y) as $\lambda = 0$ in a limiting sense. This extension does not apply to non-Normal NEDs, and instead we use the simple power family, y^λ, which excludes log(y). While the log transformation is still of interest for non-Normal NEDs (provided $y > 1$), the transformation is distinct from the power family. In terms of the AIC profile, AIC(log) is an isolated point off the smooth AIC(λ) profile, which contrasts with the Normal case where AIC(log) is a point on the profile at $\lambda = 0$. We emphasize the new role played by log(y) because of the striking result that all AIC profiles in our study coincide at $\lambda = 0$, despite the fact that LogNormal, LogPoisED, LogGamma, and LogWald distributions are certainly not the same thing.

We now consider log-likelihoods for y which correspond to the transformation $z = y^\lambda$, where $y^\lambda \sim \text{NED}(\mu,\varphi)$. The log-likelihood for $z = y^\lambda$, $l_z(\mu,\varphi; z)$, is given by equation (1) or equation (2) above. It follows that the log-likelihood for $y = z^{1/\lambda}$, $l(\mu,\varphi; y)$, is obtained by substituting y^λ for z in $l_z(\mu,\varphi; z)$ and adding a log-jacobian term based on dz/dy, as shown in equation (3).

$$l(\mu,\varphi;y) = l_z(\mu,\varphi;y^\lambda) + \log\left(\frac{\lambda}{y^{1-\lambda}}\right) \tag{3}$$

We use nomenclature in keeping with the long-established relationship between the Normal and LogNormal. In this case, $z = \log(y) \sim$ Normal, and $y = \exp(z) \sim \exp(\text{Normal})$ is said to follow a LogNormal distribution, with log-likelihood $l(\mu,\varphi; y)$ given by equation (4). Similarly, if $z = y^{1/2} \sim$ Gamma, then $y = z^2 \sim (\text{Gamma})^2$ is said to follow a SqrtGamma distribution, and more generally, if $z = y^\lambda \sim$ NED, then $y = z^{1/\lambda} \sim (\text{NED})^{1/\lambda}$ is said to follow a λ-NED distribution.

$$l(\mu,\varphi;y) = l_z(\mu,\varphi;\log(y)) + \log\left(\frac{1}{y}\right) \tag{4}$$

The log-jacobian term contains no parameters, so estimates from $y^\lambda \sim$ NED and $y \sim (\text{NED})^{1/\lambda}$ are identical even though their likelihoods differ. For Box-Cox purposes, however, this distinction is critical because models can be compared by AIC only if their responses are identical. For example, $y \sim$ Gamma and $y^{1/2} \sim$ Gamma cannot be compared directly, but only via comparison of $y \sim$ Gamma and $y \sim (\text{Gamma})^2$.

4. Example 1: AIC(λ) profiles: Four NED distributions

In this section we apply the generalized Box-Cox method using blink rate (BR) as the response. The scatter plot of BR against group (0 = Healthy, 1 = Dryeye) is shown in Figure 1. For each NED, we fit a sequence of models $y \sim (\text{NED})^{1/\lambda}$ corresponding to transformations $y^\lambda \sim$ NED over a grid of λ in $[-1, 1]$. For reasons noted previously, $\lambda = 0$ was excluded, except in the Normal case where $\lambda = 0$ represents log(y). The resulting AIC profiles are depicted in Figure 2.

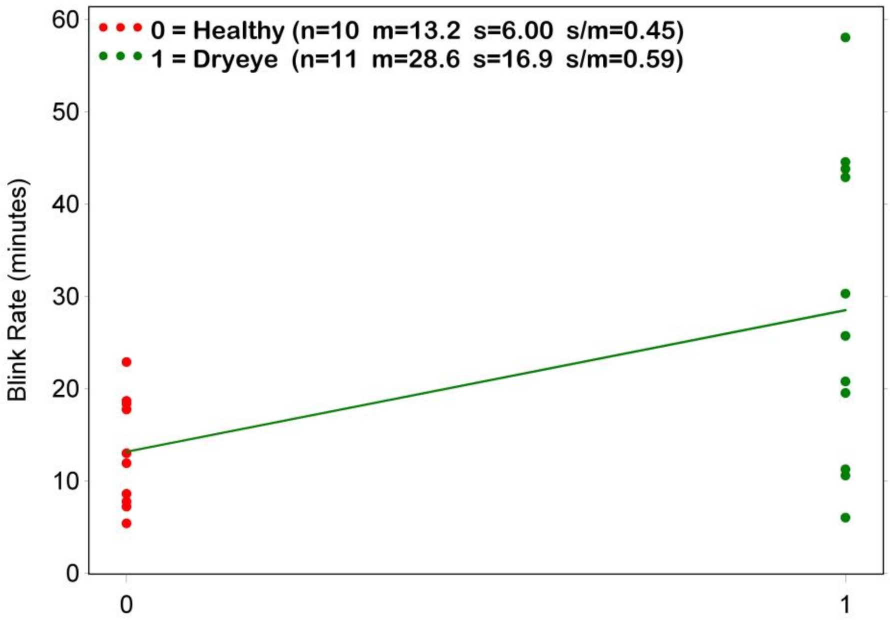

Figure 1. Scatter plots for Blink Rate against group. Groups comprise 10 Healthy (x=0) subjects and 11 Dryeye(x=1) subjects

Figure 1 suggests two mildly right-skewed distributions with variances increasing with means. As a result, the Normal-based Box-Cox method favors λ = 0 and use of log(y) ~ Normal(μ,φ), where μ = 2.5 + 0.67x. The generalized Box-Cox method finds little to choose between log(y) ~ Normal, $y^{1/4}$ ~ PoisED, y ~ Gamma, and y ~ Wald, with AICs of 162.1, 161.9, 160.7, and 160.5, respectively. However, we are inclined to remain on the original scale and select y ~ Gamma(μ,φ), where μ = 13.2 + 15.3x.

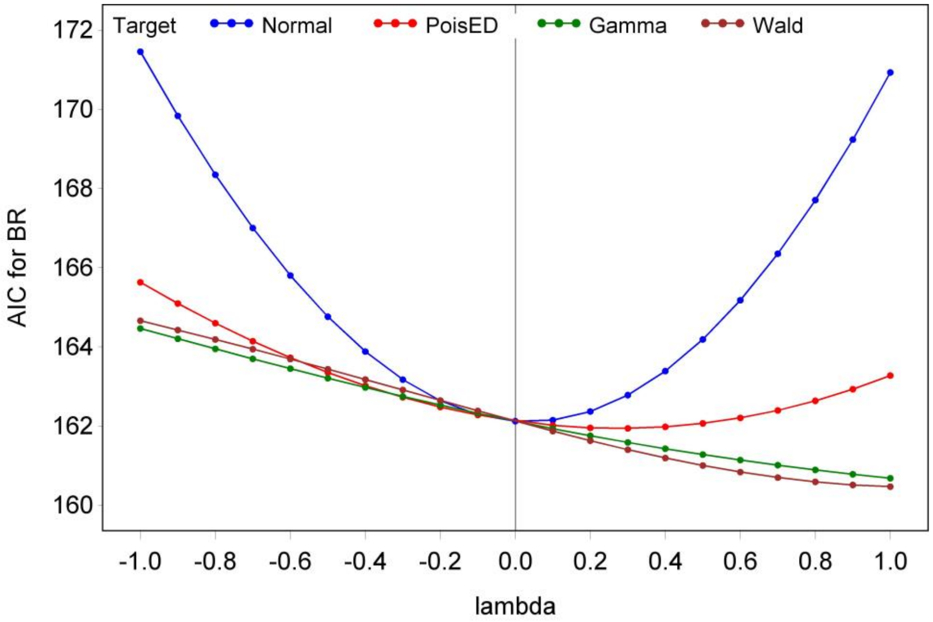

Figure 2. AIC(λ) vs λ for BRλ following a GLM based on the Normal, PoisED, Gamma, and Wald

5. Example 2: AIC(λ) profiles: Four Gamma(p) distributions

We now illustrate a variant of the above for the Gamma with various noncanonical variance functions. This illustrates the use of the Box-Cox method for a class of distributions wider than the standard NEDs, and also sheds light on our main NED application above.

We consider Gamma models with variances proportional to 1, μ, μ^2, and μ^3, and refer to the new distributions as Gamma0, Gamma1, Gamma2, and Gamma3, respectively. These variance functions imply non-constant dispersions of ψ/μ^2, ψ/μ, ψ, and ψμ respectively (ψ is constant, but μ depends on *i*), and thus GLMs based on these distributions are not robust, with the exception of Gamma2 (the standard Gamma).

The four AIC profiles are depicted in Figure 3. For λ > 0, Figure 3 maintains the order of Figure 2 in terms of variance functions. Thus Gamma models with v(μ) = 1 and $v(\mu) = \mu^3$ are the worst and best models in Figure 3 in a similar way as Normal and Wald models are the worst and best models in Figure 2. This reflects the fact that, given sufficiently different means, the fit of a NED is strongly influenced by its variance function. But fit also depends on distributional shape, and the fact that AIC profiles in Figure 3 for the Gamma2 and Gamma3 are further apart (160.7 – 159.5 = 1.2 at λ = 1) than AIC profiles for the Gamma and Wald in Figure 2 (160.7 – 160.5 = 0.2 at λ = 1) suggests that BR has a Wald-like variance function and a Gamma-like shape.

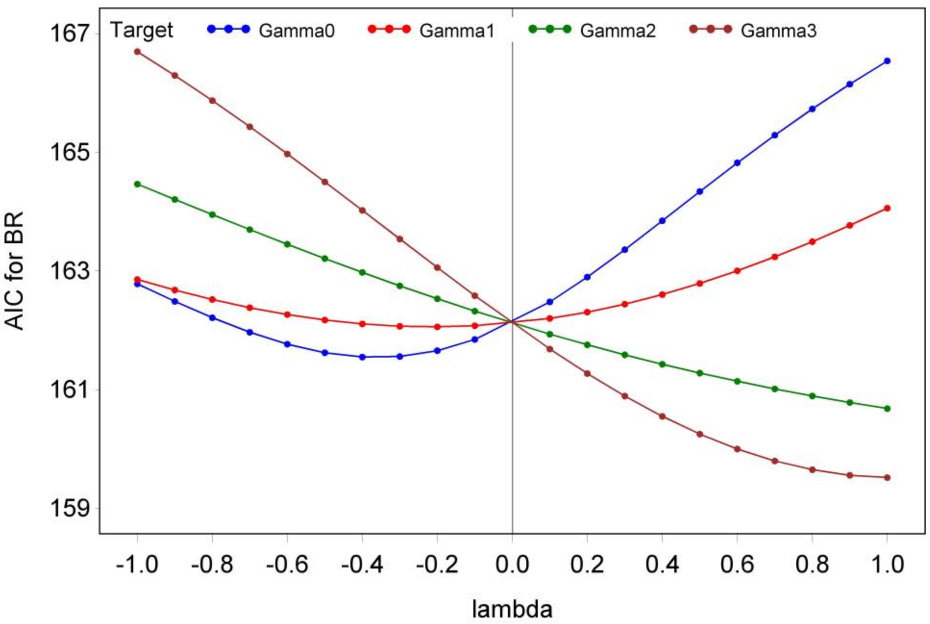

Figure 3. AIC(λ) vs λ for BR^λ following a GLM based on Gamma(p), p = 0,1,2,3

6. Example 3: Choosing BR vs IBI on original scale

In this section we discuss a problem where the choice of response involves the choice of transformation. The particular setting is for a study involving “rates” and “paces”, in our case blink rate, BR = N(i)/T(i), and interblink interval, IBI = T(i)/N(i). Both are directly interpretable (as blinks per minute, and minutes per blink), and are equivalent as data but not as responses. Scatter plots of (a) BR against group, and (b) IBI against group are shown in Figure 4.

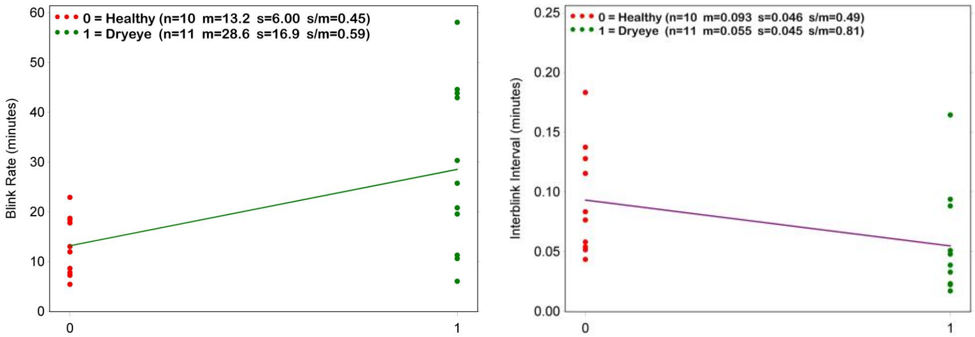

Figure 4. Scatter plots for 10 Healthy (x=0) and 11 Dryeye(x=1) subjects: (a)Blink Rate (BR), (b)Interblink Interval (IBI)

From the viewpoint of the Box-Cox method, special features of this problem are (a) only two transformations are considered, y and z = 1/y, (b) both y and z are directly interpretable, and (c) the intention is to analyze either y or z on the raw scale. For simplicity we focus on the Gamma distribution. Figure 5 shows the AIC profile for IBI (purple line, right axis) along with the AIC profile for BR (green line, left axis) shown previously.

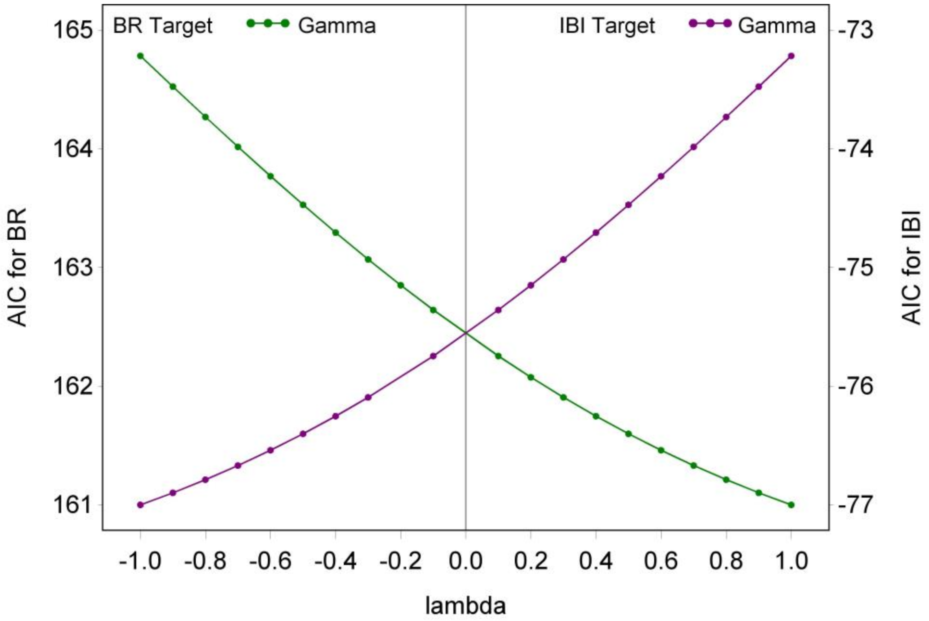

Figure 5. AIC(λ) vs λ for BR$^{\lambda}$ and IBI$^{\lambda}$ following a GLM based on the Gamma

A direct comparison of BR (AIC = 160.7) and IBI (AIC = -73.4) is meaningless because the responses are different. However, a legitimate comparison can be made by comparing BR for $\lambda = -1$ and $\lambda = 1$, which corresponds to comparing IBI and BR as responses. The difference in AICs is 164.5 – 160.7 = 3.8 units in favor of BR. The same result obtains if IBI is used as the reference response. In this case, a comparison of IBI for $\lambda = 1$ and $\lambda = -1$ corresponds to a comparison of IBI and BR as responses, again giving a difference in AICs of -73.4 – (-77.2) = 3.8 units in favor of BR. These and other details of the group comparison are provided in Table 1.

			Dryeye	Healthy				
Response	Model	n	mean1	mean0	diff	se	p-value	AIC
BR	Gamma	21	28.55	13.21	15.34	5.17	0.003	160.7
BR	$(\text{Gamma})^{-1}$	21	0.055	0.093	-0.038	0.020	0.048	164.5
IBI	Gamma	21	0.055	0.093	-0.038	0.020	0.048	-73.4
IBI	$(\text{Gamma})^{-1}$	21	28.55	13.21	15.34	5.17	0.003	-77.2

Table 1. Dryeye vs Healthy: BR and IBI assuming Gamma and Inverse Gamma models

7. Conclusion

Box and Cox introduced a useful method based on transforming the response variable of a regression model to follow a Normal linear model with constant variance. The usefulness and popularity of the method stemmed from the accessibility of the Normal linear model in years following 1964. Ryan and Woodall (2005) list Box and Cox (1964) as the 19th most-cited paper in statistics in the post-war era (approximately 1945 through 2004).

The main aim of our paper has been to generalize the Box-Cox method while retaining its robustness (Example 1). To this end, we generalize the target model from a linear model based on a Normal with constant variance to a GLM based on a NED with constant dispersion. As a particular example, we applied the method to the choice between reciprocal responses (y and 1/y) in a context in which both are far from Normal (Example 3). This example, a direct generalization of the first example in Box and Cox (1964), actually motivated our paper.

In addition, we explored an extension of Box-Cox methods to a larger class of distributions, namely GLMs based on NEDs with noncanonical variance functions (Example 2). This class of GLMs is of some interest in its on right.

References

Akaike H. 1974. A new look at the statistical model identification, IEEE Transactions on Automatic Control.

Box GEP and Cox DR. 1964. An analysis of transformations. JRSS(B).

Johnston PR et al. 2013. The interblink interval in normal and dry eye subjects. Clinical Ophthalmology.

Jorgensen B. 1987. Exponential dispersion models. JRSS(B).

Jorgensen B. 1997. The Theory of Dispersion Models. Chapman and Hall.

McCullagh P and Nelder JA. 1983 (edition 1), 1989 (edition 2). Generalized linear models. Chapman and Hall.

Morris C. 1982. Natural exponential families with quadratic variance functions. Annals of Statistics.

Nelder JA and Wedderburn RWM. 1972. Generalized linear models. JRSS(A).

Nelson and Granger. 1979. Experience with using the Box-Cox transformation when forecasting economic time series. Journal of Econometrics.

Ryan TP and Woodall WH. 2005. The Most-Cited Statistical Papers. Journal of Applied Statistics.

SAS Institute Inc. 2011. SAS/STAT® 9.3 User's Guide.