Transformed Generalized Linear Models

Gauss M. Cordeiro[a,*] Marinho G. de Andrade[b]

[a]*Departamento de Estatística e Informática,
Universidade Federal Rural of Pernambuco,
52171-900, Recife, PE, Brazil*

[b]*Departamento de Matematica Aplicada e Estatística,
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo,
C.P.668, 13560-970, São Carlos, SP, Brazil*

Abstract

The estimation of data transformation is very useful to yield response variables satisfying closely a normal theory linear model. Generalized linear models enable the fitting of models to a wide range of data types. These models are based on exponential dispersion models. We propose a new class of transformed generalized linear models to extend the Box and Cox models and the generalized linear models. We use the generalized linear model framework to fit these models and discuss maximum likelihood estimation and inference. We give a simple formula to estimate the parameter that index the transformation of the response variable for a subclass of models. We also give a simple formula to estimate the rth moment of the original dependent variable. We explore the possibility of using these models to time series data to extend the generalized autoregressive moving average models discussed by Benjamin et al. (2003). The usefulness of these models is illustrated in a simulation study and in applications to three real data sets.

Key words: Dispersion parameter, Exponential family, Family of transformations, Generalized linear model, Generalized ARMA model, Likelihood ratio, Profile likelihood.

1 Introduction

The use of transformations in regression analysis is very common and may be helpful when the original model does not satisfy the usual assumptions. The power transformation family proposed by Box and Cox (1964) is often used for transforming to a normal linear model. The Box and Cox models add useful tools for applied statistics, pertaining to the separable aspects of

[*] Corresponding Author, *Email address*: gausscordeiro@uol.com.br

variance homogeneity, model additivity and normal distribution. This article considers the problem of extending the Box-Cox models to a non-Gaussian framework with heteroskedasticity and a possible non-linear function of regression parameters. We include some well-known models such as the Box and Cox (1964) and the generalized linear models, first introduced by Nelder and Wedderburn (1972), as special cases. We work with a general parametric family of transformations from the response variable Y to

$$Y^{(\lambda)} = \Lambda(Y, \lambda), \tag{1}$$

where λ is a scalar parameter defining a particular transformation. We further assume that for each λ, $Y^{(\lambda)}$ is a monotonic function of Y. Usually, we consider the Box and Cox (1964) power transformation, $Y^{(\lambda)} = (Y^\lambda - 1)/\lambda$ when $\lambda \neq 0$ or $Y^{(\lambda)} = \log(Y)$ when $\lambda = 0$, and assume that there is a λ value for the response variable such that $Y^{(\lambda)}$ follows a linear regression model $\mu = X\beta$ with normal error and constant variance. In practice this will rarely be true. It is frequently assumed in connection with the power transformation that Y is positive; if Y could be negative many values of λ would be clearly inadmissible. Manly (1976) proposed the exponential transformation to be used with negative $Y's$ of the form: $Y^{(\lambda)} = (e^{\lambda Y} - 1)/\lambda$ when $\lambda \neq 0$ or $Y^{(\lambda)} = Y$ when $\lambda = 0$. This transformation seems to be effective at turning skew unimodal distributions into nearly symmetric normal-like distributions. Alternative transformations to the power transformation are reviewed by Sakia (1992).

The Box-Cox type of power transformations have generated a great deal of interest, both in theoretical work and in practical applications. Inference procedures for the regression coefficients and transformation parameter under this model setting have been studied extensively. Clearly not all data could be power-transformed to normal. While Draper and Cox (1969) have shown that the estimation of λ is fairly robust to non-normality as long as the variable has a reasonably symmetric distribution, this may not be the case when skewness is encoutered. They studied this problem and conclude in one example that if the raw data follow an exponential distribution, values of λ close to its estimate will, in fact, yield transformed distributions which are, in fact, Weibull but they look very like symmetric distributions. Then, power transformations can be useful even in situations where they cannot produce normality exactly. Bickel and Doksum (1981) studied the joint estimation of (λ, β) and proved that the asymptotic marginal (unconditional) variance of the maximum likelihood estimate (MLE) of β could be inflated by a very large factor over the conditional variance for fixed λ. Although there does not appear to be any definite result, most researchers agree that while there is an effect on not knowing the true value of λ, its cost may not be large enough to discredit the conventional application based on conditioning. Lawrence (1987) gave an expression for the estimated variance of the MLE of λ.

Guerrero and Johnson (1982) suggested a power transformation applied to the odds ratio to generalize the logistic model. For continuous proportions their transformation is defined by using $log\{Y/(1-Y)\}$ in place of Y in the Box and Cox transformation. Another transformation proposed by Aranda-Ordaz (1981) for continuous proportions is defined by

$$Y^{(\lambda)} = \frac{2\{Y^{\lambda} - (1-Y)^{\lambda}\}}{\lambda\{Y^{\lambda} + (1-Y)^{\lambda}\}},$$

which reduces to the logistic transformation when $\lambda = 0$ and to the linear transformation when $\lambda = 1$.

Generalized linear models (GLMs) are based on distributions that are exponential dispersion models, discussed in great detail in Jorgensen (1997). GLMs extend the normal theory linear model, include a general algorithm for computing MLEs and enable the fitting of different types of models to a wide range of data types. Although the power transformation has been widely used, one thing is clear: that seldom does this transformation fulfill the basic assumptions of linearity, normality and homoscedasticity simultaneously. This transformation has found more practical utility in the empirical determination of functional relationships in a variety of fields, especially in econometrics. In view of this, we work with a general family of monotonic transformations (1) (which is data based) and combine the idea of transforming the response variable with the framework of the GLM.

Transformed generalized linear models (TGLMs) assume that there exists some value of λ such that the transformed random variables $Y_1^{(\lambda)}$, ..., $Y_n^{(\lambda)}$ can be treated as independently distributed following the basic assumptions of the GLMs. The exactness of these assumptions may not be important in the applications. We therefore consider the possibility of a heteroscedastic variance, a more general family for the distribution of the response variable and a nonlinear function for the regression parameters. The optimal value of λ may lead to a more nearly GLM fitted to the transformed data. The strong assumptions within the Box-Cox models that the power transformation yields a more nearly linear model, stabilizing the error variance with a normally distributed error, are then relaxed.

In Section 2 we define the TGLMs and give a summary of key results. The maximum likelihood estimation is discussed in Section 3 and some special models are considered in Section 4. In Section 5 we discuss the model inference. In Section 6 we consider the problem of extending our approach to deal with non-Gaussian time series models by proposing an extension of the generalized autoregressive moving average (GARMA) models defined by Benjamin et al. (2003). In Section 7, we present simulation studies to illustrate the methodology of fitting the TGLMs. In Section 8, we analyze three real data sets. The article ends with some concluding remarks in Section 9.

2 Model Definition

Let $y = (y_1, \ldots, y_n)^T$ be the vector of observations and by using (1) we obtain the transformed observations $y^{(\lambda)} = (y_1^{(\lambda)}, \ldots, y_n^{(\lambda)})^T$. Assume the transformed random variables $Y_1^{(\lambda)}, \ldots, Y_n^{(\lambda)}$ in $Y^{(\lambda)}$ be independent and each $Y_i^{(\lambda)}$ following a continuous exponential dispersion model with probability density function (with respect to Lebesgue measure) of the form

$$\pi(y_i^{(\lambda)}; \theta_i, \phi) = \exp\left[\frac{1}{\phi}\left\{y_i^{(\lambda)}\theta_i - b(\theta_i)\right\} + c(y_i^{(\lambda)}, \phi)\right], \tag{2}$$

where $b(x)$ and $c(x, \phi)$ are known appropriate functions. Some of the most useful statistical distributions are within form (2). The parameter ϕ is called the dispersion parameter and is the same for all observations, although possibly unknown. The idea of an exponential dispersion model goes back to Tweedie (1947), who noticed many of the important mathematical properties and special cases of exponential dispersion models. A systematic study of the properties of the exponential dispersion models was presented by Jorgensen (1997).

We do not consider members of (2) which are discrete distributions such as Poisson, binomial, negative binomial and compound Poisson distributions, but we can work with continuous proportions. The mean and variance of $Y_i^{(\lambda)}$ are, respectively, $E(Y_i^{(\lambda)}) = \mu_i = db(\theta_i)/d\theta_i$ and $Var(Y_i^{(\lambda)}) = \phi\, V_i$, where $V = V(\mu) = d\mu/d\theta$ is the variance function. The parameter $\theta = \int V^{-1} d\mu = q(\mu)$ is a known one-to-one function of μ. The exponential dispersion model (2) is uniquely characterized by its variance function V, which plays a key role in the study of its mathematical properties and in estimation. For gamma models, the dispersion parameter ϕ is the reciprocal of the index; for normal and inverse Gaussian models, ϕ is the variance and $Var(Y_i^{(\lambda)})/E(Y_i^{(\lambda)})^3$, respectively. These are the most important continuous models in (2).

Our aim is to make a parametric transformation $Y^{(\lambda)}$ of a response variable Y so that $Y^{(\lambda)}$ satisfies the usual assumptions of the GLMs. Our generalized form (1) is used to determine the specific form within a particular class of transformation functions which is optimal by reference to a maximum likelihood criterion. We define the TGLM by the families of transformations (1) and distributions (2) and the systematic component

$$g(\mu) = \eta = X\beta, \tag{3}$$

where $g(\cdot)$ is a known one-to-one continuously twice-differentiable function, X is a specified $n \times p$ model matrix of full rank $p < n$ and $\beta = (\beta_1, \ldots, \beta_p)^T$ is a set of unknown linear parameters to be estimated. The link function is assumed to be monotonic and differentiable. The parameters of the TGLMs are then the vector β and the scalars ϕ and λ. We have $p + 2$ parameters to be estimated. The TGLM formalizes the notion that a certain form of the GLM would be appropriate for some transformation of the response, where the transformation necessary to achieve the GLM form is not known before the data are collected. The aim of the transformation (1) is to ensure that the usual assumptions (2) and (3) for the GLMs hold for the transformed variable $Y^{(\lambda)}$. To fit the transformed gamma and inverse Gaussian models to some types of data, it is sometimes necessary to consider λ within some limit values to guarantee that $y^{(\lambda)}$ is positive.

We may thus summarize the TGLMs in the form of three components of structural importance: a general family of transformations, a more general form for the distribution of the transformed response and a possible nonlinear link function for the regression parameters. TGLMs are then an extension of the GLMs and have some important special cases: the Box and Cox models for which the transformation (1) is the Box and Cox power family, the distribution in (2) is normal and the systematic component is $\mu = \eta = X\beta$; the classical GLMs for which the transformation function is independent of λ given by $\Lambda(Y, \lambda) = Y$; and the power generalized linear models (PGLMs) defined here when (1) is the Box-Cox transformation or the simple power transformation $Y^{(\lambda)} = Y^{\lambda}$ in addition to the equations (2) and (3).

The function $c(x, \phi)$ plays a fundamental role in the process of fitting the TGLMs. It does not have simple closed-form expressions for several exponential dispersion models; see, the generalized hyperbolic secant (GHS) model and the continuous exponential dispersion models with power variance functions discussed by Jorgensen (1997). However, when (2) is a two-parameter full exponential model with canonical parameters $1/\phi$ and θ/ϕ, $c(x, \phi)$ has the following decomposition

$$c(x, \phi) = \frac{1}{\phi}\, a(x) + d(\phi) + d_1(x). \tag{4}$$

Equation (4) holds for normal, gamma and inverse Gaussian models but does not hold in general for exponential dispersion models.

3 Model Fitting

We observe the model matrix X and the raw data y and assume that the transformed response $Y^{(\lambda)}$ for some unknown transformation parameter λ in (1) satisfies the usual assumptions (2) and (3) for the GLMs. The model parameters are then (λ, β, ϕ). The main objective in the analysis of the TGLMs is to make likelihood inference on the model parameters. The maximum likelihood method is used since it is conceptually easy although the profile log-likelihood for λ could be difficult to compute in some cases.

Let $J(\lambda, y)$ be the Jacobian of the transformation from Y to $Y^{(\lambda)}$. The log-likelihood for the model parameters can be expressed in terms of the vector of transformed observations $y^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})^T$ by

$$l(\beta, \phi, \lambda) = \frac{1}{\phi} \sum_{i=1}^{n} \left\{ y_i^{(\lambda)} \theta_i - b(\theta_i) \right\} + \sum_{i=1}^{n} \left[c(y_i^{(\lambda)}, \phi) + \log \left\{ J(\lambda, y_i) \right\} \right], \tag{5}$$

where

$$J(\lambda, y_i) = \left| \frac{d\Lambda(y_i, \lambda)}{dy_i} \right|.$$

For maximizing the log-likelihood (5), we assume first that λ is fixed and then obtain the likelihood equations for estimating β and ϕ. The vector β can be estimated without knowledge of ϕ. Let $\widehat{\beta}^{(\lambda)}, \widehat{\eta}^{(\lambda)} = X\ \widehat{\beta}^{(\lambda)}$, $\widehat{\mu}^{(\lambda)} = g^{-1}(\widehat{\eta}^{(\lambda)})$ and $\widehat{\phi}^{(\lambda)}$ be the MLEs of β, η, μ and ϕ, respectively, for given λ. The estimate $\widehat{\beta}^{(\lambda)}$ can be obtained easily from the fitting of the GLM (2)-(3) to $y^{(\lambda)}$ by iteratively reweighted least squares. The iteration is

$$\widehat{\beta}^{(\lambda)} = (X^T\widehat{W}^{(\lambda)}X)^{-1}X^T\widehat{W}^{(\lambda)}\widehat{z}^{(\lambda)}, \tag{6}$$

where $W = diag\{w_1, \ldots, w_n\}$ is a diagonal matrix with $w_i = V_i^{-1}(d\mu_i/d\eta_i)^2$ and $z^{(\lambda)} = (z_1^{(\lambda)}, \ldots, z_n^{(\lambda)})^T$ is the working vector with components

$$z_i^{(\lambda)} = \eta_i^{(\lambda)} + (y_i^{(\lambda)} - \mu_i^{(\lambda)})\left(\frac{d\eta_i}{d\mu_i}\right)^{(\lambda)}.$$

An initial approximation $\widehat{\beta}^{(\lambda)(1)}$ is used to evaluate $z^{(\lambda)}$ and $W^{(\lambda)}$ from which equation (6) can be used to obtain the next estimate $\widehat{\beta}^{(\lambda)(2)}$. This new value can update $z^{(\lambda)}$ and $W^{(\lambda)}$, and so the iterations continue until convergence is observed.

Estimation of the dispersion parameter ϕ is a more difficult problem than the estimation of β and the complexity depends on the functional form of $c(x, \phi)$. In principle, ϕ could also be estimated by maximum likelihood although there may be practical difficulties associated with this for some members of (2). The MLE $\widehat{\phi}^{(\lambda)}$ of ϕ for fixed λ is

$$\widehat{\phi}^2 \sum_{i=1}^{n} \frac{dc(y_i^{(\lambda)}, \phi)}{d\phi}\bigg|_{\phi=\widehat{\phi}} = \sum_{i=1}^{n}\left\{y_i^{(\lambda)}\widehat{\theta}_i^{(\lambda)} - b(\widehat{\theta}_i^{(\lambda)})\right\}, \tag{7}$$

where $\widehat{\theta}^{(\lambda)} = q(g^{-1}(X\widehat{\beta}^{(\lambda)}))$.

Given the variance function $V(x)$ we can easily obtain $q(x) = \int V(x)^{-1}dx$ and $b(x) = \int q^{-1}(x)dx$, and then the deviance $D^{(\lambda)}$, conditioning on λ, of the TGLM defined as twice the difference of the maximum log-likelihood corresponding to the saturated model and the maximum of the log-likelihood of the model under investigation. This statistic for given λ depends only on the data and not on any unknown parameters and can be written in the form

$$D^{(\lambda)} = 2\sum_{i=1}^{n} D_i^{(\lambda)}(y_i^{(\lambda)}, \widehat{\mu_i}^{(\lambda)}), \tag{8}$$

where

$$D_i^{(\lambda)}(y_i^{(\lambda)}, \widehat{\mu_i}^{(\lambda)}) = e(y_i^{(\lambda)}) - \left\{y_i^{(\lambda)}q(\widehat{\mu}_i^{(\lambda)}) - b(q(\widehat{\mu}_i^{(\lambda)}))\right\}, \tag{9}$$

with $e(x) = x\,q(x) - b(q(x))$, is the deviance component for the *ith* observation. Examples of deviance functions for some exponential dispersion models are given by Jorgensen (1997).

The MLE $\widehat{\phi}^{(\lambda)}$ is a function of the deviance (8) of the model. Using (7) we obtain

$$\widehat{\phi}^{(\lambda)2} \sum_{i=1}^{n} \frac{dc(y_i^{(\lambda)}, \phi)}{d\phi}\bigg|_{\phi=\widehat{\phi}} = \sum_{i=1}^{n} e(y_i^{(\lambda)}) - \frac{D^{(\lambda)}}{2}. \tag{10}$$

Equation (10) is in general nonlinear except for the normal and inverse Gaussian models and requires the use of a nonlinear numerical algorithm for estimating ϕ. Substituting the MLEs $\widehat{\beta}^{(\lambda)}$ and $\widehat{\phi}^{(\lambda)}$ in (5) yields the profile log-likelihood for λ

$$l_P(\lambda) = \frac{1}{\widehat{\phi}^{(\lambda)}} \sum_{i=1}^{n} \left\{ y_i^{(\lambda)} \widehat{\theta}_i^{(\lambda)} - b(\widehat{\theta}_i^{(\lambda)}) \right\} + \sum_{i=1}^{n} \left[c(y_i^{(\lambda)}, \widehat{\phi}^{(\lambda)}) + \log \left\{ J(\lambda, y_i) \right\} \right]. \tag{11}$$

The resulting expression $l_P(\lambda)$ in terms of the deviance of the TGLM is

$$l_P(\lambda) = \frac{1}{\widehat{\phi}^{(\lambda)}} \left\{ \sum_{i=1}^{n} e(y_i^{(\lambda)}) - \frac{D^{(\lambda)}}{2} \right\} + \sum_{i=1}^{n} \left[c(y_i^{(\lambda)}, \widehat{\phi}^{(\lambda)}) + \log \left\{ J(\lambda, y_i) \right\} \right]. \tag{12}$$

To operationalize equations (10) and (12) for any TGLM we need the functions $e(x)$ and $c(x, \phi)$, the deviance $D^{(\lambda)}$ and the Jacobian. The plot of the profile log-likelihood $l_P(\lambda)$ in (12) against λ for a trial series of values determines numerically the value of the MLE $\widehat{\lambda}$. Once the MLE $\widehat{\lambda}$ is obtained from the plot, it can be substituted in the algorithm (6) and in the equation (10) to produce the unconditional estimates $\widehat{\beta} = \widehat{\beta}^{(\hat{\lambda})}$ and $\widehat{\phi} = \widehat{\phi}^{(\hat{\lambda})}$. The process of estimating β, ϕ and λ can be carried out by standard statistical software such as MATLAB, S-PLUS, R and SAS.

For some exponential dispersion models, the MLE of the dispersion parameter in (10) could be very complicated and we can use a method of moments estimator to obtain a consistent estimate of ϕ directly from the MLE $\widehat{\beta}^{(\lambda)}$. We have the Pearson estimate of ϕ

$$\tilde{\phi}^{(\lambda)} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i^{(\lambda)} - \widehat{\mu_i}^{(\lambda)})^2}{V(\widehat{\mu_i}^{(\lambda)})}.$$

This estimate could be inserted into (12) to produce a modified profile log-likelihood for λ which is maximized in the usual way. Another simple alternative estimate for ϕ which could be used in (12) is the deviance based estimate obtained from $\tilde{\phi}^{(\lambda)} = D^{(\lambda)}/(n-p)$ on the grounds that the expected value of $D^{(\lambda)}/\phi$ is approximately $n-p$. For transformed gamma models, the MLE of ϕ should be preferred.

For two-parameter full exponential family distributions, the decomposition of the function $c(x, \phi)$ in (4) and (10) yields the equation for $\widehat{\phi}^{(\lambda)}$

$$n\, \widehat{\phi}^{(\lambda)(2)} d'(\widehat{\phi}^{(\lambda)}) = \sum_{i=1}^{n} t(y_i^{(\lambda)}) - \frac{D^{(\lambda)}}{2}, \tag{13}$$

where $t(x) = x\,q(x) - b(q(x)) + a(x)$. Now using (7) in (11), gives

$$l_P(\lambda) = n\,v(\widehat{\phi}^{(\lambda)}) + \sum_{i=1}^{n} \left[d_1(y_i^{(\lambda)}) + \log\{J(\lambda, y_i)\} \right], \tag{14}$$

where $v(x) = xd'(x) + d(x)$. It is very easy to work with the equations (13) and (14). In Table 1 we give the functions $d(x)$, $t(x)$, $v(x)$ and $d_1(x)$ for some TGLMs which enable us to compute $\widehat{\phi}^{(\lambda)}$ in (13) and the profile log-likelihood (14). For transformed normal models, (14) is identical to the equation (8) given by Box and Cox (1964) and can be viewed as a generalization of this equation for some other continuous models.

Table 1: Some Special Transformed Models

Model	$d(x)$	$t(x)$	$v(x)$	$d_1(x)$
Normal	$-\frac{1}{2}\log(x)$	0	$-\frac{1}{2}\{1+\log(x)\}$	$-\frac{1}{2}\log(2\pi)$
Gamma	$-\frac{\log(x)}{x} - \log\Gamma(\frac{1}{x})$	-1	$\frac{1}{x}\Psi(\frac{1}{x}) - \frac{1}{x} - \log\Gamma(\frac{1}{x})$	$-\log(x)$
I.G.	$-\frac{1}{2}\log(x)$	0	$-\frac{1}{2}\{1+\log(x)\}$	$-\frac{1}{2}\log(2\pi x^3)$

We now estimate the mean of the untransformed dependent variable Y_i by using a method analogous to the small-θ method given in Draper and Cox (1969). When $\lambda \neq 0$ we can write

$$Y = (1+\lambda\mu)^{1/\lambda}\,\{1 + \theta(Y^\lambda - \mu)\}^{1/\lambda},$$

where $\theta = \frac{\lambda}{1+\lambda\mu}$. From the binomial expansion we obtain

$$\{1 + \theta(Y^\lambda - \mu)\}^{1/\lambda} = 1 + \sum_{i=1}^{\infty} \frac{\theta^i}{i!} \prod_{j=0}^{i-1} (\frac{1}{\lambda} - j)(Y^\lambda - \mu)^i.$$

We also have

$$E(Y) = (1+\lambda\mu)^{1/\lambda} \left\{ 1 + \sum_{i=2}^{\infty} \frac{a_i\,\mu_i}{i!\,(1+\lambda\mu)^i} \right\}, \tag{15}$$

where $a_i = \prod_{j=0}^{i-1}(1 - j\lambda)$ and μ_i is the ith central moment of $Y^{(\lambda)}$. The central moments of the exponential dispersion model are easily obtained from the recurrence relation of their cumulants. We have $\mu_2 = \phi V$, $\mu_3 = \phi^2 V\,V^{(1)}$, $\mu_4 = \phi^2(\phi V^{(2)} + 3)V^{(2)} + \phi^3 V\,V^{(1)2}$, $\mu_5 = \phi^4\{V^3 V^{(3)} + 4V^2 V^{(1)} V^{(2)} + V V^{(1)3} + (10\phi)^{-1} V^2 V^{(1)}\}$, and so on, where $V^{(r)} = d^r V/d\mu^r$. Equation (15) generalizes the expansion given by Pankratz and Dudley (1987) for the non-biasing factor obtained when λ^{-1} is a positive integer and the transformed data is normal ($V = 1$). If we consider only the first term in (15) we obtain a generalization of the expressions given in Taylor (1986) and Guerrero (1993)

$$E(Y) = (1 + \lambda\mu)^{1/\lambda} \left\{ 1 + \frac{(1-\lambda)\phi\, V}{2\,(1+\lambda\mu)^2} \right\},$$

which are valid only for transformed normal data. The correction factor in braces is larger than one if $\lambda < 1$ and less than one if $\lambda > 1$.

Further, we can obtain the rth ordinary moment of Y by expanding the binomial $\{1 + \theta(Y^\lambda - \mu)\}^{r/\lambda}$ in the same way. We have

$$E(Y^r) = (1 + \lambda\mu)^{r/\lambda} \left\{ 1 + \sum_{i=2}^{\infty} \frac{b_i^{(r)}\, \mu_i}{i!\,(1+\lambda\mu)^i} \right\}, \qquad (16)$$

where $b_i^{(r)} = \prod_{j=0}^{i-1}(r - j\lambda)$. Clearly, $b_i^{(1)} = a_i$. Combining (16) and (15) we can obtain all cumulants of Y up to any order of $(1+\lambda\mu)^{-v}$ for $v \geq 2$. In special, the variance of Y to order $(1+\lambda\mu)^{-4}$ is given by

$$\begin{aligned} Var(Y) = (1+\lambda\mu)^{2/\lambda} \Bigg[& \frac{\mu_2}{(1+\lambda\mu)^2} + \frac{(1-\lambda)\mu_3}{(1+\lambda\mu)^3} \\ & + \frac{(1-\lambda)\{(7-11\lambda)\mu_4 - 3(1-\lambda)\mu_2^2\}}{12\,(1+\lambda\mu)^4} \Bigg]. \end{aligned}$$

An obvious estimate of $E(Y^r)$ follows by using the MLEs of the parameters λ, μ and ϕ. The adequacy of this expression in terms of λ and ϕ should be investigated in Monte Carlo simulations. When $\lambda = 0$ we can obtain from $E(Y^r) = e^{r\mu} E\{e^{r\,(Y^{(0)} - \mu)}\}$

$$E(Y^r) = e^{r\mu} \left\{ 1 + \sum_{i=2}^{\infty} \frac{r^i \mu_i}{i!} \right\}.$$

For well fitted models, the quantities $\widehat{\mu}_i$ for $i > 2$ will usually be small. The variance of Y follows as

$$Var(Y) = e^{2\mu} \left[\sum_{i=2}^{\infty} \left\{ \frac{(2^i - 2)\mu_i}{i!} + \frac{\mu_i^2}{i!^2} \right\} - 2 \sum_{i \neq j = 2}^{\infty} \frac{\mu_i\, \mu_j}{i!\, j!} \right].$$

For a general transformation (1) let $Y = F(Y^{(\lambda)}, \lambda)$ be its inverse. By expanding F in Taylor series we obtain

$$E(Y) = F(\mu, \lambda) + \sum_{i=2}^{\infty} \frac{F^{(i)}(\mu, \lambda)\, \mu_i}{i!},$$

where $F^{(i)}(\mu, \lambda)$ is the ith derivative of $F(\mu, \lambda)$ with respect to μ. Analogously, the rth moment of Y follows from

$$E(Y^r) = F(\mu,\lambda)^r + \sum_{i=2}^{\infty} \frac{G^{(i)}(\mu,\lambda)\,\mu_i}{i!},$$

where $G^{(i)}(\mu,\lambda)$ is the ith derivative of $G(\mu,\lambda) = F(\mu,\lambda)^r$ with respect to μ.

4 Special Models

For transformed normal and inverse Gaussian models, (13) yields

$$\widehat{\phi}^{(\lambda)} = \frac{D^{(\lambda)}}{n}, \tag{17}$$

and the profile log-likelihood for λ from (14) reduces to

$$l_P(\lambda) = -\frac{n}{2}\log(\widehat{\phi}^{(\lambda)}) - \frac{n}{2}\{1+\log(2\pi)\} + \sum_{i=1}^{n}\log\left\{\frac{J(\lambda,y_i)}{\sqrt{V(y_i^{(\lambda)})}}\right\}. \tag{18}$$

To maximize the profile log-likelihood (18), we only need to find a λ value that minimizes the ratio below

$$\hat{\lambda} = \arg\min_{\lambda} \frac{\left\{D^{(\lambda)}\widetilde{V}(y^{(\lambda)})\right\}^{1/2}}{\widetilde{J}(\lambda,y)}, \tag{19}$$

where $\widetilde{V}(y^{(\lambda)})$ and $\widetilde{J}(\lambda,y)$ are the geometric means of $V(y_i^{(\lambda)})$ and $J(\lambda,y_i)$ for $i=1,\cdots,n$, respectively. For PGLMs with the Box-Cox traansformation, $\widetilde{J}(\lambda,y) = \widetilde{y}^{\lambda-1}$, where $\widetilde{y}$ is the geometric mean of the original data and, in particular, for the Box-Cox models ($V=1$), the equation (19) yields a known result given by Yang and Abeysinghe (2002).

For transformed gamma models, (13) reduces to a result given by Cordeiro and McCullagh (1991)

$$\log\left(\widehat{\phi}^{(\lambda)\,-1}\right) - \Psi\left(\widehat{\phi}^{(\lambda)\,-1}\right) = \frac{D^{(\lambda)}}{2n}. \tag{20}$$

An approximate solution for $\widehat{\phi}^{(\lambda)}$ in (20) for small ϕ is

$$\widehat{\phi}^{(\lambda)} \approx \frac{2D^{(\lambda)}}{n\left\{1+\left(1+\dfrac{2D^{(\lambda)}}{3n}\right)^{1/2}\right\}}.$$

The sum of the first two terms in (18) is substituted by $n\,h(\widehat{\phi})$, where

$$h(\phi) = \frac{1}{\phi}\left\{\Psi\left(\phi^{-1}\right) - 1 - \phi\log\Gamma\left(\phi^{-1}\right)\right\}.$$

When ϕ is sufficiently small, we can obtain to order $O(\phi^3)$

$$h(\phi) = -\frac{1}{2} - \frac{\phi}{6} - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\phi),$$

which gives

$$l_P(\lambda) = -\frac{n}{2}\left\{\log(\widehat{\phi}) + \frac{\widehat{\phi}}{3}\right\} - \frac{n}{2}\left\{1 + \log(2\pi)\right\} + \sum_{i=1}^{n}\log\left\{\frac{J(\lambda, y_i)}{\sqrt{V(y_i^{(\lambda)})}}\right\}. \quad (21)$$

Clearly, (21) converges to the form (18) when $\phi \to 0$. In fact, the profile log-likelihood for λ for all TGLMs have the same form (18) for very small dispersion parameter values. This fact follows since, when ϕ tends to zero, the exponential dispersion model (2) can be written in the limit as

$$\pi(y_i^{(\lambda)}; \theta_i, \phi) \approx \left\{2\pi\phi V(y_i^{(\lambda)})\right\}^{-1/2} \exp\left\{\frac{-D_i^{(\lambda)}(y_i^{(\lambda)}, \mu_i)}{2\phi}\right\},$$

where $D_i^{(\lambda)}(y_i^{(\lambda)}, \mu_i)$ is the *ith* true deviance component that comes from (9). For this asymptotic case, $\widehat{\phi}^{(\lambda)}$ is just obtained from (17). This expression can be justified to some extent as a saddlepoint approximation for (2) provided that ϕ and all higher-order cumulants are sufficiently small and is exact only for the normal and inverse Gaussian models. It is clear that the equation (19) for the MLE of λ holds for any TGLM with very small dispersion parameter.

5 Model Inference

We essentially make inference about β and ϕ conditioning on $\lambda = \widehat{\lambda}$. Then, the estimated $\widehat{\lambda}$ is viewed as known, and confidence intervals for the parameters β, η, μ and ϕ, hypothesis tests, analysis of the deviance, residuals and diagnostics can be carried out routinely in the usual context of GLMs from the fitted values $\widehat{\beta}$, $\widehat{\eta}$, $\widehat{\mu}$ and $\widehat{\phi}$. The approximate covariance matrix of $\widehat{\beta}$ is given by $\phi\,(X^T\widehat{W}X)^{-1}$. The approximate variance of $\widehat{\phi}$ is $Var(\widehat{\phi}) = n^{-1}\,\phi^4\,\{\psi'(\phi^{-1}) - \phi\}^{-1}$ for gamma models and $Var(\widehat{\phi}) = 2\,\phi^2/n$ for normal and inverse Gaussian models, where ψ' is the trigamma function.

It is frequently of interest to test whether the parameter of the transformation family (1) conforms to a hypothesized value. We can easily obtain from (12) a likelihood ratio (LR) statistic $w = 2\,\{l_P(\widehat{\lambda}) - l_P(\lambda^{(0)})\}$ for testing $\lambda = \lambda^{(0)}$ which has the asymptotic χ_1^2 distribution and construct a large sample confidence interval for λ by inverting the LR test. Approximate confidence limits for λ can then be readily found from $\{\lambda \mid l_P(\lambda) > l_P(\hat{\lambda}) - \frac{1}{2}\chi_1^2(\alpha)\}$ and the accuracy of this approximation follows from the fact that $Pr\,\{w \geq \chi_1^2(\alpha)\} = \alpha + O(n^{-1/2})$. We can also work with $sign(\hat{\lambda} - \lambda^{(0)})\,w^{1/2}$ to make inference about λ, which is asymptotically standard normal with absolute error typically of order $n^{-1/2}$.

The scaled deviance in TGLMs is defined conditioning on $\widehat{\lambda}$ as twice the difference between the log-likelihood achieved under the model and the maximum attainable value, namely $S^{(\hat{\lambda})} = \phi^{-1} D^{(\hat{\lambda})}$, and depends on a known or consistently estimated dispersion parameter ϕ, and in either case we can take $S^{(\hat{\lambda})}$ as distributed as χ^2_{n-p} approximately. However, the chi-square approximation may not be effective because the dimension of the saturated model is n and the usual asymptotic argument does not apply. If we are testing two nested TGLMs, the χ^2 distribution may be a good approximation for the difference of scaled deviances. Indeed, suppose that X_A $(n \times p_A)$ and X_B $(n \times p_B)$ represent two different choices of X, and that they are nested $X_A < X_B$ say, meaning that all columns of X_A are contained in the linear span of the columns of X_B. After fitting the two models conditioning on $\widehat{\lambda}$, the scaled deviances are $S_A^{(\hat{\lambda})}$ and $S_B^{(\hat{\lambda})}$. The LR statistic $w^{(\hat{\lambda})} = S_A^{(\hat{\lambda})} - S_B^{(\hat{\lambda})}$ to test X_A against X_B has an asymptotic χ^2 distribution with $p_B - p_A$ degrees of freedom with an error of order n^{-1}. Consider now a set $J = A, \cdots, I$ of arbitrary TGLMs, the model J with log-likelihood $\hat{l}_J$ obtained by maximizing (5) with respect to all $p_J + 2$ parameters, namely p_J parameters in the systematic component and the scalar parameters ϕ and λ. Evaluation and selection among the models A, ..., I may be based on Akaike information criterion (AIC) defined for the Jth TGLM by $AIC_J = 2\,(p_J + 2 - \hat{l}_J)$.

6 TGARMA Models

In this section, we work with the family of transformations (1) and incorporate the idea of transforming the response variable to follow the framework of the GARMA model introduced by Benjamin et al. (2003). For time series data $\{y_t, t = 1, \ldots, n\}$ conditional rather than marginal distributions are modeled. We assume that the conditional distribution of the transformed response $\{Y_t^{(\lambda)}, t = 1, \ldots, n\}$ given the past history of the process belongs to the continuous exponential dispersion model (2). The conditional density function of the transformed response $Y_t^{(\lambda)}$ is defined given the set $H_t = \{x_t, \ldots, x_1, y_{t-1}^{(\lambda)}, \ldots, y_1^{(\lambda)}, \mu_{t-1}, \ldots, \mu_1\}$ that represents past values of the transformed series and their means and past and possibly present values (when known) of the covariates, meaning all that is known except for λ to the observer at time t, where x_t is a specified $1 \times m$ $(m < n)$ vector of explanatory variable. The conditional mean μ_t and variance ϕV_t of $Y_t^{(\lambda)}$ are expressed as in Section 2, but we take the systematic component with an extra part τ_t, as proposed by Benjamin et al. (2003), that includes additively autoregressive moving average (ARMA) terms by conditioning on the first r transformed observations, where $r = \max\{p, q\}$. We have

$$g(\mu_t) = \eta_t = x_t\beta + \sum_{j=1}^{p} \varphi_j \left\{ g(y_{t-j}^{(\lambda)}) - x_{t-j}\beta \right\} + \sum_{j=1}^{q} \psi_j \left\{ g(y_{t-j}^{(\lambda)}) - \eta_{t-j} \right\}, \quad (22)$$

for $t = r+1, \ldots, n$, where $\beta = (\beta_1, \ldots, \beta_m)^T$. Equations (1), (2) and (22) define the TGARMA (p, q, λ) model. The aim of the transformation (1) is to ensure that the usual assumptions for GARMA models hold for the transformed series $Y_t^{(\lambda)}$. GARMA model is a special case of the TGARMA model when (1) is $\Lambda(Y_t) = Y_t$ independent of λ. The power GARMA model, termed here PGARMA model, is another special case when $\Lambda(Y_t)$ is the Box-Cox transformation or the simple power family.

We can write the systematic component (22) of the TGARMA model for the observations $y_{r+1}^{(\lambda)}, \ldots, y_n^{(\lambda)}$ conditioning on the first r transformed observations, in matrix notation, by $\eta = M\gamma$, where $M = [X\ A\ B]$ is the local model matrix of order $(n-r) \times (m+p+q)$, X is the matrix formed by the rows x_t for $t = r+1, \ldots, n$, $\gamma = (\beta^T, \varphi^T, \psi^T)^T$, $\varphi = (\varphi_1, \ldots, \varphi_p)^T$, $\psi = (\psi_1, \ldots, \psi_q)^T$ and the matrices A and B of orders $(n-r) \times p$ and $(n-r) \times q$ are functions of the model parameters given by

$$A = \begin{bmatrix} g(y_r^{(\lambda)}) - x_r\beta & \cdots & g(y_{r+1-p}^{(\lambda)}) - x_{r+1-p}\beta \\ g(y_{r+1}^{(\lambda)}) - x_{r+1}\beta & \cdots & g(y_{r+2-p}^{(\lambda)}) - x_{r+2-p}\beta \\ \vdots & \ddots & \vdots \\ g(y_{n-1}^{(\lambda)}) - x_{n-1}\beta & \cdots & g(y_{n-p}^{(\lambda)}) - x_{n-p}\beta \end{bmatrix}_{(n-r)\times p}$$

and

$$B = \begin{bmatrix} g(y_r^{(\lambda)}) - \eta_r & \cdots & g(y_{r+1-q}^{(\lambda)}) - \eta_{r+1-q} \\ g(y_{r+1}^{(\lambda)}) - \eta_{r+1} & \cdots & g(y_{r+2-q}^{(\lambda)}) - \eta_{r+2-q} \\ \vdots & \ddots & \vdots \\ g(y_{n-1}^{(\lambda)}) - \eta_{n-1} & \cdots & g(y_{n-q}^{(\lambda)}) - \eta_{n-q} \end{bmatrix}_{(n-r)\times q}.$$

The $m+p+q+2$ parameters of the TGARMA model to be estimated are then the vector γ and the scalars ϕ and λ. The main objective in the analysis of the TGARMA models is to make partial likelihood inference on the model parameters. The model fitting procedure described herein is valid only for continuous time series and exclude count time, binary and categorical time series. The log-likelihood for the parameter vector γ and scalars ϕ and λ expressed in terms of the transformed series $y^{(\lambda)} = (y_{r+1}^{(\lambda)}, \ldots, y_n^{(\lambda)})^T$ and conditioned on the first r transformed observations, has the same form of the equation (5), except that the sum is over $y_{r+1}^{(\lambda)}, \cdots, y_n^{(\lambda)}$. As a matter of fact we are working with a log-partial likelihood in the sense of Cox (1975), which continues to be very important in a great many areas of applications such as time series. For maximizing the log-likelihood (5), we proceed as in Section 3 by assuming first that the transformation parameter λ is fixed and obtain the likelihood equations for estimating γ and ϕ. Let $\widehat{\gamma}^{(\lambda)}$, $\widehat{\eta}^{(\lambda)} = \widehat{M}^{(\lambda)}\widehat{\gamma}^{(\lambda)}$ and $\widehat{\phi}^{(\lambda)}$ be the MLEs of γ, η and ϕ, respectively, for given λ. The MLE $\widehat{\gamma}^{(\lambda)}$ does not depend on the dispersion parameter ϕ and can be obtained from the fitting of the model defined by (2) and (22) to $y^{(\lambda)}$ by iteratively reweighted least squares

$$\widehat{\gamma}^{(\lambda)} = \left(\widehat{M}^{(\lambda)T}\widehat{W}^{(\lambda)}\widehat{M}^{(\lambda)}\right)^{-1}\widehat{M}^{(\lambda)T}\widehat{W}^{(\lambda)}\widehat{z}^{(\lambda)},$$

where the weight matrix and the working variate are $W = diag\,\{w_{r+1}, \ldots, w_n\}$ and $z^{(\lambda)} = (z_{r+1}^{(\lambda)}, \ldots, z_n^{(\lambda)})^T$. For $t = r+1, \ldots, n$, the current estimates $\widehat{\eta}_t^{(\lambda)}$ and the adjusted means of the transformed series are easily obtained from (22) and $\widehat{\mu}_t^{(\lambda)} = g^{-1}(\widehat{\eta}_t^{(\lambda)})$.

The MLEs of the parameters ϕ and λ follow the general formulae (10) and (12) given in Section 3. All equations presented in Sections 3 and 4 hold here except that the sum is over $y_{r+1}^{(\lambda)}, \cdots, y_n^{(\lambda)}$ and n and p should be replaced by $n-r$ and $m+p+q$, respectively.

We can make inference about ϕ and the parameters in γ conditioning on the transformed parameter $\lambda = \hat{\lambda}$ as in the preceding discussion in Section 5. Confidence intervals for the parameters γ, η_t, μ_t and ϕ, analysis of the deviance, LR tests, residuals and diagnostics for the TGARMA models follow the usual context of GARMA models conditioning on $\widehat{\lambda}$. The test of the transformation parameter is performed in the same way of the TGLMs.

We can estimate the mean of the untransformed depend variable Y_t by using a Taylor series expansion of $Y_t = F(Y_t^{(\lambda)}, \lambda)$, where $F(.)$ is the inverse transformation $\Lambda^{-1}(.)$ of (1) indicated at the end of Section 3. Conditioning on the set H_t we obtain

$$\widehat{E}(Y_t) \approx F(\widehat{\mu}_t, \widehat{\lambda}) + \frac{\widehat{\phi}\,\widehat{V}_t}{2} F^{(2)}(\widehat{\mu}_t, \widehat{\lambda}).$$

Additional terms can be easily included in this equation since the central moments of $Y_t^{(\lambda)}$ are just given in terms of the derivatives of the variance function. For the Box-Cox power transformation $F^{(2)}(\mu_t, \lambda) = (1-\lambda)(1+\lambda\mu_t)^{(1-2\lambda)/\lambda}$.

7 Simulation results

We now present simulation results comparing the performance of the algorithms discussed in Sections 3 and 4 to estimate the parameters of the TGLMs. We simulated power gamma and inverse Gaussian models with canonical link functions by using the Box-Cox transformation with three specific values for $\lambda = 0, 0.5$ and 2, and a two-parameter linear component $\eta = \beta_0 + \beta_1 x$, where x is the explanatory variable and β_0 and β_1 are parameters to be estimated. The dispersion parameter was fixed at $\phi = 1/10$ and $1/20$. Our aim is to illustrate the use of the profile log-likelihood to estimate the transformation parameter λ. The number of observations was set at $n = 30, 60$ and 90 and the number of replications at $10,000$. The values of the covariate x were obtained as random draws from the uniform distribution on the interval $(0,1)$ and were held constant throughout the simulations with equal sample sizes. The sample means of the MLEs with their respective standard errors in parentheses obtained out of $10,000$ simulations from the fitted power gamma and inverse

Gaussian models are given in Tables 2 and 3, respectively.

For each combination of the true parameters given in Tables 2 and 3, the transformed dependent variable $y^{(\lambda)}$ was generated following gamma and inverse Gaussian distributions and, in each simulation, we fitted both power gamma and inverse Gaussian models to $y^{(\lambda)}$ and computed the MLEs of β_0, β_1, ϕ and λ. Then, we repeated this process $10,000$ times. The convergence was not achieved for all the simulated samples and then we eliminated those samples for which the iterative fitting algorithm (6) failed to converge. The figures in Tables 2 and 3 convey important information. First, the simulations show that the MLEs of the TGLMs can be computed through the algorithms described in Sections 3 and 4 with minimal effort. Second, the sample means of the MLEs are usually much closer to the corresponding true values when the sample size n increases according to the first-order asymptotic theory.

Table 2: Sample means and standard errors of the MLEs for the power gamma model with true parameters $\beta_0 = 0.3, \beta_1 = 0.7$

Gamma		$\lambda = 0$		$\lambda = 0.5$		$\lambda = 2.0$	
n		$\phi = 1/10$	$\phi = 1/20$	$\phi = 1/10$	$\phi = 1/20$	$\phi = 1/10$	$\phi = 1/20$
30	$\widehat{\lambda}$	0.0911	0.1129	0.5537	0.5592	2.0392	2.0645
		(0.2865)	(0.3232)	(0.4025)	(0.4139)	(0.4647)	(0.7130)
	$\widehat{\phi}$	0.1130	0.0596	0.1026	0.0524	0.0963	0.0503
		(0.0551)	(0.0309)	(0.0473)	(0.0256)	(0.0339)	(0.0226)
	$\widehat{\beta}_0$	0.2807	0.2742	0.3096	0.3064	0.3102	0.3175
		(0.1523)	(0.1635)	(0.1565)	(0.1534)	(0.1163)	(0.1593)
	$\widehat{\beta}_1$	0.6735	0.6699	0.6753	0.6759	0.6885	0.6824
		(0.1161)	(0.0888)	(0.1210)	(0.0992)	(0.1432)	(0.1291)
60	$\widehat{\lambda}$	0.0729	0.0865	0.5493	0.5556	2.0464	2.0431
		(0.2389)	(0.2822)	(0.3604)	(0.3927)	(0.6723)	(0.6798)
	$\widehat{\phi}$	0.1120	0.0579	0.1055	0.0534	0.1029	0.0514
		(0.0435)	(0.0249)	(0.04076)	(0.0217)	(0.0393)	(0.0201)
	$\widehat{\beta}_0$	0.2814	0.2788	0.3045	0.3044	0.3199	0.3187
		(0.1276)	(0.1423)	(0.1370)	(0.1467)	(0.1514)	(0.1496)
	$\widehat{\beta}_1$	0.6793	0.6786	0.6785	0.6787	0.6826	0.6854
		(0.0958)	(0.0696)	(0.0936)	(0.0781)	(0.1281)	(0.1085)
90	$\widehat{\lambda}$	0.0552	0.0609	0.5466	0.5495	2.0347	2.0117
		(0.1988)	(0.2339)	(0.3352)	(0.3611)	(0.6469)	(0.6710)
	$\widehat{\phi}$	0.1095	0.0559	0.1059	0.0537	0.1028	0.0512
		(0.0366)	(0.0214)	(0.0364)	(0.0201)	(0.0361)	(0.0192)
	β_0	0.2846	0.2838	0.3023	0.3024	0.3195	0.3248
		(0.1050)	(0.1166)	(0.1276)	(0.1335)	(0.1489)	(0.1514)
	β_1	0.6852	0.6869	0.6819	0.6819	0.6866	0.6904
		(0.0759)	(0.0563)	(0.0791)	(0.0647)	(0.1149)	(0.0996)

Table 3: Sample means and standard errors of the MLEs for the power inverse Gaussian (I.G.) model with true parameters $\beta_0 = 1, \beta_1 = 5$

I.G.		$\lambda = 0$		$\lambda = 0.5$		$\lambda = 2.0$	
n		$\phi = 1/10$	$\phi = 1/20$	$\phi = 1/10$	$\phi = 1/20$	$\phi = 1/10$	$\phi = 1/20$
30	$\widehat{\lambda}$	0.0779	0.0744	0.5412	0.5212	2.0400	1.9923
		(0.4467)	(0.4530)	(0.4628)	(0.2861)	(0.7460)	(0.7436)
	$\widehat{\phi}$	0.1011	0.0505	0.1002	0.0502	0.0993	0.0491
		(0.0302)	(0.0150)	(0.0288)	(0.0137)	(0.0280)	(0.0138)
	$\widehat{\beta}_0$	1.0516	1.0289	1.1054	1.0323	1.1286	1.1271
		(0.6554)	(0.5814)	(0.6599)	(0.3629)	(0.6634)	(0.57644)
	$\widehat{\beta}_1$	4.8434	4.8724	4.8691	4.9473	4.9754	5.0785
		(1.1470)	(0.9709)	(1.3160)	(0.7476)	(1.4766)	(1.2173)
60	$\widehat{\lambda}$	0.0776	0.0564	0.5568	0.5247	1.9959	1.9968
		(0.4415)	(0.4379)	(0.4516)	(0.2834)	(0.7293)	(0.7250)
	$\widehat{\phi}$	0.1014	0.0503	0.1003	0.0499	0.0996	0.0493
		(0.0225)	(0.0114)	(0.0210)	(0.0098)	(0.0203)	(0.0102)
	$\widehat{\beta}_0$	1.0346	1.0378	1.0514	1.0166	1.1256	1.1200
		(0.6355)	(0.5295)	(0.5337)	(0.3336)	(0.5959)	(0.5738)
	$\widehat{\beta}_1$	4.8501	4.8957	4.8948	4.9547	5.0645	5.0556
		(0.9375)	(0.7958)	(0.9252)	(0.6084)	(1.2459)	(1.1028)
90	$\widehat{\lambda}$	0.0735	0.0403	0.5264	0.5224	2.0463	1.9953
		(0.4245)	(0.4274)	(0.2825)	(0.2834)	(0.7237)	(0.7177)
	$\widehat{\phi}$	0.1014	0.0501	0.0999	0.0501	0.0989	0.0494
		(0.0194)	(0.0097)	(0.0161)	(0.0082)	(0.0190)	(0.0085)
	$\widehat{\beta}_0$	1.0189	1.0447	1.0152	1.0169	1.09743	1.1171
		(0.5512)	(0.5067)	(0.3784)	(0.3265)	(0.5891)	(0.5638)
	$\widehat{\beta}_1$	4.8704	4.9328	4.9688	4.9589	5.0167	5.0537
		(0.8097)	(0.7271)	(0.7133)	(0.5705)	(1.1383)	(1.0431)

Third, the parameter λ is well estimated in all cases and the MLEs of the other parameters of both power models are in reasonable accordance with their corresponding true parameters for most of the cases reported. Clearly, large sample sizes are really necessary for the MLEs to become very accurate in terms of bias and mean square errors. Bias corrections based on the second-order asymptotic theory can then be derived in future research to obtain improved MLEs in TGLMs with samples of small to moderate size. We have also considered other choices of values for the covariate x but they had little impact on the final results.

The simulated data and all the calculations were performed by using a structural programming language written in Matlab. We obtain an estimated regression $t_n = \widehat{a}\, n^2 + \widehat{b}\, \log(n)/n$ (in seconds) for the computing time (t_n) of each simulation with the gamma model in terms of the sample size (n), where $\widehat{a} = 1 \times 10^{-3}$ and $\widehat{b} = 15$. This expression yields an average computing time for one simulation of order $O(n^2)$. Since each cell of Table 2 was computed from $10,000$ simulations, the average computing time for each cell of this table is about $10\, n^2$ seconds. The same analysis was done for the inverse Gaussian model in Table 3 showing that $\widehat{a} = 1.8 \times 10^{-3}$ and $\widehat{b} = 20$. The computing time for each cell of Table 3 is in average $18\, n^2$ seconds.

8 Applications to real data

We now apply the overall procedure of estimation described in Sections 3 and 4 to estimate the parameters of the TGLMs fitted to three real data sets. In the first data set the response variable (y) denotes the percentile of the number of illiterate people older than 15 years who were declared caucasian out of the adult population and the explanatory variable (x) is the logarithm of the family income. The data were obtained from the Brazilian Institute of Geography and Statistics-IBGE (2002) by sampling in the 27 brazilian states. In the second data set collected by the U.S. Navy and presented in Example 7.4 of Myers (1990), the response variable (y) is the quantity of man-hours per month devoted to surgical services at $n = 15$ Naval hospitals. Here the explanatory variable (x) is the inverse of the number of surgical cases The third data set presented in Pinheiro (2007) gives the weights (y) and the lengths (x) from the head to the tail of 184 rose shrimps (*Farfantepenaeus brasiliensis*) sampled in the coast of the state of Rio Grande do Norte, Northeast of Brazil.

We consider power gamma models with reciprocal link function $g(\mu) = \mu^{-1}$ and power inverse Gaussian models with reciprocal of the square link function $g(\mu) = \mu^{-2}$ using the Box-Cox transformation. For both power canonical models we take $g(\mu) = \eta = \beta_0 + \beta_1\, x$, where $\mu = E\{y^{(\lambda)}\}$ and x is the explanatory variable. For the three data sets, we fitted the two power models to $y^{(\lambda)}$ by fixing the transformation parameter at $\lambda = 1$ and by choosing the optimal value of λ that maximizes the profile log-likelihood (14). We give in Tables 4, 5 and 6 the MLEs of the linear parameters and the dispersion parameter (with their corresponding variances conditioning on λ fixed or estimated in parentheses) from the fitting of the models above. When the transformation parameter λ is estimated we also give in these tables approximate 95% confidence intervals for this parameter. Based on these asymptotic confidence intervals, we note that the GLM is only accepted for the case of the power gamma model fitted to the data set 1. These intervals show appreciable ranges of compatible values for λ, including zero, corresponding to the log transformation, for the power gamma model fitted to the data set 1 and for the power inverse Gaussian model fitted to the data sets 1 and 2.

Table 4: Data set 1 – ($n = 27$), $y = illiteracy\ rate$, $x = \log(family\ income)$

Parameter	Gamma		Inverse Gaussian	
λ	1	0.2720	1	−0.0660
		(-0.493; 1.250)		(-0.686; 0.582)
β_0	−0.1312	−0.2348	−1.0175e-2	−0.2022
	(2.9990e-4)	(1.8657e-3)	(2.0938e-6)	(1.2863e-3)
β_1	2.9240e-2	7.3207e-2	2.0323e − 3	5.5007e-2
	(9.3073e-6)	(5.5076e-5)	(6.9717e-8)	(3.7839e-5)
ϕ	3.1261e-2	6.6049e-3	2.0234e-3	1.0361e-3
	(1.1560e-6)	(1.1044e-8)	(3.0326e-7)	(7.9517e-8)

Table 5: Data set 2 – ($n = 15$), $y = Men\text{-}hour$, $x = (surgical)^{-1}$

Parameter	Gamma		Inverse Gaussian	
λ	1	0.4070	1	0.1260
		(0.125; 0.595)		(−0.561; 0.260)
β_0	3.3529e-7	6.7525e-3	−2.1990e-8	2.3843e-3
	(9.5172e-11)	(3.6651e-8)	(4.7474e-17)	(4.0794e-9)
β_1	0.1317	3.7176	5.1684e-5	1.1181
	(1.1376e-4)	(0.0155)	(1.4191e-10)	(1.3232e-3)
ϕ	3.3401e-2	1.5086e-3	4.1652e-5	2.2905e-5
	(2.6143e-6)	(2.4506e-10)	(2.3131e-10)	(6.9949e-11)

Table 6: Data set 3 – ($n = 184$), $y = weight\ shrimp$, $x = (length\ shrimp)^{-2}$

Parameter	Gamma		Inverse Gaussian	
λ	1	0.2580	1	−0.2230
		(0.083; 0.430)		(−0.364; −0.083)
β_0	−4.0119e-2	3.7569e-2	−1.1373e-2	−2.1765e-4
	(2.6335e-6)	(1.1511e-5)	(1.7409e-7)	(2.7627e-5)
β_1	1187.2069	2390.5943	201.7313	2690.9494
	(331.2460)	(1176.7644)	(50.3210)	(2724.4848)
ϕ	7.1893e-3	2.0877e-3	5.3119e-3	7.0552e-4
	(2.0232e-9)	(4.9673e-11)	(3.0670e-7)	(5.4104e-9)

In order to compare the fitted models we use the mean square error (MSE) and the mean absolute percentile error (MAPE) given by

$$MSE = \frac{100\%}{n\,\widehat{\sigma}^2_{y^{(\widehat{\lambda})}}} \sum_{i=1}^{n} (y_i^{(\widehat{\lambda})} - \widehat{\mu}_i)^2$$

and

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i^{(\widehat{\lambda})} - \widehat{\mu}_i}{y_i^{(\widehat{\lambda})}} \right|,$$

respectively, where $\widehat{\sigma}^2_{y^{(\lambda)}}$ is the sample variance of $y^{(\lambda)}$. In Table 7 we give these statistics for the two power canonical models fitted to the three data sets. In this table the row labeled $\widehat{l}$ gives the maximized log-likelihood for the corresponding λ value (fixed or estimated) and the row denoted by w gives the value of the LR statistic for testing $\lambda = 1$. The values of w in Table 7 show that the PGLMs should be chosen in five of the six fitted models and for these cases there is a considerable reduction in the values of the statistics MSE and $MAPE$.

Table 7: A comparison of the fitted models

Data	Model	Gamma		Inverse Gaussian	
1	λ	1	0.272	1	-0.0660
	MSE(%)	25.2380	20.3860	47.5060	21.9504
	MAPE(%)	14.3412	6.4130	16.6307	3.9665
	$\widehat{l}$	-77.8956	-77.0918	-83.9105	-80.0304
	w	1.6076		7.7602	
2	λ	1	0.4070	1	0.1260
	MSE(%)	10.6890	1.6230	50.8540	1.2870
	MAPE(%)	16.2125	3.3961	34.6076	1.5402
	$\widehat{l}$	-123.836	-113.5073	-137.5813	-114.2916
	w	20.6574		46.5794	
3	λ	1	0.2580	1	-0.223
	MSE(%)	4.7802	2.9769	19.508	5.1389
	MAPE(%)	6.7269	3.4955	380.0793	2.5561
	$\widehat{l}$	-267.4472	-262.8362	-442.7164	-331.0713
	w	9.2220		223.2902	

We now illustrate graphically some fitted PGLMs. The Figures 1, 2 and 3 refer to the power inverse Gaussian model fitted to the first data set (rate of illiteracy versus income family). Figure 1 shows the profile log-likelihood curve plotted against the transformation parameter λ. Its maximum of -80.0304 occurs near $\lambda = -0.0660$ and there is an appreciable range of compatible values for λ including zero, corresponding to the logarithmic transformation. Figure 2 shows that the power inverse Gaussian model is well fitted. Figure 3 shows that the estimated means of the original observations, predicted using only the first term in (15) are very well retransformed.

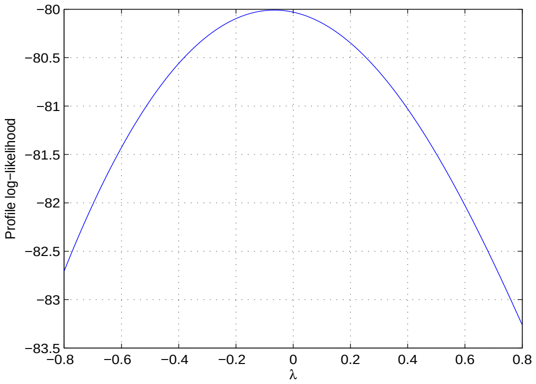

Figure 1: The profile log-likelihood curve for λ for the power inverse Gaussian model fitted to the data set 1

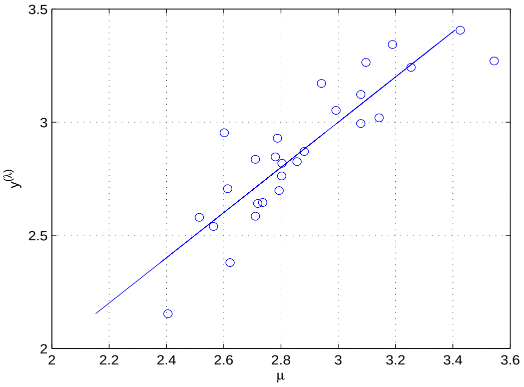

Figure 2: Plot of $y^{(\widehat{\lambda})}$ versus $\widehat{\mu}$ for the power inverse Gaussian model fitted to the data set 1

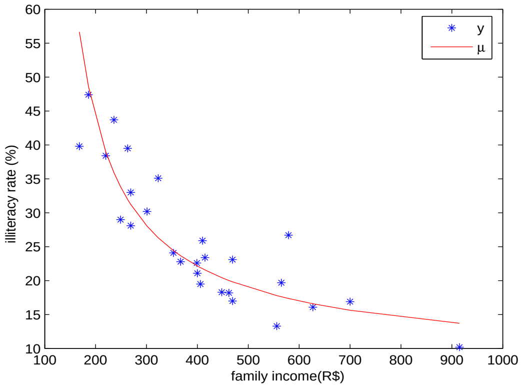

Figure 3: Untransformed values y and estimated means $\widehat{E}(Y)$ versus income for the power I.G. model fitted to the data set 1

Figures 4, 5 and 6 show the power inverse Gaussian model fitted to the second data set (quantity of man-hours versus number of surgical cases). The profile log-likelihood in Figure 4 is bimodal and the global maximizing value of $\hat{l} = -114.2916$ occurs near $\widehat{\lambda} = 0.1260$. In Figure 5 we plotted the transformed observations $y^{(\hat{\lambda})}$ against the fitted means $\widehat{\mu}$. This graphic is approximately linear and clearly gives an indicative that the power inverse Gaussian model provides a reasonably good fit to the data set 2. Figure 6 plots the original data y and the estimated expected values $\widehat{E}(y)$ versus the number of surgical cases confirming that the power inverse Gaussian model gives also a good prediction in the original scale. Figures 7, 8 and 9 refer to the power gamma model fitted to the third data set (weights and lengths of the shrimps). Figure 7 shows the profile log-likelihood plotted versus λ yielding the optimal value $\widehat{\lambda} = 0.2580$ for the transformation parameter. An approximate 95% confidence interval for λ is $(0.0830, 0.4300)$. Figure 8 shows good agreement between the transformed response and the fitted mean. In Figure 9, the original values y and the predicted means $\widehat{E}(Y)$ are plotted against the lengths (x) of the shrimps showing that the power gamma model produces accurate forecasts of the original observations.

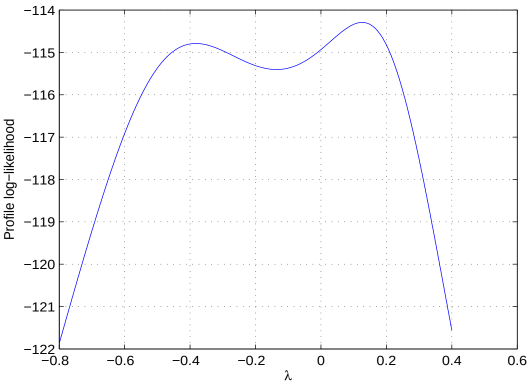

Figure 4: The profile log-likelihood curve for λ for the power inverse Gaussian model fitted to the data set 2

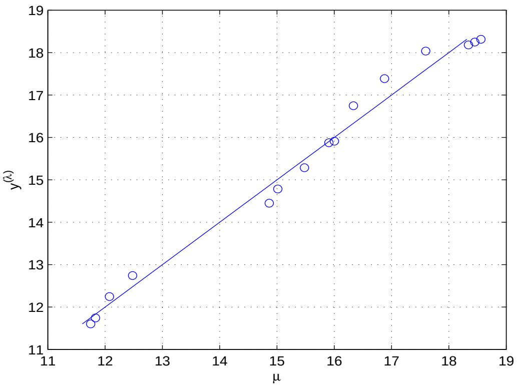

Figure 5: Plot of $y^{(\widehat{\lambda})}$ versus $\widehat{\mu}$ for the power inverse Gaussian model fitted to the data set 2

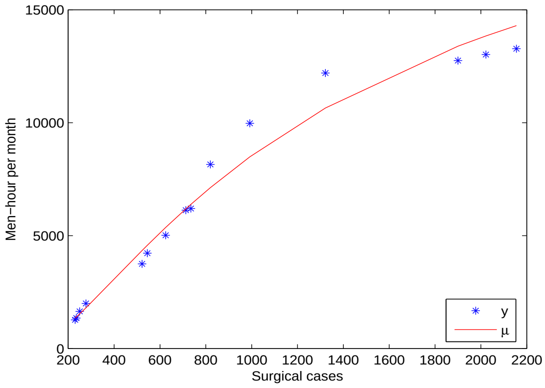

Figure 6: Plots of y and $\widehat{E}(Y)$ versus surgical cases for the power inverse Gaussian model fitted to the data set 2

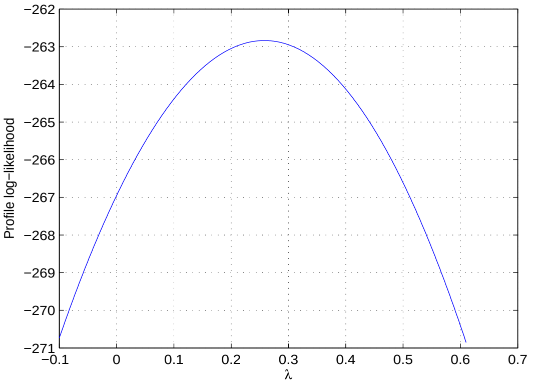

Figure 7: Profile log-likelihood for λ for the power gamma model fitted to the data set 3

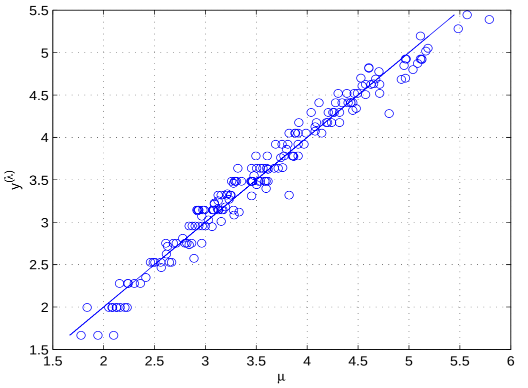

Figure 8: Plot of $y^{(\widehat{\lambda})}$ versus $\widehat{\mu}$ for the power gamma model fitted to the data set 3

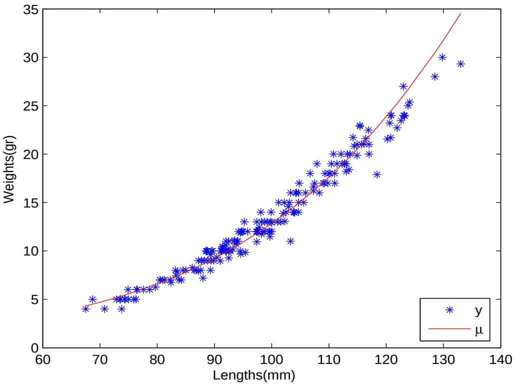

Figure 9: Plots of the untransformed values y and estimated means $\widehat{E}(Y)$ versus length for the power gamma model fitted to the data set 3

Finally, these applications showed that the TGLMs could be practical, effective and worthwhile technique for analyzing real data sets.

9 Conclusion

We define the TGLMs as an extension of the Box and Cox models and classical GLMs in order to unify these apparently diverse statistical techniques. The choice of the components for a TGLM is a very important task in the development and build of an adequate model. TGLMs are quite effective in the modelling of a mean regression response to continuous data. Many of the ideas in the GLMs carry over with little change to the whole class of TGLMs. We show that the MLEs of all parameters in this new class of models can be obtained easily. We can make inference in the TGLMs conditioning on the MLE of the transformation parameter following the inference procedures in the GLMs. We can also estimate the moments of the untransformed dependent variable by using simple formulae which generalize some previous results in the literature. We also present the idea of using the TGLMs for modelling time series data. We provide some examples of real data modelled by TGLMs.

References

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. Biometrika, 68, 357-363.

Benjamin, M.A., Rigby, R.A., Stasinopoulos, D.M. (2003). Generalized autoregressive moving average models. J. Amer. Statist. Assoc., 98, 214-223.

Bickel, P.J., Doksum, K. A. (1981). An analysis of transformations revisited. J. Amer. Statist. Assoc., 76, 296-311.

Box, G.E.P., Cox, D.R. (1964). An analysis of transformation. J. R. Statist. Soc. B, 26, 211-252.

Cordeiro G.M., McCullagh, P. (1991). Bias correction in generalized linear models. J. R. Statist. Soc. B, 53, 629-643.

Cox, D. R. (1975). Partial likelihood. Biometrika, 62, 69-76.

Draper, N. R., Cox, D. R. (1969). On distributions and their transformation to normality. J. R. Statist. Soc. B, 31, 472-476.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. Journal of Forecasting, 12, 37-48.

Guerrero, V. M., Johnson, R. A. (1982). Use of the Box-Cox transformation with binary response models. Biometrika, 69, 309-314.

IBGE (2002). Indicadores de desenvolvimento sustentável (IDS), Instituto Brasileiro de Geografia e Estatísitica-IBGE, Pesquisa Nacional por Amostra de Domicílios – PNAD, http://www.ibge.gov.br/.

Jorgensen, B. (1997). The Theory of Dispersion Models. Chapman and Hall: London.

Lawrence, A. J. (1987). A note on the variance of the Box-Cox regression transformation estimate. Appl. Statist., 36, 221-223.

Manly, B. F. (1976). Exponential data transformation. The Statistician, 25, 37-42.

Myers, R. H. (1990). Classical and Modern Regression with Applications, Second Edition, Duxbury Press (PWS-KENT Publishing Company), 299-305.

Nelder, J.A., Wedderburn, R.W.M. (1972). Generalized linear models. J. R. Statist. Soc. A, 135, 370-384.

Pankratz, A., Dudley, U. (1987). Forecasts of power-transformed series. Journal of Forecasting, 6, 239-248.

Pinheiro, A. P. (2007). Genetics and biometrics Characterization of the shrimp pink populations, *Farfantepenaeus brasiliensis*, in three places in the Rio Grande do Norte coast, Project CNPq 140229/2005-1, Program of PhD degree in Ecology and Natural Resources/UFSCar, São Carlos, SP-Brazil.

Sakia, R.M. (1992). The Box-Cox transformation technique: a review. The Statistician, 41, 168-178.

Taylor, J. M. G. (1986). The retransformed mean after a fitted power transformation. J. Amer. Statist. Assoc., 81, 114-118.

Tweedie, M. C. K. (1947). Functions of a statistical variate with given means, with special reference to Laplacian distributions. Proc. Cambridge Phil. Soc., 49, 41-49.

Yang, Z., Abeysinghe, T. (2002). An explict variance formula for the Box-Cox functional form estimator. Economics Letters, 76, 259-265.