Box–Cox response transformations for random effect models

Amani Almohaimeed[1] [2], **Jochen Einbeck**[1]

[1] Department of Mathematical Sciences, Durham University, UK

[2] Qassim University, Saudi Arabia

jochen.einbeck@durham.ac.uk

Abstract

For the linear model with random effects of unspecified distribution, we develop methodology for simultaneous response transformation and estimation of regression parameters. This is achieved by extending the "Nonparametric Maximum Likelihood" towards a "Nonparametric Profile Maximum Likelihood" technique. The methods allow to deal with overdispersion as well as two–level data scenarios.

1. Introduction

For data with a two–level structure, such as longitudinal data, correlation of responses within upper–level units can be induced by adding a random effect z_i to the linear predictor $x_{ij}^T\beta$, with the upper-level indexed by $i = 1, \ldots, r$, and the lower-level indexed by $j = 1, \ldots, n_i$, $\sum n_i = n$. Conditional on the random effect, the responses y_{ij} are independently distributed with mean function

$$E(y_{ij}|z_i) = x_{ij}^T\beta + z_i. \quad (1)$$

The objective of the Box–Cox transformation [3] is to select an appropriate parameter λ which is then used to transform the responses such that they follow a normal distribution more closely than the untransformed data. Under the scenario of model (1), this transformation can be written as

$$y_{ij}^{(\lambda)} = \begin{cases} \frac{y_{ij}^{\lambda}-1}{\lambda} & \lambda \neq 0, \\ \log y_{ij} & \lambda = 0 \end{cases}$$

for $y_{ij} > 0$, $i = 1, ..., r$ and $j = 1, ..., n_i$. It is assumed that there is a value of λ for which $y_{ij}^{(\lambda)}|z_i \sim N(x_{ij}^T\beta + z_i, \sigma^2)$, where z_i is a random effect with an unspecified mixing distribution $g(z_i)$. Taking account of the Jacobian of the transformation from $y_{ij}^{(\lambda)}$ to y_{ij}, the conditional density function of y_{ij} given z_i is

$$f(y_{ij}|z_i) = \frac{y_{ij}^{\lambda-1}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_{ij}^{(\lambda)} - x_i^T\beta - z_i)^2\right].$$

2. Estimation

Under the NPML estimation approach, the distribution of the random effect is approximated by a discrete distribution at mass points $z_1, \ldots, z_K$, which can be considered as intercepts for the different unknown subgroups on the upper level. Hence, the likelihood in relation to the original observations can be approximated as [1]

$$L(\lambda, \beta, \sigma^2, g) = \prod_{i=1}^{r} \int \left[\prod_{j=1}^{n_i} f(y_{ij}|z_i)\right] g(z_i)dz_i \approx \prod_{i=1}^{r}\sum_{k=1}^{K} \pi_k m_{ik}, \quad (2)$$

where $m_{ik} = \prod_{j=1}^{n_i} f(y_{ij}|z_k)$. Defining indicators $G_{ik} = 1$ if case i stems from cluster k and 0 otherwise, the complete log–likelihood is

$$\ell^* = \log L^* = \sum_{i=1}^{r}\sum_{k=1}^{K} [G_{ik}\log\pi_k + G_{ik}\log m_{ik}]$$

where $L^* = \prod_{i=1}^{r}\prod_{k=1}^{K}(\pi_k m_{ik})^{G_{ik}}$. Of course, ℓ^* depends on λ. For fixed λ, one proceeds via standard EM, with E-step $w_{ik} = E(G_{ik}) = \frac{\pi_k m_{ik}}{\sum_\ell \pi_\ell m_{i\ell}}$. In the M-step, the expected complete likelihood is maximized, yielding

$$\hat\beta^{(\lambda)} = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}x_{ij}^T\right)^{-1} \sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}\left(y_{ij}^{(\lambda)} - \sum_{k=1}^{K} w_{ik}\hat z_k\right),$$

$$\hat\sigma^{2(\lambda)} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{K} w_{ik}\left[\sum_{j=1}^{n_i}(y_{ij}^{(\lambda)} - x_{ij}^T\hat\beta - \hat z_k)^2\right]}{\sum_{i=1}^{r} n_i},$$

$$\hat z_k^{(\lambda)} = \frac{\sum_{i=1}^{r} w_{ik}\left[\sum_{j=1}^{n_i}(y_{ij}^{(\lambda)} - x_{ij}^T\hat\beta)\right]}{\sum_{i=1}^{r} n_i w_{ik}}, \quad \hat\pi_k^{(\lambda)} = \frac{\sum_{i=1}^{r} w_{ik}}{r}.$$

Replacing the results into m_{ik} and then into (2), we get the non–parametric profile likelihood function $L_P(\lambda)$, or its logarithmic version $\ell_P(\lambda) = \log(L_P(\lambda))$. The non–parametric profile maximum likelihood (NPPML) estimator is therefore given by

$$\hat\lambda = arg\ \max_{\lambda}\ \ell_P(\lambda),$$

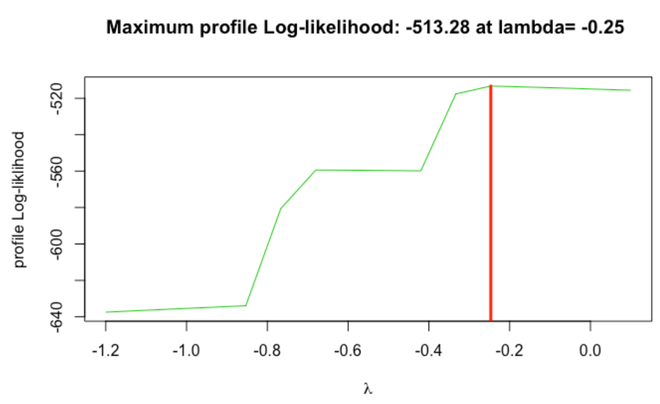
which can be found through a grid search over λ.

3. Example: Oxboys data

We consider a data set available as part of the **R** package **nlme** [4], which consists of measurements of `height` (cm) of 26 boys in Oxford at 9 standardized `age` measurements, yielding a total of 234 observations. We fitted a variance component model

$$E(y_{ij}|z_i) = \mathtt{age}_j + z_i$$

where z_i is boy–specific random effect and $\mathtt{age}_j$ is the j-th standardized age measurement (dimensionless, and equal for all boys for fixed j).



It can be seen that the best estimate of λ that maximizes $\ell_P(\lambda)$ is $\hat\lambda = -0.25$. The results before and after applying the response transformation are summarized in the table below. Comparing the Akaike Information Criterion (AIC) values of the untransformed model fit ($\lambda = 1$) and our method using $K = 5, 6$ and 7, respectively, shows a better performance of the NPPML approach. In other words, using the response after applying the transformation leads to a better fitting model than the original data.

	$K=5$		$K=6$		$K=7$	
	$\hat\lambda = -0.51$	$\lambda = 1$	$\hat\lambda = -0.25$	$\lambda = 1$	$\hat\lambda = -0.25$	$\lambda = 1$
$-2\log L$	1119.3	1132.8	1026.2	1048.3	1024.2	1132.8
AIC	1141.3	1154.9	1052.2	1074.3	1054.2	1162.9

4. Simulation Study

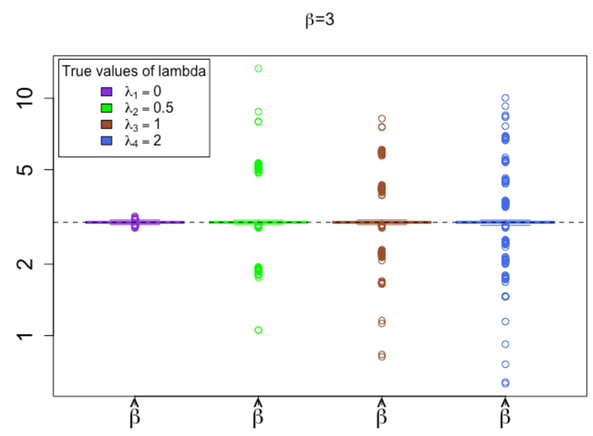
We are interested in examining the method's ability to estimate the true parameter values. Therefore, we first simulate data by applying the Box–Cox transformation 'backwards' to a dataset that follows a normal distribution using a set of λ values. Specifically, for each of four given values λ_ℓ, $\ell = 1, 2, 3, 4$, we generate 1000 datasets with 100 observations as follows,

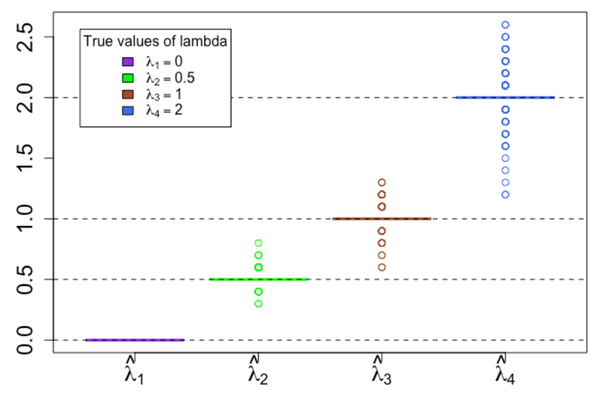
$$\begin{aligned}
\zeta_{ij\ell} &= \tilde y(\eta_{ij}, \lambda_\ell),\ \ i = 1, ..., 20,\ j = 1, ..., 5 \\
\tilde y(\eta_{ij}, \lambda_\ell) &= \begin{cases} \left(1 + \lambda_\ell\eta_{ij}\right)^{1/\lambda_\ell} & (\lambda_\ell \neq 0), \\ e^{\eta_{ij}} & (\lambda_\ell = 0) \end{cases} \\
\eta_{ij} &= 3\,x_{ij} + z_i + \varepsilon_{ij} \\
x_{ij} &\sim U(-4, 4), \quad \varepsilon_{ij} \sim N(0, 0.5^2) \\
\lambda_1 &= 0,\ \lambda_2 = 0.5,\ \lambda_3 = 1,\ \lambda_4 = 2 \\
z_i &\sim \text{Multinomial}\{1, (z_1, \ldots, z_4)|\pi_1, \ldots, \pi_4\} \\
z_k &= (15, 20, 30, 35) \text{ with masses } \pi_k = 1/4, k = 1, ..., 4.
\end{aligned}$$

Note that $\tilde y(\cdot)$ denotes the 'backward' Box–Cox–transformation, and that the generated data possess a variance component structure due to the random effect terms z_i.

We estimate λ and β simultaneously, yielding for each (true) value of λ a total of 1000 estimates of $\hat\lambda$ and $\hat\beta$. The figure below shows the boxplots for the regression and transformation parameter estimates, respectively. The reference lines in the figures indicate the actual values of the parameters. It is clear that the median of the estimated β and λ is approximately equal to the true value in each plot. There are some outliers in each of the plots; in fact the outliers in the transformation estimates cause the outliers in the regression estimates as they shift the scale of the linear predictor. The means and medians of the estimated parameters are provided in the table below; we see that the medians for the transformation parameters sit exactly at their true values, and those of the regression parameters approximately so.

We also investigate standard errors. An empirical but robust measure of spread can be obtained by computing the IQR of (the non–logarithmic version of) each of the four empirical distributions of $\hat\beta$. Via normal reference, the IQR can be mapped back to the scale of the standard deviations by division through 1.349. We call the resulting robust estimate of standard deviation RESD($\hat\beta$). We display RESD($\hat\beta$) values along with means and medians of EM–based standard errors, $SE(\hat\beta)$, obtained by extraction from the model fitted in the last M–step. It is conceptually clear that EM–based standard errors cannot be 'correct' as they ignore the variation caused by the EM algorithm itself, but we see that they are still satisfyingly close to their empirical counterparts.





True values	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
β	3	3	3	3
Mean($\hat\lambda$)	0	0.5026	1.003	2.0049
Median($\hat\lambda$)	0	0.5	1	2
Mean($\hat\beta$)	2.9996	3.0901	3.0770	3.1090
Median($\hat\beta$)	3.0003	3.0001	3.0003	3.0006
RESD($\hat\beta$)	0.0246	0.0251	0.0255	0.0335
Mean($\hat{SE}(\hat\beta)$)	0.0256	0.0267	0.0264	0.0268
Median($\hat{SE}(\hat\beta)$)	0.0214	0.0214	0.0214	0.0214

5. Implementation

The methodology is implemented in **R** package **boxcoxmix** [2] which is available on CRAN. This package features further variants and capabilities which have not been introduced here, such as a version for simple 'overdispersion' models (where $n_i \equiv 1$), and several routines to select the starting points for the EM algorithm.

References

[1] Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2009). *Statistical Modelling in R*. Oxford: University Press.

[2] Almohaimeed, A. and Einbeck, J. (2017). boxcoxmix: Response transformations for random effect and variance component models. URL `https://CRAN.R-project.org/package=boxcoxmix`

[3] Box, G.E. and Cox, D. (1964). *Journal of the Royal Statistical Society. Series B*, **26**, 211 – 252.

[4] Pinheiro, D. et al (2017). nlme: Linear and nonlinear mixed effect models. URL `https://CRAN.R-project.org/package=nlme`