Supplementary Document for Prediction, Bootstrapping and Monte Carlo Analyses Based on Linear Mixed Models with QAPE 2.0 Package

by Alicja Wolny–Dominiak and Tomasz Żądło

Abstract The paper presents a new R package **qape** for prediction, accuracy estimation of various predictors and Monte Carlo simulation studies of properties of both predictors and estimators of accuracy measures.

The Monte Carlo procedure to compute EBP

The Monte Carlo procedure to compute EBP according to Molina and Rao (2010) used in `ebpLMMne()` is presented.

1. $\boldsymbol{\psi}$ is estimated based on sample data and estimator $\hat{\boldsymbol{\psi}}$ is obtained.
2. Using the distribution function of $\mathbf{Y}_r|\mathbf{Y}_s$, whose functional form is assumed to be known, and where $\boldsymbol{\psi}$ is replaced by $\hat{\boldsymbol{\psi}}$, L vectors $\mathbf{Y}_r$ are generated of unobserved values of the dependent variable, denoted by $\mathbf{Y}_r^{(l)}$ (where $l = 1, 2, ..., L$).
3. L population vectors are built based on one subvector of the dependent variables observed in the sample and L subvectors of unobserved values of the dependent variable generated in the previous step. The result is: $\mathbf{Y}^{(l)} = \left[\mathbf{Y}_s^T \mathbf{Y}_r^{(l)T}\right]^T$, where $l = 1, 2, ..., L$.
4. The EBP value is computed as follows: $\hat{\theta}_{EBP} = L^{-1} \sum_{l=1}^{L} \theta(\mathbf{Y}^{(l)})$. If the LMM is not assumed for the original variable of interest but for its transformation $T(.)$, the back-transformation is used additionally: $\hat{\theta}_{EBP} = L^{-1} \sum_{l=1}^{L} \theta(T^{-1}(\mathbf{Y}^{(l)}))$.

It is worth nothing, that if the distribution of $\mathbf{Y}$ is multivariate normal, then the distribution of $\mathbf{Y}_r|\mathbf{Y}_s$ (used in step (ii) above) is also multivariate normal, which means the generation process of L population vectors in the algorithm presented above is very time-consuming in real-life surveys. Therefore, the EBP is considered under the special case of the LMM, which makes it possible to accelerate the algorithm by generating $\mathbf{Y}_r^{(l)}$, $l = 1, 2, ..., L$, not from the multivariate normal distribution but using the univariate normal distribution. The model, called the nested error LMM, is given by:

$$\mathbf{Y}_k = \mathbf{X}_k \boldsymbol{\beta} + v_k \mathbf{1}_{N_k} + \mathbf{e}_k, \tag{1}$$

where $k = 1, 2, ..., K$ and $\mathbf{1}_{N_k}$ is a vector of ones of size $N_k \times 1$, v_k is a random effect, such that v_k are independent for $k = 1, 2, ..., K$, $\mathbf{e}_k$ $(N_k \times 1)$ is a vector of random components. Let us additionally assume that $v_k \sim N(0, \sigma_v^2)$ and $\mathbf{e}_k \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_{N_k})$. Let the number of elements of $\mathbf{Y}_k$ observed in the sample be denoted by n_k. Under (1), step (ii) of the above procedure is as follows (cf. Molina and Rao (2010) p. 375):

- for k where $n_k > 0$ we generate $\mathbf{Y}_r^{(l)} = \left[\mathbf{Y}_{r1}^{(l)T} \quad ... \quad \mathbf{Y}_{rk}^{(l)T} \quad ... \quad \mathbf{Y}_{rK}^{(l)T}\right]^T$, $l = 1, 2, ..., L$, based on the following model:
$$\mathbf{Y}_{rk} = \boldsymbol{\mu}_{rk|sk} + u_k \mathbf{1}_{N_k - n_k} + \boldsymbol{\varepsilon}_{rk}, \tag{2}$$
where $\boldsymbol{\mu}_{rk|sk} = \mathbf{X}_{rk}\hat{\boldsymbol{\beta}} + \hat{\sigma}_v^2 \mathbf{1}_{N_k - n_k} \mathbf{1}_{n_k}^T \left(\hat{\sigma}_v^2 \mathbf{1}_{n_k} \mathbf{1}_{n_k}^T + \hat{\sigma}_e^2 \mathbf{I}_{n_k}\right)^{-1} (\mathbf{Y}_{sk} - \mathbf{X}_{sk}\hat{\boldsymbol{\beta}})$, u_k and $\boldsymbol{\varepsilon}_{rk}$ are independent, $\boldsymbol{\varepsilon}_{rk} \sim N(\mathbf{0}, \hat{\sigma}_e^2 \mathbf{I}_{N_k - n_k})$, $u_k \sim N(0, \hat{\sigma}_v^2(1 - \omega_k))$, $\omega_k = \hat{\sigma}_v^2 (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 n_k^{-1})^{-1}$,
- for k where $n_k = 0$ we generate $\mathbf{Y}_r^{(l)}$, $l = 1, 2, ..., L$, based on (1), where unknown parameters are replaced with estimates.

The correction procedure in residual bootstrap

This appendix presents the correction procedure according to Carpenter et al. (2003), Chambers and Chandra (2013) and Thai et al. (2013), which can be used in the residual bootstrap to avoid the problem of underdispersion of the classic residual bootstrap distribution.

Let us consider the lth vector of random effects in Equation (5) given by Equation (6), both presented in the paper. Let $\boldsymbol{G}_l$ be the variance-covariance matrix of size $K_l \times K_l$ defined as $\boldsymbol{G}_l = Var\left(\left[v_{l1j} \dots v_{lkj} \dots v_{lK_lj}\right]^T\right)$, where v_{lkj} is the jth element of $\mathbf{v}_{lk}$. Let the estimated (e.g. using restricted maximum likelihood method) matrix $\boldsymbol{G}_l$ be denoted by $\hat{\boldsymbol{G}}_l$ and the empirical covariance matrix of size $K_l \times K_l$ be defined as follows $\boldsymbol{G}_{(emp)l} = J_l^{-1} \begin{bmatrix} \hat{\mathbf{v}}_{l1}^T \\ \dots \\ \hat{\mathbf{v}}_{lk}^T \\ \dots \\ \hat{\mathbf{v}}_{lK_l}^T \end{bmatrix} \begin{bmatrix} \hat{\mathbf{v}}_{l1}^T \\ \dots \\ \hat{\mathbf{v}}_{lk}^T \\ \dots \\ \hat{\mathbf{v}}_{lK_l}^T \end{bmatrix}^T$, where $\hat{\mathbf{v}}_{lk}$ are the estimated best linear unbiased predictors of $\mathbf{v}_{lk}$.

Let us write the estimated and the empirical covariance matrices using the Cholesky decomposition, in terms of a lower triangular matrix, as $\hat{\boldsymbol{G}}_l = \mathbf{L}_{(est)l}\mathbf{L}_{(est)l}^T$ and $\boldsymbol{G}_{(emp)l} = \mathbf{L}_{(emp)l}\mathbf{L}_{(emp)l}^T$. Let $\mathbf{A}_l = (\mathbf{L}_{(est)l}\mathbf{L}_{(emp)l}^{-1})^T$. Let us define the corrected estimates of $\mathbf{v}_l$ as (Carpenter et al. (2003), Thai et al. (2013)): $\hat{\mathbf{v}}_{(cor)l} = \hat{\mathbf{v}}_l\mathbf{A}_l$, where $\hat{\mathbf{v}}$ is the empirical best linear unbiased predictor of $\mathbf{v}$. Let us additionally assume that $Var(\mathbf{e}) = \mathbf{R} = \sigma_e^2 diag_{1\leq i\leq N}(d_i)$, where d_i values are known weights. The corrected residuals are as follows (Chambers and Chandra (2013)): $\hat{e}_{(cor)i} = \hat{\sigma}_e\sqrt{d_i}\hat{e}_i(n^{-1}\sum_{k=1}^n \hat{e}_i)^{-0.5}$, where $i = 1, 2, ..., n$, $\hat{\sigma}_e^2$ is the estimate (e.g. REML estimate) of σ_e^2, $\hat{e}_i$ are residuals computed under the model given by Equation (5) in the paper.

Replacing $\hat{\mathbf{v}}_l$ and $\hat{e}_i$ by specified above $\hat{\mathbf{v}}_{(cor)l}$ and $\hat{e}_{(cor)i}$, respectively, the corrected version of the residual bootstrap procedure is obtained.

Bibliography

J. R. Carpenter, H. Goldstein, and J. Rasbash. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):431–443, 2003. [p2]

R. Chambers and H. Chandra. A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics*, 22(2):452–470, 2013. [p2]

I. Molina and J. Rao. Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3): 369–385, 2010. [p1]

H.-T. Thai, F. Mentré, N. H. Holford, C. Veyrat-Follet, and E. Comets. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical statistics*, 12(3):129–140, 2013. [p2]

Alicja Wolny–Dominiak
Department of Statistical and Mathematical Methods in Economics
University of Economics in Katowice
50, 1 Maja Street
40–287 Katowice
Poland
alicja.wolny-dominiak@uekat.pl
web.ue.katowice.pl/woali/

Tomasz Żądło
Department of Statistics, Econometrics and Mathematics
University of Economics in Katowice
50, 1 Maja Street
40–287 Katowice
Poland
tomasz.zadlo@uekat.pl
web.ue.katowice.pl/zadlo/