Universität Trier

Data-driven transformations and survey-weighting for linear mixed models

Patricia Dörr
Jan Pablo Burgard

Research Papers in Economics
No. 16/19

Data-driven transformations and survey-weighting for linear mixed models

Patricia Dörr* and Jan Pablo Burgard*

*Economic and Social Statistics Department, Trier University, Germany

December 2019

Abstract

Many variables that social and economic researchers seek to analyze through regression analysis violate normality assumptions. A standard remedy in that case is the logarithmic transformation. However, taking logarithms is not always sufficient to restablish model assumptions. A more general approach is to determine a family of transformations and to estimate the adequate parameter of such a transformation. This can also be done in mixed effects models, which can account for unobserved heterogeneity in grouped data.

When the analyzed data is gathered from a complex survey whose design is informative for the model - which is difficult to exclude a priori - a bias on the transformed linear mixed models can occur. As the bias affects the transformation parameter, too, the distortion to the parameters in the population is even more problematic than in standard regression.

In standard regression, survey weights are used to account for the design. To the best of our knowledge, none of the existing algorithms allows to include survey weights in these transformed linear mixed models. This paper adapts a recently suggested algorithm to include survey weights to Box-Cox or dual transformed mixed models. A simulation study demonstrates the need to account for informative survey design.

1 Introduction

Already in the 1960s, Box and Cox [1964] found that transformations of dependent variables help to meet the assumptions of regression modeling, in particular the conditional normality assumption. They introduced the so called *Box-Cox transformations*, a family of functions from which one is chosen for the transformation of the dependent variable. Using maximum likelihood (ML) estimation, the optimal choice of the transformation parameter is an optimization problem.

Gurka et al. [2006] extended that framework from linear models to linear mixed models (LMMs). Rojas-Perilla et al. [2017] used the methodology in a popular

application of mixed models, the small area estimation (SAE) framework. Both authors demonstrate the applicability of the data transformations on mixed models on real world examples. While Gurka et al. [2006] use a data from a clinical study of pulmonary disease, which might not be subject to extrem survey designs, Rojas-Perilla et al. [2017] use data of the consumer and expenditure survey in Mexico. Even after conditioning on the explanatories, income and wealth data are seldomly normally distributed, thus requiring transformation for an adequate regression analysis. In addition, national population surveys rarely rely on a simple random sample of the population, there is rather a complex survey design with oversampling of subgroups of interest. For the estimation of summary statistics, the design is easily accounted for by the use of survey weights. In regression analysis, this is more difficult as we will outline in the next section. On the other hand, the assumptions, under which the design can be ignored [Pfeffermann, 1993], are often too restrictive to be assumed valid in real data application. Therefore, there is need to develop an algorithm that can take into account the survey design, especially in the case of (non-normal) mixed models. Burgard and Dörr [2018] have introduced a MCEM-algorithm that does so for generalized linear mixed models. This algorithm can be adapted to data-driven transformations of the dependent variable as well. However, problems due to the transformation occur and must be accounted for.

In the next section, the statistical problem is formulated, both under classical statistical modeling and later on accounting for the survey design as well. Then, the optimization problem is outlined and the adjusted MCEM algorithm described. Following this section, the esitmation of standard errors is briefly discussed. The algorithm is then checked under several simulation scenarios: First, its reliability under the statistical model is confirmed replicating the simulation study from Gurka et al. [2006]. The need for survey-weighting will be demonstrated using another simulation study leaned on Burgard and Dörr [2018]. The final section concludes.

2 Statistical formulation of the model under survey sampling

To each unit i of a finite population $U = \{1, \dots, N\}$, a vector of characteristics $(y_i, \mathbf{x}_i^T, \mathbf{z}_i^T)^T \in \mathbb{R}^{1+p+q}$ is attributed. The value y_i is considered to be a realization of the following statistical model:

$$\tilde{Y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T G + \varepsilon_i \tag{2.1}$$

$$Y_i = h(\tilde{Y}_i; \lambda) \tag{2.2}$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad \forall i \in U \tag{2.3}$$

$$G \sim N(0, \Sigma) \quad . \tag{2.4}$$

Like Box and Cox [1964], Gurka et al. [2006] and Rojas-Perilla et al. [2017], we assume that the family of (back-)transformations, $\{h(\cdot, \lambda), \lambda \in \Lambda\}$ is known a priori although the specific value of λ is not. Note that this set-up is to some

extent more general than the set-up given in Gurka et al. [2006] and Rojas-Perilla et al. [2017] because we do not require the random variable G to have a diagonal covariance matrix Σ. Furthermore - though in a different context - this data generating process (DGP) - is also more general than that given in Burgard and Dörr [2018]: There, Y_i was assumed to follow a law from the exponential family and the link function h is explicitly known. G is referred to as random effect, ε is the idiosyncratic error and $\boldsymbol{\beta}$ is the fixed effects vector. Common families of transformation are the Box-Cox transformation [Box and Cox, 1964]

$$h^{-1}(\cdot;\lambda) : (0,\infty) \to \mathbb{R}, \quad h^{-1}(y;\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, \text{ if } \lambda \neq 0 \\ \log y, \text{ if } \lambda = 0 \end{cases} \qquad (\lambda \in \mathbb{R}) \ . \tag{2.5}$$

Rojas-Perilla et al. [2017] also considers the dual transformation [Yang, 2006]

$$h : (0,\infty) \to \mathbb{R}, \quad h^{-1}(y,\lambda) = \begin{cases} \frac{y^\lambda - y^{-\lambda}}{2\lambda}, \text{ if } \lambda \neq 0 \\ \log y, \text{ if } \lambda = 0 \end{cases} \qquad (\lambda \in \mathbb{R}) \ . \tag{2.6}$$

However, both λ and $-\lambda$ yield the same transformation. We can thus restrict the transformation to $\lambda \geq 0$.

For ease of notation, we summarize $Y := (Y_1, \ldots, Y_N)$ and $\mathbf{y}$ describes a realization of Y. For a random effects realization $G = \boldsymbol{\gamma}$, we define $\eta_i := \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{\gamma}$. In addition, denote by $\boldsymbol{\rho}$ the distinct elements of the covariance matrix Σ. For some realizations $G = \boldsymbol{\gamma}$, $\tilde{Y}_i = \tilde{y}_i$ (and in consequence, $Y_i = y_i$, $y_i = h(\tilde{y}_i, \lambda)$), $i \in U$, the statistical model formulated by (2.1) to (2.4) leads thus to the conditional density of $\tilde{y}_i$ and y_i:

$$f_{\tilde{Y}}(\tilde{y}_i, \boldsymbol{\gamma}; \boldsymbol{\psi}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{1}{2\sigma^2}(\tilde{y}_i - \eta_i)^2\right) \cdot f_G(\boldsymbol{\gamma}; \boldsymbol{\rho}) \tag{2.7}$$

and

$$f_Y(y_i, \boldsymbol{\gamma}; \boldsymbol{\psi}, \lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{1}{2\sigma^2}(\tilde{y}_i - \eta_i)^2\right) \cdot \frac{\mathrm{d}\, h^{-1}(y_i, \lambda)}{\mathrm{d}\, y_i} \cdot f_G(\boldsymbol{\gamma}; \boldsymbol{\rho}) \tag{2.8}$$

The joint log-likelihood of the observed data and the random effect realization is thus

$$\begin{aligned} \mathcal{LL}(\mathbf{y}, \boldsymbol{\gamma}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) = & \underbrace{-N \log\left(\sqrt{2\pi}\right)}_{=:A} - \frac{N}{2}\log(\sigma^2) \underbrace{- \log\left(\det \Sigma\right) - \frac{1}{2}\boldsymbol{\gamma}^T\Sigma^{-1}\boldsymbol{\gamma}}_{=:B} + \\ & \underbrace{\sum_{i=1}^{N} \frac{-(\tilde{y}_i - \eta_i)^2}{2\sigma^2} + \log\left(\frac{\mathrm{d}\, h^{-1}(y_i, \lambda)}{\mathrm{d}\, y_i}\right)}_{=:C} \end{aligned} \tag{2.9}$$

Note that term C has the shape of a total of the finite population. If only a sample $s \subset U$ of the units is available, where s is the realization of a random sample S with probability law $S \sim P_D$ ('survey design'), a common estimator for a finite population total is the Horvitz-Thompson estimator (HT) [Horvitz

and Thompson, 1952]. If all groups, to whom an element of $\boldsymbol{\gamma}$ is attributed, are represented in s, the HT of the joint likelihood (2.9) is

$$\widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) = -\frac{\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i}{2} \log(\sigma^2) - \log\left(\det \Sigma\right) - \frac{1}{2}\boldsymbol{\gamma}^T \Sigma^{-1} \boldsymbol{\gamma} + \sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \left(\frac{-(\tilde{y}_i - \eta_i)^2}{2\sigma^2} + \log\left(\frac{\partial\, h^{-1}(y_i, \lambda)}{\partial\, y_i} \right) \right) + A \quad . \tag{2.10}$$

where w_i are the design-weights, i.e. the inverse inclusion probability $w_i = P_D(i \in S)^{-1}$. The HT is known to be design-consistent under fairly general conditions on P_D [Chauvet, 2014]. If not every element of $\boldsymbol{\gamma}$ has an attributed unit in the sample s, the HT estimator in (2.9) cannot be evaluated.

In that case, if a plug-in version of (2.9) is used, the term $\hat{C}$ (cf. Equation 2.12) scales up to a population estimator whereas the subvector $\boldsymbol{\gamma}_s$ and the corresponding covariance matrix Σ_s are only representants of a hypothetic population of size $|s| =: n$. This means, the plug-in version for B, B_s (Equation 2.11), has not an adequate size relative to $\hat{C}$. In that case, the design weights should be rescaled such that they sum up to n instead of N. In that way, the size relation of terms

$$B_s := -\log\det \Sigma_s - \frac{1}{2}\boldsymbol{\gamma}_s^T \Sigma_s^{-1} \boldsymbol{\gamma}_s \tag{2.11}$$

and

$$\hat{C} := \sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \left(\frac{-(\tilde{y}_i - \eta_i)^2}{2\sigma^2} + \log\left(\frac{\partial\, h^{-1}(y_i, \lambda)}{\partial\, y_i} \right) \right) \tag{2.12}$$

for the S-sample joint log-likelihood is the same as the relation between B and C in the finite population log-likelihood. In the following, we assume that the design weights w_i are appropriately scaled and do not differentiate between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_s$.

3 Maximization of the Likelihood

3.1 EM-Algorithm

Remember that $\boldsymbol{\gamma}$ is not observable. A common solution for only partially observed data is the EM-algorithm [Dempster et al., 1977], that is optimization of $\mathrm{E}_G\left(\widehat{\mathcal{LL}}(\mathbf{y}, S; \boldsymbol{\rho}, \lambda) | \mathbf{y}, S, \boldsymbol{\rho}_k, \lambda_k\right)$ and then iteratively setting the maximizer $(\boldsymbol{\beta}_{k+1}, \sigma^2_{k+1}, \boldsymbol{\rho}_{k+1}, \lambda_{k+1})$ instead of its predecessor. We get the expected HT

log-likelihood conditional on some $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\tilde{\boldsymbol{\rho}}, \tilde{\lambda})$ and $\mathbf{y}$ as

$$\begin{aligned}
&\mathrm{E}_G\left(\widehat{\mathcal{LL}}(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)|\mathbf{y}, S, \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \tilde{\boldsymbol{\rho}}, \tilde{\lambda}\right) = \\
&-\frac{1}{2}\int_{\mathbb{R}^q}\left(\left(\sum_{i=1}^{N}\mathbb{1}_S(i)\cdot w_i\right)\log(\hat{\sigma}^2_\lambda) + \boldsymbol{\gamma}^T\Sigma^{-1}\boldsymbol{\gamma}\right)\cdot m(\boldsymbol{\gamma}|\mathbf{y}, S; \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \tilde{\boldsymbol{\rho}}, \tilde{\lambda})\,\mathrm{d}\,\boldsymbol{\gamma} + \\
&\sum_{i=1}^{N}\mathbb{1}_S(i)\cdot w_i\cdot\log\left(\frac{\partial h^{-1}(y_i, \lambda)}{\partial y_i}\right) - \frac{1}{2}\log\left(\det\Sigma(\boldsymbol{\rho})\right) + a \quad , \qquad (3.1)
\end{aligned}$$

where m denotes the conditional density of $\boldsymbol{\gamma}$ given the observable data $\mathbf{y}$ in sample S. The EM-algorithm for the set-up of linear mixed models with data transformation is summarized in Algorithm 3.1. Note that the (expected) con-

Algorithm 3.1 EM-Algorithm for LMMs under Transformation

Require: Start values $\boldsymbol{\psi}_0$, λ_0, $k = 0$
  while Convergence criterion is not met **do**
    Calculate $\mathrm{E}_G\left(\widehat{\mathcal{LL}}(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2\boldsymbol{\rho}, \lambda)|\mathbf{y}, \boldsymbol{\beta}_k, \sigma^2_k, \boldsymbol{\rho}_k, \lambda_k\right)$
    Maximize $\mathrm{E}_G\left(\widehat{\mathcal{LL}}(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)|\mathbf{y}, \boldsymbol{\beta}_k, \sigma^2_k, \boldsymbol{\rho}_k, \lambda_k\right)$ and set

$$(\boldsymbol{\beta}_{k+1}, \sigma^2_{k+1}, \boldsymbol{\rho}_{k+1}, \lambda_{k+1}) = \arg\max \mathrm{E}_G\left(\widehat{\mathcal{LL}}(\mathbf{y}, S; \boldsymbol{\rho}, \lambda)|\mathbf{y}, S, \boldsymbol{\beta}_k, \sigma^2_k, \boldsymbol{\rho}_k, \lambda_k\right)$$

    $k \leftarrow k + 1$
  end while

centrated log-likelihood is separable in the parameters $\boldsymbol{\rho}$ and λ, which simplifies the maximization.

The EM-algorithm is known to converge to a stationary point of the log-likelihood. Neither the finite population log-likelihood nor its sample analogue (2.10) are strictly concave in λ, thus there are several local minima possible. The EM-algorithm thus must be started with several different initial vectors $(\boldsymbol{\rho}_0, \lambda_0)$ in order to assure global optimality. Furthermore, Gurka et al. [2006] and Sugasawa et al. [2015] state for the Box-Cox transformation that h^{-1} only achieves approximate normality (even without recurring to a sampling process P_D) because the transformation does not map into the complete real line unless $\lambda = 0$. Instead of the ML estimator, Sugasawa et al. [2015] recommends a moment estimator that reduces the skewness of the residuals which should equal zero under the model assumptions. Other alternative estimators to maximum likelihood (ML) for λ are discussed in Rojas-Perilla et al. [2017]. However, their software implementation becomes more difficult (in contrast to that of Sugasawa et al. [2015]).

3.2 Monte-Carlo Integration

The integral in Equation (3.1) is hard to evaluate analytically. Therefore, an alternative is Monte-Carlo (MC) integration: In each E-step, the random effect

G is drawn from the distribution M with density m several times, plugged into the joint HT log-likelihood (2.10) and then the outcomes of $\widehat{\mathcal{LL}}$ are averaged. This procedure has been proposed for generalized LMMs in McCulloch [1997], Booth and Hobert [1999] and – for the case of survey sampling – in Burgard and Dörr [2018]. This leads to the MCEM Algorithm 3.2. The number of

Algorithm 3.2 MCEM-Algorithm for LMMs under Transformation

Require: Start values $\boldsymbol{\psi}_0$, λ_0, $k = 0$, $B_0 \in \mathbb{N}$
  while Convergence criterion is not met **do**
    Sample B_k times $G \sim M(\cdot|\mathbf{y}, S; \boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{\rho}_k, \lambda_k)$,
    denote by $\boldsymbol{\gamma}_b$ the b-th realization.
    Calculate

$$\widehat{\mathcal{LL}}_k^{MC}\left(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right) := \frac{1}{B_k} \sum_{b=1}^{B_k} \widehat{\mathcal{LL}}\left(\mathbf{y}, \boldsymbol{\gamma}_b, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right)$$

    Maximize $\widehat{\mathcal{LL}}_k^{MC}\left(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right)$ and set

$$(\boldsymbol{\beta}_{k+1}, \sigma_{k+1}^2, \boldsymbol{\rho}_{k+1}, \lambda_{k+1}) = \arg\max \widehat{\mathcal{LL}}_k^{MC}\left(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right)$$

    Increase $B_{k+1} > B_k$
    $k \leftarrow k + 1$
  end while

MC samples B_k must increase during the estimation process in order to assure convergence of the MCEM toward the traditional EM algorithm [Booth and Hobert, 1999, Neath et al., 2013] Sampling from M is not trivial, though. Given Y, the random effects are not normally distributed anymore. The conditioning on S complicates this step even more. Hence, we discuss MC-sampling in order to approximate M in the following.

3.3 Importance Sampling

According to Bayes' Theorem, we have that m equals the joint density of Y and G divided by the marginal density of Y. As the random effects are not observable, any sampling design P_D in practice can only depend indirectly on $\boldsymbol{\gamma}$ via $\mathbf{y}$. In addition, the design may also depend on auxiliary information $\mathbf{x}$, $\mathbf{z}$ or others. These are omitted here for brevity. Theoretically, however, the survey sampling can also depend on $\boldsymbol{\gamma}$, which is a common way to design informativity in simulation studies [Pfeffermann et al., 1998, Rabe-Hesketh and Skrondal, 2006]. In that case, the following analysis can only be considered to be approximative. We note for $P_D(S|\mathbf{y}, \boldsymbol{\gamma}) = P_D(S|\mathbf{y})$ that the marginal density

of Y and S,

$$f(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) = P_D(S|\mathbf{y}) \cdot \int_{\mathbb{R}^q} \frac{1}{\sqrt{2\pi\sigma^2}^N} \cdot \frac{1}{\sqrt{2\pi \det \Sigma(\boldsymbol{\rho})}}$$
$$\cdot \exp\left(-\frac{1}{2}\boldsymbol{\gamma}^T \Sigma(\boldsymbol{\rho})^{-1}\boldsymbol{\gamma}\right) \cdot \exp\left(\sum_{i=1}^{N} \frac{-(\tilde{y}_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{z}_i^T\boldsymbol{\gamma})^2}{2\sigma^2} + \log \frac{\partial h^{-1}(y_i, \lambda)}{\partial y_i}\right) \mathrm{d}\,\boldsymbol{\gamma}$$
$$=: P_D(S|\mathbf{y}) \cdot f_Y(\mathbf{y}, \boldsymbol{\gamma}, U; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) \tag{3.2}$$

can be splitted like above and thus $P_D(S|\mathbf{y})$ cancels out in

$$m(\boldsymbol{\gamma}|\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) = \frac{f_Y(\mathbf{y}, \boldsymbol{\gamma}, U; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)}{f(\mathbf{y}, U; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)} \tag{3.3}$$
$$\propto \exp\left(\log f_Y(\mathbf{y}, \boldsymbol{\gamma}, U; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)\right) \quad . \tag{3.4}$$

The exponent in Equation (3.4) is estimated design-consistently through the HT log-likelihood (2.10). Up to a normalizing constant,

$$\hat{m}(\boldsymbol{\gamma}|\mathbf{y}, S, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) \propto \exp\left(\widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)\right) \tag{3.5}$$

is a design-consistent estimator of m. This estimator allows us to contrast random realizations $\boldsymbol{\gamma}_b$, $b = 1, \ldots, B$ from a proposal distribution to be discussed in the next paragraph with an estimator of the true sampling distribution m.

Furthermore, we can approximate the density of $\boldsymbol{\gamma}$ around its mode $\boldsymbol{\gamma}_0$ due to the second order Taylor approximation

$$\widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) \approx \widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_0, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) -$$
$$\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T(-H^{-1})^{-1}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \tag{3.6}$$

where H is the Hessian of $\widehat{\mathcal{LL}}$ evaluated at $\boldsymbol{\gamma}_0$. Thus, around $\boldsymbol{\gamma}_0$, G is approximately normal with mean $\boldsymbol{\gamma}_0$ and covariance matrix H. This means that a possible proposal distribution for the importance sampling is $N(\boldsymbol{\gamma}_0, -H^{-1})$ [Pinheiro and Bates, 1995]. Note that the mode and the Hessian for the proposal need to be updated in each E-step k due to the dependence on the current parameter estimates $(\boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{\rho}_k, \lambda_k)$, see the sample distribution in Algorithm 3.2.

The importance weight ω_b^{MC} of random sample b can then be approximated by

$$\tilde{\omega}_b^{MC} = \frac{\exp\left(\widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_b, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)\right)}{\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_b - \boldsymbol{\gamma}_0)^T(-H)(\boldsymbol{\gamma}_b - \boldsymbol{\gamma}_0)\right)} \tag{3.7}$$

$$\omega_b^{MC} = \frac{\tilde{\omega}_b^{MC}}{\sum_{b=1}^{B} \tilde{\omega}_b^{MC}} \tag{3.8}$$

These are the self-weighted importance weights in order to circumvent the calculus of the normalizing constant [Owen, 2013]. Self-weighting yields cancellation of $\widehat{\mathcal{LL}}$-components that do not depend on $\boldsymbol{\gamma}$, such as $\det \Sigma$ and $\sum_{i=1}^{N} \mathbb{1}_S(i) w_i \frac{\partial h^{-1}(y_i;\lambda)}{\partial y_i}$, which reduces the computational effort.

The MC-estimator of Equation (3.1) is thus

$$\hat{E}_G\left(\widehat{\mathcal{LL}}(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)|\mathbf{y}, S, \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \tilde{\boldsymbol{\rho}}, \tilde{\lambda}\right) = \frac{1}{B}\sum_{b=1}^{B} \omega_b^{MC}\, \widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_b, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) \tag{3.9}$$

The MCEM algorithm under importance sampling is summarized in Algorithm 3.3. For the optimization with respect to $\boldsymbol{\gamma}_{0,k}$ via the Newton algorithm, though,

Algorithm 3.3 MCEM-Algorithm using Importance Sampling for LMMs under Transformation

Require: Start values $\boldsymbol{\beta}_0$, σ_0^2, $\boldsymbol{\rho}_0$, λ_0, $k = 0$, $B_0 \in \mathbb{N}$
while Convergence criterion is not met **do**
Find

$$\boldsymbol{\gamma}_{0,k} = \arg\max_{\boldsymbol{\gamma}} \widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}, S; \boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{\rho}_k, \lambda_k)$$

Calculate H_k – the Hessian of $\widehat{\mathcal{LL}}$ evaluated at $\boldsymbol{\gamma}_{0,k}$ and $\boldsymbol{\beta_k}, \sigma_k^2, \boldsymbol{\rho}_k, \lambda_k$
Sample B_k times $G \sim N(\boldsymbol{\gamma}_{0,k}, -H_k)$,
denote by $\boldsymbol{\gamma}_b$ the b-th realization.
Calculate the importance weights

$$\tilde{\omega}_b^{MC} = \frac{\exp\left(\widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_b, S; \boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{\rho}_k, \lambda_k)\right)}{\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_b - \boldsymbol{\gamma}_{0,k})^T(-H_k)(\boldsymbol{\gamma}_b - \boldsymbol{\gamma}_{0,k})\right)}$$
$$\omega_b^{MC} = \frac{\tilde{\omega}_b^{MC}}{\sum_{b=1}^{B} \tilde{\omega}_b^{MC}}$$

Calculate

$$\hat{E}_G\left(\widehat{\mathcal{LL}}_k^{MC}\left(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right)\right) := \frac{1}{B_k}\sum_{b=1}^{B_k} \omega_b^{MC}\, \widehat{\mathcal{LL}}\left(\mathbf{y}, \boldsymbol{\gamma}_b, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right)$$

Maximize $\widehat{\mathcal{LL}}_k^{MC}\left(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right)$ and set

$$(\boldsymbol{\beta}_{k+1}, \sigma_{k+1}^2, \boldsymbol{\rho}_{k+1}, \lambda_{k+1}) = \arg\max \widehat{\mathcal{LL}}_k^{MC}\left(\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda\right)$$

Increase $B_{k+1} > B_k$
$k \leftarrow k + 1$
end while

the inverse Hessian in each optimization step would be needed. In addition, for the sampling of G, a square root of the Hessian would be necessary to multiply a standard normal vector by this matrix in order to get the correct covariance structure. The calculus of both is a computational burden. Therefore, we suggest like in Burgard and Dörr [2018], to use the approximate BFGS algorithm suggested in Powell [1987] that generates as a by product a matrix C such that $CC^T = -H^{-1}$.

The traditional EM-algorithm increases the likelihood in each iteration step. However, this property gets lost by the stochastic approximation of the integral in the E-step. An implementation of the suggested algorithm thus needs to track the estimates of $\hat{\tilde{E}}_G(\mathcal{LL})$ in order to assure that the random noise does not disturb to much the optimization process. Burgard and Dörr [2018] found that algorithm 3.3 (adapted to generalized linear mixed models) works reasonably well for the linear case. However, tracking of the iterations became necessary under the mixed logit regression due to noise.

3.4 Maximization

In general, maximization must be simulatneously over the complete parameter vector $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda)$. However, note that estimated expectation $\hat{\mathrm{E}}_G(\widehat{\mathcal{LL}})$ is separable into a random effect part and a fixed effects part. Thus, optimization over $\boldsymbol{\rho}$ can be separated from that of the other model parameters and we have

$$\hat{\Sigma}^{MC} = \sum_{b=1}^{B} \omega_b^{MC} \boldsymbol{\gamma}_b \boldsymbol{\gamma}_b^T \tag{3.10}$$

in any M-step. For the fixed effects component, we can use a concentrated likelihood approach that was discussed in Spitzer [1982] and Hyde [1999] because it might happen that the full likelihood estimation problem is not well conditioned. For the first order conditions for maxmimizing $\hat{\mathrm{E}}_G(\widehat{\mathcal{LL}})$, we get

$$\frac{\partial \hat{E}_G(\widehat{\mathcal{LL}})}{\partial \boldsymbol{\beta}} = \sum_{b=1}^{B} \omega_b^{MC} \sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \frac{\tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{z}_i^T \boldsymbol{\gamma}_b}{\sigma^2} \cdot \mathbf{x}_i = 0 \tag{3.11}$$

$$\begin{aligned} \frac{\partial \hat{E}_G(\widehat{\mathcal{LL}})}{\partial \sigma^2} = \sum_{b=1}^{B} \omega_b^{MC} \sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \frac{(\tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{z}_i^T \boldsymbol{\gamma}_b)^2}{2\,\sigma^4} - \\ \frac{\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i}{2} \frac{1}{\sigma^2} = 0 \end{aligned} \tag{3.12}$$

$$\begin{aligned} \frac{\partial \hat{E}_G(\widehat{\mathcal{LL}})}{\partial \lambda} = \sum_{b=1}^{B} \omega_b^{MC} \sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \frac{-(\tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{z}_i^T \boldsymbol{\gamma}_b)}{\sigma^2} \cdot \frac{\partial h^{-1}(y_i, \lambda)}{\partial \lambda} + \\ \sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \frac{\partial^2 h^{-1}(y_i; \lambda)}{\partial y_i \partial \lambda} = 0 \end{aligned} \tag{3.13}$$

Solving Equations (3.11) and (3.12) yield for given λ (noting that $\tilde{y}_i = h^{-1}(y_i; \lambda)$)

$$\hat{\boldsymbol{\beta}}_\lambda = \left(\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \mathbf{x}_i \cdot \left(\tilde{y}_i - \mathbf{z}_i^T \sum_{b=1}^{B} \omega_b^{MC} \boldsymbol{\gamma}_b \right) \right) \tag{3.14}$$

and

$$\hat{\sigma}_\lambda^2 = \sum_{b=1}^{B} \omega_b^{MC} \frac{\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot (\tilde{y}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda - \mathbf{z}_i^T \boldsymbol{\gamma}_b)^2}{\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i} \quad . \tag{3.15}$$

This gives the concentrated MC-integrated log-likelihood

$$\hat{\mathrm{E}}_G\left(\widehat{\mathcal{LL}}(\mathbf{y}, S; \lambda)\right) = -\frac{\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i}{2} \log \hat{\sigma}^2_\lambda + \sum_{i=1}^{N} \mathbf{1}_S(i) w_i \log\left(\frac{\partial h^{-1}(y_i;\lambda)}{\partial y_i}\right) + \text{const} \quad , \tag{3.16}$$

which is maximized by the solution $\hat{\lambda}$ to

$$-\sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \frac{(\tilde{y}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda - \mathbf{z}_i^T \sum_{b=1}^{B} \omega_b^{MC} \boldsymbol{\gamma}_b)}{\hat{\sigma}^2_\lambda} \cdot \frac{\partial h^{-1}(y_i;\lambda)}{\partial \lambda} + \sum_{i=1}^{N} \mathbb{1}_S(i) \cdot w_i \cdot \frac{\partial^2 h^{-1}(y_i;\lambda)}{\partial y_i \partial \lambda} = 0 \quad . \tag{3.17}$$

Note the similarity between Equation (3.16) and the concentrated log-likelihood derived in Hyde [1999]: Hyde [1999] goes beyond the concentrated likelihood derived in Spitzer [1982], but does not analyze mixed models under transformation so that he needs not use the expected log-likelihood and therefore, his definition of the fixed effects estimator (3.14) and (3.15) differ slightly.

For an iterative solution to (3.17) via the Newton algorithm, derivatives with respect to λ in Equation (3.17) must also be taken for $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\sigma}^2_\lambda$ in the second order differentiation, i.e. the derivative of (3.17). The log-likelihood is not strictly concave in λ and thus an iterative estimation process must evaluate $\hat{E}_G(\widehat{\mathcal{LL}})$ in order to assure maximization.

4 Standard Error Estimation

In order to make inference on the superpopulation model, standard error estimators of $\hat{\boldsymbol{\psi}} := (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\boldsymbol{\rho}}, \hat{\lambda})$ are required. Standard inference on the estimated model parameters given $\hat{\lambda}$ underestimates the true variance because the variability of $\hat{\lambda}$ is ignored [Gurka et al., 2006]. A large sample variance estimator is based on the information matrix of the log-likelihood and provides the Cramér-Rao lower bound [Rao, 1992].

The observed data Fisher information I_ψ is the expected difference between the expected complete data Hessian and the expected outer product of the complete data gradient [Louis, 1982]. Although Louis [1982] gives this lower bound for the traditional EM-algorithm without MC-integration, Booth and Hobert [1999] use that result for a stopping rule of their MCEM-algorithm. Having the importance sample from the final EM-step, the evaluation of the MC-estimator

$$\hat{I}_\psi = \hat{E}_G\left(\nabla^2_\psi \widehat{\mathcal{LL}} - \nabla_\psi \widehat{\mathcal{LL}} (\nabla_\psi \widehat{\mathcal{LL}})^T\right) \quad , \tag{4.1}$$

where ∇^2_ψ indicates the matrix of second order derivatives of the HT log-likelihood with respect to the model parameters, is not so difficult: For each

simulated $\boldsymbol{\gamma}_b$, $b = 1, \ldots, B$, the derivatives of the complete data log-likelihood is evaluated, multiplied by the importance weight ω_b^{MC} and cumulated. In addition, note that the matrix of second order derivatives is block-diagonal for $(\hat{\lambda}, \hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2)^T$ and $\hat{\rho}$.

Burgard and Dörr [2018] show that the provided MC-standard error estimators by the Fisher information underestimate the standard errors observed in Monte-Carlo studies on generalized LMMs. Our simulation studies in Section 6 show that sometimes already the point estimation of LMMs under transformation is so problematic that standard error estimators that do not account for bias in the point estimators can only fail.

5 Model Predictions under Survey Sampling

In some cases, the interest of the analyst will not lie on the inference from the superpopulation model but on prediction from the statistical model. Such an example is small area estimation [Rao, 2003]. Whilst the empirical best predictor in the linear case can be derived from the mode $\hat{\boldsymbol{\gamma}}$ of the joint log-likelihood given the estimator $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\boldsymbol{\rho}}, \hat{\lambda})$ [Rao, 2003, chapter 5], this is not necessarily the case when the link function h in (2.1) does not equal the identity. Such an example are binary data where the logit link is applied and a linear regression is run on the linear predictor. Rao [2003, chapter 9] shows that in this case, the expectation conditional on the observed data does not equal the mode and the former is appropriate for prediction.

The Best Prediction (BP) for an unobserved unit i is the expectation of Y_i conditional the realizations for units in U and the sample S that returns the minimal mean squared error:

$$\mathrm{E}_Y \left((Y_i - h(\eta_i(\hat{\boldsymbol{\gamma}})); \lambda)^2 | Y = \mathbf{y}, S \right) \tag{5.1}$$

which means that the conditional expectation $E_Y(Y_i|Y = \mathbf{y}) = E_Y(h(\tilde{Y}_i; \lambda)|Y = \mathbf{y}, S)$ is required. If $h = \mathrm{id}$, this is the EBLUP of the classical LMM. Otherwise, note that by Equation (3.3), the density m includes the Jacobian of the transformation h^{-1} for the already observed data and thus is not necessarily normal. Hence, by the definition of the variance, it is known that (5.1) is minimal if and only if

$$\mathrm{E}_Y(Y_i|Y = \mathbf{y}, S) = h(\eta_i(\hat{\boldsymbol{\gamma}})) \quad , \tag{5.2}$$

and the solution $\hat{\boldsymbol{\gamma}}$ to the implicit function (5.1) is not trivial. It is easier to calculate directly the conditional prediction (5.2):

$$\begin{aligned} \mathrm{E}_Y(Y_i|Y = \mathbf{y}, S) &= \mathrm{E}_{\tilde{Y}} \left(h(\tilde{Y}_i) | Y = \mathbf{y}, S \right) \\ &= \int_{\mathbb{R}^q} h\left(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma}; \lambda \right) \cdot m(\boldsymbol{\gamma}|\mathbf{y}, S; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \lambda) \, \mathrm{d}\, \boldsymbol{\gamma} \quad . \end{aligned} \tag{5.3}$$

Sakia [1990] suggests for an ANOVA Box-Cox model without survey sampling a second order Taylor approximation of Y_i around the marginal predictor of $\tilde{Y}_i$,

namely $\mathbf{x}_i^T\boldsymbol{\beta}$. Sakia [1990] employs $m(\boldsymbol{\gamma}|\tilde{y})$, which is normal in that case, and therefore simplifies calculation. However, if the integration dimension q is large, integration approximation becomes inaccurate and thus the approximation of $\mathrm{E}_Y(Y_i|Y=\mathbf{y},U)$ gets worse. Furthermore, normality of m is not assured under a complex survey design.

Rojas-Perilla et al. [2017] use a parametric bootstrap to get an estimator of the conditional expectation. Their simplified algorithm is only applicable to random intercepts models and under survey designs that maintain the normality of $\boldsymbol{\gamma}|\tilde{Y}=\tilde{\mathbf{y}}$.

On the other hand, Monte-Carlo realizations of G and importance weights from the last E-step are available and thus yield as empirical Monte-Carlo best predictor (in our case, also conditional on the survey realization $S=s$)

$$\hat{E}_G(Y_i|Y=\mathbf{y},S=s)=\hat{h}(\eta_i;\hat{\boldsymbol{\gamma}})=:\frac{1}{B}\sum_{b=1}^{B}h(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\mathbf{z}_i^T\boldsymbol{\gamma}_b;\hat{\lambda})\cdot\omega_b^{MC} \quad . \tag{5.4}$$

Noting that the importance sampling would also be possible if $\boldsymbol{\beta},\sigma^2,\boldsymbol{\rho},\lambda$ were known and/or $S=U$, the suggestion yields a more flexible predictor. However, under a non-informative survey design, the predictor is less efficient than the parametric bootstrap estimator suggested in Rao [2003] and Rojas-Perilla et al. [2017].

6 Application to Simulated Data

In a first step, we mimic the simulation outlined in Burgard and Dörr [2018] in order to study the estimation algorithm's behaviour under Box-Cox and dual transformations (rather than GLMMs).

6.1 Simulation Study

6.1.1 Replication of Gurka et al. [2006]

Data Generating Process As the suggested Algorithm 3.3 is novel in the context of mixed models under data transformations, we first validate its performance in a model-based simulation study. We therefore replicate the simulation scenario described in Gurka et al. [2006]. Gurka et al. [2006] have extended the known Box-Cox transformations in linear regression to the case of mixed models. Furthermore, they suggest a scaling procedure of the input data in order to apply standard mixed model estimation programs given a transformation parameter λ. They use these scaled input data and a restricted ML (REML) approach combined with a line search for the optimal λ in their simulation study.

The DGP of Gurka et al. [2006] is described as follows:

$$\tilde{Y}_{d,i} = 5 + 2 \cdot x_{1,d,i} + x_{2,d,i} + G_d + \varepsilon_i, \qquad i = 1, \ldots, 5 \tag{6.1}$$

$$G_d \sim N(0, 0.5), \quad d = 1, \ldots, 100 \tag{6.2}$$

$$\varepsilon_i \sim N(0, 0.5) \tag{6.3}$$

$$Y_{d,i} = \begin{cases} \frac{\tilde{Y}_{d,i}^{\lambda} - 1}{\lambda}, \ \lambda \in \{0, 0.5, 1\} \\ \log \tilde{Y}_{d,i}, \ \lambda = 0 \end{cases} . \tag{6.4}$$

Actually, the data have a panel structure, d is interpreted as one unit that is observed at five equally spaced time intervals named $x_{2,d,i} = \frac{i-1}{4}$ and $x_{1,d,i} = x_{1,d}$ is a once realized Bernoulli variable with success probability 0.5.

Gurka et al. [2006] assume in addition a missing completely at random (MCAR) nonresponse mechanism and delete in average about 20 % of the observations. The nonresponse mechanism is orthogonal to the DGP and thus does not affect the estimation procedure. Thus, the nonresponse mechanism only reduces the sample size which impacts the variability of the point estimators but not their expectation. We therefore do not implement the nonresponse mechanism. We simulate 1,500 Monte-Carlo runs. Deviations from the reported results in Gurka et al. [2006] are attributed to the not implemented nonresponse, different numbers of MC-replications and other software implementations.

Results As implementation of the estimation method introduced in Gurka et al. [2006], we estimate the transformed model (6.2) with the `R`-package `lme4` [Bates, 2011] using restricted ML. Like Gurka et al. [2006] and Rojas-Perilla et al. [2017], we rescale the dependent variable by the geometric mean of the jacobians in order to apply standard statistical software. The resulting restricted ML criterion is a function of the transformation parameter and optimized using `optimize()` (the `R`-package `emdi` [Kreutzmann et al., 2018] uses the same procedure). The results of the replication are summarized in Figures 6.1 to 6.4.

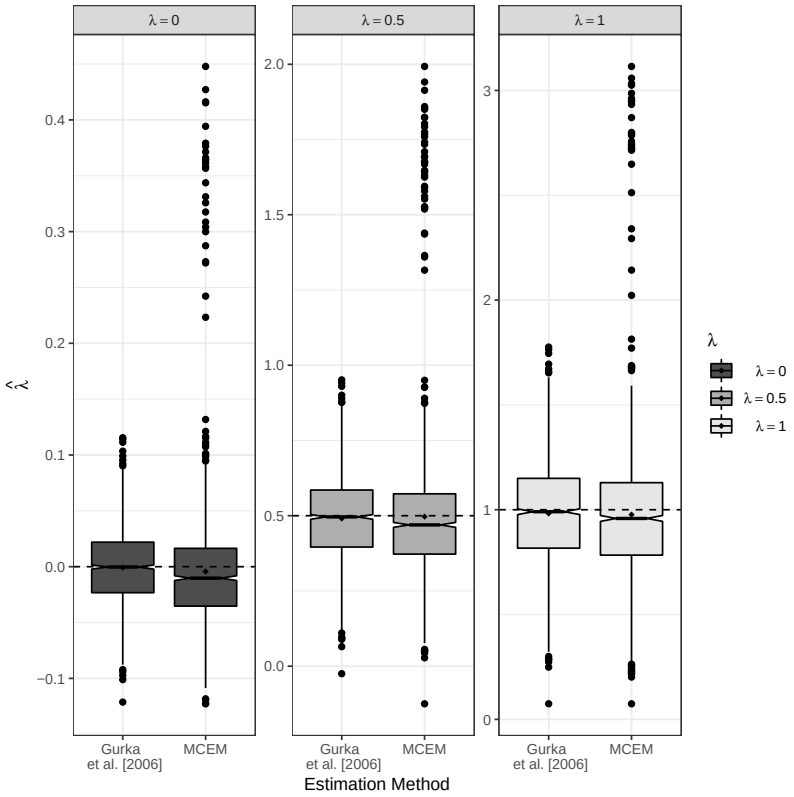

Figure 6.1: Estimated Box-Cox Transformation Parameter – Replication of Gurka et al. [2006]

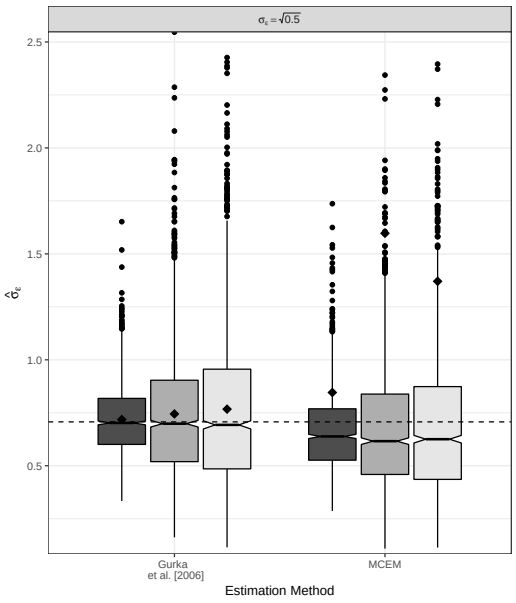

Figure 6.2: Estimated Residual Variance – Replication of Gurka et al. [2006]

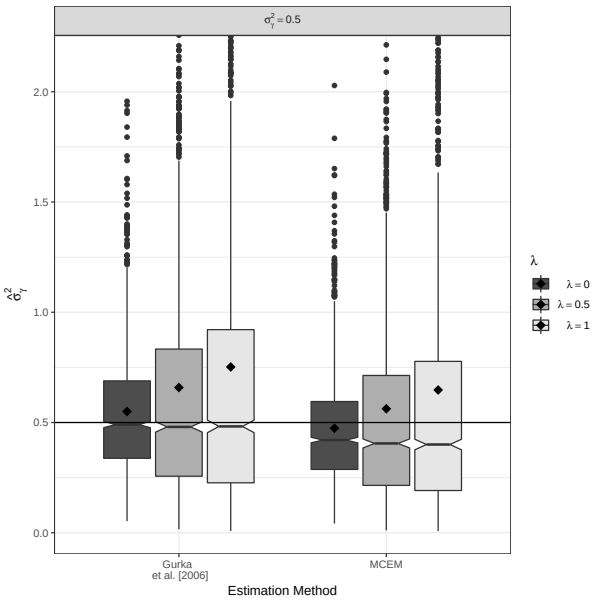

Figure 6.3: Estimated Random Effects Variance – Replication of Gurka et al. [2006]

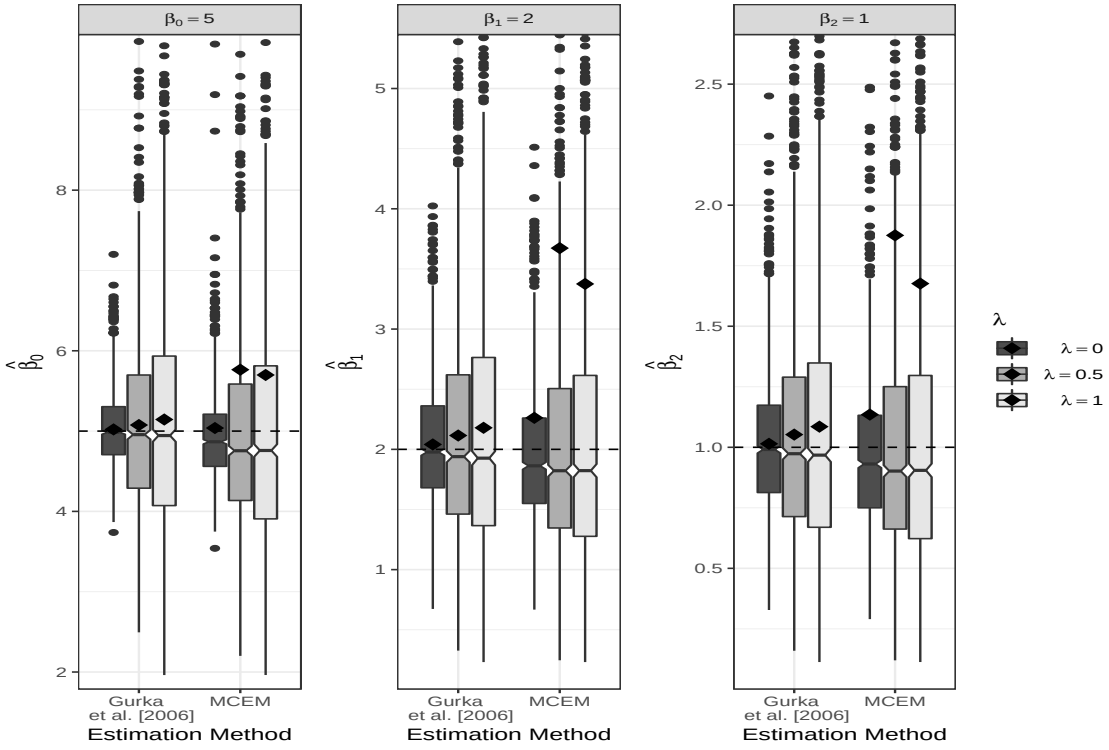

Figure 6.4: Estimated Fixed Effects – Replication of Gurka et al. [2006]

We find that the proposed Algorithm 3.3 yields similar results like the method of Gurka et al. [2006] for the transformation parameter. The random effects variance is estimated with a smaller bias in average. However, there is a bias in the fixed effects estimator that increases with λ. This bias translates to a biased residual variance, too. Note, however, that the quantiles for the fixed effects are similar to that of Gurka et al.'s method. Therefore, the biased mean seems to be a result of outliers – those outliers do not affect the estimation of λ and σ^2_γ because the latter two are directly estimated from the likelihood whereas – due to the concentrated likelihood approach – the fixed effects and residual variance estimators are functions of $\hat{\lambda}$.

6.1.2 Simulation Study under Sample Randomization

Validity under the DGP In a second step, we generate data from the process described from Equations (6.5) to (6.9) in oder to assess the consistency under another statistical model, before going on to part three of our simulation study, that seeks to underline the importance of survey weights to account for the sampling randomization and uses the same DGP, too. We generate $k = 1, \ldots, 1500$ finite populations under the following statistical model:

$$\tilde{Y}_{d,i} = (\beta_0 + G_{d,0}) + x_i \cdot (\beta_1 + G_{d,1}) + \varepsilon_i, \quad i = 1, \ldots, 200 \tag{6.5}$$

$$\begin{pmatrix} G_{0,d} \\ G_{1,d} \end{pmatrix} \sim \text{MVN}\left(\mathbf{0}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right), \quad d = 1, \ldots, 25 \tag{6.6}$$

$$\varepsilon_i \sim N(0, \sigma^2), \quad \sigma^2 = 1 \tag{6.7}$$

$$Y_{d,i} = h(\tilde{Y}_{d,i}, \lambda), \quad \lambda \in \{0.3, 1.5\} \tag{6.8}$$

$$h(\cdot, \lambda) \in \left\{ \frac{(\cdot)^\lambda - 1}{\lambda}, \frac{(\cdot)^\lambda - (\cdot)^{-\lambda}}{2\lambda} \right\} \quad . \tag{6.9}$$

The variance components of the random effects are thus $\rho = (\sigma_{11}, \sigma_{12}, \sigma_{22})^T = (3, 1, 4)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T = (20, 2)^T$. We then apply Algorithm 3.3 to the generated data without any sub-sampling from the finite population (that is, in this step, we do not have a sampling randomization P_D and $s = U$ and $w_i = 1$ for all $i \in U$). After the Algorithm's performance under the DGP is validated, we introduce sampling randomization.

Results under the DGP We summarize again the results of the simulation study in Boxplots 6.5 to 6.8. Results for the dual transformation are similar and presented in the appendix. Interestingly, the bias in the fixed effects does not appear in this simulation study – another indicator that for there were outliers in the previous simulation study. Rather, the introduced estimators perform better than those of Gurka et al. [2006]. This is especially striking for the random effects variance components. The tendency to yield better estimates for $\boldsymbol{\rho}$ can already be found in the previously discussed replication study.

Validity under the Design A common simulation set-up for sample informativity is to make the inclusion probabilities a function of the random effects

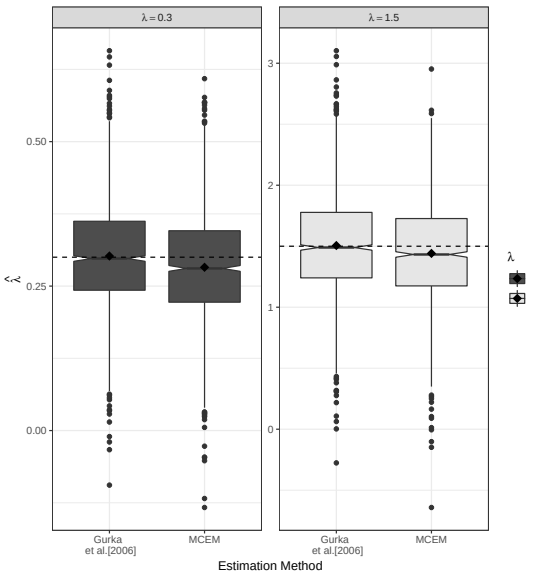

Figure 6.5: Estimated Box-Cox Transformation Parameter

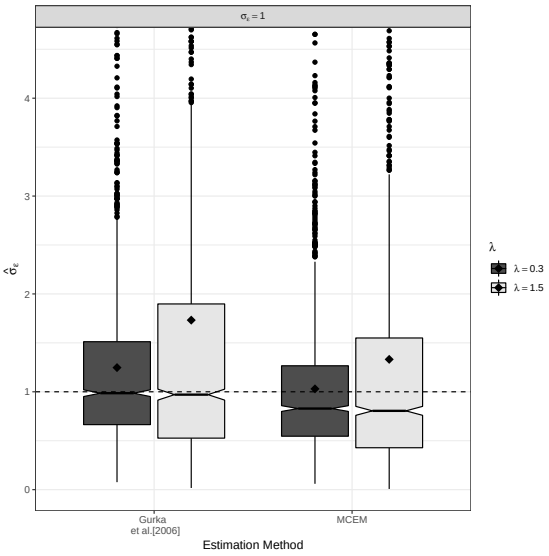

Figure 6.6: Estimated Residual Variance – Box-Cox Transformation

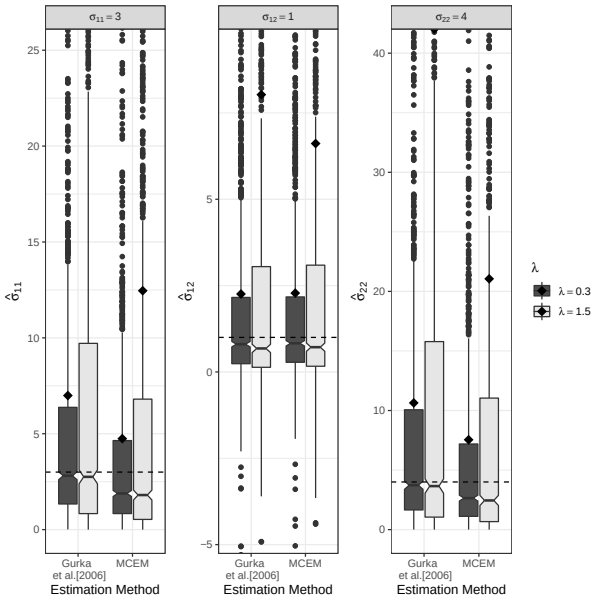

Figure 6.7: Estimated Random Effects Variance – Box-Cox Transformation

[Pfeffermann et al., 1998, Rabe-Hesketh and Skrondal, 2006]. However, given the assumption $P_D(S|\mathbf{y}, \boldsymbol{\gamma}) = P_D(S|\mathbf{y})$ argued for in Equation (3.2), this set-up would only be approximatively correct. We thus recur to the following sampling mechanisms.

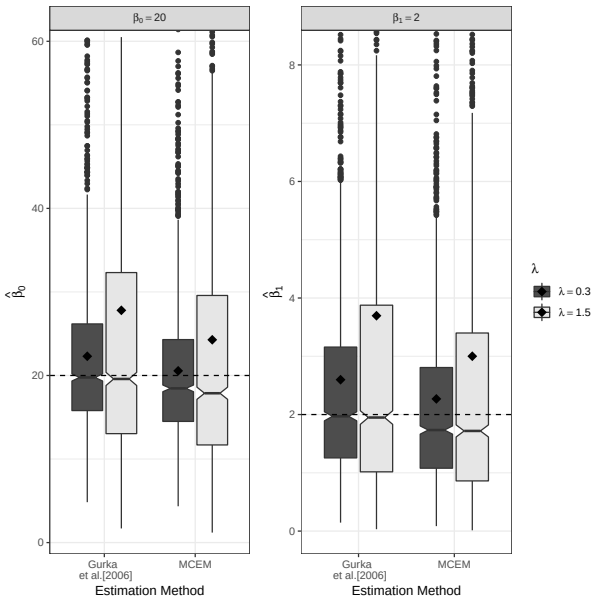

Figure 6.8: Estimated Fixed Effects – Box-Cox Transformation

The randomization process is similar to that described in Burgard and Dörr [2018]: The $X_i > 0$ $i \in U$ are only realized once and used throughout in each simulation run. As a non-informative sampling under the model, we choose a πps sampling design with the inclusion probability for unit i proportional to the auxiliary, $\pi_i \propto x_i$. For the informative sampling, we set $\pi_i \propto f(\varepsilon_i)$ where f is a monotonic decreasing function in its argument. The sample size is fixed and equals 200, and is drawn from a finite population U, $|U| = 2000$ generated like in Equations (6.5) to (6.9). That is, we oversample units with negative residuals. Precisely, we have

$$f : \mathbb{R} \to \mathbb{R}, \quad f(\varepsilon_i) = \begin{cases} 16, \text{ if } \varepsilon_i < -2 \\ 4, \text{ if } \varepsilon_i \in [-2, 0) \\ 1, \text{ if } \varepsilon_i \in [0, 2) \\ 0.25, \text{ else} \end{cases} \quad . \tag{6.10}$$

Without accounting for the sampling design, the estimator $\hat{\lambda}$, which aims at the restablishment of normality assumptions, will correct for the small values that are oversampled even for an underlying skewed distribution. Thus, the estimator is assumed to be biased, though the sign is a priori unclear and depends on the other parameters of the DGP. $\hat{\lambda}$'s bias in turn has an impact on the estimation of both fixed effects $(\beta_0, \beta_1) = (20, 2)$ and the variance components $\boldsymbol{\rho} = (3, 1, 4)^T$ and $\sigma^2 = 1$. Hence, without accounting for the (under the model) informative survey-design, the restricted ML-estimators are presumably biased.

In order to underpin the importance of survey weights under informative sampling, we estimate the statistical model with and without survey weights. The unweighted estimation is an application of Gurka et al. [2006].

Results under the Sample Randomization

Non-informative Design Like it was outlined in the previous section, the survey weights do not add any additional information to the statistical model of the finite population. Hence, it is not surprising that the results resemble those under the DGP, see Figures 6.9 to 6.12. Again, results for the dual transformation are similar and reported in the appendix.

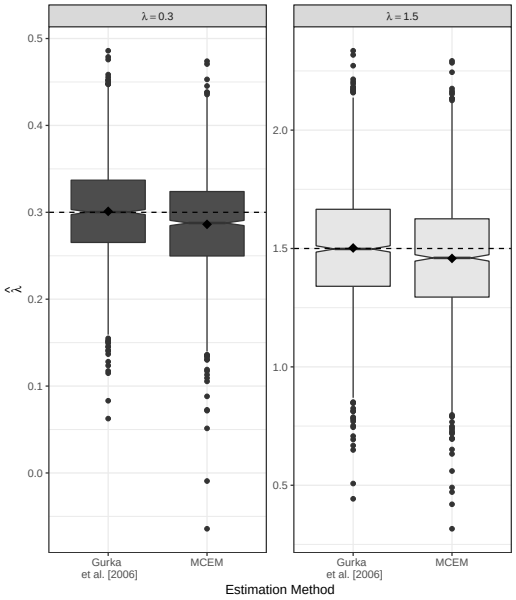

Figure 6.9: Estimated Box-Cox Transformation Parameter – Non-informative Design

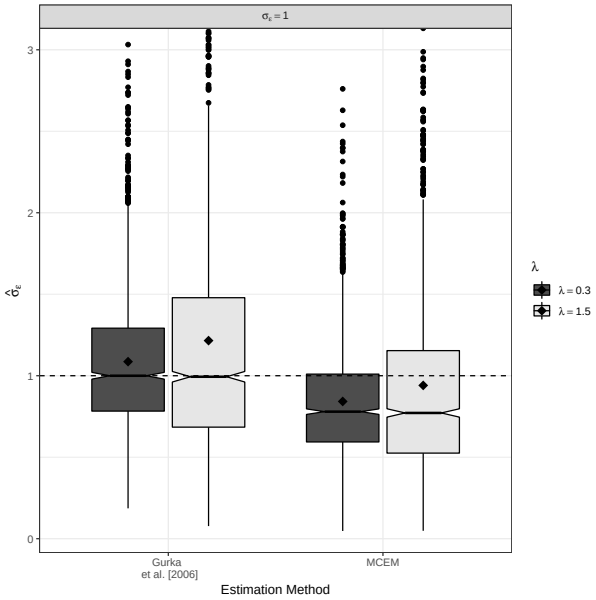

Figure 6.10: Estimated Residual Variance – Box-Cox Transformation – Non-informative Design

Informative Design Finally, we present results for the regression parameter estimators when the data is gathered from a survey under a complex and informative design. The corresponding boxplots are Figures 6.13 to 6.16. Interestingly, we find that the estimated transformation parameter is estimated unbiasedly for both, the weighted and unweighted regression (cf. Figure 6.13). For the variance components, in contrast, the bias already found in the non-informative design (for $\lambda = 1.5$) has aggravated. However, the bias is more extreme when the survey design is not accounted for.

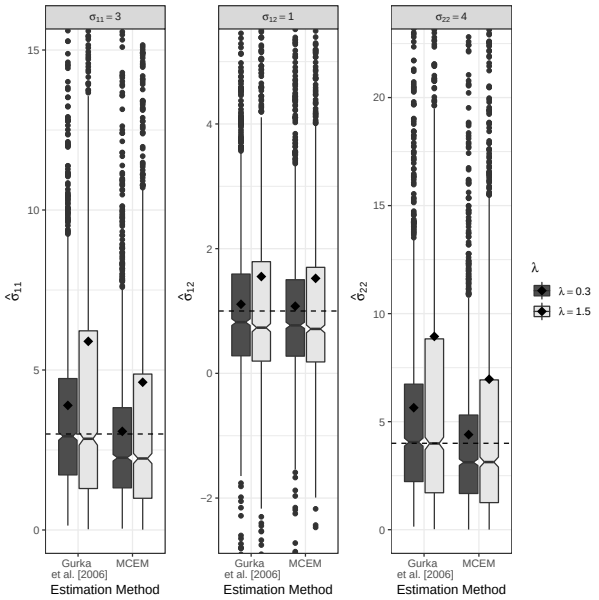

Figure 6.11: Estimated Random Effects Variance – Box-Cox Transformation – Non-informative Design

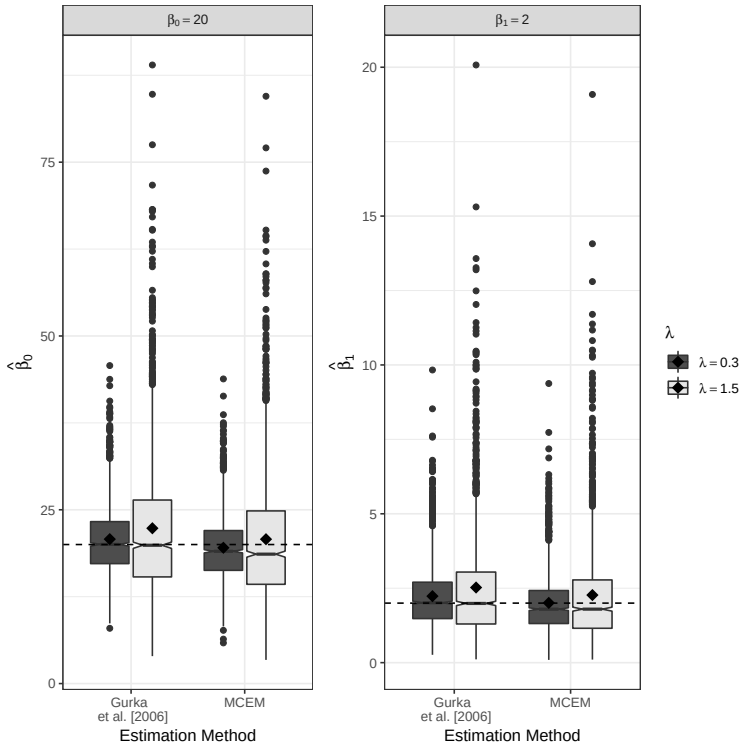

Figure 6.12: Estimated Fixed Effects – Box-Cox Transformation – Non-informative Design

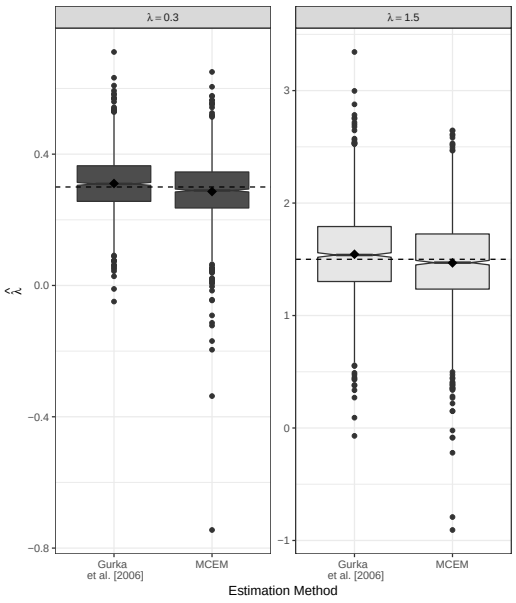

Figure 6.13: Estimated Box-Cox Transformation Parameter – Informative Design

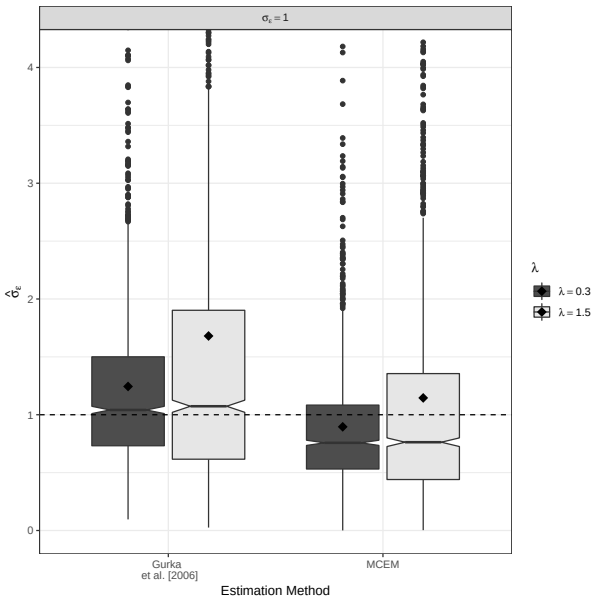

Figure 6.14: Estimated Residual Variance – Box-Cox Transformation – Informative Design

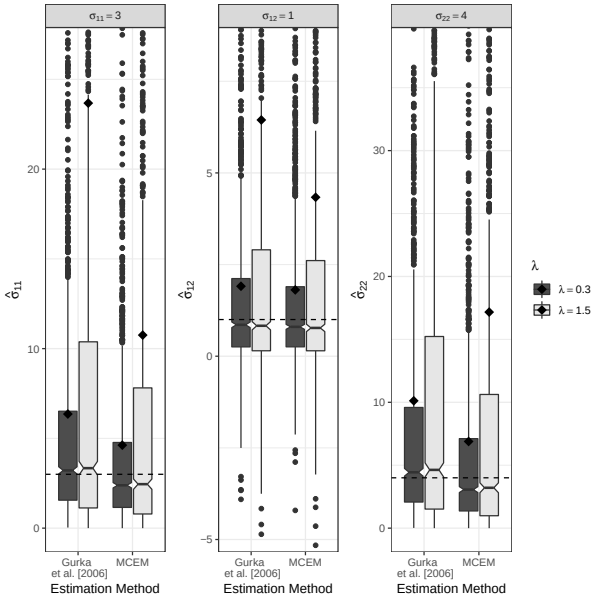

Figure 6.15: Estimated Random Effects Variance – Box-Cox Transformation – Informative Design

The impact of survey weighting, though, is obvious for the estimation of the fixed effects and the residual variance - parameters that were estimated unbi-

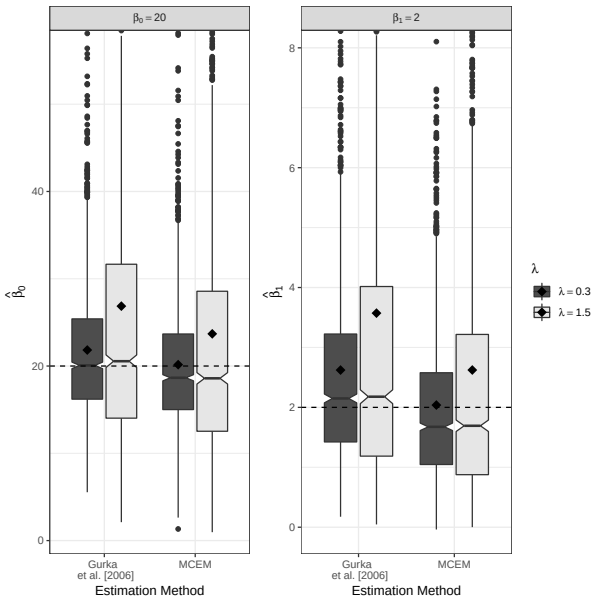

Figure 6.16: Estimated Fixed Effects – Box-Cox Transformation – Informative Design

asedly under a non-informative design (cf. Figures 6.12 and 6.10). For these estimators, bias can at least be reduced using the suggested survey weights, for $\lambda = 0.3$, it is even removed. This is a striking argument to account for survey design in regression analysis, be it for causal inference, where especially the slope parameter β_1 matters, or predictive analysis.

7 Numerical Aspects

The MCEM-algorithm does not necessarily increase the log-likelihood in each iteration step [Booth and Hobert, 1999]. In the simulation studies, we noticed that the MCEM algorithm is much unstabler for the LMM under Box-Cox and dual transformations than for GLMMs. Like for the binary case in Burgard and Dörr [2018], we keep track of the estimated expected log-likelihood in order to prevent the algorithm to exit the environment around an optimum.

Furthermore, we observed that a good choice of starting values is more important than for the GLMM case because the log-likelihood is not globally concave. This also holds for the inital guess of the square root of $-H^{-1}$ and the optimal $\hat{\boldsymbol{\gamma}}$ in Powell's [1987] algorithm. Therefore, we first optimize the log-likelihood without survey weights and use the resulting model parameter estimates as starting values for the MCEM-algorithm.

For the square root of $-H^{-1}$ and the mode $\hat{\boldsymbol{\gamma}}$, we start in each MCEM-step

Powell's [1987] gradient descent algorithm with

$$\boldsymbol{\gamma}_0 = \sum_{b=1}^{B} \omega_b^{MC} \boldsymbol{\gamma}_b \tag{7.1}$$

and

$$\left(-H^{-1}\right)_0^{1/2} = \frac{1}{\sqrt{\|\nabla_\gamma \widehat{\mathcal{LL}}(\boldsymbol{\gamma}_0)\|}} \cdot \mathrm{I}_q \quad , \tag{7.2}$$

where $\|\nabla_\gamma \widehat{\mathcal{LL}}(\boldsymbol{\gamma}_0)\|$ is the Eucledian norm of the gradient of $\widehat{\mathcal{LL}}$ with respect to the random effects $\boldsymbol{\gamma}$ given the current parameter estimates. Note that the conditional expectation (7.1) depends on the importance sample of random effects in a given Monte-Carlo E-step and that the importance weights also depend on the current parameter estimates, please confer Algorithm 3.3.

8 Discussion

In this paper, we have modified the Monte-Carlo EM-algorithm elaborated in Burgard and Dörr [2018] for application to Box-Cox and dual transformations in LMMs under survey sampling. In a simulation study, we studied the feasability of the introduced algorithm and showed the importance to incorporate survey design in regression analysis. At least for a small to moderate number of random effects, we found that the computational effort is manageable and the results are competitive to the established, unweighted estimators.

There are still numerical challenges in the algorithm and with a look on optimization, it would be desirable to find methods that turn the proposed MCEM algorithm more stable and less sensitive to the starting values. Perhaps, a better (and computationally feasible) proposal density for the Monte-Carlo integration could improve the estimation results.

We think that the proposed algorithm has the potential to become an important estimation method in regression analysis: Many variables of high (micro-) economic interest such as income, wealth or returns of sales are skewed and this property needs an adequate handling in regression analysis. Furthermore, most of these variables are gathered in surveys with a complex sample design such as the German Socio-Economic Panel, the Italian Survey on Household Income and Wealth, the American Consumer Expenditure Survey or the international Household Finance and Consumption Survey. Our simulation study suggests that weighting could be a good way to account for such designs.

References

Douglas Bates. Computational methods for mixed models. *Vignette for lme4*, 2011.

James G Booth and James P Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Methodology)*, 61(1):265–285, 1999. doi: 10.1111/1467-9868.00176. URL https://doi.org/10.1111/1467-9868.00176.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodology)*, pages 211–252, 1964. doi: 10.2307/2984418. URL http://www.jstor.org/stable/2984418.

Jan Pablo Burgard and P Dörr. Survey-weighted generalized linear mixed models. 2018.

Guillaume Chauvet. A note on the consistency of the narain-horvitz-thompson estimator. *arXiv preprint arXiv:1412.2887*, 2014.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (Methodology)*, pages 1–38, 1977. doi: 10.2307/2984875. URL http://www.jstor.org/stable/2984875.

Matthew J Gurka, Lloyd J Edwards, Keith E Muller, and Lawrence L Kupper. Extending the box–cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):273–288, 2006. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2005.00391.x.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446.

Scott Hyde. Likelihood based inference on the box-cox family of transformations: Sas and matlab programs. Master's thesis, Department of Mathematical Sciences, Montana State University, 1999.

Ann-Kristin Kreutzmann, Soeren Pannier, Natalia Rojas-Perilla, Timo Schmid, Matthias Templ, and Nikos Tzavidis. *emdi: Estimating and Mapping Disaggregated Indicators*, 2018. URL https://cran.r-project.org/package=emdi. R package version 1.1.4.

Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodology)*, pages 226–233, 1982. doi: 10.2307/2345828. URL http://www.jstor.org/stable/2345828.

Charles E McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997. doi: 10.1080/01621459.1997.10473613. URL https://doi.org/10.1080/01621459.1997.10473613.

Ronald C Neath et al. On convergence properties of the monte carlo em algorithm. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, pages 43–62. Institute of Mathematical Statistics, 2013.

Art B Owen. Monte carlo theory, methods and examples, 2013. http://statweb.stanford.edu/~owen/mc/.

Danny Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337, 1993. doi: 10.2307/1403631. URL http://www.jstor.org/stable/1403631.

Danny Pfeffermann, Chris J Skinner, David J Holmes, Harvey Goldstein, and Jon Rasbash. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 60(1):23–40, 1998. doi: 10.1111/1467-9868.00106. URL https://doi.org/10.1111/1467-9868.00106.

José C Pinheiro and Douglas M Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35, 1995.

MJD Powell. Updating conjugate directions by the bfgs formula. *Mathematical Programming*, 38(1):29, 1987. doi: 10.1007/BF02591850. URL https://doi.org/10.1007/BF02591850.

Sophia Rabe-Hesketh and Anders Skrondal. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827, 2006. doi: 10.1111/j.1467-985X.2006.00426.x. URL https://doi.org/10.1111/j.1467-985X.2006.00426.x.

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics*, pages 235–247. Springer, 1992.

JNK Rao. *Small Area Estimation*. Wiley, New York, 1 edition, 2003. doi: 10.1002/0471722189.

Natalia Rojas-Perilla, Sören Pannier, Timo Schmid, and Nikos Tzavidis. Data-driven transformations in small area estimation. Technical report, Discussion Paper, School of Business & Economics: Economics, 2017.

RM Sakia. Retransformation bias: a look at the box-cox transformation to linear balanced mixed anova models. *Metrika*, 37(1):345–351, 1990.

John J Spitzer. A primer on box-cox estimation. *The Review of Economics and Statistics*, pages 307–313, 1982.

Shonosuke Sugasawa, Tatsuya Kubokawa, et al. Box-cox transformed linear mixed models for positive-valued and clustered data. Technical report, University of Tokyo, 2015.

Zhenlin Yang. A modified family of power transformations. *Economics Letters*, 92(1):14–19, 2006.

A Additional Simulation Results under the DGP

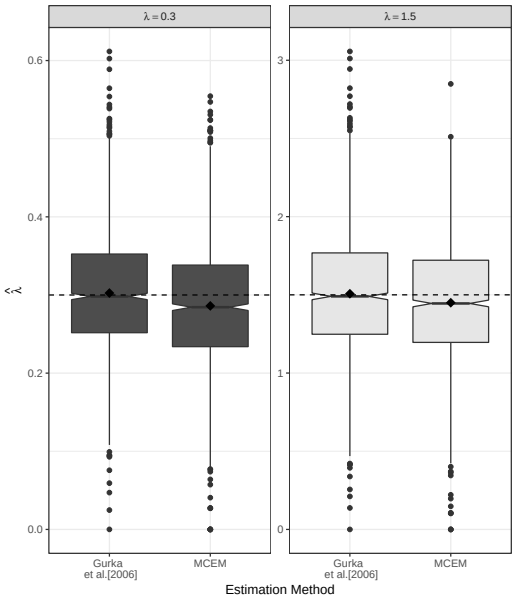

Figure A.1: Estimated Dual Transformation Parameter

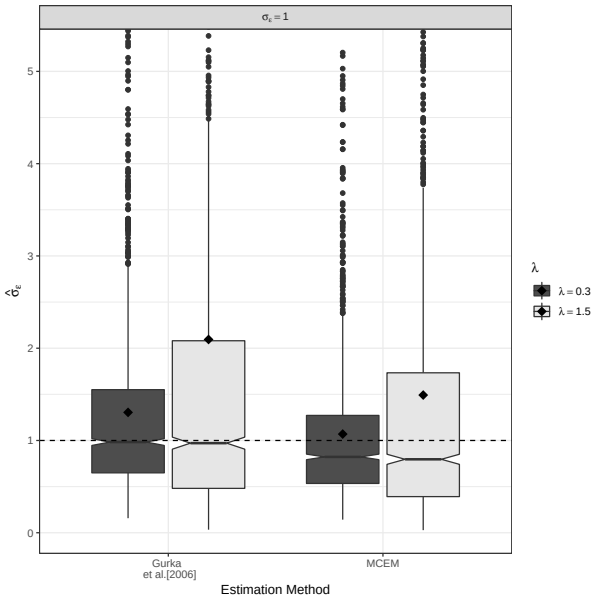

Figure A.2: Estimated Residual Variance – Dual Transformation

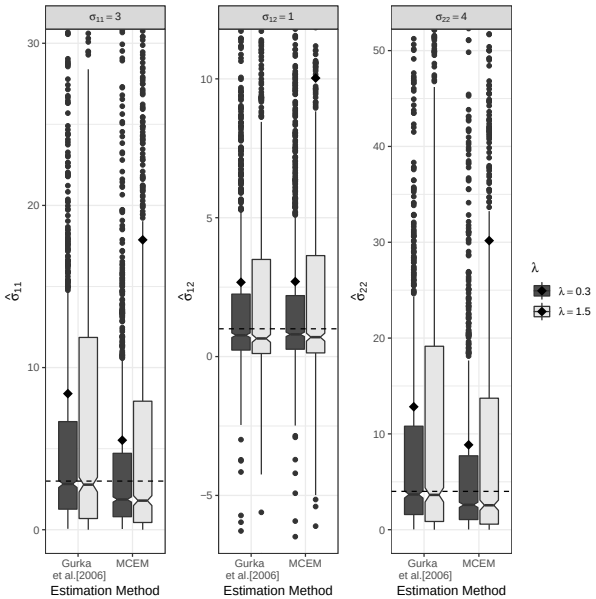

Figure A.3: Estimated Random Effects Variance – Dual Transformation

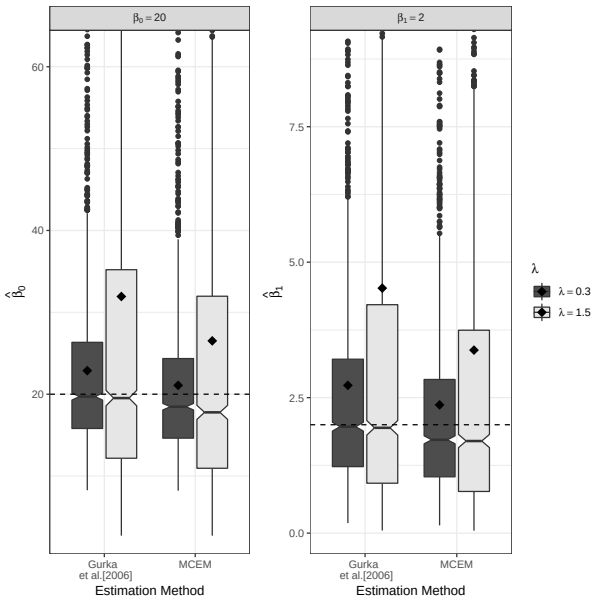

Figure A.4: Estimated Fixed Effects – Dual Transformation

B Additional Simulation Results under the Non-informative Design

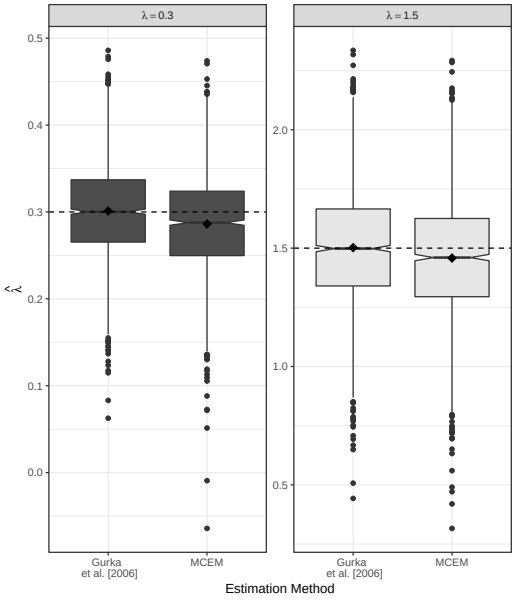

Figure B.1: Estimated Dual Transformation Parameter – Non-informative Design

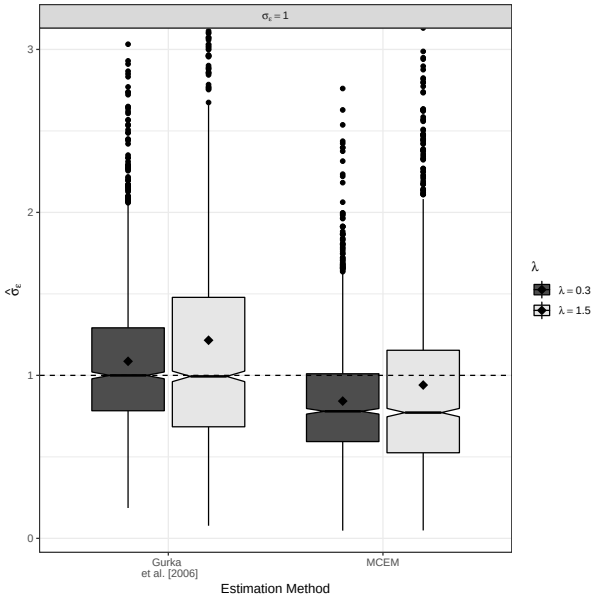

Figure B.2: Estimated Residual Variance – Dual Transformation – Non-informative Design

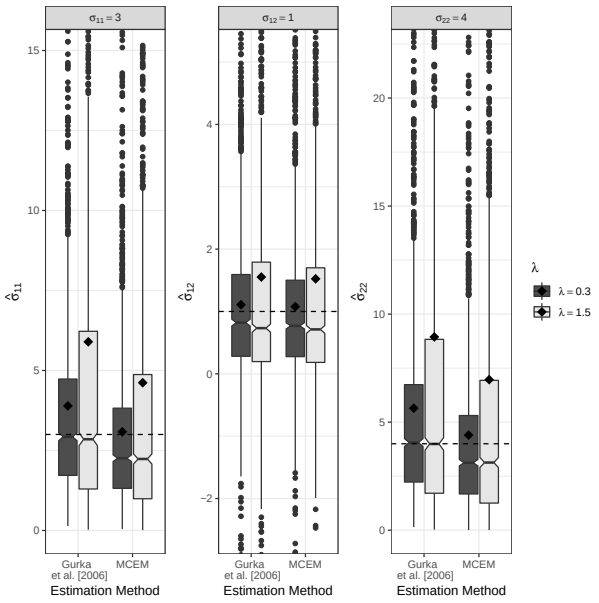

Figure B.3: Estimated Random Effects Variance – Dual Transformation – Non-informative Design

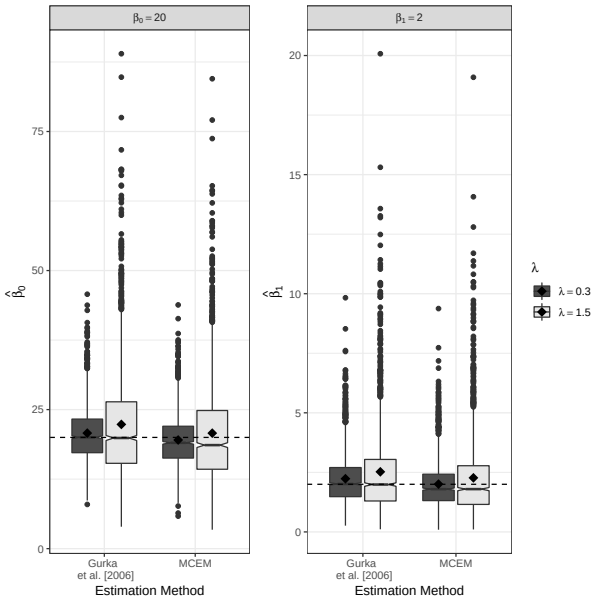

Figure B.4: Estimated Fixed Effects – Dual Transformation – Non-informative Design

C Additional Simulation Results under the Informative Design

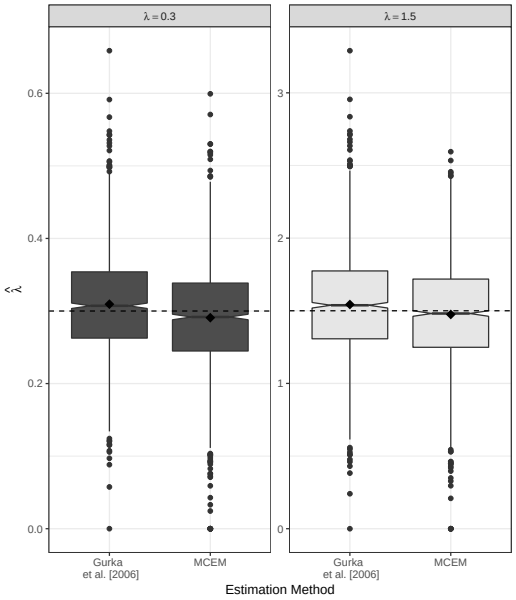

Figure C.1: Estimated Dual Transformation Parameter – Informative Design

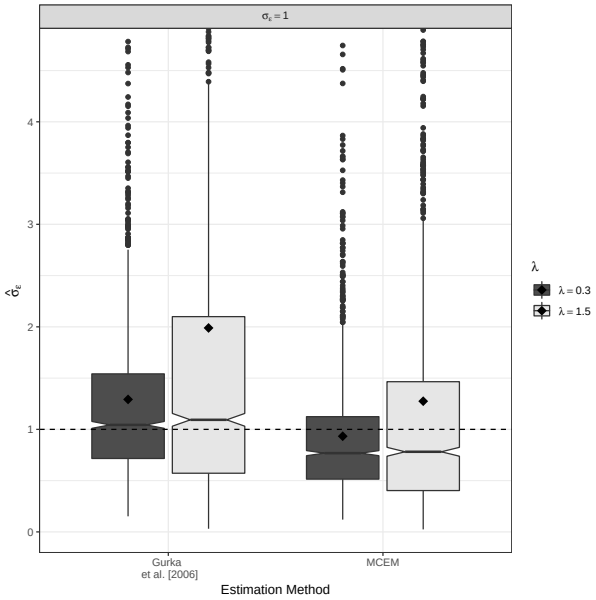

Figure C.2: Estimated Residual Variance – Dual Transformation – Informative Design

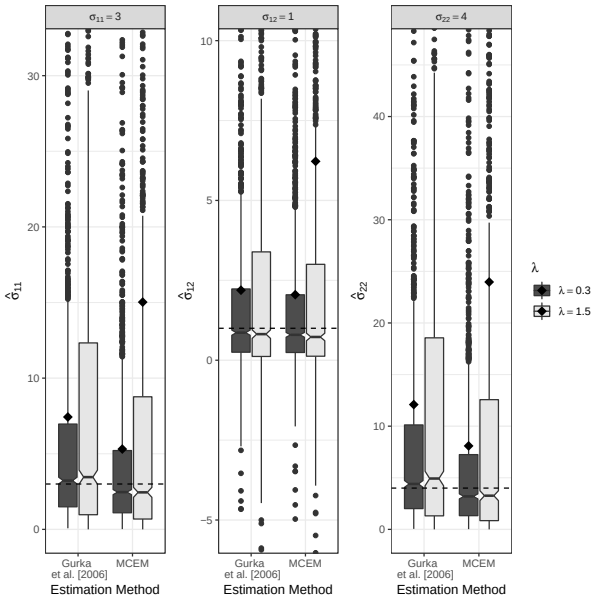

Figure C.3: Estimated Random Effects Variance – Dual Transformation – Informative Design

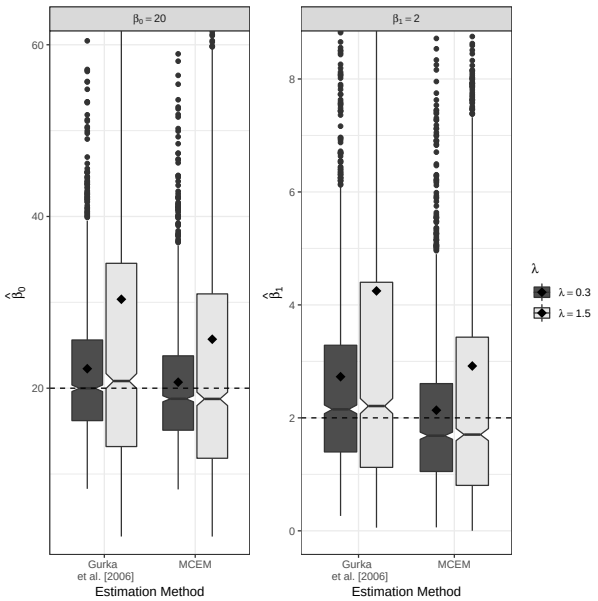

Figure C.4: Estimated Fixed Effects – Dual Transformation – Informative Design