Freie Universität Berlin

A Unit-level Quantile Nested Error Regression Model for Domain Prediction with Continuous and Discrete Outcomes

Beate Weidenhammer
Timo Schmid
Nicola Salvati
Nikos Tzavidis

School of Business & Economics
Discussion Paper

Economics

2016/12

A Unit-level Quantile Nested Error Regression Model for Domain Prediction with Continuous and Discrete Outcomes

Beate Weidenhammer[*], Timo Schmid[*], Nicola Salvati[**], and Nikos Tzavidis[***]

[*]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany

[**]Economic Department, University of Pisa, Pisa, Italy

[***]Statistical Sciences Research Institute, University of Southampton, Southampton, UK

Abstract

In this paper we will present recent work on a new unit-level small area methodology that can be used with continuous and discrete outcomes. The proposed method is based on constructing a model-based estimator of the distribution function by using a nested-error regression model for the quantiles of the target outcome. A general set of domain-specific parameters that extends beyond averages is then estimated by sampling from the estimated distribution function. For fitting the model we exploit the link between the Asymmetric Laplace Distribution and maximum likelihood estimation for quantile regression. The specification of the distribution of the random effects is considered in some detail by exploring the use of parametric and non-parametric alternatives. The use of the proposed methodology with discrete (count) outcomes requires appropriate transformations, in particular jittering. For the case of discrete outcomes the methodology relaxes the restrictive assumptions of the Poisson generalised linear mixed model and allows for what is potentially a more flexible mean-variance relationship. Mean Squared Error estimation is discussed. Extensive model-based simulations are used for comparing the proposed methodology to alternative unit-level methodologies for estimating a broad range of complex parameters.

Key words: Asymmetric Laplace Distribution; Generalized linear mixed model; Jittering; Non-parametric estimation; Small area estimation.

1 Introduction

The use of unit-level models is now considered to be standard practice in small area estimation. An important application of unit-level small area models is in estimating non-linear parameters. The seminal paper by Molina and Rao (2010) proposes the use of Empirical Best Prediction (EBP) under a nested error regression model for estimating income related indicators for example the incidence of poverty and

the poverty gap in small areas. The Molina and Rao (2010) methodology is implemented by assuming normality for the unit-level residuals and the domain random effects of the model that is fitted on the logarithmically transformed outcome. What if the model assumptions do not hold even after transformation?

In this paper we will present recent work on a new unit-level small area methodology that can be used with continuous and discrete, in particular count, outcomes Weidenhammer et al. (2014), Tzavidis et al. (2015), Tzavidis and Schmid (2015). The proposed method is based on constructing a model-based estimator of the distribution function by using a nested-error regression model for the quantiles of the target outcome. A general set of domain-specific parameters that extends beyond averages is then estimated by sampling from the estimated empirical distribution function. For fitting the model we exploit the link between the Asymmetric Laplace Distribution and maximum likelihood estimation for quantile regression. The specification of the distribution of the random effects is considered in some detail by exploring the use of parametric and non-parametric alternatives. The use of the proposed methodology with discrete (count) outcomes requires appropriate transformations, in particular jittering. For the case of discrete outcomes the methodology relaxes the restrictive assumptions of the Poisson generalised linear mixed model and allows for what is potentially a more flexible mean-variance relationship that can also capture the presence of over-dispersion.

The paper is structured as follows. In Section 2 we review linear mixed models. Section 3 presents linear quantile mixed effects regression following Geraci and Bottai (2007) and Section 4 presents an extension of the linear model to the case of count outcomes by using jittering following Machado and Silva (2005). Using the models presented in Sections 3 and 4 Section 5 proposes novel methodology for domain prediction, hereafter referred to as Microsimulation via Quantiles (MvQ). Bootstrap-based Mean Squared Error (MSE) estimation is studied in Section 6 using a modified version of the semi-parametric bootstrap proposed by Carpenter et al. (2003) and a modified version of the wild bootstrap proposed by Feng et al. (2011). In Section 7 we empirically evaluate the proposed methodology separately for count and continuous outcomes by using a Monte-Carlo simulation under a range of scenarios for linear and non-linear target parameters. We conclude the paper by summarising our main findings and by providing some ideas for further research.

2 The Linear Mixed Model

Linear mixed models are in common use in statistics. One main application is longitudinal data, where D objects are each observed at different times. Another one is the Small Area Estimation (SAE), where D areas have each a within sample size of n_i, $i = 1, 2, \ldots, D$ of individuals or units. Both have in common that dependencies within observations, may they come from the same object or the same area,

are captured in a random effect V_i. This leads to the linear mixed model,

$$Y_{ij} = x_{ij}^T\beta + V_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i, \tag{1}$$

where Y_{ij} is the observation and x_{ij} is a p-dimensional vector of independent variables of the time or individual j in object or area i, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the unknown p-dimensional parameter vector, V_i is the random effect, and ε_{ij} is the individual error. Since the paper focuses mainly on SAE, we will name all properties in area and individual terms keeping in mind that they are exchangeable for other applications of mixed models.

So far there are no distribution assumptions on the error terms V_i and ε_{ij} but that they are centred, thus

$$E[V_i] = 0 \quad \text{and} \quad E[\varepsilon_{ij}] = 0, \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i,$$

and have each a finite variance

$$Var(V_i) = \sigma_V^2 < \infty \quad \text{and} \quad Var(\varepsilon_{ij}) = \sigma_\varepsilon < \infty,$$
$$i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i.$$

Additionally they are independently distributed of each other. Thus V_{i_1} is independently distributed from $V_{i_2}$for all $i_1 \neq i_2$, $\varepsilon_{i_1 j_1}$ is independently distributed from $\varepsilon_{i_2 j_2}$ for all $(i_1, j_1) \neq (i_2, j_2)$, and V_{i_1} is independently distributed from $\varepsilon_{i_1 j}$ for all $i_1, i_2 = 1, 2, \ldots, D$ and $j = 1, 2, \ldots n_{i_1}$. The sample size in area i is n_i leading to an overall sample size of

$$n = \sum_{i=1}^{D} n_i. \tag{2}$$

Of common use is a normal assumption on the random effect

$$V_i \overset{iid}{\sim} N(0, \sigma_V^2), \quad i = 1, 2, \ldots, D.$$

Other distributions are possible but have not been as widely-used. A normal assumption on the individual error terms is also in common use, especially in the SAE approach:

$$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2), \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i.$$

We can rewrite model (1) in matrix form as follows

$$Y = \mathbf{X}\beta + \mathbf{Z}V + \varepsilon, \tag{3}$$

where $Y = (Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{D,n_D})^T$ is the vector of the observations, the matrix

$$\mathbf{X} := \left(x_{1,1}^T, x_{1,2}^T, \cdots, x_{1,n_1}^T, x_{2,1}^T, \cdots, x_{D,n_D}^T\right)^T \tag{4}$$

is the design matrix, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the unknown p-dimensional parameter vector, the matrix

$$\mathbf{Z} := \begin{pmatrix} \mathbf{1}_{n_1} & & & \\ & \mathbf{1}_{n_2} & & \\ & & \ddots & \\ & & & \mathbf{1}_{n_D} \end{pmatrix}, \tag{5}$$

where $\mathbf{1}_{n_i}$ ia an n_i-dimensional vector of ones, is the design matrix for the random vector $V = (V_1, V_2, \ldots, V_D)^T$, and $\varepsilon = (\varepsilon_{1,1}, \varepsilon_{1,2}, \ldots, \varepsilon_{1,n_1}, \varepsilon_{2,1}, \ldots, \varepsilon_{D,n_D})^T$ is the vector of the individual errors. The assumption of normal distributions of the random effect and the individual errors can now be rewritten as

$$V \sim N\left(0_D, \sigma_V^2 I_D\right) \quad \text{and} \quad \varepsilon \sim N\left(0_n, \sigma_\varepsilon^2 I_D\right).$$

This and the independence of V and ε leads to a normal distribution for the observation vector

$$Y \sim N\left(\mathbf{X}\beta, \Sigma\right) \tag{6}$$

with $\Sigma := \sigma_\varepsilon^2 I_n + \sigma_V^2 \mathbf{Z}\mathbf{Z}^T$. With the assumption of known variances σ_V^2 and σ_ε^2 this leads directly to the best linear unbiased estimator (BLUE) by the Gauß-Markov Theorem

$$\hat{\beta}(Y) = \left(\mathbf{X}^T\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\Sigma^{-1}Y \tag{7}$$

and the best linear unbiased predictor for the random effect

$$\hat{V}(Y) = \sigma_V^2\mathbf{Z}^T\Sigma^{-1}\left(Y - \mathbf{X}\hat{\beta}(Y)\right). \tag{8}$$

This leads to the best linear unbiased estimator for Y_{ij} given x_{ij} as follows

$$\hat{Y}_{ij} = x_{ij}^T\hat{\beta} + \hat{V}_i, \tag{9}$$

where $\hat{\beta} = \hat{\beta}(Y)$ from Equation 7 and $\hat{V}_i$ is the i^{th} entry of $\hat{V}$ from Equation 8.

Normally the variance parameters are not known and need to be estimated first. This leads to the empirical best linear unbiased estimator and predictor (EBLUE & EBLUP), where the variance parameters are replaced in Equations 7 and 8 by their estimators $\hat{\sigma}_V^2$ and $\hat{\sigma}_\varepsilon^2$ (Rao (2003), Chapter 6.2.3). This approach is a two-stage method, where the variance parameters are estimated first and then set into the BLUE and

BLUP equations 7 and 8.

Another way of dealing with unknown variance parameters is a maximum likelihood approach. From (6) follows directly the density and thus the log-likelihood density of the observation Y. The unknown parameters are $\theta = (\sigma_V, \sigma_\varepsilon, \beta^T)^T$. By differentiation of the log-likelihood density and setting this derivative to zero the maximum likelihood estimator $\hat{\theta} = (\hat{\sigma}_V, \hat{\sigma}_\varepsilon, \hat{\beta}^T)^T$ can be derived. In a last step the BLUP for V is obtained as in Equation 8 by replacing the variance components σ_V and σ_ε by their estimates $\hat{\sigma}_V$ and $\hat{\sigma}_\varepsilon$. In the end the best linear unbiased estimator for Y_{ij} given x_{ij} is given as in (9).

This maximum likelihood approach is also a two-stage method and is similar to the one we are going to employ for the quantile estimator in mixed models.

3 The Linear Quantile Mixed Model for Continuous Outcomes

One may be interested in estimating target parameters beyond the mean, e.g. the median or quantiles of the target distributions. For quantile estimation in linear mixed models the idea of quantile regression in linear models needs to be adapted.

3.1 The Model

Similar like the quantile model (Koenker and Bassett, 1978) without random effects for a fixed $\tau \in (0, 1)$ the linear quantile mixed model (Geraci and Bottai, 2007) is defined as follows

$$Q_{Y_{ij}|x_{ij}}(\tau) = x_{ij}^T \beta_\tau + V_{\tau,i}, \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i, \tag{10}$$

where $Q_{Y_{ij}|x_{ij}}(\tau)$ stands for the τ-quantile of Y_{ij} given x_{ij}. Thus the linear quantile model was extended by adding the random effect $V_{\tau,i}$. This linear quantile mixed model (10) only needs to be employed whenever the distribution term of the error term in the linear mixed model (1) in unknown. For a known error distribution with distribution function F_ε the τ-quantile of Y_{ij} given x_{ij} is then

$$Q_{Y_{ij}|x_{ij}}(\tau) = x_{ij}^T \beta + V_i + F_\varepsilon^{-1}(\tau), \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i,$$

where β is the same parameter vector as in the linear mixed model (1). In practice one may assume that the distribution of ε is unknown giving more flexibility to the model. This is how we proceed in the following.

In contrast to the linear mixed model (1) the random effect $V_{\tau,i}$ carries now τ in a footnote implying that for different τ the random effect may be different. A further discussion about this approach can be found in Weidenhammer (2016). For reasons of simplification we will drop the τ in the subscript in the

following keeping in mind the dependence to τ. In matrix form (10) can be rewritten as

$$Q_{Y|\mathbf{X}}(\tau) = \mathbf{X}\beta_\tau + \mathbf{Z}V, \tag{11}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the same matrices as defined in (4) and (5), respectively. We can state the equivalence of model (10) to

$$Y_{ij} = x_{ij}^T\beta_\tau + V_i + \varepsilon_{\tau,ij}, \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i \tag{12}$$

with

$$\varepsilon_{\tau,ij} \overset{iid}{\sim} ALD(0, \sigma, \tau), \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i.$$

In matrix form this model can be rewritten as

$$Y = \mathbf{X}\beta_\tau + \mathbf{Z}V + \varepsilon_\tau, \tag{13}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the same matrices as defined in (4) and (5), respectively. The error term ε_τ is the vector of the individual error terms $\varepsilon_{\tau,ij}$ in (12). Its distribution is an n-dimensional asymmetric Laplace distribution as discussed in (Geraci and Bottai (2007))

$$\varepsilon_\tau \overset{iid}{\sim} ALD_n(0_n, \sigma, \tau).$$

Hence the asymmetric Laplace distribution serves also as the distribution of the individual error term $\varepsilon_{\tau,ij}$ here. For reasons of simplification we will drop the τ in the subscript of the error term in the following keeping in mind that its distribution is dependent on τ. As in the linear quantile model of Koenker and Bassett (1978) we assume that the scale parameter σ is unknown. Thus it gives a measure of the variance of the individual error term in the linear mixed model (1) whose distribution is assumed to be unknown. Whenever we mention the linear quantile mixed model in the further investigation we mean the latter model (12). Due to the equivalence of the two models this choice is a matter of taste. We prefer model (12) because it has a regular appearance in linear modelling with error terms on the right hand side and the observations on the left hand side. On the other hand model (10) carries the error distribution within the quantile expression on the left hand side and there is no direct exposure of the observation Y_{ij} in this model.

The dependence of the random effects on τ is discussed in detail in Weidenhammer (2016). Following Weidenhammer (2016) we use a random effect which depends on τ. This makes the model more flexible in terms of the distribution of Y within the areas, which cannot be explained by the independent data $\mathbf{X}$. Nevertheless the footnote on V_τ will be dropped in the future appearances for reasons of clarity keeping in mind the dependence on the choice of τ.

3.2 The Quantile Estimation in Linear Mixed Models

For the quantile estimator in the linear mixed model we need an estimator for the parameter β_τ and a predictor for the random vector V leading to the quantile estimator for a fixed $\tau \in (0,1)$

$$\hat{Q}_{Y_{ij}|x_{ij}}(\tau) = x_{ij}^T \hat{\beta}_\tau + \hat{V}_i, \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots n_i. \tag{14}$$

This estimation is fulfilled in two steps, which will be described in the following Sections 3.2.1 and 3.2.2.

3.2.1 Step 1: Maximum Likelihood Estimation

From the linear quantile mixed model (13) we know the conditional distribution of Y given V

$$Y|V \sim ALD_n\left(\mathbf{X}\beta_\tau + \mathbf{Z}V, \sigma, \tau\right).$$

Thus the joint distribution of the observation vector Y and the random effect vector V is given as

$$(Y, V) \sim ALD_n\left(\mathbf{X}\beta_\tau + \mathbf{Z}V, \sigma, \tau\right) \times N_D\left(0_D, \sigma_V^2 I_D\right).$$

It follows that the density of the joint distribution is given as

$$f_{(Y,V)}(y, v) = f_{ALD_n(\mathbf{X}\beta_\tau + \mathbf{Z}V, \sigma, \tau)}(y|v) \cdot f_{N_D\left(0_D, \sigma_V^2 I_D\right)}(v).$$

This can be simplified as in (Weidenhammer (2016)) to the joint distribution

$$f_{(Y,V)}(y, v) = \prod_{i=1}^{D} \left(\prod_{j=1}^{n_i} f_{ALD\left(x_{ij}^T \beta_\tau + v_i, \sigma, \tau\right)}(y_{ij}|v_i) \right) f_{N\left(0, \sigma_V^2\right)}(v_i). \tag{15}$$

The density and thus the distribution of the observation vector Y is then given as the marginal density of the joint density in (15)

$$\begin{aligned} f_Y(y) &= \int_{\mathbb{R}^D} \prod_{i=1}^{D} \left(\prod_{j=1}^{n_i} f_{ALD\left(x_{ij}^T \beta_\tau + v_i, \sigma, \tau\right)}(y_{ij}|v_i) \right) f_{N\left(0, \sigma_V^2\right)}(v_i) dv \\ &\overset{(\star)}{=} \prod_{i=1}^{D} \int_{\mathbb{R}} \left(\prod_{j=1}^{n_i} f_{ALD\left(x_{ij}^T \beta_\tau + v_i, \sigma, \tau\right)}(y_{ij}|v_i) \right) f_{N\left(0, \sigma_V^2\right)}(v_i) dv_i, \end{aligned} \tag{16}$$

where $(\star)$ follows by application of the Theorem of Fubini. A closed form solution of this integral is not calculable. Thus (16) is the simplified expression of the density of the observation Y. The unknown parameters in this density are σ_V, σ, and β_τ. From the density in Equation 16 we can derive the log-

likelihood density $\ell(\theta|y)$ and find the maximum likelihood estimator

$$\hat{\theta} := \text{argmax}_{\theta\in\Theta}\ell(\theta|y).$$

for $\theta = (\sigma_V, \sigma, \beta_\tau^T)^T \in \Theta := \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^p$ as the roots of the equations

$$\frac{\partial}{\partial\theta}\ell(\theta|y) = 0_{p+2}. \tag{17}$$

Since there is no analytical solution to (17), numerical approaches are needed. Geraci and Bottai (2007) first introduced an EM algorithm and later Geraci and Bottai (2014) made use of a Gaussian quadrature. The latter procedure is faster and more stable than the EM algorithm. The user is able to assume different distributions on the random effect and the number of knots in the Gaussian quadrature. As a result the maximum likelihood estimator $\hat{\theta} = (\hat{\sigma}_V, \hat{\sigma}, \hat{\beta}_\tau^T)^T$ can be calculated.
The existence and consistency of this maximum likelihood estimation is proven in Weidenhammer (2016).

3.2.2 Step 2: Prediction of Random Effect

In a second step a prediction for the random effect is calculated using the maximum likelihood estimator $\hat{\theta} = (\hat{\sigma}_V, \hat{\sigma}, \hat{\beta}_\tau^T)^T$ from Step 1 introduced in Section 3.2.1. As in linear mixed models (1) Geraci and Bottai (2014) stated that the predictor for the random effect can be written in the linear quantile mixed model (12) as

$$\hat{V}(Y) = \hat{\sigma}_V^2 \mathbf{Z}^T \hat{\Sigma}^{-1} \left(Y - \mathbf{X}\hat{\beta}_\tau - \hat{E}\left[\varepsilon\right]\right) \tag{18}$$

with the estimated covariance matrix of Y

$$\hat{\Sigma} = \hat{\sigma}_V^2 \mathbf{Z}\mathbf{Z}^T + \widehat{Var}(\varepsilon)$$

and the estimated expected value and variance of ε are

$$\hat{E}\left[\varepsilon\right] = \frac{\hat{\sigma}(1-2\tau)}{\tau(1-\tau)}\mathbf{1}_n \quad \text{and}$$
$$\widehat{Var}(\varepsilon) = \frac{\hat{\sigma}^2(1-2\tau+2\tau^2)}{\tau^2(1-\tau)^2} I_n.$$

These are the expected value and the variance of an n-dimensional asymmetric Laplace distribution with parameters $\mu = 0$, $\hat{\sigma}$, and τ (Weidenhammer (2016)). Note that the estimated covariance matrix can also be rewritten as $\hat{\Sigma} = \hat{\sigma}_V^2 \mathbf{Z}\mathbf{Z}^T + \widehat{Var}(\varepsilon_{1,1}) I_n$ with $\widehat{Var}(\varepsilon_{1,1}) = \frac{\hat{\sigma}^2(1-2\tau+2\tau^2)}{\tau^2(1-\tau)^2}$.
As a result the quantile estimator given in 14 can be calculated by inserting $\hat{\beta}_\tau$ from the maximum likelihood estimation in Step 1 and $\hat{V}_i$ given in Equation 18.

4 Linear Quantile Mixed Models for Count Outcomes

The quantiles of count data must be integers due to the fact that counts themselves are integers. Since the linear quantile mixed model (12) is a model for continuous data, it is not directly applicable on counts. The generalized linear mixed model for a discrete random variable is Y_{ij} given x_{ij} is given as

$$\exp(x_{ij}^T\beta + V_i), \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots, n_i \tag{19}$$

with

$$V_i \overset{iid}{\sim} N(0, \sigma_V^2).$$

This mean model needs to be improved in order to estimate quantiles of Y_{ij} given x_{ij} for a fixed $\tau \in (0, 1)$, $Q_{Y_{ij}|x_{ij}}(\tau)$. This will be fulfilled by jittering the data as discussed in the following Section 4.1. The main idea is the same as for count data in linear models, where Machado and Silva (2005) already showed the consistency of the quantiles of counts. Here, the consistency of quantile estimators in linear mixed models as proved in Weidenhammer (2016) implies the consistency of the quantiles of counts.

4.1 Jittering the Count Data

The observations Y_{ij} ($i = 1, 2, \ldots, D; j = 1, 2, \ldots, n$) are discrete and in linear models Machado and Silva (2005) had the idea of jittering in order to get continuous data. This method also works in the linear mixed model. By adding a standard uniform random variable U_{ij} independent from Y_{ij}, x_{ij}, and V_i we get a continuous observation Z_{ij}:

$$Z_{ij} := Y_{ij} + U_{ij}. \tag{20}$$

On this continuous random variable Z_{ij} we can apply the linear quantile mixed model 12.

Theorem:

For a fixed $\tau \in (0, 1)$ the quantile of Z_{ij} as defined in (20) is said to be

$$Q_{Z_{ij}|x_{ij}}(\tau) = \exp(x_{ij}^T\beta + V_i) + \tau.$$

Proof. Let $\tau \in (0, 1)$ be fixed. For a continuous random variable $Y_{ij} + U(-\tau, 1 - \tau)$, where the mean model (19) holds for Y_{ij} the τ-quantile is said to be

$$\begin{aligned} & Q_{Y_{ij}+U(-\tau,1-\tau)|x_{ij}}(\tau) = \exp(x_{ij}^T\beta + V_i) \\ \iff \quad & Q_{Y_{ij}+U(-\tau,1-\tau)+\tau|x_{ij}}(\tau) = \exp(x_{ij}^T\beta + V_i) + \tau \\ \iff \quad & Q_{Y_{ij}+U(0,1)|x_{ij}}(\tau) = \exp(x_{ij}^T\beta + V_i) + \tau. \end{aligned}$$

□

4.2 Transformation of the Jittered Data

In order to be able to apply the quantile estimation approach of linear quantile mixed models 12 there is need to transform the jittered data Z_{ij}. This is similar to the approach in the linear model discussed in Machado and Silva (2005) and is for a fixed $\tau \in (0,1)$ fulfilled as follows

$$T(Z_{ij},\tau) := \begin{cases} \log(\zeta), Z_{ij} \leq \tau \\ \log(Z_{ij}-\tau), Z_{ij} > \tau \end{cases}$$

with a small value ζ. This transformation is almost a continuous function and $\log(\zeta)$ is just the function value for negative values for $(Z_{ij} - \tau$ since the logarithm is not defined for negative values. Therefore it follows for the transformed jittered data

$$T^{-1}(Z_{ij},\tau) \approx \exp(Z_{ij}) + \tau$$

and therefore we can state the following corollary.

Corollary:

The quantile of the transformed jittered data is given as

$$Q_{T(Z_{ij},\tau)|x_{ij}}(\tau) = x_{ij}^T\beta_\tau + V_i.$$

Proof. The transformation T is almost continuous and thus it holds that

$$Q_{T(Z_{ij},\tau)|x_{ij}}(\tau) = T\left(Q_{Z_{ij}|x_{ij}}(\tau)\right).$$

In Theorem 4.1 was shown that

$$Q_{Z_{ij}|x_{ij}}(\tau) = \exp(x_{ij}^T\beta + V_i) + \tau,$$

which implies that

$$\begin{aligned} Q_{T(Z_{ij},\tau)|x_{ij}}(\tau) &= T\left(\exp(x_{ij}^T\beta + V_i) + \tau, \tau\right) \\ &= \exp(x_{ij}^T\beta + V_i). \end{aligned}$$

□

4.3 Applying Quantile Estimation in the Linear Mixed Model on the Transformed Jittered Data

The transformed jittered data

$$Y_{ij}^{\star} := T(Z_{ij}, \tau)$$

is now continuous and we can now apply the quantile estimation in linear mixed models as introduced in Section 3.2. There we estimate β_τ and predict V. In order to average out the error, which is based in the jittering, we apply an averaged jittering. That means we jitter our data M times and repeat the estimation of β_τ and V in each step. In the end we take the averaged estimators

$$\hat{\beta}_\tau = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_{\tau,m} \quad \text{and} \quad \hat{V} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m.$$

This leads to the quantile estimator of $Y_{ij}^{\star}$

$$\hat{Q}_{Y_{ij}^{\star}|x_{ij}}(\tau) = x_{ij}^T\hat{\beta}_\tau + \hat{V}_i, \quad i = 1, 2, \ldots, D; j = 1, 2, \ldots, n_i. \tag{21}$$

4.4 Back-Transformation and Count Quantile

From the τ-quantile of $Y_{ij}^{\star}$ we can calculate the τ-quantile of the observed counts Y_{ij} by the following theorem.

Theorem:

For a fixed $\tau \in (0, 1)$ the estimator for the τ-quantile of the observed counts Y_{ij} given x_{ij} is given by

$$\begin{aligned}\hat{Q}_{Y_{ij}|x_{ij}}(\tau) &= \lceil T^{-1}(\hat{Q}_{Z_{ij}|x_{ij}}(\tau) - 1\rceil \\ &= \lceil \exp(x_{ij}^T\hat{\beta}_\tau + \hat{V}_i) + \tau - 1\rceil\end{aligned}$$

for $i = 1, 2, \ldots, D$ and $j = 1, 2, \ldots, n_i$.

Proof. Following the ideas of Machado and Silva (2005) the transformation T is almost continuous and bijective and thus it hold that

$$\hat{Q}_{Z_{ij}|x_{ij}}(\tau) = T^{-1}\left(\hat{Q}_{Y_{ij}^{\star}|x_{ij}}(\tau)\right).$$

Because of $Y_{ij} = Z_{ij} + U_{ij}$ with $U_{ij} \sim U(0, 1)$ it also holds that

$$Y_{ij} - 1 \leq Z_{ij} - 1 \leq Y_{ij}.$$

Because the quantile function is nondecreasing this implies

$$\hat{Q}_{Y_{ij}|x_{ij}}(\tau) - 1 \leq \hat{Q}_{Z_{ij}|x_{ij}}(\tau) - 1 \leq \hat{Q}_{Y_{ij}|x_{ij}}(\tau).$$

The result now follows because $\hat{Q}_{Y_{ij}|x_{ij}}(\tau)$ is an integer. □

We discussed that the idea of jittering count data also works in linear mixed models. Thus one is able to estimate quantiles of count data by applying the quantile estimation in linear mixed models described in Section 3.2. This method works on continuous data, which is why the count data needed to be made continuous by the jittering and transformed in order to have a linear quantile mixed model as in (12). After the estimation a back-transformation of the quantile estimators of the transformed jittered data gives the quantiles of the counts. Additional details are provided in Weidenhammer (2016).

5 Quantile Nested Error Regression Model for Domain Prediction

In practice there are parameters of interest in one area or overall observations, which are beyond mean estimation. In the linear mixed model (1) the predictor of Y given x as given in (9) is a predictor for the mean for the j^{th} unit in area i. The area mean $\hat{\bar{Y}}_i$ can then be given as the averaged means

$$\hat{\bar{Y}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{Y}_{ij}$$

or for the samples units ($j \in S_n$) and the unsampled units ($j \in R_n$)

$$\hat{\bar{Y}}_i = \frac{1}{N_i} \left(\sum_{j \in S_n} Y_{ij} + \sum_{j \in R_n} \hat{Y}_{ij} \right).$$

Thus the mean of an area is the mean of all mean predictors. Similar the overall mean can be given as

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{i=1}^{D} \sum_{j=1}^{N_i} \hat{Y}_{ij}.$$

This is totally different in quantile estimation. Equation 14 gives the conditional τ-quantile for the j^{th} unit in area i, from which one cannot derive the τ-quantile of the whole area $\hat{Q}_{Y_i|x_i}(\tau)$ nor the overall τ-quantile $\hat{Q}_{Y|x}(\tau)$. The mean of quantiles is no quantile

$$\hat{Q}_{Y_i|x_i}(\tau) \neq \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{Q}_{Y_{ij}|x_{ij}}(\tau).$$

Nevertheless there is a way of estimating area quantiles and more parameters of interest, which is called Microsimulation via Quantiles (MvQ) (Weidenhammer, 2016).

5.1 The Idea of Microsimulation via Quantiles

Between quantiles and the distribution of a random variable Y exists a natural relationship. The distribution function F_Y can be rewritten as

$$F_Y(y) = \min\{\tau | Q_Y(\tau) \geq y\}.$$

Thus the empirical distribution function can be rewritten as

$$\hat{F}_Y(y) = \min\left\{\tau | \hat{Q}_Y(\tau) \geq y\right\},$$

where $\hat{Q}_Y(\tau)$ are the empirical quantiles.
In linear mixed models the quantiles can be estimated as given in (14). This estimation is fulfilled on a fixed τ. Let us now estimate quantile estimators on a increasing grid of τ's $T_K := (\tau_1, \tau_2, \ldots, \tau_K)^T$ with $\tau_k < \tau_{k+1}$ for all $k = 1, 2, \ldots K$. This leads to an empirical distribution function of Y_{ij}, the outcome for the j^{th} unit in area i as follows

$$\hat{F}_{Y_{ij}|x_{ij}}(y) = \min\left\{\tau_k | \hat{Q}_{Y_{ij}|x_{ij}}(\tau_k) \geq y, k = 1, 2, \ldots, K\right\}, \tag{22}$$

which is also dependent on the choice of the grid T_K. Thus we are able to estimate the whole distribution of the j^{th} unit in area i by (22). This even gives us the distribution of Y within one area or the over all distribution, from which we are able to estimate every parameter of interest by Monte Carlo simulation.

5.2 The Implementation of Microsimulation via Quantiles

For a given grid of τ, $T_K = (\tau_1, \tau_2, \ldots, \tau_K)$, e.g. $T_{99} = (.01, .02, \ldots, .99)$, we estimate the quantiles as described in Section 3.2. This gives us an $N \times K$-dimensional matrix

$$\begin{pmatrix} \hat{Q}_{y_{11}|x_{11}}(\tau_1) & \hat{Q}_{y_{11}|x_{11}}(\tau_2) & \ldots & \hat{Q}_{y_{11}|x_{11}}(\tau_K) \\ \hat{Q}_{y_{12}|x_{12}}(\tau_1) & \hat{Q}_{y_{12}|x_{12}}(\tau_2) & \ldots & \hat{Q}_{y_{12}|x_{12}}(\tau_K) \\ \vdots & \vdots & & \vdots \\ \hat{Q}_{y_{DN_D}|x_{DN_D}}(\tau_1) & \hat{Q}_{y_{DN_D}|x_{DN_D}}(\tau_2) & \ldots & \hat{Q}_{y_{DN_D}|x_{DN_D}}(\tau_K) \end{pmatrix}.$$

Each row of this matrix gives us an estimation of the distribution function of Y_{ij} given x_{ij} as given in (22). From each $\hat{F}_{y_{ij}}$ we draw a Monte Carlo sample of size MC

$$\tilde{y}_{ij} = (\tilde{y}_{ij}^{(1)}, \tilde{y}_{ij}^{(2)}, \ldots, \tilde{y}_{ij}^{(MC)})^T. \tag{23}$$

This represents a microsimulation of the outcome Y_{ij} of the j^{th} unit in area i. For the whole area i the Monte Carlo sample

$$\tilde{y}_i = (\tilde{y}_{i1}^{(1)}, \tilde{y}_{i1}^{(2)}, \ldots, \tilde{y}_{i1}^{(MC)}, \ldots, \tilde{y}_{iN_i}^{(1)}, \tilde{y}_{iN_i}^{(2)}, \ldots, \tilde{y}_{iN_i}^{(MC)})^T.$$

is a microsimulation of size $N_i \cdot MC$. This sample is just the combination of all microsimulations given in (23) and gives an estimated distribution of the outcome of Y in area i. Similar to this approach one could draw a microsimulation of all units $\tilde{y}$ and areas by glueing the samples given in (23) for all j and i together.

From $\tilde{y}_i$ or $\tilde{y}$ we can estimate now every parameter of interest. This can be fulfilled by taking the empirical version of this parameter from $\tilde{y}_i$ or $\tilde{y}$. Say we want to know the area mean the estimator would be

$$\widehat{mean}_i = mean(\tilde{y}_i)$$

and the τ-quantile estimator in area i is

$$\hat{Q}_{Y_i|x_i}(\tau) = q_\tau(\tilde{y}_i),$$

where $q_\tau(\tilde{y}_i)$ is defined as the empirical τ-quantile of the vector $\tilde{y}_i$. In the same matter other parameters can be estimated from the microsimulated data $\tilde{y}$. This approach can also performed for linear models by setting the quantile estimators of Koenker and Bassett (1978) in the empirical distribution function $F_{Y_i|x_i}$.

Microsimulation via Quantiles (MvQ) provides good tools for estimating parameter, which are beyond the mean like quantiles. Since there is the empirical distribution function estimated, we get the distribution of the observation Y and may get any parameter of interest from that. This can be easily fulfilled by a Monte Carlo simulation. Then even parameters like the Gini coefficient or poverty rates are possible. Furthermore the MvQ method can be combined with the jittering introduced in Section 4. Hence parameters of interest of count data may also be estimated. Therefore the quantile estimators of the count data, which can be estimated as described before serve as the inverse of the empirical distribution function. From there everything else can be obtained by a Monte Carlo simulation.

6 Mean squared error estimation for the MvQ

Molina and Rao (2010) have already mentioned that mean squared error estimation (MSE) is a difficult problem in the case of non-linear indicators and analytic solutions are hard to obtain. In this section we introduce two bootstrap procedures for estimating the MSE of the proposed MvQ approach we presented in Section 5. In particular, the first bootstrap scheme generates bootstrap populations in the case of con-

tinuous outcomes. In contrast, the second approach can be applied for count outcomes and incorporates the additional uncertainty due to the jittering.

MSE estimation for Continuous Outcomes:

The steps of the bootstrap are as follows:

1. We select τ at random by using a uniform distribution $U(0,1)$.
2. For given $\hat{\sigma}_V$ estimated with the original sample generate V_i^* from $N(0, \hat{\sigma}_V)$. An alternative is to generate V_i^* non-parametrically but by using centering and rescaling to adjust for shrinkage following Carpenter et al. (2003).
3. $\varepsilon^*_{\tau,ij}$ are re-sampled from the empirical distribution of residuals appropriately centered and rescaled (Carpenter et al., 2003). An alternative option is to use a wild bootstrap (Feng et al., 2011) in the case of quantile mixed models to accommodate the non-id case.
4. For given $\hat{\beta}_\tau$ estimated with the original sample, V_i^* and $\varepsilon^*_{\tau,ij}$ generate the bootstrap population according to model 13 by
$$Y^* = \mathbf{X}\hat{\beta}_\tau + \mathbf{Z}V^* + \varepsilon^*_\tau. \tag{24}$$
5. Construct B bootstrap populations.
6. For each population b compute the population target indicators, z_i^{*b}.
7. From each bootstrap population select a bootstrap sample according to the sampling scheme of the original sample.
8. Implement the MvQ presented in Section 5 with the bootstrap sample, get $\hat{z}_i^{*b}$

$$\widehat{MSE}(\hat{z}_i) = B^{-1}\sum_{b=1}^{B}(\hat{z}_i^{*b} - z_i^{*b})^2.$$

MSE estimation for Count Outcomes:

The steps of the bootstrap are as follows:

1. For given $\hat{\sigma}_V$ estimated with the original sample generate V_i^* from $N(0, \hat{\sigma}_V)$ at $\tau = 0.5$. An alternative is to generate V_i^* non-parametrically but by using centering and rescaling to adjust for shrinkage following Carpenter et al. (2003). Note that it is also possible to use a quantile τ that is randomly selected from a uniform distribution $U(0,1)$.
2. Calculate the linear predictor η^*_{ij} by
$$\eta^*_{ij} = x_{ij}^T\hat{\beta}_\tau + V_i^*.$$

3. Match η_{ij}^* and $\hat{\eta}_t = \mathbf{x}_t^T \hat{\beta}_\tau + \hat{V}_i$ $(t \in N)$ by

$$\min_{t \in N} \left| \eta_{ij}^* - \hat{\eta}_t \right|$$

and define t^* as the corresponding index.

4. Select

$$Y_{ij}^* \sim \hat{F}_{Y_{t^*}}(y),$$

where $\hat{F}_{Y_{t^*}}(y)$ is defined in 22.

5. Construct B bootstrap populations.

6. For each population b compute the population target indicators, z_i^{*b}.

7. From each bootstrap population select a bootstrap sample according to the sampling scheme of the original sample.

8. Implement the MvQ for count data presented in Section 4 and 5 with the bootstrap sample, get $\hat{z}_i^{*b}$

$$\widehat{MSE}(\hat{z}_i) = B^{-1} \sum_{b=1}^{B} (\hat{z}_i^{*b} - z_i^{*b})^2.$$

The properties of both bootstrap schemes for the count and continuous data we describe in this section are empirically evaluated in Section 7.

7 Model-based evaluations

In this section, we present results from Monte-Carlo simulations that we carried out for assessing the performance of the proposed MvQ approach from Section 5. This estimator is compared against alternative methodology like the empirical best prediction (EBP) approach introduced by Molina and Rao (2010) for continuous outcomes in Section 7.1. We further evaluate the performance of the MSE estimators for continuous and count outcomes discussed in Section 6.

7.1 Continuous Outcomes

We generated population data for $D = 50$ small areas with $N_i = 200$ using a nested error regression model as follows

$$Y_{ij} = 4500 - 400x_{ij} + V_i + \varepsilon_{ij}. \tag{25}$$

The covariates were generated from a normal-distribution with $x_{ij} \sim N(\mu_i, 3^2)$ with $\mu_i \sim U(-3, 3)$ and the random effects were generated by $V_i \sim N(0, 500^2)$. The unit level errors ε_{ij} under three different settings. In particular, we focus on

- Normality: $\varepsilon_{ij} \sim N(0, 1000^2)$
- Contamination: $\varepsilon_{ij} \sim 0.98N(0, 1000^2) + 0.02N(0, 6000^2)$
- Heteroscedasticity: $\varepsilon_{ij} = (1 + 0.1x_{ij})e_{ij}$ with $e_{ij} \sim N(0, 1000^2)$.

Note that additional simulations results are available from the authors on request. The studies cover, for instance, scenarios where the unit level errors are generated by Pareto, log-normal or extreme value distributions to mimic characteristics of income data.

The samples were selected from the population by simple random sampling without replacement within each area leading to a sample size of $n = 921$ ($min = 8$, $mean = 18.4$, $max = 29$). The population and sample sizes were held fixed for all areas. Each setting was repeated independently $R = 100$ times. Three estimators of the small area population indicators are evaluated. These are the EBP approach of Molina and Rao (2010), the proposed MvQ estimator introduced in Section 5 and the direct estimator which only relies on sample information from the particular small area. We focus here on non-linear indicators; in particular, the Gini coefficient (gini), the head count ratio (hcr), poverty gap (pgap) and the 25%, 50% and 90% quantiles. For a detailed definition of the indicators we refer to Foster et al. (1984).

The following quality measures, over Monte-Carlo simulations R, are used to evaluate the performance of an estimator of the target indicator in area i, $\hat{\kappa}_i$,

- Absolute bias

$$Bias(\hat{\kappa}_i) = \frac{1}{R}\sum_{r=1}^{R} \hat{\kappa}_{i,r} - \kappa_{i,r}.$$

- Relative bias [%]:

$$RB(\hat{\kappa}_i) = \frac{1}{R}\sum_{r=1}^{R} \frac{\hat{\kappa}_{i,r} - \kappa_{i,r}}{\kappa_{i,r}} \cdot 100$$

- Root mean squared error:

$$RMSE(\hat{\kappa}_i) = \sqrt{\frac{1}{R}\sum_{r=1}^{R} (\hat{\kappa}_{i,r} - \kappa_{i,r})^2},$$

$\hat{\kappa}$ is a generic notation used to denote an estimator of the small area target parameter and κ is the corresponding true value. Note that we report relative bias for the 25%, 50% and 90% quantiles and absolute bias for the indicators Gini coefficient, head count ratio and poverty gap.

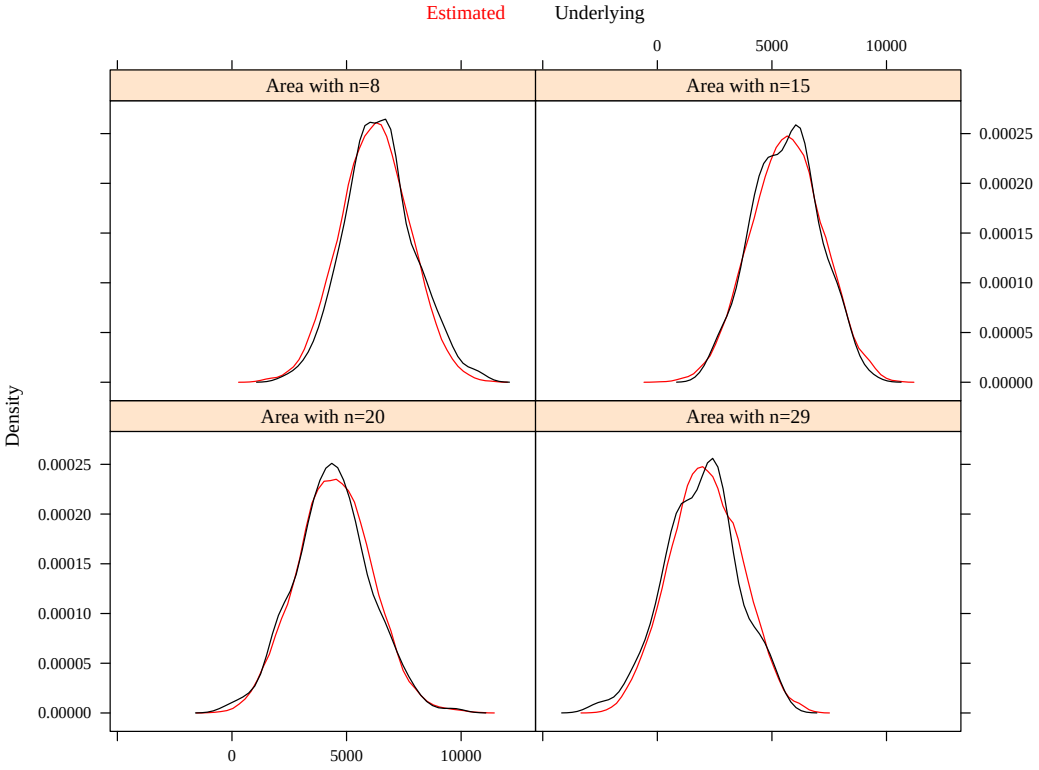

Figure 1: Estimated and true (underlying) area distribution: Normality

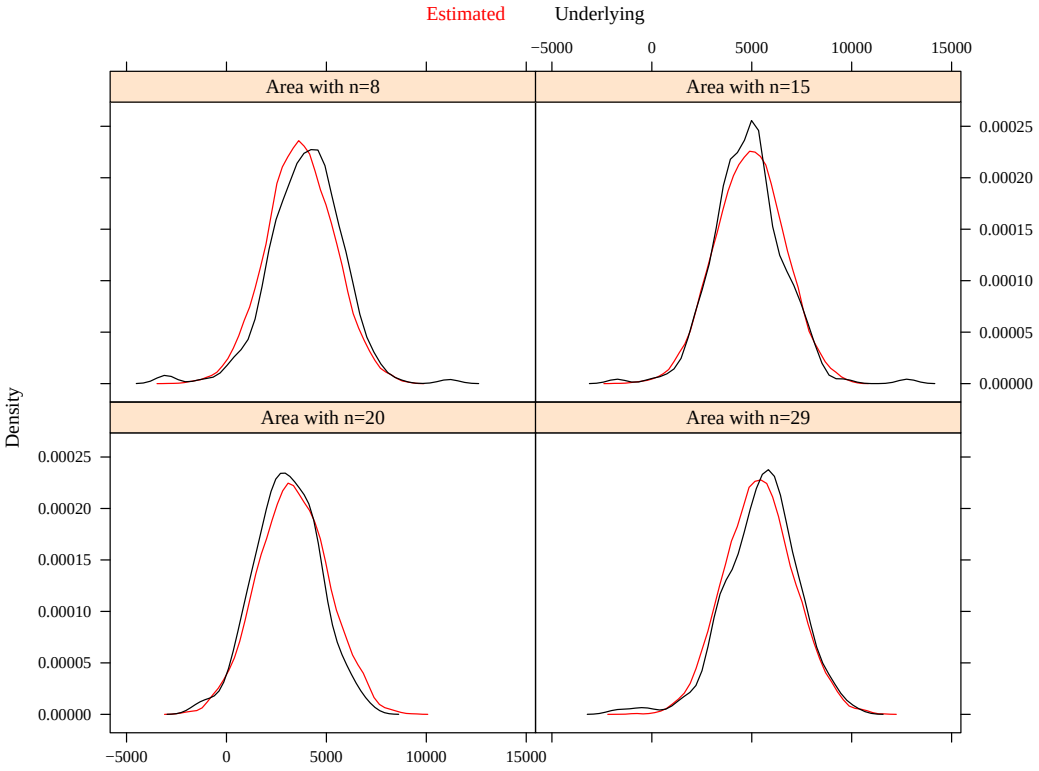

Figure 2: Estimated and true (underlying) area distribution: Contamination

Before discussing the performance of the MvQ compared to the competitors, we have a closer look to the estimated area distributions based on the MvQ approach. In particular, Figures 1 and 2 show the true (black) and estimated (red) area distributions for four small areas of a particular Monte-Carlo simulation run. It can be observed that the MvQ approach can construct the underlying (true) distribution for different area sample sizes.

Table 1: Mean values of RMSE and bias of predictors over small areas.

Normality							
Indicator	Estimator	hcr	pgap	gini	25%	50%	90%
RMSE	Direct	0.0801	0.0363	0.0403	487.1	448.1	599.1
	MvQ	0.0375	0.0180	0.0150	236.0	231.1	278.9
	EBP	0.0364	0.0170	0.0146	232.7	225.5	254.7
Bias	Direct	-0.0008	-0.0009	-0.0116	2.6376	0.1536	-3.4458
	MvQ	-0.0050	-0.0036	0.0003	0.8086	1.2682	1.7924
	EBP	-0.0013	-0.0005	-0.0007	0.7607	0.3283	0.0577
Contamination							
		hcr	pgap	gini	25%	50%	90%
RMSE	Direct	0.0815	0.0560	0.0652	490.9	453.3	638.5
	MvQ	0.0451	0.0248	0.0228	293.2	283.0	405.1
	EBP	0.0499	0.0267	0.0282	320.2	280.7	394.1
Bias	Direct	-0.0016	-0.0003	-0.0086	2.7866	0.2005	-2.8206
	MvQ	0.0048	-0.0041	0.0032	-0.6070	1.6448	4.2329
	EBP	0.0195	0.0067	0.0130	-2.9022	0.6553	3.7867
Heteroscedasticity							
		hcr	pgap	gini	25%	50%	90%
RMSE	Direct	0.0824	0.0486	0.0499	542.4	453.7	483.5
	MvQ	0.0370	0.0237	0.0218	254.6	238.2	282.8
	EBP	0.0380	0.0256	0.0228	271.4	272.3	304.1
Bias	Direct	-0.0016	-0.0012	-0.0105	2.3455	-0.1863	-3.0831
	MvQ	-0.0042	-0.0079	-0.0031	1.0262	1.1119	1.4098
	EBP	0.0024	-0.0124	-0.0011	-1.3808	-2.7965	1.7127

Table 1 reports the average bias (over areas) and the average RMSE (over areas) of the small area estimators for different target indicators. More detailed results regarding the performance are available from the authors on request. As expected the EBP approaches leads to more efficient results (in terms of RMSE) compared to the other estimators in the case of normality. In contrast, the MvQ approach doesn't rely on normality for the unit level errors and lead to more efficient results compared to the EBP for most of the indicators in the scenarios with contamination and heteroscedasticity. We now turn to the bias results in Table 1. Although the biases are small, we notice that the EBP has a smaller bias compared to the direct estimator and the MvQ approach under normality. On the other hand the MvQ has a slightly smaller bias in the case of contamination for the poverty and inequality indicators.

MSE estimation for continuous outcomes for the MvQ approach is implemented with the bootstrap

approaches discussed in Section 6 with $B = 100$ bootstrap replicates. MSE results for the 25% and 90% quantiles are excluded but are available from the authors on request. Table 2 presents the results for the MSE estimators and shows the mean values of their area-specific relative bias (RB) and relative RMSE (RRMSE). Note that the empirical MSE (over Monte-Carlo replications) is treated as the *true* MSE. We denote by *Semi* the semi-parametric bootstrap of Carpenter et al. (2003) where the residuals are re-sampled from the empirical distributions. *Wild* labels the wild bootstrap for quantile mixed models following the ideas of Feng et al. (2011).

We observe that both bootstrap methods work well for the scenario under normality. Under contamination and heteroscedasticity, the *Wild* bootstrap method lead a slightly smaller bias compared the to *Semi* bootstrap at the price of a higher variability in terms of RRMSE. More detailed results are available from the authors on request.

Table 2: Performance of MSE estimators in model-based simulations: Mean values of relative RMSE (RRMSE) and relative bias (RB) over small areas.

Normality					
Indicator	MSE	hcr	pgap	gini	50%
RRMSE	Semi	33.87	51.49	35.63	14.71
	Wild	36.22	53.87	38.53	17.13
RB	Semi	6.49	14.02	14.87	10.97
	Wild	8.43	16.59	16.89	11.81
Contamination					
	MSE	hcr	pgap	gini	50%
RRMSE	Semi	35.09	47.61	38.51	15.66
	Wild	35.77	45.20	35.07	15.77
RB	Semi	-15.53	-24.05	-29.43	-12.79
	Wild	-6.17	-7.85	-8.97	-5.63
Heteroscedasticity					
	MSE	hcr	pgap	gini	50%
RRMSE	Semi	36.41	44.23	32.53	12.92
	Wild	38.79	55.90	45.61	17.37
RB	Semi	11.46	-12.29	-18.39	7.35
	Wild	5.30	5.60	5.87	7.36

7.2 Discrete Outcomes

As in Section 7.1 we generated population data for $D = 50$ small areas with $N_i = 200$ following the generalized linear mixed model 19

$$\eta_{ij} = \exp(0.8 + 2x_{ij} + V_i), \tag{26}$$

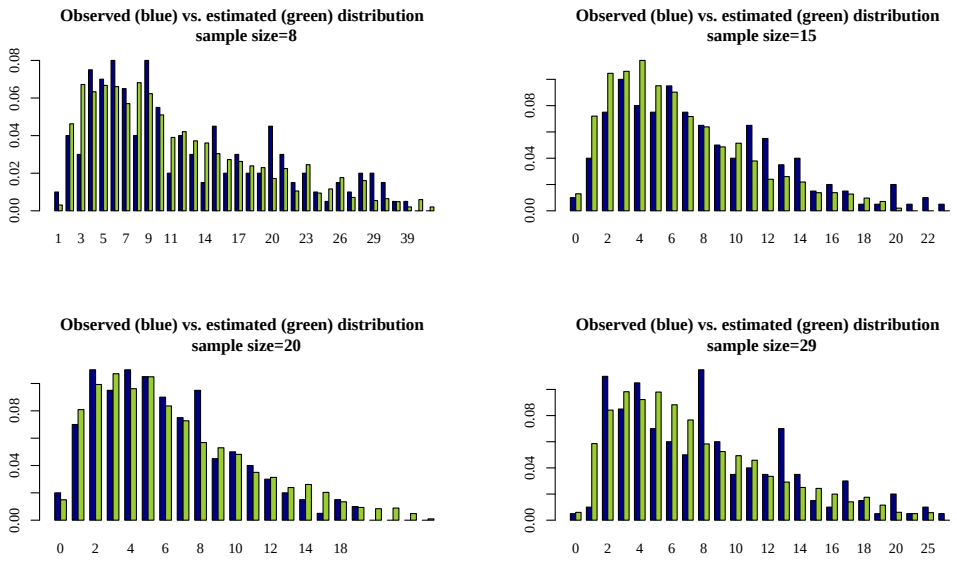

Figure 3: Estimated (green) and true (observed, blue) area distribution: Poisson.

where the covariates were generated from an uniform distribution with $x_{ij} \sim U(0,1)$ and the random effects were generated by $V_i \sim N(0, 0.3^2)$. We investigate two different types of distributions:

- Poisson: $Y_{ij}|V_i \sim Pois(\eta_{ij})$
- Negative binomial: $Y_{ij}|V_i \sim NB(\eta_{ij}, s)$ and $\text{Var}(Y_{ij}|V_i) = \eta_{ij} + \frac{\eta_{ij}^2}{s}$, where s denotes the scale parameter $s = 1, 2, 3, 5$.

According to the continuous case in Section 7.1, the samples were selected from the population by simple random sampling without replacement within each area leading to a sample size of $n = 921$ ($min = 8$, $mean = 18.4$, $max = 29$). The population and sample sizes were held fixed for all areas. Each setting was repeated independently $R = 200$ times.

Three estimators of the small area population indicators are evaluated. These are the Poisson predictor based on a generalized linear mixed model (Glmer), the proposed MvQ estimator for count outcomes introduced in Section 5 and the direct estimator. Note that we focus here on only on the domain means. Additional target parameters like the median or quantiles are available from the authors on request. We used the same quality measures to evaluate the performance of the estimators like in Section 7.1.

Before assessing the performance of the MvQ for count outcomes compared to the competitors, we have a closer look to the estimated area distributions based on the MvQ approach. Figures 3 and 4 show the true (blue) and estimated (green) area distributions for four small areas of a particular Monte-Carlo simulation run for two scenarios. It can be observed that the MvQ approach rebuilds the true (observed) distribution for different area sample sizes.

Figure 5 and 6 present the relative bias and the RMSE of the small area estimators for domain means. As expected the Glmer approaches leads to more efficient results (in terms of RMSE) compared to the direct estimator and the MvQ approach in the case of the Poisson scenario. In contrast, the MvQ approach

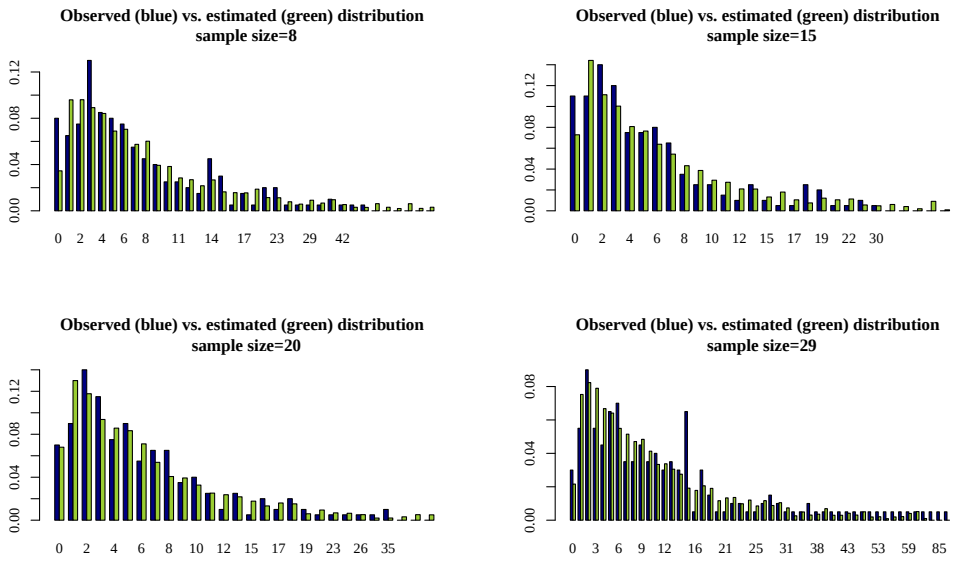

Figure 4: Estimated (green) and true (observed, blue) area distribution: Negative binomial (s=2).

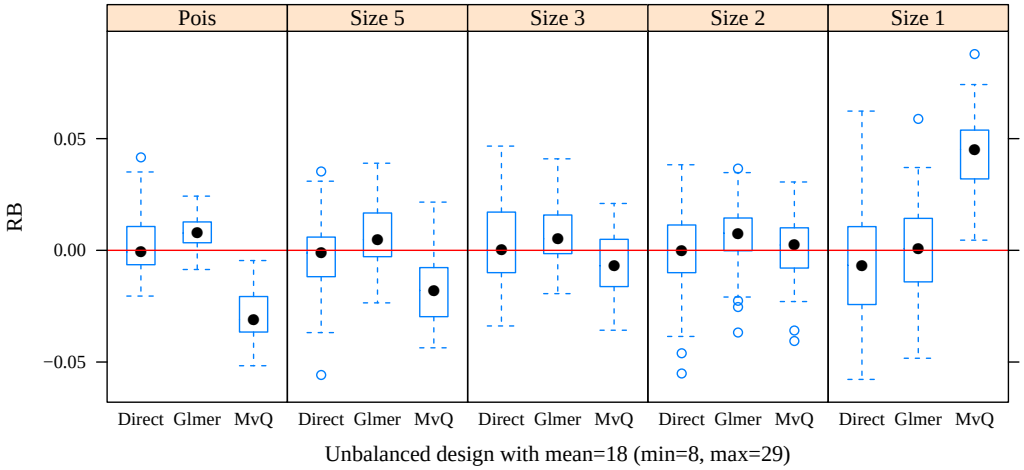

Figure 5: Relative bias of the predictors for the estimation of domain averages.

doesn't rely on the Poisson assumption and is able to adapt on different distributions. This leads to more efficient results compared to the Glmer approach in the context of more skewed distributions (NB size 2 and NB size 1). We now turn to the bias results in Figure 5. Although the biases are small, we notice that the direct estimator and the Glmer approach have a smaller bias compared to the MvQ approach in most of the settings.

MSE estimation for count outcomes for the MvQ approach is implemented with the bootstrap approach discussed in Section 6 with $B = 100$ bootstrap replicates. MSE results for the negative binomial distribution with $s = 1$ and $s = 3$ as well as further evaluations for different target parameters, like the median or other quantiles, are excluded but are available from the authors on request. Table 3 reports the empirical RMSE (over Monte-Carlo replications) and the estimated RMSE. We observe that the proposed bootstrap approach works quite well in these particular scenarios and is able to track the empirical

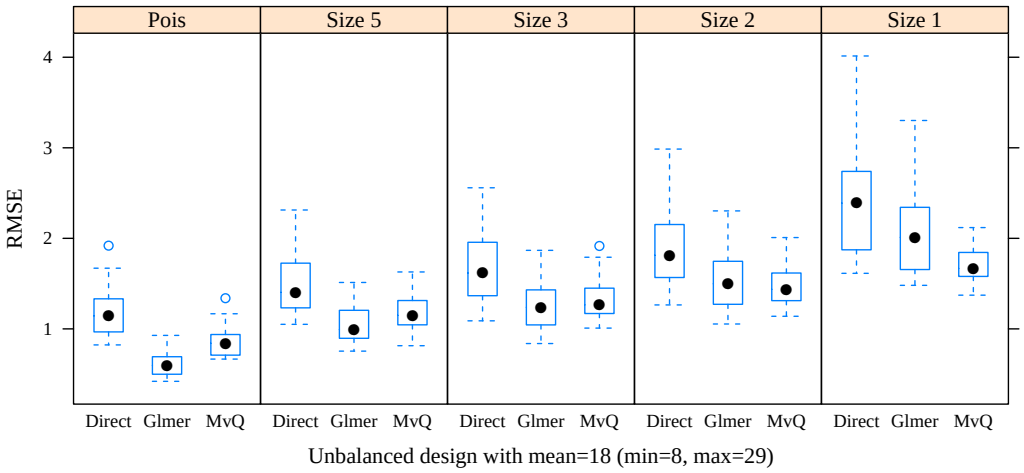

Figure 6: RMSE of the predictors for the estimation of domain averages.

RMSE.

Table 3: Performance of MSE estimators in model-based simulations over small areas.

Poisson						
RMSE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
empirical	0.581	0.662	0.782	0.802	0.923	1.201
estimated	0.618	0.691	0.788	0.827	0.944	1.178
NB $s = 5$						
RMSE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
empirical	0.843	0.986	1.110	1.139	1.314	1.476
estimated	0.908	0.992	1.130	1.157	1.293	1.545
NB $s = 2$						
RMSE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
empirical	1.158	1.278	1.401	1.442	1.621	1.865
estimated	1.189	1.291	1.438	1.465	1.616	1.852

8 Concluding remarks

The paper proposes a new approach for small area estimation. The method is based on constructing a model-based estimator of the distribution function. The gains offered by the proposed methodology are twofold. First, a general set of domain target parameters that extends beyond domain averages can be estimated from the distribution function. Second, the methodology allows for modelling continuous and count outcomes. In particular, the approach allows for more flexible mean-variance relationships in the case of count outcomes and it does not rely on normality of the error term in the case of continuous outcomes. MSE estimation is performed by using different bootstrap schemes. The results from the

model-based simulations indicate that the proposed methodology is a promising alternative to existing unit-level methodologies.

Currently, estimation relies on the normality assumption of the random effects. We are currently investigating alternative specifications of the distribution of the random effects by exploring parametric and non-parametric alternatives. This provides one avenue for future research. Another line for further work could be to investigate the impact of a constrained fitting of the quantiles for constructing the distribution function. Although the implementation of the proposed methodology is facilitated by the availability of a computationally efficient algorithm using C++ in R, its application in practice is challenging. Developing and providing an easy-applicable R package including the proposed methodology offers an additional avenue for future research.

Acknowledgements

Tzavidis and Schmid gratefully acknowledge support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from the UK Economic and Social Research Council.

References

Carpenter, J. R., H. Goldstein, and J. Rasbash (2003). 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.

Feng, X., X. He, and J. Hu (2011). Wild bootstrap for quantile regression. *Biometrika 98*(4), 995–999.

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. *Econometrica 52*(3), 761–766.

Geraci, M. and M. Bottai (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics 8 (1)*, 140–154.

Geraci, M. and M. Bottai (2014). Linear quantile mixed models. *Statistics and Computing 24 (3)*, 461–479.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Machado, J. and J. M. C. S. Silva (2005). Quantiles for counts. *Journal of the American Statistical Association 100 (472)*, 1226–1237.

Molina, I. and J. N. K. Rao (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics 38 (3)*, 369–385.

Rao, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.

Tzavidis, N. and T. Schmid (2015). Applications of mixed models methodology for small area estimation in mexico. In *ISI 2015 Conference*, Rio de Janeiro, Brasil.

Tzavidis, N., T. Schmid, B. Weidenhammer, and N. Salvati (2015). A unit-level quantile nested error regression model for domain prediction with continuous and discrete outcomes. In *ISI 2015 Conference*, Rio de Janeiro, Brasil.

Weidenhammer, B. (2016). *Consistency of Quantile Regression in Linear Mixed Models*. Ph. D. thesis, Freie Universität Berlin.

Weidenhammer, B., N. Tzavidis, T. Schmid, and N. Salvati (2014). Domain prediction for counts using microsimulation via quantiles. In *Small Area Estimation 2014 Conference*, Poznan, Poland.

Corresponding author: Timo Schmid, Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany.
E-mail: timo.schmid@fu-berlin.de

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2016 erschienen:

2016/1 BARTELS, Charlotte und Maximilian STOCKHAUSEN
Children's opportunities in Germany – An application using multidimensional measures
Economics

2016/2 BÖNKE, Timm; Daniel KEMPTNER und Holger LÜTHEN
Effectiveness of early retirement disincentives: individual welfare, distributional and fiscal implications
Economics

2016/3 NEIDHÖFER, Guido
Intergenerational Mobility and the Rise and Fall of Inequality: Lessons from Latin America
Economics

2016/4 TIEFENSEE, Anita und Christian WESTERMEIER
Intergenerational transfers and wealth in the Euro-area: The relevance of inheritances and gifts in absolute and relative terms
Economics

2016/5 BALDERMANN, Claudia; Nicola SALVATI und Timo SCHMID
Robust small area estimation under spatial non-stationarity
Economics

2016/6 GÖRLITZ, Katja und Marcus TAMM
Information, financial aid and training participation: Evidence from a randomized field experiment
Economics

2016/7 JÄGER, Jannik und Theocharis GRIGORIADIS
Soft Budget Constraints, European Central Banking and the Financial Crisis
Economics

2016/8 SCHREIBER, Sven und Miriam BEBLO
Leisure and Housing Consumption after Retirement: New Evidence on the Life-Cycle Hypothesis
Economics

2016/9 SCHMID, Timo; Fabian BRUCKSCHEN; Nicola SALVATI und Till ZBIRANSKI
Constructing socio-demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal
Economics

2016/10 JESSEN, Robin; ROSTAM-AFSCHAR, Davud und Sebastian SCHMITZ
How Important is Precautionary Labor Supply?
Economics

2016/11 BIER, Solveig; Martin GERSCH, Lauri WESSEL, Robert TOLKSDORF und Nina KNOLL
Elektronische Forschungsplattformen (EFP) für Verbundprojekte: Bedarfs-, Angebots- und Erfahrungsanalyse
Wirtschaftsinformatik