Unit level model for small area estimation with count data under square root transformation

Kelly C. M. Gonçalves\textsuperscript{a,1}, Malay Ghosh\textsuperscript{b}

\textsuperscript{a}Department of Statistical Methods, Universidade Federal do Rio de Janeiro
\textsuperscript{b}Department of Statistics, University of Florida

Abstract. In recent years, the demand for small area statistics has greatly increased worldwide. Small area models are formulated with random area-specific effects assumed to account for the between-area variation that is not explained by auxiliary variables. The unit level models relate the unit values of a study variable to unit-specific covariates. The main aim of this paper is to consider small area estimation under unit level models based on count data. In particular, instead of modelling the variables assuming the Poisson distribution, which is a usual choice, we consider the square root transformation of the original data. One practical advantage is that the proposed transformation achieves approximate homoscedasticity of the error variances, reducing one layer of estimation problem. Inference for the model is carried out under the hierarchical Bayes approach. The square root transformation is evaluated under a simulation study and two design-based studies with real datasets.

1 Introduction

Small area estimation is gaining increasing popularity in recent years. The need for such estimates is felt both in the public and private sectors. By necessity, small area inference is model-based owing to the lack of sufficient samples in individual local areas. Consequently, the direct estimates for such areas are met with high variances and high coefficients of variation. Improved estimates for these areas are obtained by “borrowing strength” from similar other areas. Rao and Molina (2015) provided an excellent account of a wide range of topics on small area estimation.

Small area models can be broadly classified as (i) area level models and (ii) unit level models. The former is used oftener than the latter because unit level models require information for the individual units in the sample. This is typically unavailable for secondary users of surveys. Fay and Herriot (1979) is a classic reference for normal theory area level models.

There exists some literature on area level small area estimation based on count data. One may refer to Tsutakawa (1988) and Papageorgiou and Ghosh (2012). These papers considered an empirical Bayes approach under a log-linear model.

One of the classical unit level models is due to Battese, Harter and Fuller (1988) (BHF). Their objective was to estimate areas under corn and soybean for twelve counties in North Central Iowa. The BHF model is a normal mixed effects model comprising of two variance components, one corresponding to the random area effect, the other being the error variance at the unit level. The BHF estimators of small area means are essentially empirical best linear unbiased predictor (EBLUP) or equivalently empirical Bayes (EB) predictors. In contrast, Datta and Ghosh (1991) proposed a hierarchical Bayes approach in a more general framework involving multiple variance components under normality assumption.

\textsuperscript{1}K. C. M. Gonçalves was supported by a scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível (CAPES)

Keywords and phrases. posterior predictive moments, normal approximation, Gibbs Sampling, NHANES, Prova Brasil.
The objective of the present paper is small area estimation under unit level models based on count data. We assume the usual Poisson distribution, but use a square root transformation (Anscombe, 1952) of the original data. Such transformation is one of the variance-stabilizing transformations and serves two useful purposes. First, a normal approximation is usually more appropriate for the transformed data. Second, this transformation achieves approximate homoscedasticity of the error variances, thus reducing one layer of estimation problem. For modeling count data, there are other options, such as the Binomial or the Negative-Binomial model, the second one being more popular to account for overdispersion. However, variance-stabilizing transformation for the latter uses more complex functions, namely the arcsin and inverse hyperbolic sine transformation, respectively (Yu, 2009). The present paper extends the work of Ghosh, Ghosh and Hirose (2020) who considered also square root transformation of count data, but only at the area level model.

The outline of the remaining sections is as follows. In Section 2, we introduce the square root transformation as well as the hierarchical Bayesian model. We also discuss the Markov chain Monte Carlo implementation of our proposed procedure. Section 3 consists of a simulation study, while Section 4 consists of the analysis of two real datasets. One was taken from US National Health and Nutrition Examination Survey (NHANES) and pertains to sports, fitness and recreational activity of individuals in fifty small domains cross-classified by race and ethnicity. The other was taken from the national population-based assessment in Mathematics of public schools called Prova Brasil, conducted by Instituto Nacional de Estudos Educacionais Anísio Teixeira (INEP) in Brazil. We show via simulation that the proposed square root transformation performs better than a logarithmic transformation of the original data. The latter has also the potential disadvantage of disregarding zero counts which happens quite often in small area estimation, since the accompanying auxiliary information can be quite pertinent for overall inference. In most cases the square root transformation also performs better than or similar as the Poisson log-linear model, which is a natural choice in this case. Finally, Section 5 summarizes our main findings.

2 Square Root Transformation and the Bayesian Model

Consider $m$ local areas with population size $N_i$ for the $i$th local area. A sample of size $n_i$ is drawn from the $i$th local area. Let $y_{ij}$ denote the count for the $j$th unit in the $i$th local area. In our first example, the count refers to the number of days per week a person does some physical activity. Our objective is to predict

$$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij},$$

(1) $i = 1, \ldots, m$, based on the sample $y_{ij}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. This we accomplish by finding the posterior predictive distribution of the unobserved $y_{ij}$ ($j = n_i + 1, \ldots, N_i$; $i = 1, \ldots, m$).

We begin with $y_{ij} \sim \text{Poisson}(\lambda_{ij})$ and the transformation $z_{ij} = y_{ij}^{1/2}$. An application of the Delta Method yields $z_{ij}$ approximately normally distributed with mean $\theta_{ij}$ and variance $1/4$, where $\theta_{ij} = \lambda_{ij}^{1/2}$. This well-known transformation, also called variance-stabilizing transformation (Anscombe, 1952), is aimed to achieve homoscedasticity with the added advantage of approximate normality. In the process, we eliminate one extra layer of variance estimation in the BHFS model. Moreover, this approximation is more accurate at lower values of the mean parameter than others (Thacker and Bromiley, 2001).

The present paper does not address the issue of overdispersion. However, we have compared our proposed method with other two approaches: the normal model assumed for the
transformation $z_{ij} = \log(y_{ij})$ and the Poisson distribution assumed for $y_{ij}$ with a log-linear model for the Poisson parameter.

We consider the Bayesian model

\[ z_{ij} = \theta_{ij} + e_{ij} \]
\[ \theta_{ij} = x_{ij}^T \beta + u_i, \] (2)

where $\beta$ is the $p$-vector of regression coefficients, $x_{ij} = (1, x_{ij1}, \ldots, x_{ij,p-1})^T$ ($p < \sum_{i=1}^m n_i$), $u_i$ is the $i$-th area effect, $e_{ij}$ is the random effect associated with the $j$-th unit within the $i$-th area, which are assumed mutually independent with $u_i \sim N(0, A)$ and $e_{ij} \sim N(0, 1/4)$.

It is convenient to rewrite the model in vector and matrix notations. For this, let $X_i^s = (x_{i1}, \ldots, x_{in_i})^T$, $X_i^g = (x_{i n_i+1}, \ldots, x_{i N_i})^T$, $z_i^s = (z_{i1}, \ldots, z_{in_i})^T$, $z_i^g = (z_{in_i+1}, \ldots, z_{i N_i})^T$, for $i = 1, \ldots, m$.

With these notations, $(z_i^s, z_i^g)$ has the following multivariate normal distribution conditional on $\beta$ and $A$:

\[ \begin{pmatrix} z_i^s \\ z_i^g \end{pmatrix} \sim N \left( \begin{pmatrix} X_i^s \beta \\ X_i^g \beta \end{pmatrix}, \begin{pmatrix} I_{n_i} + AJ_{n_i} & A1_{N_i-n_i}^T\\ A1_{N_i-n_i} & 1/4I_{N_i-n_i} + AJ_{N_i-n_i} \end{pmatrix} \right), \]

where $I_{n_i}$ is the identity matrix of dimension $n_i$, $J_{n_i} = 1_{n_i}1_{n_i}^T$ and $J_{N_i-n_i} = 1_{N_i-n_i}1_{N_i-n_i}^T$.

From properties of the multivariate normal distribution, we get the conditional predictive distribution $z_i^s \mid z_i^g, \beta, A \sim N(\mu_i, \Sigma_i)$, where

\[ \mu_i = X_i^s \beta + A1_{N_i-n_i}(1/4I_{n_i} + AJ_{n_i})^{-1}(z_i^g - X_i^g \beta) \]
\[ \Sigma_i = (1/4)I_{N_i-n_i} + A1_{N_i-n_i}(1/4I_{n_i} + AJ_{n_i})^{-1}A1_{N_i-n_i}^T, \] (3)

By the Woodbury identity (Woodbury, 1950) we have $(1/4I_{n_i} + AJ_{n_i})^{-1} = 4I_{n_i} - 16AJ_{n_i}1_{1+4An_i}$. Thus, the mean $\mu_i$ and the variance $\Sigma_i$ in equation (3) simplify to

\[ \mu_i = X_i^s \beta + 4An_i(\bar{z}_i^g - \bar{x}_i^g \beta) \]
\[ \Sigma_i = \frac{1}{4}I_{N_i-n_i} + \frac{A}{(1 + 4An_i)}J_{N_i-n_i}, \]

where $\bar{z}_i^g = n_i^{-1} \sum_{j=1}^{n_i} z_{ij}$ and $\bar{x}_i^g = n_i^{-1} \left( \sum_{j=1}^{n_i} x_{1ij}, \ldots, \sum_{j=1}^{n_i} x_{1ip} \right)$.

In particular, for any $j, j' = n_i+1, \ldots, N_i$, such that $j \neq j'$, conditional on $z_i^g$, $\beta$, and $A$, $(z_{ij}, z_{ij'})^T$ has the following bivariate normal distribution:

\[ \begin{pmatrix} z_{ij} \\ z_{ij'} \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_{ij} \\ \mu_{ij'} \end{pmatrix}, \begin{pmatrix} \sigma_{ij}^2 & \sigma_{ij'} \\ \sigma_{ij'} & \sigma_{ij''} \end{pmatrix} \right), \]

where

\[ \begin{align*}
\mu_{ij} &= x_{ij}^T \beta + 4An_i(\bar{z}_i^g - \bar{x}_i^g \beta), \\
\mu_{ij'} &= x_{ij'}^T \beta + 4An_i(\bar{z}_i^g - \bar{x}_i^g \beta), \\
\sigma_{ij}^2 &= \sigma_{ij'}^2 = \frac{4}{1 + 4An_i}, \\
\sigma_{ij'} &= \frac{A}{1 + 4An_i}. \\
\end{align*} \]

We are interested in finding posterior predictive means, variances and the covariances of $y_{ij} = z_{ij}^2$, $j = n_i + 1, \ldots, N_i$, $i = 1, \ldots, m$. To this end, we need the following lemma.

**Lemma 1.** Suppose $T_1$ and $T_2$ have a joint bivariate Normal distribution with means $\mu_k$ and variances $\sigma_k^2$, $k = 1, 2$, respectively, and covariance $\sigma_{kk'}$, $k, k' = 1, 2$. Then

\[ E(T_2^2) = \sigma_k^2 + \mu_k^2, \quad V(T_2) = 2\sigma_k^4 + 4\sigma_k^2\mu_k^2, \]
\[ Cov(T_k^2, T_k') = 2\sigma_{kk'} + 4\sigma_{kk'}\mu_k\mu_k', \quad k \neq k'. \]
The proof of Lemma 1 is given in Appendix A.

By this result, for \( j, j' = n_t + 1, \ldots, N_t \) and \( i = 1, \ldots, m \), we get

\[
E(y_{ij} | z_{ij}^s, \beta, A) = \sigma_{ij}^2 + \mu_{ij}^2, \quad E(y_{ij} | z_{ij}^s, \beta, A) = \sigma_{ij}^2 + \mu_{ij}^2.
\]

\[
V(y_{ij} | z_{ij}^s, \beta, A) = 2\sigma_{ij}^4 + 4\sigma_{ij}^2\mu_{ij}^2, \quad V(y_{ij} | z_{ij}^s, \beta, A) = 2\sigma_{ij}^4 + 4\sigma_{ij}^2\mu_{ij}^2,
\]

\[
\text{Cov}(y_{ij}, y_{ij'} | z_{ij}^s, \beta, A) = 2\sigma_{ij}^2\mu_{ij} + 4\sigma_{ij}^2\mu_{ij'},
\]

for \( \mu_{ij}, \mu_{ij'}, \sigma_{ij}^2, \sigma_{ij'}^2 \) and \( \sigma_{ijj'} \) given in equation (4). Substituting equations (4) into equations (5), we get

\[
E(y_{ij} | z_{ij}^s, \beta, A) = \frac{1}{4} + \frac{A}{1+4A_m} + \left[ x_{ij}^T \beta + \frac{4A_n}{1+4A_n} (z_{ij}^s - \bar{x}_i^s) \right]^2
\]

\[
V(y_{ij} | z_{ij}^s, \beta, A) = 2 \left( \frac{1}{4} + \frac{A}{1+4A_m} \right)^2 + 4 \left( \frac{1}{4} + \frac{A}{1+4A_m} \right) \left[ x_{ij}^T \beta + \frac{4A_n}{1+4A_n} (z_{ij}^s - \bar{x}_i^s) \right]^2
\]

\[
\text{Cov}(y_{ij}, y_{ij'} | z_{ij}^s, \beta, A) = 2 \left( \frac{A}{1+4A_m} \right)^2 + 4 \left( \frac{A}{1+4A_m} \right) \left[ x_{ij}^T \beta + \frac{4A_n}{1+4A_n} (z_{ij}^s - \bar{x}_i^s) \right] \times \left[ x_{ij'}^T \beta + \frac{4A_n}{1+4A_n} (z_{ij'}^s - \bar{x}_i^s) \right]
\]

Now, let \( Z^s \) represent the column vector of the observed sample units for the \( m \) small areas, that is, \( Z^s = (z_1^T, \ldots, z_m^T)^T \). The main aim is to predict the population mean for each small area, as described in equation (1). We have

\[
E \left[ \sum_{j=1}^{N_t} y_{ij} | Z^s \right] = \sum_{j=1}^{N_t} y_{ij} + \sum_{j=n_t+1}^{N_t} E \left[ y_{ij} | Z^s \right]
\]

\[
V \left[ \sum_{j=1}^{N_t} y_{ij} | Z^s \right] = V \left[ \sum_{j=n_t+1}^{N_t} y_{ij} | Z^s \right].
\]

In particular, we can get improved estimates with the Rao-Blackwellization (Gelfand and Smith, 1990) for the quantities in equation (7), specifically,

\[
E \left[ y_{ij} | Z^s \right] = E \left[ E \left( y_{ij} | z_{ij}^s, \beta, A \right) | Z^s \right]
\]

\[
V \left[ y_{ij} | Z^s \right] = V \left[ E \left( y_{ij} | z_{ij}^s, \beta, A \right) | Z^s \right] + \text{Cov} \left[ E \left( y_{ij} | z_{ij}^s, \beta, A \right) | Z^s \right] + \text{Cov} \left[ E \left( y_{ij} | z_{ij}^s, \beta, A \right), E \left( y_{ij'} | z_{ij'}^s, \beta, A \right) | Z^s \right].
\]

The previous results are all obtained conditional on the parameter vector \( \phi = (\beta, A^{-1}) \) which need to be estimated. In particular, to get the posterior predictive moments described in equation (8), we consider a hierarchical Bayes approach using the Gibbs sampling method.

### 2.1 Hierarchical Bayes approach

Inference procedure for the parametric vector \( \phi = (\beta, A^{-1}) \) is performed under the Bayesian paradigm. So model specification is completed after assigning a prior distribution, \( \pi(\phi) \). An advantage of the Bayesian paradigm is that the inference procedure is performed in a single framework, and uncertainty about parameter estimation is naturally accounted for in the prediction. The posterior distribution contains all information about the parametric vector given the data, which leads to a direct evaluation of a predictive distribution via an integral.

We assume the components of \( \phi \) to be independent, a priori. It is assumed that a priori, \( \pi(\beta) \propto k \), where \( k \) is a constant and \( A^{-1} \sim \text{Gamma}(a_0/2, b_0/2) \), where \( \text{Gamma}(\alpha, \theta) \) denotes the Gamma distribution with mean \( \alpha/\theta \) and variance \( \alpha/\theta^2 \). Let \( u = (u_1, \ldots, u_m) \).

The joint conditional distribution of \( (\beta, u, A^{-1}) \) given \( Z^s \) is

\[
p(\beta, u, A^{-1} | Z^s) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^{m} \left( z_i^s - X_i^s \beta - u_i 1_{n_i} \right)^T \left( z_i^s - X_i^s \beta - u_i 1_{n_i} \right) \right\} \times A^{-m/2} \exp \left\{ \frac{1}{2A} \sum_{i=1}^{m} u_i^2 \right\} \times A^{-a_0/2+1} \exp \left\{ -\frac{b_0}{2} A^{-1} \right\}.
\]
The prior is improper. Propriety of the posterior distribution is proved in Appendix B. The posterior distribution does not have analytical closed-form. Thus we use MCMC methods to obtain samples from it. Gibbs sampling is particularly simple here since all the full conditional distributions have a closed form.

- \( \beta \mid . \sim N(\mu_\beta, \Sigma_\beta) \), where
  \[ \Sigma_\beta^{-1} = 4 \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij}x_{ij}^T \]
  and \( \mu_\beta = \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij}x_{ij}^T \right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} (z_{ij} - u_i) \).
- \( A^{-1} \mid . \sim \text{Gamma}(a_1/2, b_1/2) \), where
  \[ a_1 = a_0 + m \]
  and \( b_1 = b_0 + \sum_{i=1}^{m} u_i^2 \).
- \( u_i \mid . \sim N(\mu_{u_i}, \Sigma_{u_i}) \), where
  \[ \Sigma_{u_i}^{-1} = 4n_i + A^{-1} \]
  and \( \mu_{u_i} = \left( \frac{4n_iA}{4n_iA+1} \right) \left( \bar{z}_i - \bar{x}_i \beta \right) \), for \( i = 1, \ldots, m \).

We obtain estimates for the posterior mean and variance in equation (8) using the posterior samples.

3 Simulation study

In order to evaluate the model’s performance, we generated several samples and compared the square root transformation model fit in equation (2) with the logarithmic one, another usual Gaussian transformation. However, the logarithmic transformation differs from the model in equation (2) by \( e_{ij} \)’s distribution. It is assumed that \( e_{ij} \mid \sigma^2 \sim N(0, \sigma^2) \) and \( \sigma^{-2} \sim \text{Gamma}(a_\sigma/2, b_\sigma/2) \) a priori. Thus, while the square root root transformation achieves approximate homoscedasticity of the error variances, reducing one layer of estimation problem, in the log transformation we should estimate the variance \( \sigma^2 \).

The inference procedure for the log model is very similar to the one presented in Section 2 for the square root transformation, expect that \( y_{ij} = \exp(z_{ij}) \). To get the posterior moments, Lemma 1 should be replaced by the following one.

Lemma 2. Suppose \( T_1 \) and \( T_2 \) have a joint bivariate normal distribution with means \( \mu_k \) and variances \( \sigma_k^2 \), \( k = 1, 2 \), respectively and covariance \( \sigma_{kk'} \), \( k, k' = 1, 2 \). Then

- \( E(\exp(T_k)) = \exp(\mu_k + \sigma_k^2/2) \)
- \( V(\exp(T_k)) = \exp(2\mu_k + \sigma_k^2) \left[ \exp(\sigma_k^2) - 1 \right] \) and
- \( \text{Cov}(\exp(T_k), \exp(T_{k'})) = \exp(\mu_k + \mu_{k'} + \sigma_k^2/2 + \sigma_{kk'}/2) \left[ \exp(\sigma_{kk'}) - 1 \right] \).

The proof of Lemma 2 is also given in Appendix A. By this result, for \( j, j' = n_i + 1, \ldots, N_i \), \( i = 1, \ldots, m \), we get

- \( E(y_{ij} \mid z_{ij}^s, \beta, A) = \exp(\mu_{ij} + \sigma_{ij}^2/2) \), \( E(y_{ij} \mid z_{ij}^s, \beta, A) = \exp(\mu_{ij} + \sigma_{ij}^2/2) \),
- \( V(y_{ij} \mid z_{ij}^s, \beta, A) = \exp(2\mu_{ij} + \sigma_{ij}^2) \left[ \exp(\sigma_{ij}^2) - 1 \right] \)
- \( \text{Cov}(y_{ij}, y_{ij} \mid z_{ij}^s, \beta, A) = \exp(2\mu_{ij} + \sigma_{ij}^2) \left[ \exp(\sigma_{ij}^2) - 1 \right] \)
- \( \text{Cov}(y_{ij}, y_{ij} \mid z_{ij}^s, \beta, A) = \exp(\mu_{ij} + \mu_{ijj'} + \sigma_{ij}^2/2 + \sigma_{ijj'}^2/2) \left[ \exp(\sigma_{ijj'} - 1) \right] \),

for \( \mu_{ij}, \sigma_{ij}^2 \) and \( \sigma_{ijj'} \) given in equation (4).

In the simulation study, we generated 100 datasets from both models and fitted each of one to the samples generated. We considered \( p = 4 \) and \( n = 2,319 \) units divided in
\( m = 50 \) small areas, with sizes between 18 and 142. The parameters were fixed as \( \beta = (1.820, -0.004, -0.002, 0.004) \) and \( A = 0.009 \) for the square root transformation model given in equation (2) and \( \beta = (1.150, -0.006, -0.002, -0.001) \), \( A = 0.010 \) and \( \sigma^2 = 0.320 \) for the logarithmic transformation. These values were chosen in conformity with the results obtained in the application to the real data which will be presented in Subsection 4.1 The auxiliary variables were also extracted from this application.

We assigned vague priors to model parameters. In particular, to assess the sensitivity to the choice of hyperparameters values, we considered two scenarios for the variance parameters: \( a_\sigma = b_\sigma = a_0 = b_0 = 0.01 \) and \( a_\sigma = b_\sigma = a_0 = b_0 = 0.1 \). For each sample, fitted model and prior considered, we ran two parallel chains starting from very different values, we let each chain run for 60,000 iterations, discarded the first 10,000 as burn in, and stored every 50th iteration, resulting in a sample size 1,000. We used the diagnostic tools available in the package CODA (Plummer et al., 2006) to check convergence of the chains.

From each of the 100 populations generated, we selected a simple random sample of 10\% of units and fitted the Normal model to all the samples assuming the square root and logarithmic transformations on the response variable. The results presented in this simulation study are primarily based on the analysis of frequentist properties of the posterior summaries. Basically, we present the measures of: mean square error (MSE), mean absolute error (MAE), coverage (Cov) of the 95\% highest posterior density (HPD) intervals and the mean of the width of the 95\% HPD intervals for the parameters for each model. The MSE and MAE are computed as averages of the squared and absolute deviations, respectively, for the 100 samples. Sometimes, we use the relative MSE and MAE, dividing them by true values of the parameters, in order to make these scale free. Let us abbreviate them by RMSE and RMAE, respectively. Moreover, Cov represents the percentage among the 100 samples for which the 95\% HPD intervals contain the true value and, the mean of the width means the average of the size of the intervals for the 100 samples.

Table 1 presents the MSE, MAE, the coverage of the 95\% HPD intervals and the width of the 95\% HPD intervals for the parameters for each model. Under the same prior considered, the parameters in model (2) present smaller MSE and MAE than in the log transformation. Moreover, although in both cases the coverage of the HPD intervals is close to the nominal level, the width of the intervals under the model given in equation (2) are smaller. Moreover, the results under the prior for the variances \( \text{Gamma}(0.1, 0.1) \) are slightly worse than the other prior, mainly, the coverage of the parameter \( A \).
Table 1  Mean square error, mean absolute error, coverage of the 95% HPD intervals and the mean of the width for the parameters for each model fitted.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>β₀</th>
<th>β₁</th>
<th>β₂</th>
<th>β₃</th>
<th>A</th>
<th>σ²</th>
<th>β₀</th>
<th>β₁</th>
<th>β₂</th>
<th>β₃</th>
<th>A</th>
<th>σ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square root transformation</td>
<td>7.55</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>15.95</td>
<td>1.10</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Log transformation</td>
<td>12.55</td>
<td>0.87</td>
<td>0.14</td>
<td>0.03</td>
<td>15.95</td>
<td>1.10</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

From now on, let us abbreviate square root transformation by SRT and logarithmic transformation by LT. Figure 1 displays RMSE, RMAE, the empirical coverage of the 95% HPD intervals and the mean of the width for the random effects $u_i$, $i = 1, \ldots, m$, under each fitted model. The random effects are better estimated under the square root transformation given in equation (2) than under the log transformation and under the prior $\text{Gamma}(0.01, 0.01)$ than $\text{Gamma}(0.1, 0.1)$.

Next, we consider prediction of the unobserved units $y_{ij}$ using equation (8) under each model considered. As we have the full data, we can also compare the results obtained with the true values of the variables. In particular, Figure 2 presents boxplots with the RMSE, the RMAE for the predicted values for each one of the small areas considered in the analysis, under each fitted model and prior. Again the square root transformation presents smaller RMSE and RMAE for predictions than the logarithmic one. Different from the previous analysis, the two choices of hyperparameters did not significantly affect the results.
Figure 1 Boxplots of the MSE, MAE, coverage of the 95% HPD intervals and the mean of the relative width for the area random effects $u_i$, $i = 1, \ldots, m$ for each model.
Finally, Figure 3 displays the relative bias (RB) for predictions of the population mean under each model considered and for each generated dataset. Different from the log transformation, the RB under the square root transformation is centered at zero and the variability among the samples is reduced. Moreover, the population mean under log transformation
seems to be underestimated for most of the samples. As expected from the previous analysis, the results here are not significantly affected by the choice of hyperparameters.

![Figure 3](image-url) Boxplots of the relative bias for the mean population under each model considered among the 100 generated datasets.

## 4 Data analysis

The main aim of both examples is to evaluate the prediction of the unobserved units, under each model and different sample sizes. In particular, we compared the population mean prediction for each case. As we have the full data we can compare the results obtained with the true population values.

In these applications, we selected 100 simple random samples of units in each small area for two different sample sizes. For each sample, besides fitting the normal model assuming the square root and logarithmic transformations on the response variable, denoted by SRT and LT model, respectively, we fitted the Poisson log-linear model to the data in the original scale, without transformation, denoted by NT model. Under the Poisson model, we obtained samples of the posterior distribution of the parameter vector and for the posterior predictive distribution of $y_{ij}$, for $j = n_i+1, \ldots, N_i$, $i = 1, \ldots, m$, using MCMC algorithm.

We assigned vague priors to model parameters, more specifically, we only assumed $a_\sigma = b_\sigma = a_0 = b_0 = 0.01$, since this specification performed better than the other one evaluated in Section 3. For each sample and fitted model, we ran two parallel chains starting from very different values. We let each chain run for 60,000 iterations, discarded the first 10,000 as burn in, and stored every 50th value to avoid possible autocorrelations within the chains. We also used the diagnostic tools available in the package CODA (Plummer et al., 2006) to check convergence of the chains.

### 4.1 Dataset on sports, fitness and recreational activities

The approach is applied to predict the number of days in a week that a person does moderate-intensity sports, fitness or recreational activities. The data is provided by the 2015-2016 US National Health and Nutrition Examination Survey (NHANES).

We considered 50 small domains (demographic subgroups) classified by race and ethnicity (Mexican American, Other Hispanic, White non-Hispanic, Black non-Hispanic and Other), by age groups (18–29, 30–39, 40–49, 50–59 and 60–84), and by gender. Figure 4 shows the distribution of the variable number of days in a week that a person does moderate-intensity
sports, fitness or recreational activities for 9 different small areas arbitrarily chosen. The distribution of this variable varies according to the small area and the overall median is 3.

We used the following individual auxiliary variables also provided by the NHANES website: the body mass index calculated as weight/height$^2$ in kg/m$^2$ ($x_1$), if the person smoked at least 100 cigarettes in his entire life ($x_2$) and the time (in hours) that the person usually sleeps at night on weekdays or workdays ($x_3$).

The 2015–16 NHANES dataset has variable values for 2,319 persons in the demographic subgroups of interest, with domain sample sizes between 18 and 142.

We selected 100 simple random samples of 10% and 20% of units in each small area and fitted the normal model assuming the square root and logarithmic transformations on the response variable, and the Poisson log-linear model to the data in the original scale, without transformation.

Firstly, Table 2 shows the estimates obtained for the models parameters, under each approach for one of the selected samples assuming the sample fraction of 20%. The parameter $\sigma^2$ from LT model was estimated as 0.32 with 95% HPD interval equals to (0.28, 0.37). Different from the intercept, the other regression coefficients are not statistically significant.

The main aim is to evaluate prediction for the unobserved units. Figure 5 presents boxplots with the RMSE, the RMAE and the RB for the predicted mean. These results are obtained under the normal model applied for each transformation considered and the Poisson model applied to the data in the original scale without any transformation.
The normal model along with the square root transformation generates smaller RMSE, RMAE and RB than the logarithmic transformation for both sample sizes. LT seems to underestimate the mean population for both cases, while SRT seems to overestimate it when the 20% sample is considered. On the other hand, the RMSE and RMAE under SRT and NT models are very similar. When the 10% sample is considered, the median of the RB under the SRT seems to be closer to zero, than under NT. The opposite happens when the 20% sample is considered.

4.1.1 Model selection

In order to study how the data support each model, we also calculated the cross-validation predictive densities for each observation $y_{ij}$ and for each sample considered. The cross-validation density for $y_{ij}$ is the conditional density $f(y_{ij} | Y_{(-ij)})$, where $Y_{(-ij)}$ denotes all data except $y_{ij}$. We looked at the value of $f(y_{ij} | Y_{(-ij)})$ at the observed data point, the so called conditional predictive ordinate (CPO) for each of the three models. Since CPOs are nothing but the observed likelihoods, models with larger CPOs provide better fit to the observed data. By using the output from the Gibbs sampler, we can calculate the CPOs for all data points. Noting that the $y_{ij}$’s are conditionally independent, i.e., $f(y_{ij} | Y_{(-ij)}, \beta, u_i) = f(y_{ij} | \beta, u_i)$ and, assuming there is a sample of size $G$ from the posterior distribution of the parameters, the CPO values are calculated as:

$$\hat{CPO}_{ij} = \left[ \frac{1}{G} \sum_{k=1}^{G} \frac{1}{f(y_{ij} | \beta^{(k)}, u_i^{(k)})} \right]^{-1},$$

where $f(y_{ij} | \beta, u_i)$ is the original Poisson density function before the transformation is applied. More detailed discussion can be found in Gelfand (1996).
In this case, since the CPO values are very similar for all the 3 models considered, we calculated the ratio of CPO values under SRT and NT models and under LT and NT models and counted the percentage of observations for which the numerator was greater than the denominator. Figure 6 shows the boxplots with these percentages for all the observations and for all the 100 samples considered. A majority of CPO values for the SRT and LT models are larger than those for the NT model. However, this percentage is even higher when the square root transformation is considered. It suggests that the SRT performs the best among the three models considered.

To evaluate the overall fit of a Bayesian model, we also used the posterior predictive p-value (Meng, 1994, PPP), which is the tail posterior probability for a statistic generated from the model compared to the statistic observed in the data. It is defined as

\[ p = P\left(T(Y^{rep}, \psi) \geq T(Y^{s}, \psi) \mid Y^{obs}\right), \]

where \( Y^{rep} \) is a vector with \( G \) independent replicates from the posterior predictive distribution and \( T(Y, \psi) \) is a test statistic that depends on the data \( Y \) and possibly the parameter \( \psi = (\beta, u) \). It is easily estimated from MCMC samples as

\[ p = \frac{1}{G} \sum_{k=1}^{G} I\left[T(Y^{rep(k)}, \psi(k)) \geq T(Y^{s}, \psi(k))\right], \]

where \( I[.] \) is the indicator function. If a model fits well the observed data, then \( T(Y^{s}, \psi) \) should be near the central part of the distribution of \( T(Y^{rep}, \psi) \). Consequently, the posterior predictive p-value is expected to be near 0.5. Extreme p-values (near 0 or 1) suggest poor fit. To carry out the posterior predictive model checking, we need to specify a test quantity. Because a model can fail to reflect the process that generated the data in any number of ways, we should compute the posterior predictive p-values for a variety of test quantities in order to evaluate any possible model failure. In our study the mean seemed to be a good choice for this test quantity.

We calculated the difference between PPP and 0.5 for each of the 100 samples and each sample size. Figure 7 presents boxplots with these differences. The values are very similar. However, both square root and logarithmic transformations performed slightly better than the Poisson log-linear model, which for most of the samples generated PPP less than 0.5.
In addition to facilitating analytical manipulations and inference, the square root transformation performs better than the logarithmic transformation in most cases, and slightly better than or equal to the Poisson log-linear model, according to the criteria used. Thus, the square root transformation could be a good alternative to deal with count data.

4.2 Dataset on students’ proficiency in Prova Brasil

Prova Brasil is a national population-based assessment conducted by INEP (Instituto Nacional de Estudos Educacionais Anísio Teixeira) in Reading and Mathematics of public schools. It includes pupils in 5th grade of primary education and students in 9th grade of lower secondary school. It consists of standardized tests and socio-economic questionnaires. The tests cover Portuguese language, with a focus on reading, and Mathematics, with a focus on problem solving. Teachers and school principals also receive questionnaires that gather demographic data, professional profiles of staff, and working conditions. According to the number of points obtained in Prova Brasil, students are distributed in 4 main levels on a proficiency scale: Insufficient, Basic, Proficient and Advanced.

We are particularly interested in predicting the number of students per school in 5th grade of primary education who achieved adequate performance, that is, who performed Proficient or Advanced levels, in the state of Rondônia, in 2011. Rondônia is located in the northern subdivision of Brazil and is divided into 52 municipalities, which were treated here as small domains. The state is composed of 431 schools (units) distributed by 52 municipalities, with sizes varying between 1 and 99.

Figure 8 shows the distribution of the variable number of students per school who achieved adequate performance for 9 different small areas arbitrarily chosen. The distribution of this variable varies according to the small area and the overall median is 10.

We used the following auxiliary variables: a dummy indicating if the school is administered by the state or the municipality ($x_1$), the school’s average Portuguese proficiency ($x_2$) and a dummy indicating if it is a rural or urban school ($x_3$).

We selected 100 simple random samples of 20% and 50% of schools in each small area. For the small areas with less than 4 schools we did not take a sample and used all the information. Similar to the previous example, for each sample we fitted the normal model assuming the square root transformation on the response variable and the Poisson log-linear model to the data in the original scale, without transformation. We did not consider the logarithmic transformation in this case, since the response variable contains the value zero.
Figure 8 Histograms with the distribution of the number of students per school in 5th grade of primary education who achieved adequate performance in Prova Brasil for 9 municipalities in the state of Rondônia in Brazil, in 2011.

Table 3 shows the estimates obtained for the models parameters, under each approach considered for one of the samples selected assuming the sample fraction of 50%. Under both models, when the school belongs to the Municipality administration the number of students with adequate performance in the exam increases significantly, on the other hand, when the school belongs to a rural area, this number decreases significantly. Finally, adding one unit to the school’s average Portuguese proficiency increases the number of students per school who achieved adequate performance by 0.05 and 0.04, under SRT and NT models, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRT</td>
<td>-6.59 (-8.46,-4.70)</td>
<td>0.31 (0.06,0.57)</td>
<td>0.05 (0.04,0.07)</td>
<td>-0.77 (-1.08,-0.50)</td>
<td>0.35 (0.14,0.56)</td>
</tr>
<tr>
<td>NT</td>
<td>-4.32 (-5.33,-3.24)</td>
<td>0.23 (0.08,0.37)</td>
<td>0.04 (0.03,0.04)</td>
<td>-0.53 (-0.75,-0.33)</td>
<td>0.12 (0.06, 0.19)</td>
</tr>
</tbody>
</table>

Table 3 Posterior mean and 95% HPD interval of the parameters for each model fitted to a 50% sample selected from the real dataset.

Figure 9 presents boxplots with the RMSE, the RMAE and the RB for the predicted means obtained under the normal model applied to the square root transformation and the Poisson model applied to the data in the original scale. SRT generates smaller RMSE and RMAE than the Poisson model for the 20% sample, however, when the 50% sample is considered, the opposite happens. With respect to the RB, the SRT seems to underestimate slightly the population mean in both cases.
4.2.1 Model selection

As in the previous application, since the CPO values are very similar for the two models considered, we calculated the percentage of observations for which the CPO value under SRT was greater than the CPO value under NT. Figure 10 shows the boxplots with these percentages for all the observations and for all the 100 samples considered. For the 20% sample we are not able to suggest a best model among the two considered. However, for the 50% sample a majority of CPO values for the SRT model are larger than those for the NT model.

Finally, Figure 11 presents boxplots with the differences between PPP and 0.5. Although, the values are very similar, the square root transformation performed slightly better than the Poisson log-linear model, which for most of the samples generated PPP less than 0.5.
From these examples with real datasets, we conclude that SRT could be a good model for count data. In some cases it performs very similar to the Poisson log-linear model, which could be a natural choice in this case. However the normal assumption and approximate homoscedasticity, which eliminates one extra layer of variance estimation, provides inference facilities, that can be advantageous.

5 Final Remarks

The demand for reliable small area estimates derived from survey data has greatly increased in recent years. A major direction considers the use of explicit small area models that “borrow strength” from related areas through auxiliary information, supposed to be correlated to the variable of interest. In particular, unit level models relate the unit values of the study variable to unit-specific auxiliary variables, our main interest in this paper.

In particular, we considered modelling of count data, assuming normality under square-root transformation of the original data. Besides practical advantages, the normal assumption facilitates analytical manipulations and the square root transformation achieves approximate homoscedasticity of the error variances. Simulation studies, presented in this paper, illustrate that the transformation performs similar to or even better than the original Poisson model. Moreover, when compared to the usual logarithmic transformation, the square root transformation produces estimators with better frequentist properties in most of the cases, in addition to not having a variance parameter to estimate.

Appendix A: Proof of Lemmas 1 and 2

Lemma 1

Let $T_1$ and $T_2$ have the following bivariate Normal distribution:

$$
\begin{pmatrix}
T_1 \\
T_2
\end{pmatrix}
\sim N \left( \begin{pmatrix}
\mu_1 \\
\mu_2
\end{pmatrix}, \begin{pmatrix}
\sigma_1^2 & \sigma_{12} \\
\sigma_{12} & \sigma_2^2
\end{pmatrix} \right).
$$

For $i = 1, 2$, we have that $E(T_i^2) = \sigma_i^2 + \mu_i^2$ and $V(T_i^2) = E(T_i^4) - E^2(T_i^2)$.
Moreover,

\[\text{Cov}(T^2_i, T^2_j) = \text{Cov}\left((T_1 - \mu_1 + \mu_1)^2, (T_2 - \mu_2 + \mu_2)^2\right)\]

\[= \text{Cov}\left((T_1 - \mu_1)^2 + 2(T_1 - \mu_1)\mu_1 + \mu_1^2, (T_2 - \mu_2)^2 + 2(T_2 - \mu_2)\mu_2 + \mu_2^2\right)\]

\[= \text{Cov}((T_1 - \mu_1)^2, (T_2 - \mu_2)^2) + 4\mu_1\mu_2\text{Cov}((T_1 - \mu_1), (T_2 - \mu_2))\]

\[= \sigma^2_i \sigma^2_j \text{Cov}(Z^2_i, Z^2_j) + 4\mu_1\mu_2 \sigma_{12}, \quad (10)\]

for \(Z_i = (T_i - \mu_i)/\sigma_i, i = 1, 2,\)

Note that, for \(\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2},\) we have

\[
\begin{pmatrix}
Z_1 \\
Z_2
\end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).
\]

From the independence of \(Z_1\) and \(Z_2 - \rho T_1,\) we get

\[\text{Cov}(Z^2_1, Z^2_2) = \text{Cov}\left((Z_2 - \rho Z_1)^2, Z^2_1\right)\]

\[= \text{Cov}((Z_2 - \rho Z_1)^2 + 2\rho Z_1(Z_2 - \rho Z_1) + \rho^2 Z^2_1, Z^2_1)\]

\[= \rho^2 \text{V}(Z^2_1) = \frac{\sigma^2_{12}}{\sigma^2_1 \sigma^2_2}.\]

Hence, from (10), we have

\[\text{Cov}(T^2_i, T^2_j) = \sigma^2_{12} + 4\mu_1\mu_2 \sigma_{12},\]

and Lemma 1 is proved.

**Lemma 2**

Let \(T_1\) and \(T_2\) have the following bivariate Normal distribution:

\[
\begin{pmatrix}
T_1 \\
T_2
\end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2_1 & \sigma_{12} \\ \sigma_{12} & \sigma^2_2 \end{pmatrix}\right).
\]

Thus, \(T_i \sim N(\mu_i, \sigma^2_{ii}), i = 1, 2,\) and \(\exp(T_i)\) is lognormal with mean

\[E[\exp(T_i)] = \mu_i + \frac{\sigma_i^2}{2}\]

and variance

\[V[\exp(T_i)] = \exp\left(2\mu_i + \sigma^2_i\right) \left[\exp(\sigma^2_i) - 1\right].\]

Moreover,

\[\text{Cov}[\exp(T_1), \exp(T_2)] = E[\exp(T_1) \exp(T_2)] - E[\exp(T_1)] E[\exp(T_2)]\]

\[= E[\exp(T_1 + T_2)] - \exp(\mu_1 + \frac{\sigma_1^2}{2}) \exp(\mu_2 + \frac{\sigma_2^2}{2}). \quad (11)\]
As, \( T_1 + T_2 \sim N (\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\sigma_{12}) \), \( \exp (T_1 + T_2) \) is also lognormal distributed with mean

\[
E [\exp (T_1 + T_2)] = \exp \left[ \mu_1 + \mu_2 + \frac{1}{2} (\sigma_1^2 + \sigma_2^2 + 2\sigma_{12}) \right] = \exp (\mu_1 + \sigma_1^2/2) \exp (\mu_2 + \sigma_2^2/2) \exp (\sigma_{12}) .
\]

(12)

Thus, substituting (12) into (11), we get

\[
Cov [\exp (T_1), \exp (T_2)] = \exp (\mu_1 + \sigma_1^2/2 + \sigma_2^2/2) [\exp (\sigma_{12}) - 1] .
\]

Appendix B: Proof of propriety of the posterior distribution

Recall the posterior distribution in (9)

\[
p(\beta, u, A^{-1} \mid Z^s) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^{m} (z_i^s - X_i^s \beta - u_i 1_{n_i})^T (z_i^s - X_i^s \beta - u_i 1_{n_i}) \right\} \times
\]

\[
A^{-m/2} \exp \left\{ -\frac{1}{2} A \sum_{i=1}^{m} u_i^2 \right\} \times A^{-a_0/2+1} \exp \left\{ -\frac{b_0}{2} A^{-1} \right\}.
\]

We can rewrite \( \sum_{i=1}^{m} (z_i^s - X_i^s \beta - u_i 1_{n_i})^T (z_i^s - X_i^s \beta - u_i 1_{n_i}) \) as:

\[
\sum_{i=1}^{m} (z_i^s - X_i^s \beta - u_i 1_{n_i})^T (z_i^s - X_i^s \beta - u_i 1_{n_i}) = \beta^T \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right) \beta
\]

\[
- 2 \sum_{i=1}^{m} (z_i^s - u_i 1_{n_i})^T X_i^s \beta + \sum_{i=1}^{m} (z_i^s - u_i 1_{n_i})^T (z_i^s - u_i 1_{n_i})
\]

\[
= \begin{bmatrix}
\beta - \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} \sum_{i=1}^{m} (X_i^s)^T (z_i^s - u_i 1_{n_i})
\end{bmatrix}^T \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right) \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right) \left( \sum_{i=1}^{m} (X_i^s)^T (z_i^s - u_i 1_{n_i}) \right)
\]

\[
\times \begin{bmatrix}
\beta - \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} \sum_{i=1}^{m} (X_i^s)^T (z_i^s - u_i 1_{n_i})
\end{bmatrix}
\]

\[
- \sum_{i=1}^{m} (z_i^s - u_i 1_{n_i})^T X_i^s \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} \sum_{i=1}^{m} (X_i^s)^T (z_i^s - u_i 1_{n_i})
\]

\[
+ \sum_{i=1}^{m} (z_i^s - u_i 1_{n_i})^T (z_i^s - u_i 1_{n_i}) \quad (13)
\]

Integrating out with respect to \( \beta \), it follows from (9) and (13) that

\[
p(u, A^{-1} \mid Z^s) \propto \exp \left\{ -2 \sum_{i=1}^{m} (z_i^s - u_i 1_{n_i})^T \left[ I_{n_i} - X_i^s \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} (X_i^s)^T \right] (z_i^s - u_i 1_{n_i}) \right\}
\]
By assumption, \((\sum_{i=1}^{m} (X_i^s)^T X_i^s)^{-1}\) is a symmetric positive definite matrix. Hence, \(X_i^s \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} (X_i^s)^T\) is a symmetric non-negative definite matrix with

\[
\text{rank} \left[ X_i^s \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} (X_i^s)^T \right] \leq \text{rank} (X_i^s) \leq n_i.
\]

Thus, \(I_{n_i} - X_i^s \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} (X_i^s)^T\) is symmetric non-negative definite. This implies that

\[
\sum_{i=1}^{m} (z_i^s - u_i^1 n_i)^T \left[ I_{n_i} - X_i^s \left( \sum_{i=1}^{m} (X_i^s)^T X_i^s \right)^{-1} (X_i^s)^T \right] (z_i^s - u_i^1 n_i) \geq 0.
\]

Hence,

\[
p(u, A^{-1} | Z^s) \leq A^{-m/2} \exp \left\{ -\frac{1}{2A} \sum_{i=1}^{m} u_i^2 \right\} \times A^{-a_0/2+1} \exp \left\{ -\frac{b_0}{2A} \right\}.
\]

Next, first integrating out with respect to \(u\) and then \(A\), we have by the propriety of normal and inverse gamma distributions that

\[
\int p(\beta, u, A^{-1} | Z^s) d\beta dudA < \infty.
\]

**Acknowledgements**

The first author was supported by a scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível (CAPES). The authors wish to acknowledge a referee and the editor for very useful comments.

**References**


