Supplement to "A Unit-Level One-Inflated Beta Model for Small Area Prediction of Seat-Belt Use Rates"

Zirou Zhou[a] and Emily Berg[b]

[a, b]Department of Statistics, Iowa State University, 2438 Osborn Dr Ames, IA, USA

ARTICLE HISTORY
Compiled May 12, 2024

This supplement has 8 sections. In Section 1, we present results of a simulation where the correlation is 0.6. In Section 2, we compare the estimated root mean square errors of the model-based predictor to the standard errors of the direct estimators in the simulation. In Section 3, we present posterior predictive p-values for the application. In Section 4, we present an exploratory analysis to evaluate if the assumption of a constant ϕ parameter is reasonable. In Section 5, we check for convergence of the MCMC procedure in the data analysis. In Section 6, we define the estimation procedure for the beta regression model in more detail. In Section 7, we discuss an extension to zero-one inflated data. In Section 8, we present the results of applying the binomial model in the data analysis.

1. Simulation with correlation=0.6

Table 1. Average RMSE and improvement ratio of alternative predictors for data with correlation = 0.6

Model Method	RMSE×100	Improvement Ratio
1. Direct Estimation	1.386	1.390
2. Frequentist: Unit-level One-inflated Beta	0.997	1
3. Linear EBLUP	1.091	1.094
5. Stan-Bayesian: Unit-level One-inflated Beta	1.685	1.690
5*. Stan-Bayesian: Unit-level One-inflated Beta (remove extreme values)	1.059	1.062
6. INLA-Bayesian Unit-level One-inflated Beta	1.010	1.013

CONTACT Zirou Zhou Email: zzhou@iastate.edu

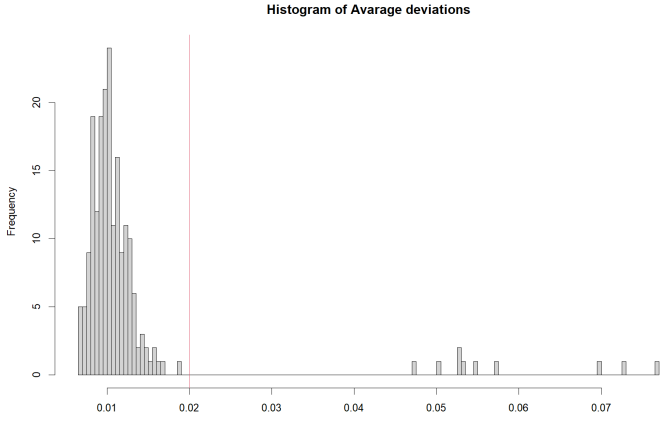

Figure 1. Average deviations for Stan model: Histogram of $\frac{1}{D}\sum_{i=1}^{D}(\hat{y}_i^{(5,m)} - \bar{y_i}^{(m)})^2$ for $m = 1, \ldots, M$. Values above 0.02 are considered extreme. True correlation is 0.6.

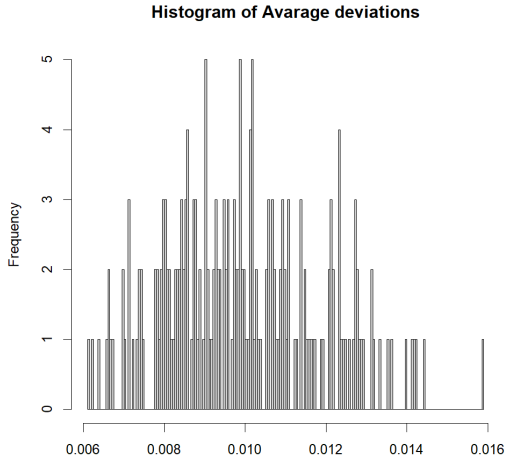

Figure 2. Average deviations for INLA model: Histogram of $\frac{1}{D}\sum_{i=1}^{D}(\hat{y}_i^{(6,m)} - \bar{y_i}^{(m)})^2$ for $m = 1, \ldots, M$. True correlation is 0.6.

2. Comparison of the estimated frequentist root MSE and estimated standard deviations of direct estimators

In the discussion of the results of the data analysis, we comment that the difference between the true RMSE and the estimated RMSE is important. The improvement ratios discussed previously reflect the true RMSE through the Monte Carlo RMSE of the alternative predictors. In this section, we consider the estimated RMSE.

The two figures below present ratios of the estimated RMSEs of the frequentist model-based predictors to the estimated standard deviations of the direct estimators. The results are based on the simulation configuration with correlation zero. For the bottom figure, values above 40 are removed to provide a clearer picture of the ratios.

What is notable from the figure is that in many cases, the estimated RMSE of the model-based predictor exceeds the estimated standard deviation of the direct estima-

tor. Although the true RMSE of the frequentist model-based predictor is below the standard deviation of the direct estimator, this is not always reflected in the estimated measures of uncertainty. The measures of uncertainty are estimates themselves that are subject to sampling variation. Therefore, the estimated RMSE of the model-based predictor can exceed the estimated standard deviation of the direct estimator.

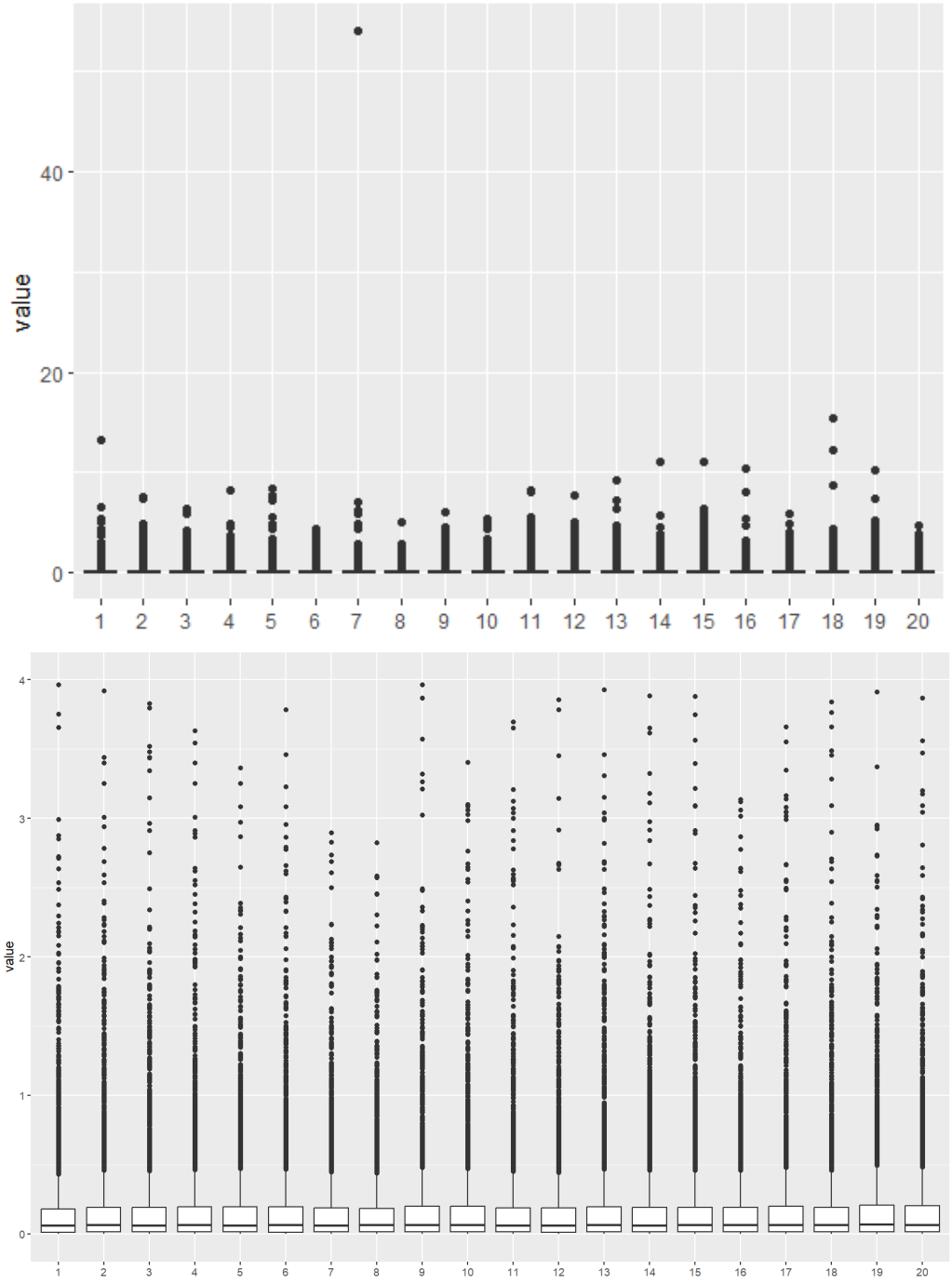

Figure 3. The box plot of the comparison ratio

3. Posterior predictive p-values for data analysis

Table 2. P-value comparing the observed value with Stan posterior

	p-value
mean	0.502
median	0.568
SD	0.080
skewness	0.001
kurtosis	0.002
Person Chi	0.382

Table 3. P-value comparing the observed value with INLA posterior

	p-value
mean	0.506
median	0.571
SD	0.093
skewness	0.001
kurtosis	0.002
Person Chi	0.499

4. Shape of different counties

The beta models shape parameters are $\phi * \mu$ and $\phi * (1-\mu)$. In the model, μ is dependent on the $Log(VMT)$. A proxy for ϕ is the square of $d_{ij} = \frac{y_{i,j}-\hat{p}_{i,j}}{\sqrt{\hat{p}_{i,j}(1-\hat{p}_{i,j})}}$. We examined scatter-plots of these squares for individual units and by county. We found no evidence of a trend. This suggests that the assumption of a constant ϕ is reasonable for our data set. As a reference, a plot of the average of d_{ij} by county is below.

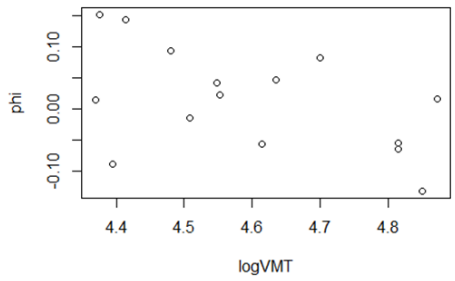

Figure 4. Average d_{ij} by county plotted against $\log(VMT)$

5. The convergence check for the MCMC procedure

For the Stan-Bayesian procedure, we need to assess the convergence of the chains to the stationary distribution. We run three chains, each with total length of 2000 and discard the first 1000 iterations as burn-in. We use the remaining 1000 iterations from

each chain to approximate the posterior distribution. To diagnose convergence of the chains to the stationary distribution, we calculate the scale reduction factor (Rhat) for every model parameter. Figure 5 contains a histogram of the scale reduction factors (constructed using built-in functions in `Rstan`). All the scale reduction factors are below 1.001. The chains appear to have converged to the stationary distribution.

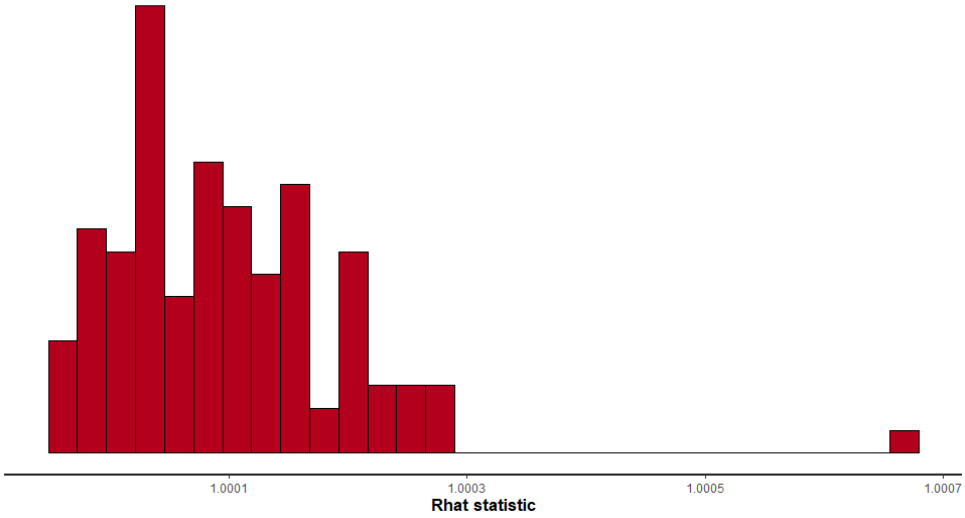

Figure 5. Rhat for the posterior chains

For the INLA-Bayesian procedure, we fit the model using the `R` function `inla()` to achieve the posterior estimations for both the fixed and random effects. We then take 3000 posterior samples from the marginal posterior distributions to be used, as in the Stan-Bayesian procedure, to approximate the posterior distribution. As INLA is not dependent on MCMC to achieve the posterior, there is no necessity to check the convergence for the posterior.

6. Procedure to obtain estimates for beta regression model

The following summary of the estimation procedure is adapted from [1]. Let $\boldsymbol{b} = (b_1, \ldots, b_D)'$. Define

$$f_i(\boldsymbol{b}, \boldsymbol{\beta}, \phi) = \prod_{\{j: y_{ij} < 1\}} f_{ij}(b_i, \boldsymbol{\beta}, \phi),$$

where $f_{ij}(b_i, \boldsymbol{\beta}, \phi)$ is the density of the beta distribution for y_{ij}. Define $f(\boldsymbol{b}, \boldsymbol{\beta}, \phi) = -\sum_{i=1}^{D} \log(f_i(\boldsymbol{b}, \boldsymbol{\beta}, \phi))$. For any $(\boldsymbol{\beta}', \phi)'$, define

$$H(\boldsymbol{\beta}, \phi) = \frac{\partial^2}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} f(\boldsymbol{b}, \boldsymbol{\beta}, \phi) \mid_{\boldsymbol{b} = \hat{\boldsymbol{b}}(\boldsymbol{\beta}, \phi)},$$

where

$$\hat{\boldsymbol{b}}(\boldsymbol{\beta}, \phi) = argmin_{\boldsymbol{b}} f(\boldsymbol{b}, \boldsymbol{\beta}, \phi),$$

and $\hat{\boldsymbol{b}}(\boldsymbol{\beta}, \phi) = (\hat{b}_1(\boldsymbol{\beta}, \phi), \ldots, \hat{b}_D(\boldsymbol{\beta}, \phi))'$.

Define the Laplace approximation for the marginal likelihood by

$$L^*(\boldsymbol{\beta}, \phi) = \sqrt{2\pi}^n \det(H(\boldsymbol{\beta}, \phi)) \exp(-f(\hat{\boldsymbol{b}}(\boldsymbol{\beta}, \phi), \boldsymbol{\beta}, \phi)).$$

Define $(\hat{\boldsymbol{\beta}}, \hat{\phi}) = argmax L^*(\boldsymbol{\beta}, \phi)$. Finally, define $\hat{\boldsymbol{b}} = \hat{\boldsymbol{b}}(\hat{\boldsymbol{\beta}}, \hat{\phi})$. If area i has no data that are not equal to one, then set $\hat{b}_i = 0$.

7. Extension to Zero-One Inflated Data

We extend the model to data with support $[0, 1]$. Assume

$$y_{ij} = \begin{cases} 1 \text{ with prob. } p_{1,ij} \\ 0 \text{ with prob. } p_{2,ij} \\ Beta(\mu_{ij}, \phi) \text{ with prob. } 1 - p_{1,ij} - p_{2,ij}, \end{cases}$$

where for $k = 1, 2$,

$$p_{k,ij} = \frac{\exp(\boldsymbol{z}'_{k,ij}\boldsymbol{\alpha}_k + u_{k,i})}{1 + \sum_{k=1}^{2} \exp(\boldsymbol{z}'_{k,ij}\boldsymbol{\alpha}_k + u_{k,i})},$$

$\text{logit}(\mu_{ij}) = \boldsymbol{x}'_{k,ij}\boldsymbol{\beta} + b_i$ and $(u_{1,i}, u_{2,i}, b_i) \overset{iid}{\sim} N(\mathbf{0}, \text{diag}(\sigma^2_{u,1}, \sigma^2_{u,2}, \sigma^2_b))$. Under this extended model,

$$E[y_{ij} \mid p_{1,ij}, p_{2,ij}, \mu_{ij}] = p_{1,ij} + (1 - p_{2,ij} - p_{3,ij})\mu_{ij}.$$

The frequentist and Bayesian inference procedures extend directly to the more general model.

8. Binomial model for the data analysis

In this Section, we evaluate the predictors based on the binomial model described in Section 4.3. Table 7 and Figure 10 compare the predictors based on the binomial model to the direct estimators. The predictors are systematically below the direct estimators, indicating that this model produces a severe, negative bias. Furthermore, the model-based root mean square errors are uniformly above the direct standard errors. Based on the results in Table 7 and Figure 10, the binomial model appears unreasonable for this data set.

Refinements to the binomial model provide avenues for future work. One refinement would be to specify a model for the M_{ij}. A second refinement would be to incorporate a one-inflated component in the binomial distribution. These possible refinements hold promise but are beyond the scope of our current work.

Table 4. Comparison between direct estimators and Binomial model

	Estimator		Root MSE	
County ID	Direct	Binomial	Direct	Binomial
1	0.9216	0.8931	0.0102	0.1451
2	0.9415	0.9147	0.0106	0.1236
3	0.8946	0.9144	0.0498	0.1220
4	0.9604	0.9238	0.0136	0.1367
5	0.9707	0.9398	0.0164	0.1212
6	0.8867	0.9015	0.0515	0.1272
7	0.9031	0.8846	0.0136	0.1305
8	0.9724	0.9438	0.0037	0.099
9	0.9737	0.9462	0.0078	0.1022
10	0.9543	0.9491	0.0249	0.1055
11	0.9631	0.9558	0.0151	0.0778
12	0.9098	0.8891	0.0165	0.1198
13	0.9481	0.9146	0.0073	0.1103
14	0.9225	0.9113	0.0195	0.1164
15	0.9737	0.9463	0.0123	0.1116

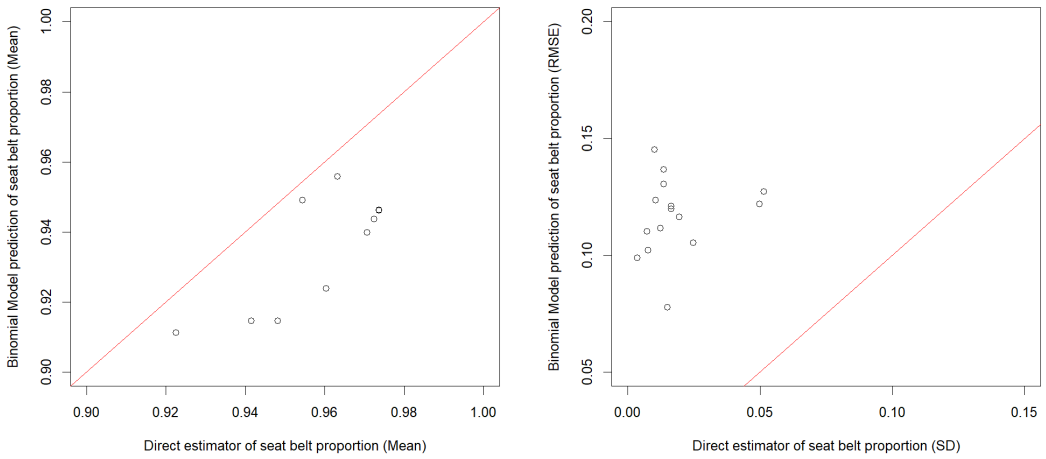

Figure 6. Comparison between direct estimation and Binomial model estimation. The red line is the (0,1) line.

References

[1] K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. Bell, *Tmb: automatic differentiation and laplace approximation*, arXiv preprint arXiv:1509.00660, (2015).