Three Extensions of the Basic Unit-Level Model

Emily Berg

Department of Statistics,
Iowa State University

November 29, 2023
Outline

- Introduction
  - Linear, unit-level model
- Extension 1: Lognormal models
  - Empirical best prediction concepts
- Extension 2: Zero-inflated lognormal models
  - Population-level covariates
- Extension 3: Informative sampling
  - Nonlinear parameters, exponential dispersion families
What is small area estimation?

- Large-scale surveys play an important role in the federal statistical system
  - National Crime Victimization Survey – criminal victimization rates for individuals ages 12 and older in the US
  - National Resources Inventory – characteristics related to natural resources and agriculture
  - Canadian labor force survey – parameters related to employment
What is small area estimation?

- Complex surveys are often designed to produce estimates for large estimation domains. Data users often request estimates for estimation domains with small sample sizes
  - National Crime Victimization Survey publishes national level estimates
    - State-level estimates are of interest
  - National Resources Inventory publishes state-level estimates
    - County-level estimates are of interest
  - Canadian labor force survey produces estimates for broad employment categories at the provincial level
    - Detailed employment categories are of interest
What is Small Area Estimation?

- The challenge
  - As a result of small sample sizes, direct estimators are unreliable
- The solution: small area estimation (Rao & Molina 2015)
  - Use models to obtain more efficient estimates
Basic Unit-Level Model: Set-up

- $i = 1, \ldots, D$ index the areas
- $j = 1, \ldots, N_i$ index the elements in the entire population for area $i$
- $j = 1, \ldots, n_i < N_i$ index the elements in the sample for area $i$
- $y_{ij}$ is the variable of interest for unit $j$ in area $i$
- The parameter of interest is the area mean defined by

$$\theta_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$$

- Data available

$$\{ y_{ij} : j = 1, \ldots, n_i \} \cup \{ x_{ij} : j = 1, \ldots, n_i \} \cup \{ \bar{x}_{N,i} : i = 1, \ldots, D \}$$

$$\bar{x}_{N,i} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$$
Basic Unit-Level Model (Battese et al. 1988)

- Model Assumption

\[ y_{ij} = \beta_0 + x'_{ij}\beta_1 + u_i + e_{ij}, \]

\[ u_i \overset{iid}{\sim} N(0, \sigma_u^2) \]

\[ e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2) \]

- Use REML (R function lmer in lme4 package, or SAE R package) to obtain estimates: \( \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_u^2, \hat{\sigma}_e^2 \)
Basic Unit-Level Model

Challenges

- Skewed response variables
- Zero-inflated data
- Informative sampling
Part 1: Unit-Level Lognormal Model (joint work with Hukum Chandra)

- Small area model for skewed data
- Illustrate basic concepts of small area prediction under nonlinear models
Response variable ($Y$) has a non-normal distribution
- Skewed
- Positive support
- Variance increases with the mean
- Nonlinear associations to covariates

Linear predictors inefficient
Unit-Level Lognormal Model: Framework

- Areas: \( i = 1, \ldots, D \); units: \( j = 1, \ldots, N_i \)
  \[
  \log(y_{ij}) = z_{ij}' \beta + u_i + e_{ij}
  \]
  \((u_i, e_{ij}) \overset{iid}{\sim} N(\mathbf{0}, \text{diag}(\sigma_u^2, \sigma_e^2))\)
  \[\theta = (\beta', \sigma_u^2, \sigma_e^2)', \quad \hat{\theta} = (\hat{\beta}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'\]

- Data available
  \[
  \{y_{ij} : j = 1, \ldots, n_i\} \cup \{z_{ij} : j = 1, \ldots, N_i\}, i = 1, \ldots, D
  \]
  \[\mathbf{y}_s = \{y_{ij} : j = 1, \ldots, n_i, i = 1, \ldots, D\}\]

- \(n_i = \text{sample size}, N_i = \text{population size}\)
- Quantity to predict: small area mean
  \[
  \bar{y}_{N_i} = \frac{1}{N_i} \left[ \sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} y_{ij} \right]
  \]
Unit-Level Lognormal Model: Predictor

- Best (Bayes) predictor (minimum MSE) – general

\[
\tilde{y}_{N_i}^B(\theta) = \frac{1}{N_i} \left\{ \sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} E[y_{ij} | \theta, y_s] \right\}
\]

- Conditional expectation for the lognormal

\[
E[y_{ij} | \theta, y_s] = \exp \left[ z_{ij}' \beta + \gamma_i (\bar{\ell}_{si} - \bar{z}_{si}' \beta) + \frac{\sigma_e^2}{2} \left( \frac{\gamma_i}{n_i} + 1 \right) \right]
\]

\[
\bar{\ell}_{si} = n_i^{-1} \sum_{j=1}^{n_i} \log(y_{ij}), \quad \gamma_i = \sigma_u^2 (\sigma_u^2 + n_i^{-1} \sigma_e^2)^{-1}
\]

- Justification for \(E[y_{ij} | \theta, y_s]\)

\[
\log(y_{ij}) | \theta, y_s \sim \mathcal{N}(z_{ij}' \beta + \gamma_i (\bar{\ell}_{si} - \bar{z}_{si}' \beta), \gamma_i \sigma_e^2 n_i^{-1} + \sigma_e^2)
\]
Empirical best predictor (general)

\[
\hat{y}_{Ni}^{EB} = \bar{y}_{Ni}^{MMSE}(\hat{\theta}) = \frac{1}{N_i} \left\{ \sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} E[y_{ij} \mid \hat{\theta}, y_s] \right\}
\]

\(\hat{y}_{Ni}^{EB}\) for the lognormal biased due to nonlinear transformation of \(\hat{\theta}\)

\[
E \left\{ E[y_{ij} \mid \hat{\theta}, y_s] - E[y_{ij} \mid \theta, y_s] \right\} \neq 0
\]

Multiplicative bias correction

Bias-corrected estimator - \(\hat{y}_{Ni}^{EB, BC}\)
General MSE of an EB predictor

\[
\text{MSE}(\hat{y}_{Ni}^{EB}) = E[(\hat{y}_{Ni}^{EB} - \bar{y}_{Ni})^2]
\]

\[
= E[(\bar{y}_{Ni}^B(\theta) - \bar{y}_{Ni})^2] + E[(\hat{y}_{Ni}^{EB} - \bar{y}_{Ni}^B(\theta))^2]
\]

\[
M_{1i}(\theta) = E[V(\bar{y}_{Ni} \mid \mathbf{y}_s)]
\]

= MSE of best predictor constructed with (unknown) true parameters

\[
M_{2i}(\theta) = \text{variance due to estimation of } \theta
\]
Unit-Level Lognormal Model: MSE Estimation

- Closed-form expression for $M_{1i}(\theta)$

$$M_{1i}(\theta) = MSE\{\tilde{y}_{Ni}^{B}(\theta)\} = \frac{\kappa_i}{N_i^2} \left[ \left( \sum_{j \in \bar{s}_i} \exp(z'_{ij} \beta) \right)^2 \xi_i + \left( \sum_{j \in s_i} \exp(2z'_{ij} \beta) \right) \psi_i \right]$$

$(\xi_i, \psi_i, \kappa_i)$ are known functions of $\sigma_u^2, \sigma_e^2, n_i$

- Taylor series approximation for $M_{2i}(\theta)$
Unit-Level Lognormal Model: MSE Estimation

- Plug-in estimator

\[ \hat{MSE}_{1i} = M_{1i}(\hat{\theta}) + \hat{M}_{2i} \]

- Biased because \( E[M_{1i}(\hat{\theta}) - M_{1i}(\theta)] \neq 0 \)

- Bias-reduced estimator:

\[ \hat{MSE}_{2i} = M_{1i}(\tilde{\theta}_i) + \hat{M}_{2i} \]

- \( \tilde{\theta}_i \) depends on Taylor expansion of \( M_{1i}(\theta) \)
- \( E[M_{1i}(\tilde{\theta}_i) - M_{1i}(\theta)] \approx 0 \)
- The bias-adjusted MSE estimator is non-negative
Unit-Level Lognormal: Simulation Models

- $N = 9990$, $D = 30$

\[
\log(y_{ij}) = \beta_0 + \beta_1 z_{ij} + u_i + e_{ij}
\]

$\begin{pmatrix} z_{ij}, u_i, e_{ij} \end{pmatrix} \sim N\{ (\mu_z, 0, 0), \text{diag}(\sigma_z^2, \sigma_u^2, \sigma_e^2) \}$

Four Parameter Sets

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_z^2$</th>
<th>$\sigma_u^2 \sigma_e^{-2}$</th>
<th>$\sigma_u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>1.6</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>1.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>1.2</td>
<td>0.5</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

$(\mu_z, \beta_0, \beta_1) = (3.253, -1.62, 0.9)$

- Mean and variance of $y_{ij}$ approximately equal to mean and variance of the number of chickens per segment in a 1960 USDA area survey (Fuller, 1991)
Simulations: Designs and Estimators

MC sample size of 2000

For each MC sample,

1. Generate a new set of $z_{ij}$
2. Select a stratified SRS with areas as strata
   - $n_i N_i^{-1} \approx 0.0375$
   - $n_i = 5, i = 1, \ldots, 15; n_i = 20, i = 16, \ldots, 30$

Estimators

- TrIP - Indirect predictor based on Karlberg (2000)
- TrMBD - Model-based direct estimator of Chandra and Chambers (2011)
- EB - empirical best predictor
- EB.BC - EB predictor with multiplicative bias correction
Simulations: Results

- Relative bias of predictor, $\hat{y}_{Ni}$
  
  $$RB_i = \frac{E_{MC}[\hat{y}_{Ni} - \bar{y}_{Ni}]}{E[\bar{y}_{Ni}]}$$

- TrIP and TrMBD unibased
- $RB_i$ of EB larger for $\sigma_z = 2$ than $\sigma_z = 1.6$
  - For $\sigma_z = 1.2$ and $n_i = 5$, the average $RB_i$ is 1.3 for $\sigma_u = 0.5$ and 1.4 for $\sigma_u = 0.2$
  - Average $RB_i$ less than 3% of MC RMSE
- EB.BC unbiased
- $RB_i$ smaller for $n_i = 20$ than $n_i = 5$
Simulations: Results

MSE of predictor, \( \hat{y}_{N_i} \), relative to MSE of EB.BC predictor

\[ \text{RelMSE}_i = \frac{\text{MSE}_{MC}(\hat{y}_{N_i})}{\text{MSE}_{MC}(\hat{y}^{EB.BC}_{N_i})} \]

<table>
<thead>
<tr>
<th>( \sigma_u^2 \sigma_e^{-2} )</th>
<th>( \sigma_z )</th>
<th>( n_i = 5 )</th>
<th>( n_i = 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1.6</td>
<td>2.2</td>
<td>6.1</td>
</tr>
<tr>
<td>0.2</td>
<td>1.6</td>
<td>1.3</td>
<td>2.5</td>
</tr>
<tr>
<td>0.5</td>
<td>1.2</td>
<td>2.2</td>
<td>5.9</td>
</tr>
<tr>
<td>0.2</td>
<td>1.2</td>
<td>1.3</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Simulations: Results

- Properties of MSE estimators
- Relative bias

\[ \text{Relative bias} \ (RB_i) = \frac{E_{MC}[\hat{MSE}_{2i}] - \text{MSE}_{MC}(\hat{y}_{Ni}^{EB, BC})}{\text{MSE}_{MC}(\hat{y}_{Ni}^{EB, BC})} \]

- Average \( RB_i \) between -2.2% and 10.7%
- Coverage of normal theory CI’s with nominal coverage of 95%
  - Empirical coverage between 94.6% and 95.3%
Unit-Level Lognormal: Take-Home Messages

- Unit-Level lognormal model
  - Extends the linear, unit-level model to handle skewed, positive response variables

- Empirical Bayes predictor
  - Closed-form expression
  - More efficient than competitors in simulations

- MSE estimator
  - Closed-form expression
  - Relative biases less than 11%
  - Empirical coverages close to the nominal level
Basic concepts of SAE for nonlinear models
- EB predictor is estimate of conditional expectation of small area mean given data
- MSE of EB predictor decomposes into a sum of two terms
  - Leading term = conditional variance of small area mean given data
  - Second term = variance due to estimation of fixed parameters
Part 2: Zero-Inflated Lognormal Model (joint work with Annie Lyu)

- Extends the lognormal to handle zero-inflated data
- Apply the method to data from an agricultural survey
- Discuss the challenges and importance of obtaining unit-level auxiliary information at the population level
Sheet and Rill Erosion

- Sheet and rill erosion (SRE) – transport of soil from thin surface layers (sheets) or small channels (rills) due to rainfall or shallow runoff

Factors Impacting Erosion

- Rainfall
- Soil properties
  - Slope length/steepness
  - Erosivity (ease of detachment)
- Crop managements
- Conservation practices

- SRE degrades agricultural land and pollutes water
- Conservation policies rely on estimates of SRE
Small Area Estimation for SRE

Conservation Effects Assessment Project (CEAP)
- Two-phase survey that quantifies water & wind erosion on cropland
  - CEAP conducts farmer interviews at a subset of locations classified as cropland in a larger survey called the National Resources Inventory (NRI)
- Survey data and auxiliary info. on soils and climate are processed through the APEX computer model
- An approximation for SRE is one APEX output
Small Area Estimation for SRE

- Estimates of average SRE in South Dakota counties are of interest
- CEAP county sample sizes are small → small area estimation
Small Area Estimation for SRE

- Exploratory analysis of CEAP SRE data

![Graph 1: SRE Distribution](image1)

![Graph 2: County Proportion vs. Log Positive](image2)
Small Area Estimation for SRE

Zero-Inflated Lognormal Model

\[
\text{SRE: } y_{ij}^* = y_{ij} \delta_{ij} \\
y_{ij} > 0; \quad \delta_{ij} \sim \text{Bernoulli}(p_{ij})
\]

\[
i = 1, \ldots, 64 \text{ (SD Counties)}; \quad j = 1, \ldots, N_i \text{ (crop field in pop.)}
\]

\[
\begin{align*}
\text{Positive Part} & \\
\log(y_{ij}) &= \beta_0 + z_{1,ij}' \beta_1 + u_i + e_{ij} \\
e_{ij} &\overset{iid}{\sim} N(0, \sigma_e^2)
\end{align*}
\]

\[
\begin{align*}
\text{Binary Part} & \\
\text{logit}(p_{ij}) &= \alpha_0 + z_{2,ij}' \alpha_1 + b_i \\
p_{ij} : &= p_{ij}(b_i)
\end{align*}
\]

Correlation b/ Positive and Binary Parts

\[
\begin{pmatrix} u_i \\ b_i \end{pmatrix} \overset{iid}{\sim} \text{BVN}(0, \Sigma_{ub}), \quad \Sigma_{ub} = \begin{pmatrix} \sigma_u^2 & \sigma_{ub} \\ \sigma_{ub} & \sigma_b^2 \end{pmatrix}
\]

\[
\sigma_{ub} = \rho \sigma_u \sigma_b
\]
Small Area Estimation for SRE

- County mean of interest: \( \bar{y}^{*}_{N_i} = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}^{*} \)
- Data for small area prediction:
  - \((y^{*}, z) = \{y_{ij}^{*} : j \in s_i\} \cup \{z_{ij} = (z'_{1,ij}, z'_{2,ij})' : j \in s_i \cup \bar{s}_i\}\)
  - \(s_i\) is sample for county \(i\) with sample size \(n_i = |s_i|\)
  - \(\bar{s}_i\) is set of nonsampled elements in county \(i\) with \(|\bar{s}_i| = N_i - n_i\)
- Minimum MSE (Bayes) predictor:
  \[
  \hat{y}^{*\text{MMSE}}_{N_i}(\theta) = N_i^{-1} \left[ \sum_{j \in s_i} y_{ij}^{*} + \sum_{j \in \bar{s}_i} \mathbb{E}\{y_{ij}^{*} | (y^{*}, z); \theta\} \right]
  \]
  \[
  \theta = (\beta_0, \beta'_1, \alpha_0, \alpha'_1, \sigma_e^2, \sigma_u^2, \sigma_b^2, \rho)'
  \]
- Challenge
  - Correlation parameter \(\rho\) introduces a need for integration over the bivariate distribution of \((u_i, b_i)\)
- Approach
  - Transform bivariate integrals to univariate integrals
Gauss-Hermite approximation to univariate integral

Empirical Bayes (EB) predictor:

\[
\hat{y}_{N_i}^{*\text{MMSE}}(\hat{\theta}) = \frac{1}{N_i} \left[ \sum_{j \in s_i} y_{ij}^* + \sum_{j \in \bar{s}_i} \mathbb{E}\{y_{ij}^* \mid (y^*, z); \hat{\theta}\} \right]
\]

Maximum likelihood estimator \(\hat{\theta}\)
Small Area Estimation for SRE: MSE Estimator

- “One-step” estimator of $M_{1i}(\theta)$
  - Replace $\theta$ with MLE $\hat{\theta}$
  - Use Gauss-Hermite to approximate univariate integral
- $M_{2i}^{\text{boot}}(\hat{\theta}) = \text{parametric bootstrap estimator of } M_{2i}(\theta)$

“Semi-boot” MSE estimator

$$\hat{MSE}_i = M_{1i}(\hat{\theta}) + M_{2i}^{\text{boot}}(\hat{\theta})$$
Small Area Estimation for SRE: Covariates

* Covariates measure factors impacting erosion and are known for the full population of cropland in South Dakota

- Rainfall
  - \( \log R \): log R-factor, a measure of long-term, average rainfall in a county

- Soils
  - \( \log S \): log of slope steepness factor at a unit’s location
  - \( \log K \): log of soil K-factor (erodibility index) at a unit’s location
    - Higher K-factors indicate greater erosivity – potential for detachment

- Crop type
  - We use crop classifications from the Cropland Data Layer (CDL), a satellite-derived landcover map with 30meter\(^2\) resolution
    - \textit{is.soybean} = 1 if location classified as soybeans; 0 otherwise
    - \textit{is.sprwht} = 1 if location classified as spring wheat; 0 otherwise
Specific covariates selected with step-wise AIC

<table>
<thead>
<tr>
<th></th>
<th>Positive Part</th>
<th>Binary Part</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (SE)</td>
<td>Estimate (SE)</td>
</tr>
<tr>
<td>logR</td>
<td>2.19 (0.36)</td>
<td>4.94 (0.72)</td>
</tr>
<tr>
<td>logK</td>
<td>0.52 (0.23)</td>
<td></td>
</tr>
<tr>
<td>logS</td>
<td>0.49 (0.08)</td>
<td>0.38 (0.21)</td>
</tr>
<tr>
<td>is.soybean</td>
<td></td>
<td>0.71 (0.33)</td>
</tr>
<tr>
<td>is.sprwht</td>
<td></td>
<td>0.98 (0.52)</td>
</tr>
<tr>
<td>Var: county</td>
<td>0.22</td>
<td>0.47</td>
</tr>
<tr>
<td>Var: residual</td>
<td>1.23</td>
<td></td>
</tr>
</tbody>
</table>

Correlation $\hat{\rho} = 0.77$ with 95% bootstrap CI of (0.21, 0.99)
Small Area Estimation for SRE: Spatial Distribution of Predictions

- Cartogram of EB predictors; fraction of shaded area inversely proportional to CV
Discussion

- An understanding of sheet and rill erosion is important for conservation efforts.
- Using CEAP data, we estimate mean SRE for South Dakota counties:
  - Zero-inflated lognormal model
  - Non-trivial correlation between $b_i$ and $u_i$
- A key challenge is deriving covariates that are available for the full population and relate to factors impacting erosion:
  - We integrate NRI, CDL, and Soil Survey to obtain covariates for the population of interest.
Part 3: Informative sampling (joint work with Abdulhakeem Eideh)

- Develop predictors for unit-level models under informative sampling
- Generalize procedures to the broad class of exponential dispersion families
- Prediction of nonlinear parameters
- Validate the methods through simulation and data analysis
Framework

- Areas: \( i = 1, \ldots, D \) (all are sampled)
- Units: \( j = 1, \ldots, N_i \)
- Sample: \( A_i \subset \{1, \ldots, N_i\} \)
- Sample inclusion indicator: \( I_{ij} = I[j \in A_i] \)
- Probability that unit \( j \) in area \( i \) is selected: \( \pi_{ij} = P_{ij}(I_{ij} = 1) \)
- Weight: \( w_{ij} = \pi_{ij}^{-1} \)
- Covariate: \( x_{ij} \) observed for \( j = 1, \ldots, N_i \)
- Response: \( y_{ij} \) observed for \( j \in A_i \)

\[
D_s = \{x_{ij} : j \in U_i\} \cup \{y_{ij} : j \in A_i\} \cup \{w_{ij} : j \in A_i\}
\]
Distributions (Pfeffermann & Sverchkov 2007)

Population Distributions

\[ y_{ij} \overset{ind}{\sim} f_p(y_{ij} \mid u_i, x_{ij}), \quad j = 1, \ldots, N_i \]

\[ u_i \overset{ind}{\sim} f_p(u_i \mid \theta_u, \phi_u), \quad i = 1, \ldots, D \]

Sample Distributions

\[ f_{si}(y_{ij} \mid x_{ij}, u_i) = f_p(y_{ij} \mid u_i, x_{ij}, l_{ij} = 1) \]

Complement Distribution

\[ f_{ci}(y_{ij} \mid x_{ij}, u_i) = f_p(y_{ij} \mid u_i, x_{ij}, l_{ij} = 0) \]

Relationships

\[ f_{ci}(y_{ij} \mid x_{ij}, u_i) \propto E_s(w_{ij} - 1 \mid x_{ij}, y_{ij}) f_{si}(y_{ij} \mid x_{ij}, u_i) \]

\[ f_p(y_{ij} \mid x_{ij}, u_i) \propto E_s(w_{ij} \mid x_{ij}, y_{ij}) f_{si}(y_{ij} \mid x_{ij}, u_i) \]
Assumptions

Exponential Dispersion Family

\[ f_{si}(y_{ij} \mid \theta_{ij}, \phi) = \exp \left[ \phi(y_{ij}\theta_{ij} - b(\theta_{ij})) + c(y_{ij}, \phi) \right] \]
\[ \theta_{ij} = g(x_{ij}, u_i) \]
\[ f(u_i \mid \theta_u, \phi_u) = \exp \left[ \phi_u(u_i\theta_u - b(\theta_u)) + c(u_i, \phi_u) \right] \]


- Mean weight model

\[ E_{si}(\pi_{ij}^{-1} \mid y_{ij}, x_{ij}) = \exp(q_i + \gamma_1 y_{ij} + \gamma_2 x_{ij} y_{ij} + \gamma_3 x_{ij}) \]

- Beta prime weight model: \( \pi_{ij} \mid l_{ij} = 1 \overset{\text{ind}}{\sim} \text{Beta}(\mu_{ij}\phi_1 + 1, (1 - \mu_{ij})\phi_1) \)

\[ E_{si}(\pi_{ij}^{-1} - 1 \mid y_{ij}, x_{ij}) = \alpha_{0,i} \exp(\alpha_1 y_{ij} + \alpha_2 x_{ij} y_{ij} + \alpha_3 x_{ij}) := \mu_{ij} - 1 \]
Implications

Theorem 1

- Under the mean weight model,

\[ f_p(y_{ij} \mid x_{ij}, u_i) = \exp[\phi(y_{ij}\theta^*_ij - b(\theta^*_ij)) + c(y_{ij}, \phi)] \]
\[ \theta^*_ij = \theta_{ij} + \frac{\gamma_2}{\phi} + x'_{1,ij}\gamma_3/\phi \]

Theorem 2

- Under the beta-prime weight model,

\[ f_{ci}(y_{ij} \mid x_{ij}, u_i) = \exp[\phi(y_{ij}\tilde{\theta}_{ij} - b(\tilde{\theta}_{ij})) + c(y_{ij}, \phi)] \]
\[ \tilde{\theta}_{ij} = \theta_{ij} + \frac{\alpha_2}{\phi} + x'_{1,ij}\alpha_3/\phi \]
Estimators

Max. Likelihood for $\psi_1 = (\beta', \phi_u, \theta_u)'$

$$\hat{\psi}_1 = \arg\max_{\psi_1} \sum_{i=1}^{D} \log \left( \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_{s_i}(y_{ij} \mid \theta_{ij}(u), \phi) \right) f_{s}(u \mid \phi_u, \theta_u) du$$

Least Squares for $\psi_2 = (\gamma_2, \gamma_3)'$

Let $(\hat{q}_1, \ldots, \hat{q}_D, \hat{\gamma}_1', \hat{\gamma}_2, \hat{\gamma}_3')$ minimize

$$\sum_{i \in s} \sum_{j \in s_i} (\log(w_{ij}) - q_i - x'_{1,ij} \gamma_1 - y_{ij} \gamma_2 - y_{ij} x'_{1,ij} \gamma_3)^2$$

Max. Likelihood for Beta-Prime Weight Model

$$(\hat{\phi}_1, \hat{\alpha}')' = \arg\max_{(\phi_1, \alpha)'} L_3(\phi_1, \alpha), \quad \alpha = (\alpha_{0,1}, \ldots, \alpha_{0,D}, \alpha_1', \alpha_2, \alpha_3')' ,$$

$$L_3(\phi_1, \alpha) = \prod_{i,j} \frac{1}{B(\mu_{ij}\phi_1 + 1, (1 - \mu_{ij})\phi_1)} \pi_{ij}^{\mu_{ij}} (1 - \pi_{ij})^{(1 - \mu_{ij})\phi_1 - 1}$$
Empirical Best Prediction Overview

- Generalizations of Molina & Rao (2010)
  - Simulate from complement distributions
- Algorithm 1: mean weight model
  - Simulate from population distribution
- Algorithm 2: beta-prime
  - Simulate from complement distribution
Algorithm 1: For $r = 1, \ldots, R$, repeat the following steps.

1. Generate $u_i^{(r)} \overset{iid}{\sim} f_i(u_i \mid D_s, \hat{\beta}, \hat{\phi}, \hat{\phi}_u, \hat{\theta}_u)$ for $i = 1, \ldots, D$, where
   \[
f_i(u_i \mid D_s, \beta, \phi, \phi_u, \theta_u) = \frac{[\prod_{j=1}^{n_i} f_{si}(y_{ij} \mid \theta_{ij}(u_i), \phi)] f_s(u_i \mid \theta_u, \phi_u)}{\int_{-\infty}^{\infty} [\prod_{j=1}^{n_i} f_{si}(y_{ij} \mid \theta_{ij}(u), \phi)] f_s(u \mid \theta_u, \phi_u) du}.
   \]

2. Generate $y_{ij}^{(r)} \overset{ind}{\sim} f_p(y_{ij} \mid \hat{\theta}_{ij}^{(r)}, \hat{\gamma}_2, \hat{\gamma}_3)$, where $\hat{\theta}_{ij}^{(r)} = g(x_{ij}^\prime \hat{\beta}, u_i^{(r)})$ for $j \in U_i$.

3. Define $\theta_i^{(r)}(\hat{\psi}) = h(y_{i1}^{(r)}, \ldots, y_{iN_i}^{(r)})$.

Empirical best predictor: $\hat{\theta}_i = \tilde{\theta}_i(\hat{\psi}) = R^{-1} \sum_{r=1}^{R} \theta_i^{(r)}(\hat{\psi})$, $\hat{\psi} = (\hat{\psi}_1, \hat{\psi}_2)'$.
Algorithm 2: For $r = 1, \ldots, R$, repeat the following steps.

1. Generate $u_i^{(r)} \overset{iid}{\sim} f_i(u_i \mid D_s, \hat{\beta}, \hat{\phi}, \hat{\phi}_u, \hat{\theta}_u)$ for $i = 1, \ldots, D$.

2. Generate $y_{ij}^{(r)} \overset{ind}{\sim} f_{ci}(y_{ij} \mid \hat{\theta}_{ij}^{(r)}, \hat{\alpha}_2, \hat{\alpha}_3)$, where $\hat{\theta}_{ij}^{(r)} = g(x_{ij}^t\hat{\beta}, u_i^{(r)})$ for $j \notin s_i$. For $j \in s_i$, set $y_{ij}^{(r)} = y_{ij}$.

3. Define $\theta_i^{(r)}(\hat{\psi}) = h(y_{i1}^{(r)}, \ldots, y_{iN_i}^{(r)})$.

Empirical best predictor: $\hat{\theta}_i = \tilde{\theta}_i(\hat{\psi}^{BP}; y_{is}) = R^{-1} \sum_{r=1}^{R} \theta_i^{(r)}(\hat{\psi}^{BP}; y_{is})$, $\hat{\psi}^{BP} = (\hat{\psi}_1', \hat{\alpha}_2, \hat{\alpha}_3)'$
MSE Estimation

\[ \text{MSE}(\hat{\theta}_i) = M_{1i} + M_{2i} \]
\[ M_{1i} = E[V(\theta_i \mid D_s)] \]
\[ M_{2i} \text{ reflects variation of parameter estimators} \]

Mean Weight Model

- Challenge: Do not specify full distribution for sampling weight → fully parametric bootstrap not apply
- Solution: Use MSE estimation procedure of Cho & Berg (2022)
  - Estimate \( M_{1i} \) as sample variance of \( \theta_i^{(r)} r = 1, \ldots, R \)
  - Estimate \( M_{2i} \) from asymptotic normal distribution of parameter estimators

Beta-Prime Weight Model

- Method 1: Cho & Berg (2022)
- Method 2: fully parametric bootstrap (González-Manteiga et al. 2008)
Simulations: Set-Up

\[ D = 50, \quad N_i = 200 \]

\[ y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij}) \]

\[ \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = -0.2 + 0.7x_{ij} + u_i \]

\[ u_i \overset{iid}{\sim} \mathcal{N}(0, 0.25) \]

**Parameters**

- **Mean** = \[ N_i^{-1} \sum_{j=1}^{N_i} y_{ij} \]
- **Var** = \[ (N_i - 1)^{-1} \sum_{j=1}^{N_i} (y_{ij} - N_i^{-1} \sum_{j=1}^{N_i} y_{ij})^2 \]
- **Odds** = \[ \left( \sum_{j=1}^{N_i} 1 - y_{ij} \right) \left( \sum_{j=1}^{N_i} y_{ij} \right)^{-1} \]
Alternative Procedures

- Inf: Proposed method
- Noninf: Proposed method with $\gamma = \alpha = 0$
- PL: Bayesian pseudo-likelihood (Parker et al. 2023)
  - Stan code provided on Github \(^1\)

\[ f(y_{si} | u, x_{si}) = \prod_{j \in A_i} f(y_{ij} | u_i, x_{ij})^{w_{ij}} \]

\[ y_{si} = \{y_{ij} : j \in A_i\} \]
\[ u = (u_1, \ldots, u_D)' \]

- MSE estimators
  - Inf-1: Cho & Berg (2022)
  - Inf-2: Fully parametric bootstrap
  - PL: posterior variance

\(^1\)https://github.com/paparker/Unit_Level_Models/blob/master/Model_1.stan
Criteria

\[
\text{AvMSE} = M^{-1}D^{-1} \sum_{i=1}^{D} \sum_{m=1}^{M} (\hat{\theta}_i^{(m)} - \theta_i^{(m)})^2
\]

\[
\text{Rel. Abs. Bias} = \frac{D^{-1}M^{-1} \sum_{i=1}^{D} | \sum_{m=1}^{M} (\hat{\theta}_i^{(m)} - \theta_i^{(m)}) |}{D^{-1}M^{-1} \sum_{m=1}^{M} \sum_{i=1}^{D} \theta_i^{(m)}}
\]

\[
\bar{MSE} = M^{-1}D^{-1} \sum_{m=1}^{M} \sum_{i=1}^{D} \hat{\text{MSE}}_i^{(m)}
\]
Case 1: Mean Weight Model

- PPS-systematic sample with

\[
\pi_{ij} = \frac{10\exp(0.25y_{ij} + \delta_{ij})}{\sum_{k=1}^{N_i} \exp(0.25y_{ik} + \delta_{ik})}
\]

<table>
<thead>
<tr>
<th></th>
<th>Rel. Abs. Bias</th>
<th>AveMSE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inf</td>
<td>Noninf</td>
<td>PL</td>
</tr>
<tr>
<td>Mean</td>
<td>0.53</td>
<td>11.63</td>
<td>1.29</td>
</tr>
<tr>
<td>Var</td>
<td>0.28</td>
<td>0.83</td>
<td>2.49</td>
</tr>
<tr>
<td>Odds</td>
<td>1.39</td>
<td>19.70</td>
<td>NA</td>
</tr>
</tbody>
</table>
Case 2: Beta Prime Weight Model

- PPS-systematic sample with

\[
\pi_{ij} = \frac{\exp(-3 + 0.25y_{ij})}{1 + \exp(-3 + 0.25y_{ij})}.
\]

<table>
<thead>
<tr>
<th></th>
<th>Rel. Abs. Bias</th>
<th>AveMSE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inf</td>
<td>Noninf</td>
<td>PL</td>
</tr>
<tr>
<td>Mean</td>
<td>0.86</td>
<td>11.75</td>
<td>1.55</td>
</tr>
<tr>
<td>Var</td>
<td>0.27</td>
<td>0.79</td>
<td>1.80</td>
</tr>
<tr>
<td>Odds</td>
<td>1.55</td>
<td>19.82</td>
<td>NA</td>
</tr>
</tbody>
</table>
Longitudinal survey of agriculture and natural resources
- We consider presence or absence of wetlands
- We use data for 2012

Multi-faceted sample design
- Foundation sample
  - Stratified 2-stage sample
- Supplemented panel design
  - 2000-present
  - Subsets of foundation sample observed each year

State estimates published
- We consider county estimation
National Resources Inventory Application

\[ D = 21 \text{ counties in New Jersey} \]

\[ y_{ij} = \begin{cases} 
1 & \text{if wetland in 2012} \\
0 & \text{otherwise} 
\end{cases} \]

Parameters: Mean, Var, Odds
Cropland data layer
Sampled elements

\[ x_{ij} = \begin{cases} 
1 & \text{if CDL any kind of wetland} \\
0 & \text{otherwise.} 
\end{cases} \]

Regard a nonsampled location to represent 100 acres

\[ x_{ij} = 1, \ j = 1, \ldots, [A_{w,i}/100] \]
\[ x_{ij} = 0, \ j = [A_{1,i}/100] + 1, \ldots, A_i \]
\[ A_{w,i} = \text{CDL wetland area of county } i \]
\[ A_i = \text{area of county } i \]
Sample Model for Application

\[ y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij}) \]

\[ \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_i \]

\[ u_i \overset{iid}{\sim} N(0, \sigma_u^2) \]

\[ E_s(w_{ij} \mid y_{ij}, x_{ij}) = \kappa_i \exp(\gamma_2 y_{ij} + \gamma_1 x_{ij}) \]
## Model Parameter Estimates

<table>
<thead>
<tr>
<th></th>
<th>Est.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-1.856</td>
<td>0.016</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>2.241</td>
<td>0.016</td>
</tr>
<tr>
<td>$\sigma_u^2$</td>
<td>0.211</td>
<td>0.004</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.070</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Predictions

- Circles: Proposed predictors against direct estimators.
- Triangles: Predictors assuming noninformative sampling against direct estimators.
Uncertainty Measures

- Average standard errors of direct estimators (Dir) and average root mean square errors of predictors (Pred).

<table>
<thead>
<tr>
<th></th>
<th>Dir</th>
<th>Pred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.028</td>
<td>0.025</td>
</tr>
<tr>
<td>Variance</td>
<td>0.014</td>
<td>0.013</td>
</tr>
<tr>
<td>Odds</td>
<td>1.678</td>
<td>0.787</td>
</tr>
</tbody>
</table>
Summary and Discussion

- Developed a small area procedure to address three issues:
  - Exponential dispersion families, informative sampling, nonlinear parameters

Mean Weight vs. Beta-Prime Weight Model

- Mean weight model: requires fewer distributions, applicable if weights are 1
- Beta-prime weight model: allows straightforward MSE estimation

Comparison to Bayesian PL

- PL method uses relatively informative priors, and specification of more diffuse priors led to computational difficulties.
- Proposed method avoids the complicated problem of prior specifications
- Estimators for proposed method can be obtained using standard software
Extensions of the Basic Unit-Level Model

- Three extensions
  - Skewed data
  - Zero-inflated data
  - Informative sampling

- Key themes
  - Concepts of empirical best prediction
  - Importance of population-level covariate information
  - Generalizability to exponential dispersion families and nonlinear parameters
Thank You


