Mixed Effects Quantile Regression Models for Small Area Estimation

Emily Berg
Iowa State University
Introduction to small area estimation (SAE)

Introduction to quantile regression (QR)

Motivation: Why consider mixed effects quantile regression for small area estimation?
  ▶ General SAE issues
  ▶ Specific Conservation Effects Assessment Project survey

Mixed effects quantile regression for SAE
  ▶ Previous approach: Asymmetric Laplace distribution
  ▶ Proposed approach: Linearly interpolated generalized Pareto density

Simulations

Conclusions and next steps
Introduction to Small Area Estimation

- Large-scale surveys are often designed for large domains
  - Canadian Labor Force Survey (Statistics Canada)
    - National unemployment rate
    - Occupational totals by province
  - National Resources Inventory (USDA-Natural Resources Conservation Service & ISU)
    - Average water erosion for NRCS regions (groups of states)
    - Cropland acres by state
  - June Area Survey (National Agricultural Statistics Service)
    - Area planted to corn nationally and in major corn producing states
Large domain estimators are often judged reliable

- One measure of reliability is the estimated coefficient of variation (CV)
- American Community Survey reliability guidelines
  - Good: CV $\leq 15\%$
  - Fair: CV between 15 & 30%
  - Use caution: CV $> 30\%$
  - Median CV for a table $\geq 60\%$ → suppress

<table>
<thead>
<tr>
<th>Year</th>
<th>Crop</th>
<th>Pasture</th>
<th>Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>0.5</td>
<td>2.5</td>
<td>3.6</td>
</tr>
<tr>
<td>2002</td>
<td>0.5</td>
<td>2.6</td>
<td>3.5</td>
</tr>
<tr>
<td>2007</td>
<td>0.4</td>
<td>2.5</td>
<td>3.2</td>
</tr>
</tbody>
</table>
Introduction to Small Area Estimation

- Estimates for small domains are often of interest
  - Small Area Income and Poverty Estimation (SAIPE): proportion of children in poverty for ≈ 13,500 school districts
    - “No Child Left Behind Act of 2001 directs the Department of Education to distribute Title I basic and concentration grants directly to school districts on the basis of the most recent Census Bureau estimates of school-age children in poverty in each school district in the U.S.” (Maples & Bell, 2007)
  - Crop area for ≈ 3,000 counties (National Agricultural Statistics Service)
    - Crop insurance rates
  - Small Area Income and Health Estimation (SAIHE): proportion with health insurance by age group & county
    - Public health policy planning and evaluation
Introduction to Small Area Estimation

- Challenge in SAE: standard survey estimators for small domains often judged unreliable
  - Small sample sizes for small domains
  - Large estimated CVs (i.e., CV > 30% or CV > 100%)

Median County-Level CVs (%) NRI Iowa
SAE approach: main ideas

1. Auxiliary information
   - Known for the whole population
   - Correlated with variable of interest
     - Example: Income tax data used in SAIPE to estimate proportions of school-aged children in poverty

2. Assumptions about the structure of the domain parameters
   - Mixed effects models
   - Allows pooling information from multiple areas to estimate the parameter for a single small area
Introduction to Small Area Estimation

SAE approach: unit-level model (Battese, Harter, & Fuller 1988)

- **Notation**
  - **Areas:** \( i = 1, \ldots, D \)
  - \((N_i, n_i) = \) (population size, sample size)
  - **Units:** \( j = 1, \ldots, N_i \)
  - \( y_{ij} = \) variable of interest
  - \( x_{ij} = \) covariate

- **Population mean**
  \[ \bar{y}_{N_i} = N_i^{-1} \sum_{j=1}^{N_i} y_{ij} \]

- **Data**
  - \((x_{ij}, y_{ij}) : j = 1, \ldots, n_i; i = 1, \ldots, D \)
  - \[ \bar{x}_{N_i} = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}; i = 1, \ldots, D \]
Introduction to Small Area Estimation

SAE approach: unit-level linear model

- Mixed effects model

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

\[(u_i, e_{ij})' \sim (\mathbf{0}, \text{diag}(\sigma_u^2, \sigma_e^2))\]

- Best linear unbiased predictor (BLUP)

\[ \hat{y}_{Ni}(\sigma_u^2, \sigma_e^2) \approx \bar{x}'_{Ni} \hat{\beta} + \gamma_i (\bar{y}_n - \bar{x}'_{ni} \hat{\beta}) \]

\[ \hat{\beta} = \left( \sum_{i=1}^{D} \mathbf{x}_i V_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{D} \mathbf{x}_i V_i \mathbf{y}_i \]

\[ V_i = 1_{n_i} 1_{n_i}' \sigma_u^2 + \text{diag}(1_{n_i} \sigma_e^2) \]

\[ \gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + n_i^{-1} \sigma_e^2} \]

- Estimated best linear unbiased predictor (EBLUP): \( \hat{y}_{Ni}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) \)

\[ \hat{\sigma}_u^2, \hat{\sigma}_e^2 \text{ from REML, ML, MOM} \]
Introduction to Small Area Estimation

SAE approach: unit-level linear model

- EBLUP estimates minimum MSE predictor for normal errors

\[
\hat{y}_{Ni} = \hat{E}[\bar{y}_{Ni} | y_{i1}, \ldots, y_{in_i}; \sigma_u^2, \sigma_e^2, \beta]
\]

- Estimated MSE of the BLUP:

\[
\text{MSE}\{\hat{y}_{Ni}(\sigma_u^2, \sigma_e^2)\} = \left[ \hat{\gamma}_i \hat{\sigma}_e^2 n_i^{-1} \right] + \left( \bar{x}_{Ni} - \hat{\gamma}_i \bar{x}_{n_i} \right)' \hat{V}\{\beta\} \left( \bar{x}_{Ni} - \hat{\gamma}_i \bar{x}_{n_i} \right)
\]

- Estimators of the MSE of the EBLUP include additional terms to account for estimation of \( \sigma_u^2 \) and \( \sigma_e^2 \)

- Replication (i.e., bootstrap) an alternative

Parameter estimation

Prediction variance
Introduction to Small Area Estimation

Unit-level model

Properties

- Closed-form expressions for predictor and MSE estimator
- Standard estimators (REML, ML, least squares) not resistant to outliers
- Predictors may not be robust to departures from model assumptions, particularly skewed error distributions
Quantile regression

- Avoid full distributional assumptions by specifying a model for the quantiles of the conditional distribution of the response given covariates
  - Encompasses heavy-tailed distributions, skewed error distributions, and non-constant error variances
Quantile Regression (QR) Idea

- \( q_{ij}(\tau) \): \( \tau \)th quantile of conditional distribution of \( y_{ij} \) given \( x_{ij} \)

\[
P(y_{ij} \leq q_{ij}(\tau) \mid x_{ij}; i) = \tau
\]

- Quantile regression relates \( q_{ij}(\tau) \) to covariates
Motivation: Why quantile regression (QR) for small area estimation?

- Specification of fully parametric models for small area estimation can be difficult
  - Non-constant variance, outliers
  - Multiple response variables
  - Properties of distributions vary across wide range of conditions (i.e., states, counties, hydrologic units)
Motivation: Why quantile regression (QR) for small area estimation?

- Small area quantile is the parameter of interest
  - Poverty and income analysis (Whitworth, et al., undated):
    - “The Office for National Statistics (ONS) currently produces estimates of mean household income...have limitations in meeting user needs though because the mean income provides little information about the distribution across households and can be inflated by the relatively small number of households with very large incomes. Estimates of the median income (together with other quantiles) are considered to be more useful and would better inform user requirements.”
  - Water quality monitoring (Pratesi, Ranalli, and Salvati, 2008)
  - Forestry (Chen and Liu, 2013)
Motivation: Conservation Effects Assessment Project (CEAP)

- Survey to measure soil and nutrient loss due to water and wind erosion on cropland
  - Supported by the Natural Resources Conservation Service (NRCS) of the United States Department of Agriculture (USDA)
- Domains: hydrologic units (HUCs)
  - Hierarchical structure: 8-digit HUCs are nested in 4-digit HUCs
  - 4-digit estimates published
  - Estimates for smaller levels of geographic detail are desired
Motivation: Conservation Effects Assessment Project (CEAP)

- Des Moines River Watershed (0710)
  - Nine 8-digit cataloging units (07100001 - 07100009)
  - 07100001 primarily in Minnesota
  - The other 8 are primarily in Iowa.
Motivation: Conservation Effects Assessment Project (CEAP)

- Response variables: ≈16 measures of soil and nutrient loss due to water and wind erosion
  - RUSLE2– sheet and rill erosion for cropped area of field
  - Sediment – edge of field sediment loss
  - Wind erosion
  - TN loss – total of surface and sub-surface losses
  - N runoff – lbs soluble in surface Nitrogen runoff
  - N sed – lbs Nitrogen attached to sediment (wind and water)
  - Ploss – lbs of Phosphorus loss, total of surface and subsurface loss
  - Soluble P – lbs soluble Phosphorus lost in runoff
  - Sed P – lbs of Phosphorus attached to sediment (wind and water)
Motivation: Conservation Effects Assessment Project (CEAP)

- Estimated means and CVs

<table>
<thead>
<tr>
<th>HUC8</th>
<th>RUSLE2</th>
<th>Sed.</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.12</td>
<td>0.12</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.25</td>
<td>0.41</td>
</tr>
<tr>
<td>2</td>
<td>0.30</td>
<td>0.42</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>0.21</td>
<td>0.30</td>
<td>0.53</td>
</tr>
<tr>
<td>4</td>
<td>0.14</td>
<td>0.27</td>
<td>0.58</td>
</tr>
<tr>
<td>5</td>
<td>0.20</td>
<td>0.34</td>
<td>0.38</td>
</tr>
<tr>
<td>6</td>
<td>0.38</td>
<td>0.54</td>
<td>0.18</td>
</tr>
<tr>
<td>7</td>
<td>0.56</td>
<td>0.77</td>
<td>0.12</td>
</tr>
<tr>
<td>8</td>
<td>0.50</td>
<td>1.14</td>
<td>0.07</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUC4</td>
<td>0.29</td>
<td>0.43</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Motivation: Conservation Effects Assessment Project (CEAP)

External Information for CEAP Estimation
- Average precipitation (8-digit HUC)
- Soil Survey
  - Census of soils
  - Map of soil characteristics
- Soil characteristics related to CEAP variables
  - Slope steepness
  - Erodibility indexes
    - K-factor for water erosion
    - I-factor for wind erosion
* Known for full population
Motivation: Conservation Effects Assessment Project (CEAP)

Preliminary analysis: unit-level linear model residuals (original scale)
Motivation: Conservation Effects Assessment Project (CEAP)

Preliminary analysis: unit-level linear model residuals (original scale)
Motivation: Conservation Effects Assessment Project (CEAP)

- Finding a single family of parametric models that adequately describes the distribution of all variables is difficult.
- Quantile regression has potential to unify analysis of multiple response variables.
Motivation: Conservation Effects Assessment Project (CEAP)

- The quantile function is a parameter of interest in CEAP.

- Reduction in wind erosion (tons/acre) due to the use of conservation practices in the Upper Mississippi River Basin

- NRCS reports contain similar plots for other variables
Motivation: Mixed effects quantile regression models

- Why Mixed effects?
  - Area (8-digit hydrologic units) sample sizes small
  - Similarities in assumed distributions across areas justify using data from multiple areas to inform the predictor for a single small area
  - Area random effects describe between-area heterogeneity, unexplained by covariates

- Why quantile regression?
  - Robust: not require specification of a fully parametric conditional distribution
  - Resistant to outliers
  - Links directly to the parameter of interest when the parameter is a quantile
Mixed Effects Quantile Regression Models

Population and data structure for small area estimation

- **Population**
  - $y_{ij} = \text{variable of interest for element } j \text{ in area } i$
  - $x_{ij} = \text{covariate}$
    - $i = 1, \ldots, D$, $j = 1, \ldots, N_i$
  - $\tau^{\text{th}} \text{ quantile: } q_{ij}(\tau)$

\[
P(y_{ij} \leq q_{ij}(\tau) \mid x_{ij}; i) = \tau
\]

- Assume $q_{ij}(\tau)$ increasing, continuous

- **Sample data**
  - $y_{ij}: i = 1, \ldots, D$, $j = 1, \ldots, n_i$
  - $x_{ij}: i = 1, \ldots, D$, $j = 1, \ldots, N_i$

- **Objective:** Use a mixed effects model for $q_{ij}(\tau)$ to predict small area parameters
Previous work: Asymmetric Laplace Distribution

- Koenker’s check function
  \[ \rho_{\tau}(u) = u(\tau - I[u < 0]) \]

- Estimating Eqn. for Quantile
  \[ q(\tau) = \text{argmin}_a R(a) \]
  \[ R(a) = E[\rho_{\tau}(y - a)] \]
  \[ \rho_{\tau}(\nu) = \nu(\tau - I[\nu \leq 0]) \]

- Asymmetric Laplace Distribution
  \[ y \sim ALD(\mu_{\tau}, \sigma_{\tau}) \]
  \[ f_Y(y \mid \mu_{\tau}, \sigma_{\tau}) \propto \sigma_{\tau}^{-1} h(y, \mu_{\tau}, \sigma_{\tau}) \]
  \[ h(y, \mu_{\tau}, \sigma_{\tau}) = \exp\left\{-\rho_{\tau}\left[\frac{y - \mu_{\tau}}{\sigma_{\tau}}\right]\right\} \]

Connection: ALD and Check Function

MLF of \( \mu_{\tau} \) under ALD is \( q(\tau) \)
Previous work: Asymmetric Laplace Distribution

- Geraci & Bottai, 2007; 2014

\[ y_{ij} \mid \alpha_i(\tau) \sim \text{ALD}(q_{ij}(\tau), \sigma^2(\tau)) \]

\[ q_{ij}(\tau) = x_{ij}' \beta(\tau) + \alpha_i(\tau), \alpha_i(\tau) \sim \mathcal{N}(0, \sigma^2_\alpha(\tau)) \]

- Monte Carlo MLE for \( \beta(\tau), \sigma^2(\tau), \sigma^2_\alpha(\tau) \)
- Best (estimated) linear predictor of \( \alpha_i(\tau) \)
  - Means and covariances based on ALD conditional distribution
- \( \text{R} \) package \text{lqmm}
Previous work: Asymmetric Laplace Distribution

- Application to SAE (Weidenhammer et al., 2016)
  - Key idea: **Grid** of $\tau \in (0, 1)$ ($\tau_k : k = 1, \ldots, K$)
  - Fit ALD model for each $\tau_k : k = 1, \ldots, K$
    
    $$(\hat{\sigma}^2(\tau_k), \beta(\tau_k), \sigma^2_\alpha(\tau_k)) : k = 1, \ldots, K$$

  - Predict $q_{ij}(\tau_k) : k = 1, \ldots, K$
    
    $$\hat{q}_{ij}(\tau_k) = \mathbf{x}_{ij}' \hat{\beta}(\tau_k) + \hat{\alpha}_i(\tau_k)$$

  - Estimate small area parameters
    - Predictor of population quantile:
      
      $$\hat{q}_{Ni}(\tau) = \tau^{th} \text{ empirical quant. of}\{\hat{q}_{ij}(\tau_k) : j = 1, \ldots, N_i; k = 1, \ldots, K\}$$
    
    - Predictor of population mean:
      
      $$\hat{y}_{Ni} = \text{avg. of}\{\hat{q}_{ij}(\tau_k) : j = 1, \ldots, N_i; k = 1, \ldots, K\}$$

- Bootstrap MSE estimation
- Extension to count data
Previous work: Asymmetric Laplace Distribution

- **Benefits**
  - Computational simplicity (R package)
  - Weidenhammer et al. (2016) demonstrate robustness to outliers & non-constant variances

**Important Characteristic**

- Model and predictor defined separately for each
  \[
  \tau_k : k = 1, \ldots, K
  \]
- Quantile functions may decrease
  - Possible for \( q_{ij}(\tau + \delta) < q_{ij}(\tau), \delta > 0 \)
  - Possible for \( \hat{q}_{ij}(\tau + \delta) < \hat{q}_{ij}(\tau), \delta > 0 \)

**Implications of Decreasing Quantile Functions for SAE**

- Empirical Bayes predictor undefined
- Appropriate bootstrap distribution unclear
Proposed approach: Linearly Interpolated Generalized Pareto Density (LIGPD)

- Objectives
  - Continuous, increasing quantile function
    - Empirical Bayes prediction
    - Bootstrap MSE estimation
  - Stable estimators in the tails
  - Computational simplicity

- LIGPD (Jang and Wang, 2015) addresses these issues
  - Fixed random effect across quantile levels
  - Main ideas:
    - Central quantiles: approximate conditional density by linearly interpolating conditional quantiles (LI)
    - Tail density: assume a generalized Pareto density (GPD)
Proposed approach: LIGPD for SAE

Mixed effects quantile regression model (Jang and Wang, 2015)

\[ q_{ij}(\tau) = q_{F,ij}(\tau) + b_i \]

- **Fixed**
  \[ q_{F,ij}(\tau) = \mathbf{x}_{ij}' \boldsymbol{\beta}(\tau) \]
  \[ q_{F,ij}(\tau) < q_{F,ij}(\tau + \delta) \]

- **Random**
  \[ b_i \sim f_b(b_i, \sigma_b) \]
  \[ E[b_i] = 0 \]

  - Bayesian inference for \( \boldsymbol{\beta}(\tau) \)
  - Normally distributed \( b_i \)

- Modifications
  - Frequentist inference for small area parameters
  - \( b_i \) has a specified mean 0 distribution
Proposed approach: LIGPD for SAE

Approximate likelihood (Jang and Wang, 2015):

\[
f(y \mid \mathbf{x}_{ij}, \beta_K, b_i) = I[y < q_{ij}(\tau_1)]\tau_1 f_{\ell ij}(y \mid \beta_K, \rho_\ell, \xi_\ell) \\
+ I[y > q_{ij}(\tau_K)]\tau_K f_{uij}(y \mid \beta_K, \rho_u, \xi_u) \\
+ \sum_{k=1}^{K-1} I[q_{ij}(\tau_k) \leq y < q_{ij}(\tau_{k+1})] \frac{\tau_{k+1} - \tau_k}{q_{ij}(\tau_{k+1}) - q_{ij}(\tau_k)}
\]

\[
\beta_K = (\beta_1(\tau_1), \ldots, \beta_K(\tau_K)), \tau_1 < \cdots < \tau_K
\]

- Motivation for central quantiles

\[
f(y \mid \mathbf{x}_{ij}, \beta_K, b_i) = \lim_{\delta \to 0} \frac{\delta}{q_{ij}(\tau + \delta) - q_{ij}(\tau)}
\]

- Generalized Pareto distributions for lower and upper tails

\[
f_s(y \mid \rho_s, \xi_s) = \begin{cases} 
\rho_s^{-1}(1 + \xi_s y/\rho_s)^{-(1+1/\xi_s)}, & \xi_s \neq 0 \\
\rho_s^{-1}\exp(-y/\rho_s), & \xi_s = 0
\end{cases}
\]
Proposed approach: LIGPD for SAE

Bayes predictor

\[
E[b_i \mid y_i; \theta] = \frac{\int_{-\infty}^{\infty} \prod_{j=1}^{ni} b_i f(y_{ij} \mid x_{ij}, \beta_K, b_i) f_b(b_i \mid \sigma_b) db_i}{\int_{-\infty}^{\infty} \prod_{j=1}^{ni} f(y_{ij} \mid x_{ij}, \beta_K, b_i) f_b(b_i \mid \sigma_b) db_i}
\]

\[
\theta = (\beta'_K, \sigma'_b)'
\]

Numerical approximation of the integral
Proposed approach: LIGPD for SAE

Estimation of $\theta$ (simple)

1. Estimate $\sigma_b = (V\{b_i\}, \gamma)$
   - Median regression of $y_{ij}$ on $x_{ij}$ and fixed area indicators
     
     $$(\hat{b}_1^{(0)}, \hat{V}_1(\hat{b}_1^{(0)})), \ldots, (\hat{b}_D^{(0)}, \hat{V}_D(\hat{b}_D^{(0)}))$$
   - Area level small area (Wang and Fuller, 2003) for $\hat{V}\{b_i\}$
   - MLE $\hat{\gamma} = \arg\max \sum_{i=1}^{D} \log[f_b(\hat{b}_i^{(0)} \mid \hat{V}\{b_i\}, \gamma)]$

2. Estimate $\beta(\tau_k): k = 1, \ldots, K$

   $$\hat{\beta}(\tau) = \arg\min \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_\tau(y_{ij} - \hat{b}_i^{(0)} - x_{ij}'\beta(\tau))$$

   - Apply isotonic regression & linear interpolation to $x_{ij}'\beta(\tau_k)$ to ensure continuous, increasing quantile function, $\hat{q}_{F,ij}(\tau)$

3. Match quantile function for $\rho_s$, & MLE for $\xi_s$ ($s = \ell, u$) using

   $y_{ij} < 0.5(\hat{q}_{F,ij}(\tau_1) + \hat{q}_{F,ij}(\tau_2)); y_{ij} > 0.5(\hat{q}_{F,ij}(\tau_{K-1}) + \hat{q}_{F,ij}(\tau_K))$

   (Jang and Wang, 2015)
Proposed approach: LIGPD for SAE

Estimation of $\theta$ (simple)

4. Update estimator of $b_i$

$$\hat{b}^{(1)}_i = E[b_i \mid y_i; \hat{\theta}]$$

5. Update $\beta(\tau_k) : k = 1, \ldots, K$

$$\hat{\beta}(\tau) = \arg_{\min} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_{\tau}(y_{ij} - \hat{b}^{(1)}_i - x'_{ij} \beta(\tau))$$
Proposed approach: LIGPD for SAE

- Empirical Bayes predictor of the quantile

\[
\hat{q}_{ij}(\tau) = \hat{q}_{F,ij}(\tau) + E[b_i \mid y_i, \hat{\theta}]
\]

\[
\hat{\theta} = (\hat{\beta}'_K, \hat{\sigma}'_b)'
\]

- Domain predictors
  - \(\tau\)th quantile:
    \[
    \hat{q}_{N_i}(\tau) = \tau^{th} \text{ empirical quantile of } \{\hat{q}_{ij}(\tau_k) : k = 1, \ldots, K; j = 1, \ldots, N_i\}
    \]
  - Mean

\[
\hat{y}_{N_i} = \frac{1}{KN_i} \sum_{i=1}^{D} \sum_{j=1}^{N_i} \hat{q}_{ij}(\tau_k)
\]
Proposed approach: LIGPD for SAE

Bootstrap distribution

- Recall: quantile regression model

\[ q_{ij}(\tau) = q_{F,ij}(\tau) + b_i \]
\[ b_i \sim f_b(b_i \mid \sigma_b) \]

- Estimate \( \hat{q}_{F,ij}(\tau) \) continuous, increasing
  - By application of isotonic regression, interpolation to \( x'_{ij} \hat{\beta}(\tau) \)

- Bootstrap distribution

\[ q_{ij}^*(\tau) = \hat{q}_{F,ij}(\tau) + b_i^* \]
\[ b_i^* \sim f_b(b_i \mid \hat{\sigma}_b) \]
Proposed approach: LIGPD for SAE

Bootstrap procedure

- Generate bootstrap population: \((b = 1, \ldots, B)\)

\[
y_{ij}^{*(b)} = \hat{q}_{F,i,j}(\tau^{(b)}) + b_i^{*(b)}, \quad j = 1, \ldots, N_i
\]

\[
b_i^{*(b)} \sim f_b(b_i \mid \hat{\sigma}_b), \quad \tau^{(b)} \sim \text{Unif}(0, 1)
\]

- Probability integral transform

- Bootstrap version of finite population parameter: \(\theta_{N_i}^{*(b)}\)
  - Quantile: \(\theta_{N_i}^{*(b)}\) is sample quantile of \(\{y_{ij}^{*(b)} : j = 1, \ldots, N_i\}\)
  - Mean: \(\theta_{N_i}^{*(b)}\) is mean of \(\{y_{ij}^{*(b)} : j = 1, \ldots, N_i\}\)

- Bootstrap version of small area predictor \(\hat{\theta}_{N_i}^{*(b)}\)
  - Select sample & implement estimation procedure

- Mean squared error estimator: \(B^{-1} \sum_{b=1}^{N} (\hat{\theta}_{N_i}^{*(b)} - \theta_{N_i}^{*(b)})^2\)
Simulations

\[ y_{ij} = -0.7 + 0.8x_{ij} + b_i + e_{ij} \]
\[ b_i \sim \mathcal{N}(0, 0.36) \]
\[ (n_i, N_i) = (5, 20) : i = 1, \ldots, 20 \]
\[ (n_i, N_i) = (20, 80) : i = 21, \ldots, 40 \]
\[ x_{ij} \sim \mathcal{N}(6, 6.25) \]

- Three distributions for \( e_{ij} \)
  - \( t_5 \sqrt{3}/\sqrt{5} \)
  - \( (\chi^2_{(5)} - 5)/\sqrt{10} \)
  - \( (\chi^2_{(2)} - 2)/\sqrt{4} \)

- Estimation procedures
  - Normal EB (NEB)
  - Predictor based on ALD (QALD)
  - LIGPD predictor

- Parameters: 25\(^{th}\) & 75\(^{th}\) percentiles, median
- \( \tau_k = 0.01, 0.02, \ldots, 0.99 \)
Simulations

- Monte Carlo (MC) MSE of predictor, relative to NEB

\[
\text{Relative MSE} = \frac{\text{MC MSE}(\hat{\theta})}{\text{MC MSE}(\text{NEB})}
\]

- \(\hat{\theta} = \text{QALD, LIGPD}\)
- MC MSE is average across areas of the same sample size

<table>
<thead>
<tr>
<th>(n_i)</th>
<th>Quartile</th>
<th>(t(5)) Error</th>
<th>(\chi^2(5)) Error</th>
<th>(\chi^2(2)) Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>QALD</td>
<td>LIGPD</td>
<td>QALD</td>
</tr>
<tr>
<td>5</td>
<td>25%</td>
<td>1.14</td>
<td>1.15</td>
<td>1.14</td>
</tr>
<tr>
<td>20</td>
<td>25%</td>
<td>1.01</td>
<td>1.00</td>
<td>1.17</td>
</tr>
<tr>
<td>5</td>
<td>50%</td>
<td>1.14</td>
<td>1.16</td>
<td>1.14</td>
</tr>
<tr>
<td>20</td>
<td>50%</td>
<td>1.15</td>
<td>1.18</td>
<td>1.35</td>
</tr>
<tr>
<td>5</td>
<td>75%</td>
<td>1.10</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>20</td>
<td>75%</td>
<td>0.99</td>
<td>1.01</td>
<td>1.17</td>
</tr>
</tbody>
</table>
Simulations

- Alternative distributions for $b_i$
  - Normal
  - Laplace
  - Skew-Normal
  - Results for Laplace and normal similar
Simulations

![Graphs of t(5) and chi-square distributions with Laplace and S-N b parameters.](image-url)
Simulations

Bias-variance trade-off

\[ \text{MSE} = (\text{Bias})^2 + \text{Variance} \]
Simulations

MC bias (%) for $e_{ij} \sim t_5$

<table>
<thead>
<tr>
<th>$n_i$</th>
<th>$\tau$</th>
<th>$b_i$ Laplace</th>
<th>$b_i$ Skew-Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>QALD</td>
<td>LIGPD</td>
</tr>
<tr>
<td>5</td>
<td>0.25</td>
<td>-1.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
<td>-3.8</td>
<td>1.6</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>-3.8</td>
<td>-1.4</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
<td>-3.3</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>-3.0</td>
<td>-1.8</td>
</tr>
<tr>
<td>20</td>
<td>0.75</td>
<td>0.6</td>
<td>-1.0</td>
</tr>
</tbody>
</table>
Simulations

Estimation of regression coefficients

Chi-square 5 Errors

- 
- 
- 
- 

\( \beta_0(\tau) \)

\( \beta_1(\tau) \)
Simulations

Relative MSE for $D = 200$

- Normal $b_i$

<table>
<thead>
<tr>
<th>$n_i$</th>
<th>$\tau$</th>
<th>$t_5$ Error</th>
<th>$\chi^2_5$ Error</th>
<th>$\chi^2_2$ Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>QALD</td>
<td>LIGPD</td>
<td>QALD</td>
</tr>
<tr>
<td>5</td>
<td>0.25</td>
<td>1.10</td>
<td>0.97</td>
<td>1.11</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
<td>0.90</td>
<td>0.88</td>
<td>1.12</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>1.14</td>
<td>1.00</td>
<td>1.16</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
<td>1.11</td>
<td>1.11</td>
<td>1.43</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>1.09</td>
<td>0.97</td>
<td>1.12</td>
</tr>
<tr>
<td>20</td>
<td>0.75</td>
<td>0.89</td>
<td>0.88</td>
<td>1.13</td>
</tr>
</tbody>
</table>
Simulations

- LIGPD bootstrap: Relative bias of bootstrap MSE estimator and empirical coverage of normal theory 95% confidence intervals
- $D = 40$

<table>
<thead>
<tr>
<th>$n_i$</th>
<th>Quartile</th>
<th>Rel. Bias (%)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$t_{(5)}$</td>
<td>$\chi^2_{(5)}$</td>
</tr>
<tr>
<td>5</td>
<td>25%</td>
<td>0.2</td>
<td>-9.0</td>
</tr>
<tr>
<td>20</td>
<td>25%</td>
<td>-6.3</td>
<td>-12.4</td>
</tr>
<tr>
<td>5</td>
<td>50%</td>
<td>-4.1</td>
<td>-9.8</td>
</tr>
<tr>
<td>20</td>
<td>50%</td>
<td>-7.7</td>
<td>-13.6</td>
</tr>
<tr>
<td>5</td>
<td>75%</td>
<td>-7.1</td>
<td>-9.8</td>
</tr>
<tr>
<td>20</td>
<td>75%</td>
<td>-9.1</td>
<td>-12.5</td>
</tr>
</tbody>
</table>
Simulations

- **Current bootstrap work**
  - Larger $N$: $(n_i, N_i) = (5, 800); (20, 2000)$
    - % of MSE due to parameter estimation greater for $(n_i, N_i) = (20, 2000)$

- **Double bootstrap**
  - For each bootstrap sample, generate 1 additional bootstrap sample
  - $\widehat{MSE}_2 = \max\{2\widehat{MSE}_1 - B^{-1} \sum_{b=1}^{B} (\bar{q}^{**}_{N_i} - \hat{q}^{**}_{N_i})^2\}$
    - $\widehat{MSE}_1 =$ single bootstrap MSE estimator
Simulations

- Relative bias of bootstrap MSE estimators (\%)
- \( N_i = (800, 2000) \)
- \( \chi^2_{(2)} \) errors
- \( D = 20 \)

<table>
<thead>
<tr>
<th>( n_i )</th>
<th>( \tau )</th>
<th>Single Bootstrap</th>
<th>Double Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.25</td>
<td>-7.4</td>
<td>1.4</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
<td>-29.0</td>
<td>-14.4</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>-9.8</td>
<td>-0.9</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
<td>-29.0</td>
<td>-14.4</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>-9.7</td>
<td>-0.7</td>
</tr>
<tr>
<td>20</td>
<td>0.75</td>
<td>-26.0</td>
<td>-11.0</td>
</tr>
</tbody>
</table>
Simulations

- Relative bias (%) % empirical coverage of bootstrap MSE estimators (%)
- $N_i = (800, 2000)$
- $\chi^2$ errors
- $D = 200$

<table>
<thead>
<tr>
<th>$n_i$</th>
<th>$\tau$</th>
<th>Bias (%)</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.25</td>
<td>-0.990</td>
<td>0.941</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
<td>-7.608</td>
<td>0.929</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>-5.931</td>
<td>0.931</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
<td>-5.361</td>
<td>0.935</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>-6.000</td>
<td>0.929</td>
</tr>
<tr>
<td>20</td>
<td>0.75</td>
<td>-9.931</td>
<td>0.925</td>
</tr>
</tbody>
</table>
Conclusions

- Conceptual: LIGPD addresses limitations of the QALD approach
  - Non-decreasing quantile function permits empirical Bayes prediction and bootstrap MSE estimation

- Empirical
  - Normal EB predictor robust to modest departures from normality, especially for median
  - LIGPD predictor is more efficient than NEB and QALD for the error distribution that is farthest from normal
  - Relative bias of bootstrap MSE estimator is typically less than 10% in absolute value
  - Empirical coverages $\approx 93\%$
  - Large $D$ improves efficiency of estimator and bootstrap variance estimator.
Current and Future Work

- Exploration of the LIGPD for a wider variety of distributions
  - Simulation configurations of Weidenhammer et al. (2016)
  - Non-normal $b_i$

- Improvements to estimators
  - Median regression instead of OLS as the basis for $V\{b_i\}$
  - Empirical Bayes for finite population parameters, instead of $b_i$
  - EM-type algorithm for parameter estimation

- Improvements to bootstrap MSE estimator
  - Double bootstrap
    - Exploit use of empirical Bayes predictor

- Apply to CEAP data
- Complex sample designs


Thank You

Acknowledgment: I thank Petrutza Caragea, Wayne A. Fuller, and Zhengyuan Zhu for helpful conversations.