Mixed Effects Quantile Regression Models for Small Area Estimation

Emily Berg
Iowa State University

Outline

- Motivation: Why consider mixed effects quantile regression for small area estimation?
- Mixed effects quantile regression
  - Previous approach: Asymmetric Laplace distribution
  - Proposed approach: Linearly interpolated generalized Pareto density
- Simulations
- Conclusions and next steps

Motivation: Why quantile regression (QR) for small area estimation?

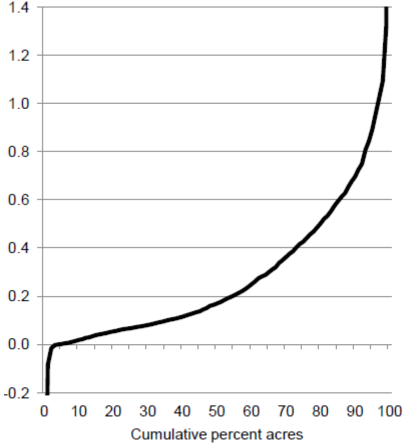
- Specification of fully parametric models for small area estimation can be difficult
  - Non-constant variance, outliers
  - Multiple response variables
  - Properties of distributions vary across wide range of conditions (i.e., states, counties, hydrologic units)
- Small area quantile is the parameter of interest
  - Poverty and income analysis (Whitworth, et al., undated)
  - Water quality monitoring (Pratesi, Ranalli, and Salvati, 2008)
  - Forestry (Chen and Liu, 2013)
- Small area estimation based on quantile regression
  - M-quantiles (Chambers and Tzavidis, 2006)
  - Asymmetric Laplace distribution (Weidenhammer et al., 2016)

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
  - 8-digit estimates desired
- Response variables: 16 measures of water and wind erosion
  - 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} = variable of interest for element j in area i
  - $\boldsymbol{x}_{ij}$ = covariate
    - $i = 1, \ldots, D$, $j = 1, \ldots, N_i$
  - τ^{th} quantile: $q_{ij}(\tau)$

$$P(y_{ij} \leq q_{ij}(\tau) \mid \boldsymbol{x}_{ij}; i) = \tau$$

    - Assume $q_{ij}(\tau)$ increasing, continuous
- Sample data
  - $y_{ij} : i = 1, \ldots, D;\ j = 1, \ldots, n_i$
  - $\boldsymbol{x}_{ij} : i = 1, \ldots, D;\ j = 1, \ldots, N_i$
- Objective: Use a mixed effects model for $q_{ij}(\tau)$ to predict small area parameters

Previous work: Asymmetric Laplace Distribution

Koenker's Check Function & ALD

- τ^{th} quantile $q(\tau)$

$$q(\tau) = argmin_a R(a)$$
$$R(a) = E[\rho_\tau(y - a)]$$
$$\rho_\tau(\nu) = \nu(\tau - I[\nu \le 0])$$

- $y \sim ALD(\mu_\tau, \sigma_\tau)$

$$f_Y(y \mid \mu_\tau, \sigma_\tau) \propto \sigma_\tau^{-1} h(y, \mu_\tau, \sigma_\tau)$$
$$h(y, \mu_\tau, \sigma_\tau) = \exp\{-\rho_\tau[\frac{(y - \mu_\tau)}{\sigma_\tau}]\}$$

MLE of μ_τ under ALD is $q(\tau)$

- Geraci & Bottai, 2007; 2014

$$y_{ij} \mid \alpha_i(\tau) \sim \text{ALD}(q_{ij}(\tau), \sigma^2(\tau)),$$
$$q_{ij}(\tau) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}(\tau) + \alpha_i(\tau), \alpha_i(\tau) \sim \text{N}(0, \sigma^2_\alpha(\tau))$$

  - Monte Carlo MLE for $\boldsymbol{\beta}(\tau)$, $\sigma^2(\tau)$, $\sigma^2_\alpha(\tau)$
  - Best (estimated) linear predictor of $\alpha_i(\tau)$
    - Means and covariances based on ALD conditional distribution
  - R package `lqmm`

Previous work: Asymmetric Laplace Distribution

- Application to SAE (Weidenhammer et al., 2016)
  - Predict $q_{ij}(\tau_k) : k = 1, \ldots, K$
  - Estimate small area parameters based on implied distribution function
  - Bootstrap MSE estimation
  - Extension to count data

Important Characteristic

- Model and predictor defined separately for each $\tau_k : k = 1, \ldots, K$

Implications for SAE

- Quantile functions may decrease
  - Empirical Bayes predictor undefined
  - Appropriate bootstrap distribution unclear

- Wiedenhammer et al. (2016) discusses essentially these issues

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	Random
$q_{F,ij}(\tau) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}(\tau)$	$b_i \sim f_b(b_i, \boldsymbol{\sigma}_b)$
$q_{F,ij}(\tau) < q_{F,ij}(\tau + \delta)$	$E[b_i] = 0$

- Jang and Wang (2015)
  - 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):

$$
\begin{aligned}
f(y \,|\, \boldsymbol{x}_{ij}, \boldsymbol{\beta}_K, b_i) = {} & I[y < q_{ij}(\tau_1)]\tau_1 f_{\ell ij}(y \,|\, \boldsymbol{\beta}_K, \rho_\ell, \xi_\ell) \\
& + I[y > q_{ij}(\tau_K)]\tau_K f_{uij}(y \,|\, \boldsymbol{\beta}_K, \rho_u, \xi_u) \\
& + \sum_{k=1}^{K-1} I[q_{ij}(\tau_k) \le y < q_{ij}(\tau_{k+1})] \frac{\tau_{k+1} - \tau_k}{q_{ij}(\tau_{k+1}) - q_{ij}(\tau_k)}
\end{aligned}
$$

$$
\boldsymbol{\beta}_K = (\boldsymbol{\beta}_1(\tau_1), \ldots, \boldsymbol{\beta}_K(\tau_K)), \tau_1 < \cdots < \tau_K
$$

- Motivation for central quantiles

$$
f(y \mid \boldsymbol{x}_{ij}, \boldsymbol{\beta}_K, b_i) = \lim_{\delta \to 0} \frac{\delta}{q_{ij}(\tau + \delta) - q_{ij}(\tau)}
$$

- Generalized Pareto distributions for lower and upper tails

$$
f_s(y \mid \rho_s, \xi_s) = \begin{cases} \rho_s^{-1}(1 + \xi_s y/\rho_s)^{-(1+1/\xi_s)}, \xi_s \neq 0 \\ \rho_s^{-1}\mathrm{exp}(-y/\rho_s), \xi_s = 0 \end{cases}
$$

Proposed approach: LIGPD for SAE

Bayes predictor

$$E[b_i \mid \boldsymbol{y}_i; \boldsymbol{\theta}] = \frac{\int_{-\infty}^{\infty} \prod_{j=1}^{n_i} b_i f(y_{ij} \,|\, \boldsymbol{x}_{ij}, \boldsymbol{\beta}_K, b_i) f_b(b_i \mid \boldsymbol{\sigma}_b) db_i}{\int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f(y_{ij} \,|\, \boldsymbol{x}_{ij}, \boldsymbol{\beta}_K, b_i) f_b(b_i \mid \boldsymbol{\sigma}_b) db_i}$$

$$\boldsymbol{\theta} = (\boldsymbol{\beta}_K', \boldsymbol{\sigma}_b')'$$

- Numerical approximation of the integral

Proposed approach: LIGPD for SAE

Estimation of $\boldsymbol{\theta}$ (simple)

1. Estimate $\boldsymbol{\sigma}_b = (V\{b_i\}, \boldsymbol{\gamma})$
   - OLS of y_{ij} on $\boldsymbol{x}_{ij}$ and *fixed* area indicators

     $$(\widehat{b}_1^{(0)}, \widehat{V}_1(\widehat{b}_1^{(0)})), \ldots, (\widehat{b}_D^{(0)}, \widehat{V}_D(\widehat{b}_D^{(0)}))$$

   - Area level small area (Wang and Fuller, 2003) for $\widehat{V}\{b_i\}$
   - MLE $\widehat{\boldsymbol{\gamma}} = \text{arg}_{max} \sum_{i=1}^{D} \log[f_b(\widehat{b}_i \mid \widehat{V}\{b_i\}, \boldsymbol{\gamma})]$
2. Estimate $\boldsymbol{\beta}(\boldsymbol{\tau}_k) : k = 1, \ldots, K$

   $$\widehat{\boldsymbol{\beta}}(\tau) = \text{arg}_{min} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_\tau(y_{ij} - \widehat{b}_i^{(0)} - \boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau))$$

   - Apply isotonic regression & linear interpolation to $\boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau_k)$ to ensure continuous, increasing quantile function, $\widehat{q}_{F,ij}(\tau)$
3. Match quantile function for ρ_s, & MLE for ξ_s $(s = \ell, u)$ using $y_{ij} < 0.5(\widehat{q}_{F,ij}(\tau_1) + \widehat{q}_{F,ij}(\tau_2)); y_{ij} > 0.5(\widehat{q}_{F,ij}(\tau_{K-1}) + \widehat{q}_{F,ij}(\tau_K))$ (Jang and Wang, 2015)

Proposed approach: LIGPD for SAE

- Empirical Bayes predictor of the quantile

$$\widehat{q}_{ij}(\tau) = \widehat{q}_{F,ij}(\tau) + E[b_i \mid \boldsymbol{y}_i, \widehat{\boldsymbol{\theta}}]$$
$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}'_K, \widehat{\boldsymbol{\sigma}}'_b)'$$

- Domain predictors
  - τ^{th} quantile:

$$\widehat{q}_{N_i}(\tau) = \tau^{\text{th}} \text{ empirical quantile of } \{\widehat{q}_{ij}(\tau_k) : k = 1, \ldots, K; j = 1, \ldots, N_i\}$$

  - Mean

$$\widehat{y}_{N_i} = \frac{1}{KN_i} \sum_{i=1}^{D} \sum_{j=1}^{N_i} \widehat{q}_{ij}(\tau_k)$$

Proposed approach: LIGPD for SAE

Bootstrap distribution

- Recall: quantile regression model

$$q_{ij}(\tau) = q_{F,ij}(\tau) + b_i$$
$$b_i \sim f_b(b_i \mid \boldsymbol{\sigma}_b)$$

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

$$q^*_{ij}(\tau) = \widehat{q}_{F,ij}(\tau) + b^*_i$$
$$b^*_i \sim f_b(b_i \mid \widehat{\boldsymbol{\sigma}}_b)$$

Proposed approach: LIGPD for SAE

Bootstrap procedure

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

$$y_{ij}^{*(b)} = \widehat{q}_{F,ij}(\tau^{(b)}) + b_i^{*(b)}, j = 1, \ldots, N_i$$
$$b_i^{*(b)} \sim f_b(b_i \mid \widehat{\boldsymbol{\sigma}}_b), \tau^{(b)} \sim \text{Unif}(0, 1)$$

  - 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 $\widehat{\theta}_{N_i}^{*(b)}$
  - Select sample & implement estimation procedure
- Mean squared error estimator: $B^{-1} \sum_{b=1}^{N} (\widehat{\theta}_{N_i}^{*(b)} - \theta_{N_i}^{*(b)})^2$

Simulations

$$
\begin{aligned}
y_{ij} &= -0.7 + 0.8x_{ij} + b_i + e_{ij} \\
b_i &\sim \mathsf{N}(0, 0.36) \\
(n_i, N_i) &= (5, 20) : i = 1, \ldots, 20 \\
(n_i, N_i) &= (20, 80) : i = 21, \ldots, 40 \\
x_{ij} &\sim \mathsf{N}(6, 6.25)
\end{aligned}
$$

- 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

Simulations

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

$$\text{Relative MSE} = \frac{\text{MC MSE}(\widehat{\theta})}{\text{MC MSE(NEB)}}$$

  - $\widehat{\theta} =$ QALD, LIGPD
  - MC MSE is average across areas of the same sample size

		$t_{(5)}$ Error		$\chi^2_{(5)}$ Error		$\chi^2_{(2)}$ Error	
n_i	Quartile	QALD	LIGPD	QALD	LIGPD	QALD	LIGPD
5	25%	1.14	1.15	1.14	1.15	1.14	0.94
20	25%	1.01	1.00	1.17	1.05	1.40	0.94
5	50%	1.14	1.16	1.14	1.08	1.16	0.94
20	50%	1.15	1.18	1.35	1.07	1.69	1.00
5	75%	1.10	1.13	1.13	1.06	1.15	0.99
20	75%	0.99	1.01	1.17	0.88	1.38	0.81

Simulations

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

		Rel. Bias (%)			Coverage (%)		
n_i	Quartile	$t_{(5)}$	$\chi^2_{(5)}$	$\chi^2_{(2)}$	$t_{(5)}$	$\chi^2_{(5)}$	$\chi^2_{(2)}$
5	25%	0.2	-9.0	0.2	93.3	92.6	93.7
20	25%	-6.3	-12.4	-6.3	92.9	92.6	93.2
5	50%	-4.1	-9.8	-4.1	93.2	92.8	93.2
20	50%	-7.7	-13.6	-7.7	92.7	92.4	93.0
5	75%	-7.1	-9.8	-7.1	93.0	92.8	93.3
20	75%	-9.1	-12.5	-9.1	92.6	92.6	92.9

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
  - LIGPD predictor is more efficient that 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 $\approx 93\%$

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
  - Exploit use of empirical Bayes predictor
- Apply to CEAP data
- Complex sample designs

References

- Chambers, R. and Tzavidis, N. (2006). "M-quantile models for small area estimation," *Biometrika*, **93**, 225–268.
- Chen, J. and Liu, Y. (2012). "Small Area Estimation under Density Ratio Model." In *JSM Proceedings*, Alexandria, VA: American Statistical Association. 5162–5173.
- Geraci, M. and M. Bottai (2007). "Quantile regression for longitudinal data using the asymmetric Laplace distribution." *Biostatistics*, **8**, 140154.
- Geraci, M. and M. Bottai (2014). "Linear quantile mixed models." *Statistics and Computing*, **24**, 461479.
- Jang, W. and Wang, J. (2015). "A Semiparameteric Bayesian Approach for Joint-Quantile Regression with Clustered Data," *Computational Statistics and Data Analysis*, **84**, 99–115.
- Pratesi, M., Ranalli, G., and Salvati, N. (2008). "Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US." *Environmetrics*, **19**, 687–701.
- Wang, J. and Fuller, W.A. (2003). "The Mean Squared Error of Small Area Predictors Constructed with Estimated Area Variances," *Journal of the American Statistical Association*, **98**, 716–723.
- Weidenhammer, B., Schmid, T., Salvati, N., and Tzavidis, N. (2016). "A Unit-Level Quantile Nested Error Regression Model for Domain Prediction with Continuous and Discrete Outcomes," Economics Discussion Paper, School of Business and Economics, Freie Universitat, Berlin.
- Whitworth, A., Martin, K., Tzavidis, N., Cruddas, M., Sexton, C., and Taylor, A. (Undated). "Small Area Estimtes of Income: Mean, Medians, and Percentiles."

Thank You

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