Estimation of Small Area Means under Semi-Parametric Measurement Error Model

Lily Wang

Department of Statistics
University of Georgia
Athens, GA 30602
email:lilywang@uga.edu

Joint with Gauri S. Datta, Peter Hall and Aurore Delaigle

Graybill 2013 Conference

June 9-12, 2013

Outline

Small Area Estimation

- Sample surveys are often designed to produce estimates at **higher** levels (Country, State or Districts).
- For micro-level planning, estimates needed at **smaller** levels.
- **Small area estimation** is generally used to estimate parameters for small sub-populations.
- **Small area:** a sub-population for which there is not enough sample to construct reliable estimates directly based on the survey sample
  - small geographical area (cities or counties)
  - small domain, demographic subgroups (race, sex)
  - area with small number of respondents or no respondent

Examples of Small Area Estimation

Some major small area estimation programs/surveys in U.S.

- **Small Area Income and Poverty Estimates** (Census Bureau): estimates of income and poverty measures for states, counties and school districts;
- **Local Area Unemployment Statistics** (Bureau of Labor Statistics): estimates of employment/unemployment for states, metropolitan areas, counties, subcounty areas;
- **County Estimates Program** (National Agricultural Statistics Service): county estimates of crop yield;
- **National Health and Nutrition Examination Survey** (National Center for Health Statistics): estimates for domains in the cross-classification of sex, race, ethnicity, age, etc.
- See more in Rao (2003).

Model-based Approach to Small Area Estimation

- Implicit models are used to develop synthetic estimates of small area characteristics (mean or proportion).
- In model-based approach, we use **indirect** estimates that "borrow strength" from related small areas.
- Model-based methods overcome problems faced with design-based methods by making assumptions that need to be tested.
- Quality of the model-based estimates depends on availability of **good auxiliary data** (administrative records or other surveys) to develop an adequate model.

Small Area Estimation Models

Two types of models for small area estimation:

- Unit level models
  - Relate the unit values of a study variable to unit-specific auxiliary data;
  - Sample values obey the assumed population model.
- Area level models
  - Relate area-specific direct survey estimates to area-specific auxiliary data;
  - Sample area values (the direct estimates) obey the assumed population model.

Area Level Model

- Suppose Y is the target variable and there are m small areas.
- Let θ_i be the unknown population mean of Y in area i.
- Let Y_i be the direct estimator of θ_i in area i.
- Let X_i be the auxiliary variables in area i.
- **Fay-Herriot Model**: a popular model for area level data

$$Y_i = \theta_i + \epsilon_i, \quad \theta_i = \beta_0 + \beta_1^{\mathrm{T}} X_i + V_i, \quad i = 1, ..., m$$

  - Model errors $V_i \sim N(0, \sigma_V^2)$ and σ_V^2 is unknown;
  - Sampling errors ϵ_is have zero mean and known variance τ_i;
  - V_i and ϵ_i are mutually independent.
- **Target:** estimation and prediction for θ_i.

Measurement Error in Covariates

Many survey data are contaminated by the mis-measured variables.

- **Prediction of Corn Yields for Counties**: Fuller (1987)
  - To measure the "nitrogen" level, we need to sample the soil of the plot and perform a lab analysis on the selected sample.
  - Due to sampling and the lab analysis, we do not observe the true available nitrogen level, but only its estimate.
  - Measurement error = true nitrogen− estimated nitrogen
- **Estimation of Mean BMI**: Ybarra and Lohr (2008)
  - In the National Heath Interview Survey, BMI is calculated using self-reported responses to the height and weight questions in the interview.
  - Measurement error occurs if a respondent did not report height or weight accurately.
  - Measurement error = true BMI−self reported BMI

Unit level Measurement Error Model

- **Functional Measurement Error model**

  Ghosh and Sinha (2007) and Datta, Rao and Torabi (2010) studied nested error linear regression model with area level covariates x_i measured with error, and true x_i are non-stochastic.

- **Structural Measurement Error model**

  Ghosh, Sinha, and Kim (2006) and Torabi, Datta and Rao (2009) studied nested error linear regression model when the true area level covariates x_i are stochastic and measured with errors.

Area Level Measurement Error Model

- Ybarra and Lohr (2008) extended Fay-Herriot model when covariate x_i is measured with error.
- **Goal:** Produce reliable estimates of mean BMI for 50 demographic categories from 2003-2004 NHANES.
  - Sample sizes in some domains are too small for the direct estimator to be sufficiently precise.
- **Auxiliary Data:** National Health Interview Survey (NHIS)
  - $\text{BMI} = \text{self reported height}/\text{self reported weight}^2$
  - Measurement error = true BMI$-$self reported BMI
- **Approach:** functional measurement error model by treating x_i as non-stochastic.

Structural Measurement Error Model

- Suppose one auxiliary variable X_i is subject to measurement error, and we observe a surrogate W_i

$$W_i = X_i + U_i,$$

where U_is have a common known distribution, $\mathrm{Var}(U_i) = \sigma_U^2$.

- Let Q_i be the remaining auxiliary p-vector measured exactly.
- **Fay-Herriot model with measurement errors**

$$Y_i = \theta_i + \epsilon_i, \quad \theta_i = \beta_0 + \beta_1 X_i + \beta_2^{\mathrm{T}} Q_i + V_i, \quad W_i = X_i + U_i$$

  - X_is have a common and unknown distribution (stochastic);
  - V_is have a common and unknown distribution with zero mean and unknown variance σ_V^2;
  - ϵ_is have zero mean and known variance τ_i;
  - Q_i, U_i, V_i, X_i and ϵ_i are completely independent.

Moment Corrected Estimators

Let $\beta = (\beta_0, \beta_1, \beta_2^{\mathrm{T}})^{\mathrm{T}}$, $\mathcal{X} = (X, Q^{\mathrm{T}})$ and $\mathcal{W} = (W, Q^{\mathrm{T}})$.

Following Buonaccorsi (2010), we have $\sqrt{m}$-consistent estimators

- $(\hat{\beta}_1, \hat{\beta}_2^{\mathrm{T}})^{\mathrm{T}} = \widehat{\Sigma}_{\mathcal{X}\mathcal{X}}^{-1}\widehat{\Sigma}_{\mathcal{X}Y}, \quad \hat{\beta}_0 = \bar{Y} - \bar{\mathcal{W}}(\hat{\beta}_1, \hat{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$
  - $\widehat{\Sigma}_{\mathcal{X}\mathcal{X}} = \begin{pmatrix} m^{-1}W^{\mathrm{T}}W - \sigma_U^2 & 0 \\ 0 & m^{-1}Q^{\mathrm{T}}Q \end{pmatrix}$
  - $\widehat{\Sigma}_{\mathcal{X}Y} = \widehat{\Sigma}_{\mathcal{W}Y} = m^{-1}\mathcal{W}^{\mathrm{T}}Y$
- $\hat{\sigma}_V^2 = \max\{0, \hat{\sigma}_Y^2 - \hat{\beta}_1^2\sigma_U^2 - \bar{\tau}\}$
  - $\hat{\sigma}_Y = (m - p - 1)^{-1}\sum_i\{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\beta}_2^{\mathrm{T}} Q_i)\}^2$
  - $\bar{\tau} = m^{-1}\sum_i \tau_i$

Weighted Moment Corrected Estimators

As described in Fuller (1987), we can also use a weighted version of the estimator, taking heteroscedacity into account.

- Let $\pi_i = 1/(\sigma_V^2 + \tau_i)$.
- We obtain the weighted estimator:

$$(\hat{\beta}_1, \hat{\beta}_2^{\mathrm{T}})^{\mathrm{T}} = \widehat{\Sigma}_{\mathcal{X}\mathcal{X},\pi}^{-1} \widehat{\Sigma}_{\mathcal{X}Y,\pi}.$$

- $\widehat{\Sigma}_{\mathcal{X}\mathcal{X},\pi} = \begin{pmatrix} m^{-1}\sum_i \pi_i (W_i^2 - \sigma_U^2) & 0 \\ 0 & m^{-1}\sum_i \pi_i Q_i Q_i^{\mathrm{T}} \end{pmatrix}$
- $\widehat{\Sigma}_{\mathcal{X}Y,\pi} = \widehat{\Sigma}_{\mathcal{W}Y,\pi} = m^{-1}\sum_i \pi_i \mathcal{W}_i Y_i$

Prediction Intervals

Assume (Y_i, W_i, Q_i) are available for all small areas.

- Let $F_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$ be the conditional C.D.F of θ_i given an observed value of (y,w,q) of (Y_i, W_i, Q_i).
- Let $t_\alpha(y,w,q)$ be the solution of

$$F_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q) = \alpha.$$

- The $100(1-\alpha)\%$ **prediction interval** for θ_i conditional on (y,w,q) is

$$[t_{\alpha/2}(y,w,q), t_{1-\alpha/2}(y,w,q)].$$

Conditional Density of θ_i

Conditional on (Y_i, W_i, Q_i), the conditional density function of θ_i is

$$f_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q) = \int f_{\theta_i|Y_i,X_i,Q_i}(t|y,x,q) f_{X_i|Y_i,W_i,Q_i}(x|y,w,q)dx,$$

$$f_{\theta_i|Y_i,X_i,Q_i}(t|y,x,q) = \frac{f_{\epsilon_i}(t-y) f_V(t-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q)}{f_{V_i+\epsilon_i}(y-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q)},$$

$$f_{X_i|Y_i,W_i,Q_i}(x|y,w,q) = \frac{f_{V_i+\epsilon_i}(y-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q) f_U(w-x) f_X(x)}{\int f_{V_i+\epsilon_i}(y-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q) f_U(w-x) f_X(x)dx}.$$

Thus, we have

$$f_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q) = \frac{f_{\epsilon_i}(t-y)\int f_V(t-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q) f_U(w-x) f_X(x)dx}{\int f_{V_i+\epsilon_i}(y-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q) f_U(w-x) f_X(x)dx}.$$

Conditional C.D.F of θ_i

Conditional on (Y_i, W_i, Q_i), the C.D.F of θ_i is

$$F_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q) = \int_{z=-\infty}^{t} f_{\theta_i|Y_i,W_i,Q_i}(z|y,w,q)dz,$$

where

$$f_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$$
$$= \frac{f_{\epsilon_i}(t-y)\int f_V(t-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}}q)f_U(w-x)f_X(x)dx}{\int f_{V_i+\epsilon_i}(y-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}}q)f_U(w-x)f_X(x)dx}.$$

Estimation of Conditional C.D.F of θ_i

We can estimate $F_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$ by

$$\hat{F}_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q) = \int_{z=-\infty}^{t} \hat{f}_{\theta_i|Y_i,W_i,Q_i}(z|y,w,q)dz,$$

where

$$\hat{f}_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$$
$$= \frac{f_{\epsilon_i}(t-y)\int \hat{f}_V(t-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}}q)f_U(w-x)\hat{f}_X(x)dx}{\int \hat{f}_{V_i} * f_{\epsilon_i}(y-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}}q)f_U(w-x)\hat{f}_X(x)dx}.$$

is the plug-in type of estimator of $f_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$.

In the Normal Case

- If ϵ_i and V_i are normally distributed and mutually independent, we have $f_{V_i+\epsilon_i}(x) = \phi(x; \sigma_V^2 + \tau_i)$.
- We can estimate $f_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$ by

$$\hat{f}_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$$

$$= \frac{\phi(t-y;\tau_i)\int \phi(t-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q;\hat{\sigma}_V^2+\tau_i) f_U(w-x)\hat{f}_X(x)dx}{\int \phi(y-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q;\hat{\sigma}_V^2+\tau_i) f_U(w-x)\hat{f}_X(x)dx}$$

- We can estimate $F_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$ by

$$\hat{F}_{\theta_i|Y_i,W_i,Q_i}(t|y,w,q)$$

$$= \frac{\int_{z=-\infty}^{t} \phi(z-y;\tau_i)\int \phi(z-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q;\hat{\sigma}_V^2+\tau_i) f_U(w-x)\hat{f}_X(x)dxdz}{\int \phi(y-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q;\hat{\sigma}_V^2+\tau_i) f_U(w-x)\hat{f}_X(x)dx}$$

Prediction Interval for θ_i

- Let $\hat{t}_\alpha(y, w, q)$ be the solution of

$$\hat{F}_{\theta_i|Y_i,W_i,Q_i}(t|y, w, q) = \alpha.$$

- The $100(1-\alpha)\%$ prediction interval for θ_i conditional on (y,w,q) is

$$\left[\hat{t}_{\alpha/2}(y, w, q), \hat{t}_{1-\alpha/2}(y, w, q)\right].$$

If Y_is are not available

- For small areas without any direct estimator, note that

$$f_{\theta|W,Q}(t|w,q) = \frac{\int f_V(t-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q) f_U(w-x) f_X(x) dx}{f_W(w)}$$

$$F_{\theta|W,Q}(t|w,q) = \frac{\int F_V(t-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q) f_U(w-x) f_X(x) dx}{f_W(w)}$$

- Assuming $V_i \sim N(0,\sigma_V^2)$, we estimate $F_{\theta|W,Q}(t|w,q)$ by

$$\hat{F}_{\theta|W,Q}(t|w,q)$$

$$= \frac{\int \phi(t-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q;\hat{\sigma}_V^2) f_U(w-x) \hat{f}_X(x) dx}{\hat{f}_W(w)}$$

Prediction Interval

- The estimated conditional C.D.F of θ given (W, Q) is

$$\hat{F}_{\theta|W,Q}(t|w,q) = \frac{\int \phi(t - \hat{\beta}_0 - \hat{\beta}_1 x - \hat{\beta}_2^{\mathrm{T}} q; \hat{\sigma}_V^2) f_U(w-x) \hat{f}_X(x) dx}{\hat{f}_W(w)}.$$

- Let $\hat{t}_\alpha(w,q)$ be the solution of

$$\hat{F}_{\theta|W,Q}(t|w,q) = \alpha.$$

- The $100(1-\alpha)\%$ prediction interval given (w,q) is

$$\left[\hat{t}_{\alpha/2}(w,q), \hat{t}_{1-\alpha/2}(w,q)\right].$$

Estimators of f_X and f_W

Let K and L be two different kernel functions.

- Deconvolution kernel estimator of f_x: Carroll and Hall (1988)

$$\hat{f}_X(x) = \frac{1}{nh_1} \sum_{i=1}^{n} K_U \left(\frac{x - W_i}{h_1} \right),$$

$$K_U(u; h) = \frac{1}{2\pi} \int \exp(-itu) K^{\mathrm{Ft}}(t) / f_U^{\mathrm{Ft}}(t/h) dt,$$

where K^{Ft} and f_U^{Ft} are the Fourier transform of K and f_U.

- Density estimator of f_w

$$\hat{f}_W(w) = \frac{1}{nh_2} \sum_{i=1}^{n} L \left(\frac{w - W_i}{h_2} \right).$$

Bandwidth Selection

- Bandwidth selection in deconvolution kernel problems
  - Cross-validation: Stefanski and Carroll (1990)
  - Bootstrap: Delaigle and Gijbels (2001)
  - Plug-in type: Delaigle and Gijbels (2004)
- To obtain the claimed $\sqrt{m}$-consistency, the bandwidths h_1 and h_2 need to be smaller than the usual bandwidths $m^{-1/5}$.
- For h_1, we used the plug-in bandwidth in Delaigle and Gijbels (2004) to get initial bandwidth. For h_2, we used the quartic (biweight) kernel with the rule of thumb bandwidth as our initial choice. We then adjust them by a factor $m^{-1/20}$.

Characteristic Function of V

- Let C_X be the characteristic function of a random variable X
- Denote $C_Y(t) = m^{-1}\sum_{i=1}^{m} C_{Y_i}(t)$, $C_\epsilon(t) = m^{-1}\sum_{i=1}^{m} C_{\epsilon_i}(t)$.
- It can be shown that

$$C_V(t) = \frac{exp(-it\beta_0)C_Y(t)}{C_X(\beta_1 t)C_{\beta_2^{\mathrm{T}}Q}(t)C_\epsilon(t)}.$$

  - $C_{\beta_2^{\mathrm{T}}Q}(t)$ can be estimated by

$$m^{-1}\sum_{j=1}^{m} exp(it\hat{\beta}_2^{\mathrm{T}}Q_j).$$

  - $C_X(\beta_1 t) = C_W(\beta_1 t)/C_U(\beta_1 t)$ can be estimated by

$$m^{-1}\sum_{j=1}^{m} exp(it\hat{\beta}_1 W_j)/C_U(\hat{\beta}_1 t).$$

Estimating f_V

Putting all together we deduce the following nonparametric estimator of f_V:

$$\hat{f}_V(v) = \frac{1}{2\pi} \int \frac{\exp(-it(v + \hat{\beta}_0)) C_Y(t) C_U(t\hat{\beta}_1)}{C_W(\hat{\beta}_1 t) C_{\hat{\beta}_2^{\mathrm{T}} Q}(t) C_\epsilon(t)} K^{\mathrm{Ft}}(t; h) dt$$

Simulation

- We generated errors $\tau_i \sim Gamma(shape = 4,\ scale = 2)$.
- We performed 500 replications for $n = 30,\ 50$ and 100 small areas. For each replication, we generated data from the following:

$$Y_i = \theta_i + \epsilon_i,\ \theta_i = 5 + 3X_i + 2Q_i + V_i,\ W_i = X_i + U_i,$$

where

- $X_i \sim N(5, 3^2)$ and $Q_i \sim Unif(0, 5)$;
- $V_i \sim N(0, 2^2)$ and $\epsilon_i \sim N(0, \tau_i)$;
- $U_i \sim N(0, (\sqrt{3}/2)^2)$ or
  $U_i \sim$ Double Exponential$(location = 0, scale = \sqrt{3/8})$.

Table : Empirical coverage rates of the prediction intervals and the root mean squared errors of the estimators.

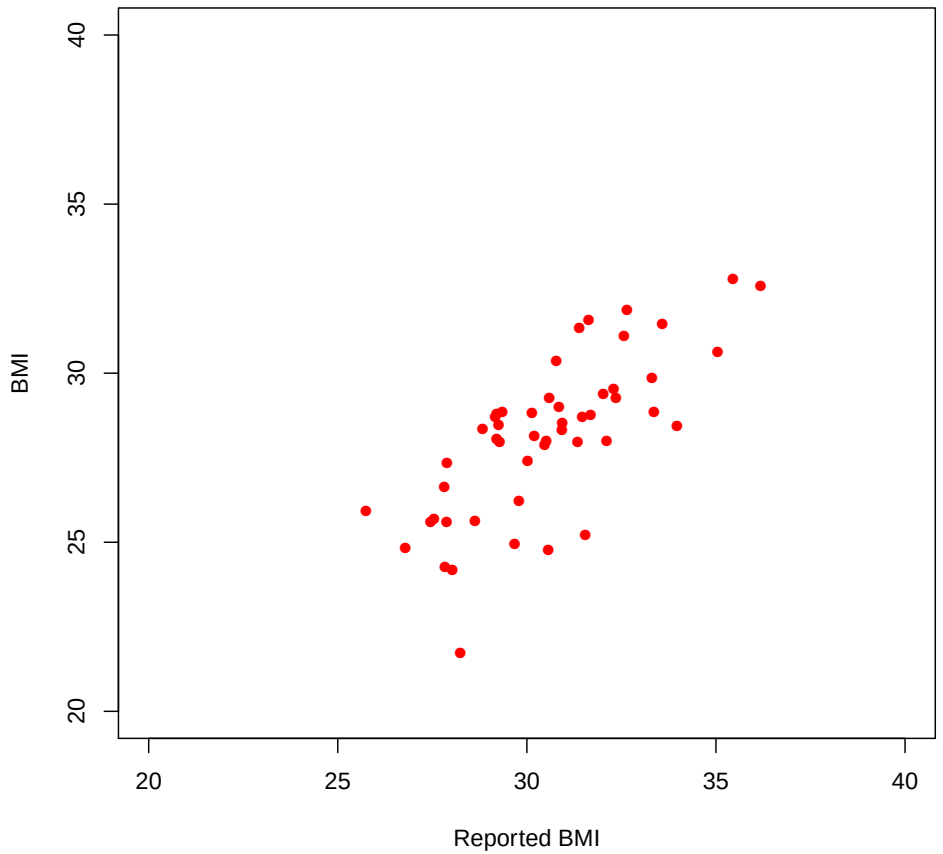
PDF of U	m	$1-\alpha$			$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	$\lvert\hat{\sigma}_V^2 - \sigma_V^2\rvert$
		0.99	0.95	0.90		
Normal	30	0.9861	0.9437	0.8906	0.768	3.556
	50	0.9852	0.9453	0.8974	0.600	2.975
	100	0.9863	0.9449	0.8993	0.277	2.019
Double Exponential	30	0.9958	0.9763	0.9443	0.921	3.846
	50	0.9949	0.9751	0.9415	0.827	2.938
	100	0.9931	0.9686	0.9310	0.386	2.337
	200	0.9875	0.9488	0.9084	0.119	1.583

Application: Prediction Interval of Mean Body Mass Index

- We consider an application to the survey data from the 2003-2004 NHANES.
- We take the 2004 National Health Interview Survey (NHIS) as the auxiliary information.
- Small areas: 50 demographic subgroups classified by race, ethnicity, age and sex.

	2003-2004 NHANES	2004 NHIS
Total sample size	4,424	29,652
Domain sample Size	8–433	95–3,087
BMI	$\frac{\text{measured height}}{\text{measured weight}^2}$	$\frac{\text{reported height}}{\text{reported weight}^2}$

- Measurement Error =reported BMI − true BMI.



Mean BMI from NHANES v.s. mean of reported BMI from NHIS.

Model

- Let θ_i be the unknown population mean BMI for the ith subgroup.
- Let Y_i be the direct estimator of the mean BMI for the ith subgroup from NHANES.
- Let X_i be the mean BMI for the ith subgroup measured exactly.
- Let W_i be the mean of reported BMI for the ith subgroup from NHIS.
- We consider the following model:

$$Y_i = \theta_i + \epsilon_i, \quad \theta_i = \beta_0 + \beta_1 X_i + V_i, \quad W_i = X_i + U_i$$

Prediction Intervals for Mean BMI

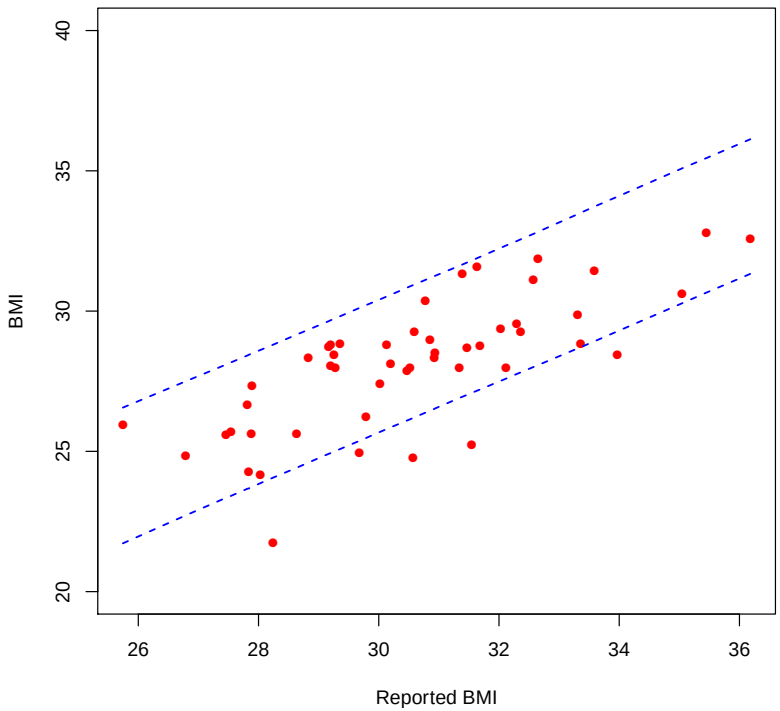
- We consider the structure measurement error model and use the reported BMI from NHIS as the auxiliary information to construct prediction intervals for the mean BMI.
- The estimated parameters in Ybarra and Lohr (2008):

$$\hat{\beta}_0 = 0.13,\ \hat{\beta}_1 = 0.93,\ \hat{\sigma}_V = 0.58$$

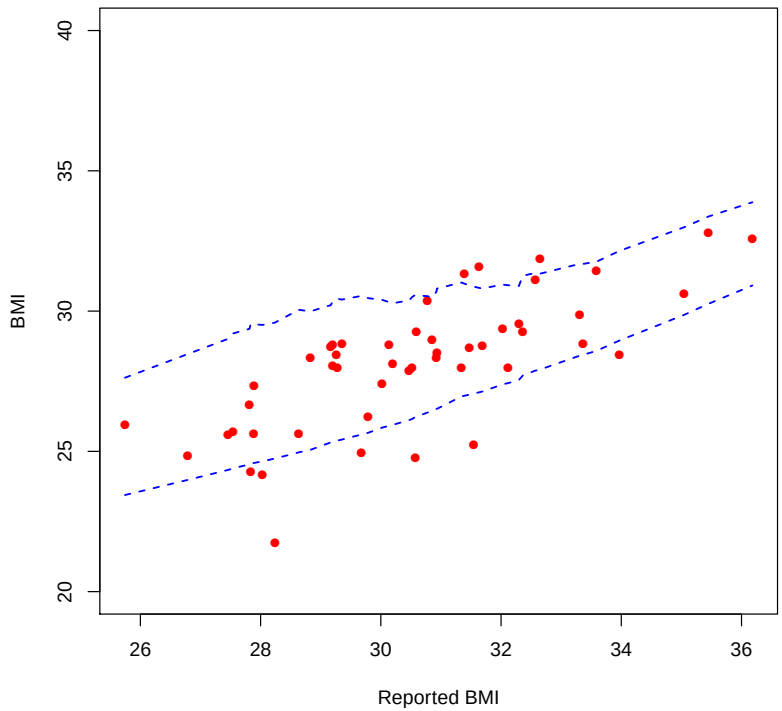
  are used to construct the prediction intervals.
- We apply two methods to the data
  - Method 1: impose the normality assumption on V_i
  - Method 2: don't impose the normality assumption on V_i

95% Prediction Intervals

(a) Method 1



(b) Method 2



Summary

- We considered estimation and prediction of small area means in presence of measurement errors in small area estimation.
- We studied structural measurement error Fay-Herriot model in the situation that auxiliary variables are available for use but some of them are measured with error.
- We proposed a nonparametric method to construct prediction intervals for small area means by using quantiles of the conditional C.D.F.
- Our method allows heavy-tailed distributions, which ensures robustness to a departure from normality.
- The preliminary simulation results confirm the superior behavior of proposed prediction intervals.

Thank You!

Email: lilywang@uga.edu

References

- Buonaccorsi, J. P. (2004). *Measurement Error: Models, Methods and Applications*. Chapman & Hall.
- Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83, 1184–1186.
- Fuller, W.A. (1987). *Measurement Error Models.* New York, NY: Wiley.
- Ghosh, M. and Sinha, K. (2007). Empirical Bayes estimation in finite population sampling under functional measurement error models. *J. Statist. Plann. Inf.*, 137, 2759–2773.
- Ghosh, M., Sinha, K. and Kim, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models. *Scand. J. Statist.*, 33, 591–608.
- Rao, J.N.K. (2003). Small Area Estimation. Hoboken, New Jersey: John Wiley & Sons.
- Torabi, M., Datta, G.S. and Rao, J.N.K. (2009). Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scand. J. Statist.*, 36, 355–368.
- Ybarra, L.M.R. and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95, 919–931.