Parametric Fractional Imputation for a Model with Error in a Covariate

Emily Berg and Jae-kwang Kim

Center for Survey Statistics and Methodology
Iowa State University

June 3, 2013

Outline

- Background
  - Models with measurement error in a covariate
  - Applications where measurement error is a concern
- Inference for a measurement error model with Parametric Fractional Impuation
- Partial measurement error, audit sample
  - National Resources Inventory link
- Simulations
- Discussion, future work

Framework

- Objective: inference for θ in $f(y|x,w;\theta)$
  - Linear:

$$y = \beta_0 + \beta_1 x + \beta_2' w + e, \quad e \sim (0, \sigma_e^2)$$

  - Exponential family:

$$\begin{aligned} f(y|x,w) &= \exp\left[\phi^{-1}(y - b(\theta)) + c(y,\phi)\right] \\ b'(\theta) &= \mu \\ g(\mu) &= \beta_0 + \beta_1 x + \beta_2' w \end{aligned}$$

- x is difficult or expensive to measure accurately
- Observe $z = x + \delta$

Examples of Measurement Error in Covariates

- Continuous response - (Fuller, 1987)
  - y = corn yield
  - x = available soil nitrogen at 11 plots on Marshall soil in Iowa
  - z = measurement of soil nitrogen – error due to subsampling and chemical analysis
- Binary response - NHANES-I (Jones et al., 1987)
  - y = presence or absence of breast cancer
  - w = age, poverty index, BMI, alcohol consumption indicator, family history of breast cancer
  - x = long-term saturated fat intake (and other similar measures of long-term average nutrition)
  - z = dietary intake from 24-hour recall – reporting error, specification error

Implications of Measurement Error

- Standard estimators of θ based on (y, z, w) instead of (y, x, w) may be biased
- Linear example ($i = 1, \ldots, n$)
  - Subject-matter model: $y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \sim (0, \sigma_e^2)$
  - Measurement error model:
    $z_i = x_i + u_i, \quad (x_i, u_i)' \sim [(\mu_x, 0)', \text{diag}(\sigma_x^2, \sigma_u^2)]$
  - OLS estimator of β_1 constructed with (z_i, y_i) biased

$$E[\hat{\beta}_{1,ols,zy}] = \kappa \beta_1, \quad \kappa = (\sigma_x^2 + \sigma_u^2)^{-1} \sigma_x^2$$

$$\hat{\beta}_{1,ols,zy} = \left[\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right]^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n) y_i$$

Solution: Extra Information, Assumptions

- External Calibration

$$\begin{aligned}(z_i, y_i) &\text{ for } i \in \text{ main sample} \\ (z_i, x_i) &\text{ for } i \in \text{ calibration sample}\end{aligned}$$

- Internal Calibration

$$\begin{aligned}(z_i, y_i) &\text{ for } i \in \text{ main sample} \\ (z_i, y_i, x_i) &\text{ for } i \in \text{ subsample of main sample}\end{aligned}$$

- Instrumental variable assumption
  1. $g(z_i | x_i = a) \neq g(z_i | x_i = b)$ for some $a \neq b$
  2. $f(y_i | x_i; \theta)$ does not depend on z_i
- Assumption 2 sometimes called non-differential measurement error
  - Necessary for identifiability in external calibration, not for internal calibration

Inference for a Measurement Error Model with Parametric Fractional Imputation

- Formalization of measurement error model
- Subject-matter model: $y_i \sim f(y_i | x_i; \theta)$
  - θ parameter of interest
- Measurement error model: $z_i \sim g(z_i | x_i; \alpha_1)$, $x_i \sim h(x_i; \alpha_2)$
  - α_1, α_2 nuisance parameters
- Calibration data structures
  - External calibration: $\{(x_i, z_i) : i \in A\}$, $\{(z_i, y_i) : i \in B\}$, $B \cap A = \phi$, sampling weights w_{iA}, w_{iB}
  - Internal calibration: $\{(x_i, z_i, y_i) : i \in A\}$, $\{(z_i, y_i) : i \in B\}$, $A \subset B$, sampling weights w_{iB}

Parametric Fractional Imputation (PFI)

- Kim (2011)
- Complete data estimating equation

$$U_{com}(\theta) = \sum_{i \in B} w_{iB} U_i(\theta; y_i, x_i)$$

- Some x_i are unobserved
- Observed estimating equation

$$U_{obs}(\theta) = \sum_{i \in B} w_{iB} E[U_i(\theta; y_i, x_i) | D_{i,obs}]$$

- Solve $U_{obs}(\theta) = 0$ by Parametric Fractional Imputation
  - Treat unobserved x_i as missing and impute

Parametric Fractional Imputation (PFI)

- EM algorithm by PFI

1. Impute by generating

$$x_i^{*(1)}, \ldots, x_i^{*(m)} \text{ from a proposal } h(x_i)$$

2. Given initial $\hat{\theta}^{(0)}$, iterate $(t = 0, 1, 2, \ldots)$
   (a.) Importance weight

$$w_{ij}(\hat{\theta}^{(t)}) \propto f(x_i^{*(j)} \mid D_{i,obs}; \hat{\theta}^{(t)}) / h(x_i^{(j)}) w_{iB}$$

$$D_{i,obs} = \text{ observed data for unit } i$$

   (b.) Update estimator of θ by solving

$$\sum_{i \in B} \sum_{j=1}^{m} w_{ij}(\hat{\theta}^{(t)}) U_i(\theta; y_i, x_i^{*(j)}) = 0$$

PFI for a Measurement Error Model

- Estimate $\alpha = (\alpha_1', \alpha_2')'$ from sample A.
- Estimating equation for θ

$$U(\theta \mid \hat{\alpha}) = \sum_{i \in B} w_{iB} E[S(\theta; y_i, x_i) | D_{i,obs}, \hat{\alpha}, \theta]$$

$$S(\theta; y_i, x_i) = \frac{\partial}{\partial \theta} \log[f(y_i | x_i; \theta)]$$

- $D_{i,obs}$ = observed data
  - Internal calibration: $D_{i,obs} = (y_i, x_i) : i \in A$ and $D_{i,obs} = (y_i, z_i) : i \in B \cap \bar{A}$
  - External calibration: $D_{i,obs} = (y_i, z_i)$
- Conditional distribution of unobserved given observed

$$f(x|y,z) \propto f(y|x;\theta) g(z|x;\alpha_1) h(x|\alpha_2)$$

PFI for a Measurement Error Model

- Consider external calibration

1. For $j = 1, \ldots, m$, $i \in B$, generate $x_i^{*(j)} \sim h(x; \hat{\alpha}_2)$
2. $\hat{\theta}^{(0)}$ initial estimate of θ
3. For $t = 0, 1, 2, \ldots,$ update the estimator of θ by solving,

$$0 = \sum_{i \in B} \sum_{j=1}^{m} w_{ij}^{*}(\hat{\theta}^{(t)}) S(\theta; x_i^{*(j)}, y_i)$$

$$w_{ij}^{*}(\hat{\theta}^{(t)}) \propto f(y_i \mid x_i^{*(j)}; \hat{\theta}^{(t)}) g(z_i \mid x_i^{*(j)}; \hat{\alpha}_1) w_{iB}$$

PFI for a Measurement Error Model

- Alternative distributions for imputation

$$x_i^{*(j)} \sim h(x_i | z_i) \rightarrow w_{ij}^*(\theta) \propto f(y | x; \theta) w_{iB}$$
$$x_i^{*(j)} \sim h(x_i | y_i, z_i) \rightarrow w_{ij}^{*(\theta)} = w_{iB}$$

- For the measurement error application $h(x_i; \hat{\alpha}_2)$ is convenient if $h(x_i | z_i)$ or $h(x_i | y_i, z_i)$ are intractable

EM Algorithm with Parametric Fractional Imputation

- Modification for internal calibration

  For $i \in A$, (y_i, x_i) observed; no need for conditional expectation

  Operationally, $w_{ij} = m^{-1} w_i$, $x_i^{*(j)} = x_i$ $(j = 1, \ldots, m; i \in A)$
- Hot-deck version

  For each $i \in B$, "imputed" values $x_i^{*(j)}$ are the n_A observed values from the calibration sample.

$$(x_i^{*(1)}, \ldots, x_i^{*(n_A)})' = (x_1, \ldots, x_{n_A})'$$

Variance Estimation and Tests

- Taylor Expansion

$$
\begin{aligned}
0 = U(\hat{\theta} \mid \hat{\alpha}) &\approx U(\theta \mid \alpha) + D_1(\hat{\alpha} - \alpha) + D_2(\hat{\theta} - \theta) \\
(D_1, D_2) &= E\left[\partial/\partial\alpha\, U(\theta \mid \alpha), \partial/\partial\theta\, U(\theta \mid \alpha)\right] \\
\hat{V}\{\hat{\theta}\} &= (\hat{D}_1^{-1})\left[\hat{V}\{U(\theta \mid \alpha)\} + \hat{D}_2 \hat{V}(\hat{\alpha})\hat{D}_2'\right](\hat{D}_1^{-1})'
\end{aligned}
$$

- Score test
  - $\theta = (\theta_1', \theta_2')'$, null hypothesis: $\theta_2 = \theta_{2,0}$
  - PFI to estimate θ_1 subject to null hypothesis
  - Test statistic based on Taylor expansion (Rao et al., 1997)
  - Computationally simpler because estimation for full θ not required

Partial Measurement Error – NRI Connection

- National Resources Inventory (NRI) - longitudinal survey, non-federal US land
  - Change in land cover and land use over time
    - Land cover/use (crop, urban, wetland), soil characteristis (slope, erodibility), measurements of erosion
  - Aerial photographs of sampled PSUs (segments, 160 acres), three points per segment (roughly)
  - Record-level data set with characteristics of sampled points from 1982,1987,1992,1997,2000-2010
- Sources of measurement error
  - Difficulty interpreting photographs of NRI segments
  - Misinterpretation of protocols
  - Errors in computer algorithms that convert collected data to measurements of erosion

Measurement Error in NRI

- Further investigation often identifies and corrects errors
  - Enhanced imagery such as Google maps
  - Subject-matter expertise
- Impractical to double-check every NRI point
- Suggests internal calibration
  - Initial sample – collected data measured with error
  - Select a subsample – check data for subsample
  - Some responses contaminated with measurement error, not all
- Connection with error in covariates
  - A response (Y) with respect to NRI estimation may be a covariate in a different context.

Partial Measurement Error, Internal Calibration

- Subject-matter model: $y_i \sim f(y_i | x_i; \theta)$
- Measurement error model

$$z_i = (1-\delta_i)x_i + \delta_i z_i^*, \quad z_i^* \sim g(z_i^* | x_i, \alpha_2), \quad x_i \sim h(x_i; \alpha_2)$$
$$\delta_i \sim \text{Bernoulli}(p_i), \quad \text{logit}(p_i) = \phi_0 + \phi_1 y_i$$

- Data structure
  - $(x_i, z_i, y_i, \delta_i) : i \in A$, $(z_i, y_i) : i \in B$
- Estimation and inference with PFI extension of methods for internal calibration

Simulation Model 1: Continuous Response

- Model from Guo and Little (2011)

$$
\begin{aligned}
y_i &= \gamma_0 + \gamma_x x_i + e_i, \quad e_i \sim \mathrm{N}(0, \tau^2) \\
z_i &= \beta_0 + \beta_1 x_i + u_i, \quad u_i \sim \mathrm{N}(0, \sigma^2 |x_i|^{2\eta}) \\
x_i &\sim \mathrm{N}(\mu_x, \sigma_x^2)
\end{aligned}
$$

- $\boldsymbol{\theta} = (\gamma_0, \gamma_x, \tau^2) = (0, 1, 1)$
- $\boldsymbol{\alpha} = (\beta_0, \beta_1, \sigma^2, \eta, \mu_x, \sigma_x^2) = (0, 0.5, 0.25, 0.4, 0, 1)$
  - $(\hat{\mu}_x, \hat{\sigma}_x^2) = (\bar{x}_{n,calib}, S^2_{x,calib})$
  - MLE for $(\beta_0, \beta_1, \sigma^2, \eta)$ based on calibration data
- Calibration data structures
  - External calibration:
    $A = \{(x_i, z_i) : i = 1, \ldots 400\}$, $B = \{(y_i, z_i) : i = 1, \ldots, 1600\}$
  - Internal calibration:
    $A = \{(y_i, z_i, x_i) : i = 1, \ldots, 400\}$, $B = \{(y_i, z_i) : i = 1, \ldots, 1600\}$

Simulation Model 1: Continuous Response

- Internal calibration with partial measurement error

$$
\begin{aligned}
x_i &\sim \mathrm{N}(\mu_x, \sigma_x^2) \\
y_i &= \gamma_0 + \gamma_x x_i + e_i, \quad e_i \sim \mathrm{N}(0, \tau^2) \\
\delta_i &\sim \mathrm{Binary}(p_i) \\
p_i &= \frac{\exp(\phi_0 + \phi_1 y_i)}{1 + \exp(\phi_0 + \phi_1 y_i)} \\
z_i &= (1 - \delta_i) x_i + \delta_i(\beta_0 + \beta_1 x_i + u_i), \quad u_i \sim \mathrm{N}(0, \sigma^2 |x_i|^{2\eta}).
\end{aligned}
$$

- $\theta = (\gamma_0, \gamma_x, \tau^2) = (0, 1, 1)$
- $\alpha = (\beta_0, \beta_1, \sigma^2, \eta, \mu_x, \sigma_x^2, \phi_0, \phi_1) = (0, 0.5, 0.25, 0.4, 0, 1)$
- Calibration data structure
  - $A = \{(x_i, \delta_i, z_i, y_i) : i = 1, \ldots, 400\}$
  - $B = \{(y_i, z_i) : i = 1, \ldots, 1600\}$

Results: Inference for γ_x

Parameter	$E_{MC}[\hat{\gamma}_x]$	$100E_{MC}[\hat{V}(\hat{\gamma}_x)]$	$100V_{MC}(\hat{\gamma}_x)$
External	0.98	0.23	0.24
Internal	1.00	0.13	0.12
Partial ME	1.00	0.08	0.08

- Variance decreases as acquire more information.
- Empirical p-value of score test of $H_0 : (\gamma_0, \gamma_x) = (0, 1)$ is 0.06 for internal and external calibration.

Simulation Model 2: Binary Response

- Many NRI variables are categorical: type of land cover (i.e., crop, pasture, urban, wetland...)
- Consider a binary response

$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(p_i), \quad && \text{logit}(p_i) = \gamma_0 + \gamma_x x_i \\
z_i &= \beta_0 + \beta_1 x_i + u_i, \quad && u_i \sim \text{N}(0, \sigma^2 |x_i|^{2\eta}) \\
x_i &\sim \text{N}(\mu_x, \sigma_x^2)
\end{aligned}
$$

- $\theta = (\gamma_0, \gamma_x) = (0, 1)$, $\alpha = (\beta_0, \beta_1, \sigma^2, \eta, \mu_x, \sigma_x^2)$
- External calibration
  - $A = \{(x_i, z_i) : i = 1, \ldots, 800\}$, $B = \{(z_i, y_i) : i = 1, \ldots, 800\}$
- Estimators: Naive, PFI, HDFI
  - Naive: logistic regression of y_i on z_i for sample B

Results: Binary Response

- MC bias, variance, and MSE of three estimators of γ_x
- True $\gamma_x = 1$

	MC Bias	MC Variance	MC MSE
Naive	-0.2241	0.0239	0.0742
PFI	0.0239	0.0386	0.0392
HDFI	0.0246	0.0387	0.0393

Results: Binary Response

	$E_{MC}[\hat{V}(\hat{\gamma}_x)]$	$V_{MC}(\hat{\gamma}_x)$	R.bias	Wald	Score
PFI	0.0386	0.0382	-0.0096	0.051	0.055
HDFI	0.0387	0.0383	-0.0093	0.061	0.062

- MC mean of estimators of the variance of $\hat{\gamma}_x$
- MC variance of estimators of γ_x
- Relative bias (R.bias) of the variance estimators
  - Ratio of MC bias of variance estimator to MC variance $\hat{\gamma}_x$
- Empirical coverages of tests of $H_0 : \gamma_x = 1$ with nominal coverage of 0.05.

Discussion, Future Work

- Parametric fractional imputation for a measurement error model
  - Compuationally simple
  - Straightforward for complex samples
- Comparison with other methods
  - Bayes (multiple imputation), regression calibration
- NRI Applications
  - Address a specific area of measurement error in the NRI
    - Consider error in response
    - Attempt to reduce variance of estimators of means, improve estimators of quantiles, improve unit-level data

Thank You

- References

- Fuller (1987). *Measurement Error Models* New York: Wiley.
- Guo, Y. and Little, R.J. (2011). Regression Analysis Involving Covariates with Heteroscedastic Measurement Error. *Statistics in Medicine*, 30, 18, 2278–2294.
- Jones, D.Y., Schatzkin, A., Green, S.B., Block, G., Brinton, L.A., Ziegler, R.G,, Hoover, R. and Taylor, P.R. (1987), Dietary fat and breast cancer in NHANES-I: Epidemiologic follow-up study. *Journal of the National cancer Institute*, 79, 465–471.
- Kim (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119–132.
- Rao, J.N.K, Scott, A.J., and Skinner, C.J. (1998). Quasi-Score Tests with Survey Data. *Statistica Sinica*. 1059–1070.