An imputation approach for analyzing mixed-mode surveys

Jae-kwang Kim [1]

Iowa State University

June 4, 2013

[1] Joint work with S. Park and S. Kim

Ouline

- Introduction
- Proposed Methodology
- Application to Private Education Expenses Survey Data
- Conclusion & Future work

Introduction

- Survey can be conducted by using several survey modes
- Survey modes : mail, internet, phone, interviewer, etc
  - Self-reported survey : mail, internet
  - Interview survey : face-to-face, telephone

Introduction

- Mixed-mode survey : A survey that uses several survey modes to collect information from a sample.
  - Advantage : help to increase survey response rates and reduce nonresponse error and data collection costs.
  - Disadvantage : mix of modes can affect the data and the estimates which are subject to biases because of different measurement errors.
- Goal : We need to calibrate the measurement bias from the difference of survey modes in order to improve the quality of survey.

Motivation: Private Education Expenses Survey Data

Data Description

- Annual survey run by Statistics Korea.
- Self-reported survey with two survey modes, mail and internet. In 2011 survey, the respondents are randomly assigned to mail or internet survey mode. But in 2012 survey, the respondents can select the survey mode.
- Survey unit: students and parents of elementary, middle, high school in Korea.
- Study variables : Time (how many hours do you have private education in a week?) and Cost (how much money do you spend for private education in a month?)
- Auxiliary variables : local level, school level, sex, age of parents, education level of parents, grade of student, and income of household.

Motivation: Private Education Expenses Survey Data

Preliminary analysis

- T-test of mean of the study variables, Time, Cost, and Cost/Time in 2011 survey data.

Variable	Mode	Mean	STD	t-value	p-value
Time	Mail	5.96	6.11	8.917	0.000
	Internet	5.44	6.21		
Cost	Mail	71.20	77.80	3.808	0.000
	Internet	68.32	82.46		
Cost/Time	Mail	3.79	3.11	-7.99	0.000
	Internet	4.12	3.80		

- Significant difference between the two survey modes
- Large standard deviations for internet survey.

Motivation: Private Education Expenses Survey Data

Preliminary analysis

- Percent of students with no private education in 2011 survey data.

School Level	Mail	Internet
Elementary School	13.9	16.6
Middle School	23.2	28.3
High School	36.3	43.3

- Significant proportion of zero values and the proportion is higher for internet survey data.

Basic Setup

- Latent variable, y : the (ideal) study variable with no measurement error.
- Auxiliary variable, $\mathbf{x}$: the variable which may explain the study variable y. Assume that $\mathbf{x}$ does not have significant measurement error.
- Observed variable: Either y_a or y_b
  - y_a : the observed variable from survey mode A.
  - y_b : the observed variable from survey mode B.
- Choice of survey mode:
  - Randomized: 2011 survey
  - Self-selected: 2012 survey

Basic Setup: Data Structure

- Data structure : $S = S_a \cup S_b$

Sample	X	Y_a	Y_b
S_a	o	o	
S_b	o		o

- The goal is to create imputed value of y_a in S_b.

Sample	X	Y_a	Y_b
S_a	o	o	
S_b	o	o	o

- Note that y_a is the counterfactual outcome for the elements in S_b

Methodology

STEP

1. Specify a measurement error model.
2. Derive prediction model using Bayes theorem.
3. Parameter estimation: EM algorithm.
4. Generating imputed values from the prediction model.

Methodology

1. Model specification : Measurement Error Model

- Measurement error model = the measurement model + the structural model.
- Measurement model : a model between latent variable y and observed variable y_a or y_b.

$$g_a(y_a|y), \quad \text{or} \quad g_b(y_b|y)$$

- Structural error model : a model between latent variable y and auxiliary variables x.

$$f\left(y \mid x\right)$$

- Choice model (or selection model) may be needed if the choice of the measurement is not random.

$$P\left(M = a \mid x, y\right)$$

where $P\left(M = a \mid x, y\right) + P\left(M = b \mid x, y\right) = 1$.

- The choice model is called ignorable if $P\left(M = a \mid x, y\right)$ does not depend on y.

Methodology

2. Imputation Model (Prediction model)

- Imputation model (=Prediction model): model for y given the realized observation.
- Assume that (y_a, y_b) is conditionally independent of x given y:

$$(y_a, y_b) \perp x \mid y.$$

  It means $f(y_b \mid \mathbf{x}, y) = f(y_b \mid y)$ and $f(y_a \mid \mathbf{x}, y) = f(y_a \mid y)$.
- Prediction model is obtained by applying the Bayes theorem.

$$f(y|y_a, x) = \frac{f(y|x)g_a(y_a|y)P(M = a \mid x, y)}{\int f(y|x)g_a(y_a|y)P(M = a \mid x, y)\, dy}$$

$$f(y|y_b, x) = \frac{f(y|x)g_b(y_b|y)P(M = b \mid x, y)}{\int f(y|x)g_b(y_b|y)P(M = b \mid x, y)\, dy}$$

Methodology

Example 1: Normal Case

Normal distribution regression model

- Structural model :

$$y_i = \beta_0 + x_i\boldsymbol{\beta}_1 + e_i, \qquad e_i \sim N(0, \sigma_e^2).$$

- Measurement model :

$$\begin{aligned} y_{ai} &= y_i + u_{ai}, \qquad u_{ai} \sim N(0, \sigma_a^2) \\ y_{bi} &= y_i + u_{bi}, \qquad u_{bi} \sim N(0, \sigma_b^2) \end{aligned}$$

To avoid non-identifiability problem, we may assume that $\sigma_a^2 = 0$.

Methodology

Example 1 (Cont'd)

- Under the normal model example (with ignorable choice mechanism),

$$\begin{aligned} y \mid (x, y_a) &\sim N\left(\tilde{y}_a, \alpha_a \sigma_e^2\right) \\ y \mid (x, y_b) &\sim N\left(\tilde{y}_b, \alpha_b \sigma_e^2\right) \end{aligned}$$

  where

$$\begin{aligned} \tilde{y}_a &= \alpha_a \left(\beta_0 + \beta_1 x_i\right) + (1 - \alpha_a) y_a \\ \tilde{y}_b &= \alpha_b \left(\beta_0 + \beta_1 x_i\right) + (1 - \alpha_b) y_b, \end{aligned}$$

  $\alpha_a = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, and $\alpha_b = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$.

- Need to estimate the parameters of the models.

Methodology

Remark

- If $f(y \mid \mathbf{x})$ is not normal, assuming ignorable choice mechanism, the prediction model becomes

$$\begin{aligned} f(y \mid \mathbf{x}, y_b) &\propto f(y \mid \mathbf{x}) g_b(y_b \mid \mathbf{x}, y) \\ &\propto f(y \mid \mathbf{x}) g_b(y_b \mid y) \end{aligned} \tag{1}$$

  The first term is structural model and the second term is measurement model.

- Two problems when generating imputed values from (1):
  1. Parameters in the models are unknown.
  2. Even if we know the parameters, sampling from (1) can be computationally challenging (often relies on MCMC method).

Idea

- Mote Carlo EM algorithm
- E-step: Generate y from $f(y|y_a, x)$ or from $f(y \mid y_b, x)$.
- M-step: Solve the imputed score equation
- Identifiability condition needs to be imposed in the parameter space. (e.g. $\sigma_a^2 = 0$)

Problem

E-step is tricky because

$$f(y \mid y_b, x) \propto f(y \mid x) f(y_b \mid y)$$

is often difficult to generate samples from.

$\rightarrow$ Parametric fractional imputation provides a useful computational tool.

Methodology

Back to Example 1: Normal case

Here we assume that $\sigma_a^2 = 0$ *i.e.* $y_a = y$ and the choice mechanism is ignorable.

$$\begin{aligned} y_i &= \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}_1 + e_i, \qquad e_i \sim N(0, \sigma_e^2) \\ y_{bi} &= y_i + u_{bi}, \qquad u_{bi} \sim N(0, \sigma_b^2). \end{aligned}$$

- Using data from mode A, S_a, we can estimate the parameter in structural error model. In regression model, β_0, β_1, and σ_e^2 can be estimated by usual method with data S_a.
- We need to estimate only σ_b^2 with data from mode B, S_b in the measurement model. In regression model,

$$\hat{\sigma}_b^2 = \frac{1}{n_b} \sum_{i \in S_b} \left\{ \frac{n}{n-p} \left(y_{bi} - \hat{\beta}_0 - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_1 \right)^2 - \hat{\sigma}_e^2 \right\},$$

where $\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1$, and $\hat{\sigma}_e^2$ is the estimate of the parameter in the structural model.

Methodology

3. Parameter estimation

Parametric fractional imputation(PFI) of Kim (2011):

1. Set $t = 0$. Estimate the parameter θ of $f(y|x;\theta)$ with data S_a. Let $\hat{\theta}_0$ be the initial value.
2. For each unit $i \in S_b$, generate M imputed values, $y_{ai}^{*(1)}, \ldots, y_{ai}^{*(M)}$, from $\hat{f}(y|x;\hat{\theta}_0)$.
3. Estimate σ_b^2:
$$\hat{\sigma}_b^2 = \frac{1}{n_b}\frac{1}{M}\sum_{i\in S_b}\sum_{k=1}^{M}\left(y_{bi} - y_{ai}^{*(k)}\right)^2$$
4. Calculate weight w_{ij}^* for each $i \in S_b$,
$$w_{ij}^* \propto \hat{g}_b(y_{bi}|y_{ai}^{*(j)})\frac{f(y_{ai}^{*(j)} \mid x_i;\hat{\theta}^{(t)})}{f(y_{ai}^{*(j)} \mid x_i;\hat{\theta}^{(0)})}$$
and $\sum_{j=1}^{M} w_{ij}^* = 1$.
5. Update $\hat{\sigma}_b^2$ by $\hat{\sigma}_b^2 = n_b^{-1}\sum_{i\in S_b}\sum_{k=1}^{M} w_{ik}^*\left(y_{bi} - y_{ai}^{*(k)}\right)^2$. Also, update $\hat{\theta}$ by solving the imputed score equations. Go to step 4. Repeat until converge.

- In data S_b, we want to generate y using the observed variable y_b and auxiliary variables $\mathbf{x}$ with the prediction model, $f(y|y_b, \mathbf{x})$.
- After estimating the parameter in the measurement model and the structural model, we can calculate the distribution $f(y|y_b, \mathbf{x})$ by Bayes theorem.
- In Example 1, we can easily generate the value $y(= y_a)$,

$$\hat{f}(y_i|y_b, x) \sim N(\hat{y}_i, \hat{\alpha}\hat{\sigma}_e^2)$$

  where $\hat{y}_i = \hat{\alpha}\tilde{y}_i + (1 - \hat{\alpha})y_{bi}$, $\hat{\alpha} = \hat{\sigma}_b^2/(\hat{\sigma}_e^2 + \hat{\sigma}_b^2)$, and $\tilde{y}_i$ is the predicted values from the structural model using only auxiliary variables $\mathbf{x}$.
- From the PFI method, we can also create single imputation by selecting one imputed value from the M fractionally imputed values using the selection probability proportional to w_{ij}^*.

Application to the Private Education Expenses Survey Data

- 2011 Survey data ($n = 45,501$ parents survey from 1,081 sample schools)
- Stratified cluster sampling (two-stage)
- Two survey modes: mail vs internet (randomized in 2011 survey)
- Bivariate $Y = (Y_1, Y_2)$: (time, cost)
- X: many demographic & socio-economic items
- Steps
  1. Editing & Outlier detection
  2. Model specification
  3. Parameter estimation (using EM + PFI)
  4. Prediction (of counterfactual outcomes)

Model specification

- Separate models for elementary, middle, high schools.
- Structural error model: Tobit regression model

$$\begin{aligned} y_{1i} &= z_{1i} I(z_{1i} > 0) \\ y_{2i} &= z_{2i} I(z_{2i} > 0) \end{aligned}$$

where

$$z_{1i} = \mathbf{x}_i' \beta + e_{1i}, \quad e_{1i} \sim N(0, \sigma_1^2)$$

and

$$z_{2i} = z_{1i} R_i, \quad R_i = \mathbf{x}_i' \gamma + e_{2i}, \quad e_{2i} \sim N(0, \sigma_2^2)$$

- Measurement error model:

$$\begin{aligned} y_{1i,b} &= z_{1i,b} I(z_{1i,b} > 0) \\ y_{2i,b} &= z_{2i,b} I(z_{2i,b} > 0) \end{aligned}$$

with $z_{1i,b} = z_{1i} + u_{1i}$ and $z_{2i,b} = z_{2i} + u_{2i}$.

Some Results

Table : Proportion of students with no private education (%)

Variable	School	Mail	Internet	Total	PFI
Time	Elementary	13.9	16.6	15.1	14.3
	Middle	23.2	28.3	25.6	24.8
	High	36.3	43.3	39.7	39.4
Cost	Elementary	13.9	16.6	15.1	14.6
	Middle	23.2	28.3	25.6	25.0
	High	36.3	43.3	39.7	39.4

Some Results

Table : Mean estimates (among Time > 0)

Variable	School	Mail	Internet	Total	PFI
Time	Elementary	9.14	9.185	9.16	8.70
	Middle	10.34	9.98	10.18	9.77
	High	7.46	7.63	7.54	7.02
Cost	Elementary	83.74	82.14	83.06	82.18
	Middle	107.92	116.26	111.70	109.69
	High	125.52	131.77	128.36	125.29

Conclusion

- Measurement error model approach to mixed-mode survey.
- EM algorithm for parameter estimation.
- Prediction by fractional imputation (Bayes theorem).
- Instead of assuming $\sigma_a^2 = 0$, one may consider $\sigma_b^2 = 0$

Future Work

- Variance estimation needs to be developed.
- We assumed ignorable choice mechanism. An extension to non-ignorable choice mechanism can be devoloped (using 2012 survey data)

The end