시장경제연구회 발표 자료

Survey Data Integration

"An imputation approach for handling mixed-mode surveys"
-Seunghwan Park, Jae-Kwang Kim, Sangun Park-

Seunghwan Park

Kangwon National University

Oct 24, 2019

Outline

- Introduction
- Methodology
  - Basic Setup
  - Measurement Error Model
  - Parameter Estimation
  - Imputation
- Application to Private Education Expenses Survey Data
- Conclusion

- Survey sample data is usually used for official statistics.
- Want to combine several sources of information to get improved an estimation.
- Auxiliary information,
  - from other survey data.
  - from census data or administrative data.
  - from big data
    - electronic health record, satellite information
    - mobile sensor data,roaming data
    - credit card purchase data, voluntary membership data, etc.

- Three situations:
  1. **Multiple samples for one target population**
  2. One sample each from multiple populations
  3. Multiple samples from multiple target populations

Table: Two data sources

	Survey data	Big data
Cost	proportional to sample size	free
measurement	study variable Y	auxillary variable X
representativeness	yes	no
Bias	Bias $\cong 0$	Bias $\neq 0$
Variance	Var $\propto (1/n)$	Var $\cong 0$

- Big data may have systematic error
  1. Selection bias : missing data, coverage error problem
  2. Information bias : measurment error

Table: Two data structure cases for data integration

Case 1	X	Y
Survey	o	
Big data	o	o

Case 2	X	Y
Survey	o	o
Big data	o	

- Parameter of interest : the population total $\theta = \sum_{i \in U} y_i$.
- In Case 1, wish to combine the two data to obtain an unbiased estimator of θ.
  1. Weighting approach : Construct weights for big data such that $\hat{\theta}_B = \sum_{i \in B} w_{i,B} y_i$ can be unbiased. (calibration weights approach, propensity model approach)
  2. Imputation approach : generate a synthetic value of Y for survey using the observations in big data.

- In Case 2, $\hat{\theta}_S = \sum_{i \in S} w_i y_i$ is available and is nearly unbiased.
- Thus, the goal of data integration is to improve the efficiency.
- Assume that there is no coverage error in big data.
- The big data may have measurement errors.
- Need to determine a level to match when combining with non-survey data.
  1. Unit(individual) level
  2. Area level

- Survey can be conducted in many different modes
- Survey modes : mail, internet, phone, interviewer, etc
  - Self-reported survey : mail, internet
  - Interview survey : face to face, telephone

- Mixed mode survey : A survey 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 error.

Introduction

- Latent variable, y : the (ideal) study variable with no measurement error.
- Auxiliary variable, $\mathbf{x}$: the variable which have correlation with 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.

- Data Structure for a mixed-mode wurvey with two survey modes

Mode	X	Y_a	Y_b
Sample A	o	o	
Sample B	o		o

- Choice of survey mode:
  - Randomized: 2011 survey
  - Self-selected: 2012 survey
- Assume that Sample A is a gold standard one ($Y_a = Y$).
- Goal : We need to calibrate the measurement bias from the difference of survey modes in order to combine several sources of information.

- Means and standard deviations of the private education expenses in 2011 PEES

School Level	Mail		Internet	
Elementary School	72.1	(60.4)	68.5	(59.5)
Middle School	82.9	(71.3)	83.4	(81.3)
High School	79.9	(92.1)	74.7	(95.8)

Standard deviations are in parentheses.

- Percent of students taking private lessons or tutoring in 2011 PEES

School Level	Mail	Internet
Elementary School	86.1	73.4
Middle School	76.8	71.7
High School	63.7	56.7

- Parameter of interest : the finite population mean of the study variable under mode A, $\psi_N = N^{-1} \sum_{i=1}^{N} y_i$.
- For a single-mode survey data (Mode A only), the Horvitz-Thompson (HT) estimator, $\hat{\psi}_{HT} = N^{-1} \sum_{i \in S} w_i y_{ai}$ is an unbiased estimator of ψ_N.
- Under the mixed-mode survey structure, a naive estimator given by

$$\hat{\psi}_{naive} = N^{-1} \left\{ \sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i y_{bi} \right\}$$

  is biased unless $E(y_{ai}) = E(y_{bi})$.

- To correct for the bias of the naive estimator, consider

$$\hat{\psi} \equiv N^{-1} \left\{ \sum_{i \in S_a} w_i y_i + \sum_{i \in S_b} E(y_i | y_{bi}, \mathbf{x}_i) \right\},$$

- The conditional expectation is computed from a prediction model

$$f(y_i | y_{bi}, \mathbf{x}_i) = f(y_i | \mathbf{x}_i) \frac{g(y_{bi} | y_i)}{\int f(y_i | \mathbf{x}_i) g(y_{bi} | y_i) dy_i}$$

  for the units in S_b.

- The prediction model will be the model for imputing the unobserved outcome.

- 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.

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

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

- Prediction model is obtained by applying the Bayes theorem.

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

Idea

- Mote Carlo EM algorithm + parametric fractional imputation
- E-step: Generate y from $f(y \mid y_b, x)$,

$$f(y_{ai}|x_i, y_{bi}) \propto f(y_{ai}|x_i)g(y_{bi}|y_{ai}).$$

- M-step: Solve the imputed score equation
- Identifiability condition needs to be imposed in the parameter space. (e.g. $\sigma_a^2 = 0$)

Suppose that θ, α and ϕ are the parameter of distributions $f(y_{ai}|\mathbf{x}_i;\theta)$, $g(y_{bi}|y_{ai};\alpha)$ and $P(m_i = a|\mathbf{x}_i, y_{ai};\phi)$, respectively. Then, the EM algorithm using the PFI method under nonignorable choice mechanism is computed by the following steps:

- [Step 1] Set $t = 0$. Calculate the estimate of the parameter θ of $f(y_{ai}|\mathbf{x}_i;\theta)$ with data S_a. Let the estimate, denoted as $\hat{\theta}^{(0)}$, be the initial value.
- [Step 2] For each unit $i \in S_b$, generate M imputed values, $y_{ai}^{*(1)}, \ldots, y_{ai}^{*(M)}$, from $f(y_{ai}|\mathbf{x}_i;\hat{\theta}^{(0)})$. Set $w_{ij(0)}^{*} = 1/M$.

- [Step 3] Update $\hat{\theta}, \hat{\alpha}$ and $\hat{\phi}$ by solving the imputed score equations:

$$\sum_{i \in S_a} w_i S_1(\theta; \mathbf{x}_i, y_{ai}) + \sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij(t)}^* S_1(\theta; \mathbf{x}_i, y_{ai}^{*(j)}) = 0$$

$$\sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij(t)}^* S_2(\alpha; y_{ai}^{*(j)}, y_{bi}) = 0$$

$$\sum_{i \in S_a} w_i S_3(\phi; m_i, \mathbf{x}_i, y_{ai}) + \sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij(t)}^* S_3(\phi; m_i, \mathbf{x}_i, y_{ai}^{*(j)}) = 0,$$

where $S_1(\theta; \mathbf{x}_i, y_{ai}) = \partial \log f(y_{ai} \mid \mathbf{x}_i; \theta)/\partial\theta$, $S_2(\alpha; y_{ai}, y_{bi}) = \partial \log g(y_{bi} \mid y_{ai}; \alpha)/\partial\alpha$ and $S_3(\phi; \mathbf{x}_i, y_{ai}) = \partial \left\{ \log I(m_i = a) \log \left(P_i/(1 - P_i) \right) + \log(1 - P_i) \right\} /\partial\phi$ with $P_i = P(m_i = a \mid \mathbf{x}_i, y_{ai}; \phi)$

- [Step 4] Calculate weight w_{ij}^* for each $i \in S_b$,

$$w_{ij(t)}^* \propto g(y_{bi}|y_{ai}^{*(j)}; \hat{\alpha}^{(t)}) \frac{f(y_{ai}^{*(j)} \mid x_i; \hat{\theta}^{(t)})}{f(y_{ai}^{*(j)} \mid x_i; \hat{\theta}^{(0)})} P(m_i = b \mid \mathbf{x}_i, y_{ai}^{*(j)}; \hat{\phi}^{(t)})$$

and $\sum_{j=1}^{M} w_{ij(t)}^* = 1$, where $\hat{\eta}^{(t)} = (\hat{\theta}^{(t)}, \hat{\alpha}^{(t)}, \hat{\phi}^{(t)})$ is the current estimate of $\eta = (\theta, \alpha, \phi)$.

- [Step 5] Set $t = t + 1$ and go to Step 3. Continue until convergence.
- The parametric fractional imputation estimator of the finite population mean is computed by

$$\hat{\psi}_{PFI} = N^{-1} \left\{ \sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij}^* y_{ai}^{*(j)} \right\}.$$

Application to the Private Education Expenses Survey Data

Data Description

- Self-reported survey with two 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 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.

- Distribution of the interested variables, Time, Cost, and Cost/Time(average expenses for private education per hour).

Variable	Mode	1st Q	Median	3rd Q	Max
Time	Mail	0.00	5.00	10.00	49.00
	Internet	0.00	4.00	9.00	48.00
Cost	Mail	0.00	60.00	110.00	900.00
	Internet	0.00	48.00	108.00	1170.00
Cost/Time	Mail	1.79	2.85	4.88	45.00
	Internet	1.82	3.12	5.20	90.00

- T-test of mean of the interested variables, Time, Cost, and Cost/Time.

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		

- Another difficulty in developing a proper imputation model for Time and Cost is the significant portion of zero values in the sample. For example, the proportion of zero values for study variable Time is more than 15% in the sample under mail mode.
- Thus, to account for the significant portion of zero-values in Time, we applied a Tobit regression model.

$$y_{a1,i} = \begin{cases} z_{a1,i} & \text{if } z_{a1,i} > 0 \\ 0 & \text{otherwise} , \end{cases} \tag{1}$$

where

$$z_{a1,i} = x_i'\beta + e_i, \quad e_i \sim N(0, \sigma_e^2).$$

Application to the Private Education Expenses Survey Data

Result

- Four estimators ofr the population mean are considered:
  - Mail : $\sum_{i \in S_a} w_i y_{ai} / (\sum_{i \in S_a} w_i)$
  - Internet : $\sum_{i \in S_b} w_i y_{bi} / (\sum_{i \in S_b} w_i)$
  - Naive : $\{\sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i y_{bi}\} / (\sum_{i \in S} w_i)$
  - PFI : $\{\sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij}^* y_{ai}^{*(j)}\} / (\sum_{i \in S} w_i)$,

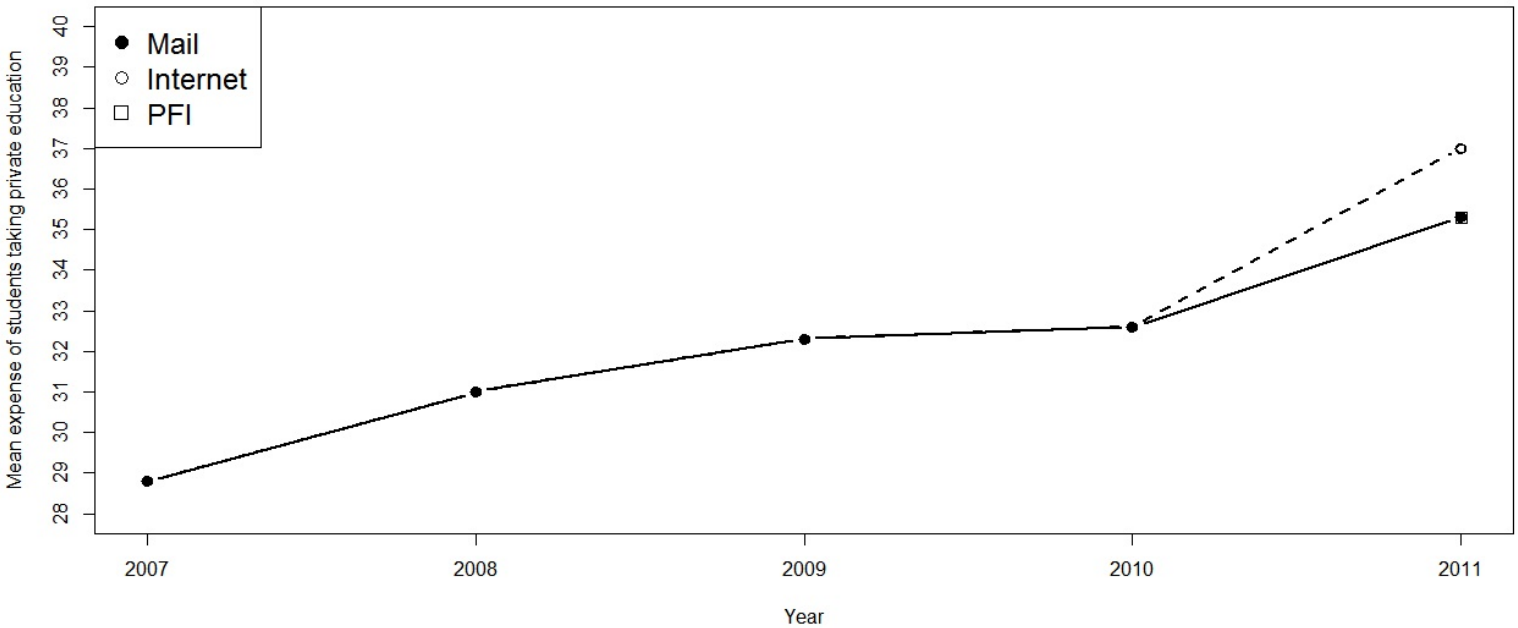
  where w_i are the sampling weights.
- Four estimate of the mean expense of students taking private education from 2011 PEES data

School	Mail	Internet	Naive	PFI
Elementary	27.91	27.38	27.69	27.45
	(0.46)	(0.53)	(0.35)	(0.21)
Middle	35.98	38.75	37.24	36.55
	(0.55)	(0.71)	(0.44)	(0.26)
High	41.84	43.93	42.79	41.76
	(0.66)	(0.77)	(0.50)	(0.30)

Application to the Private Education Expenses Survey Data

Result

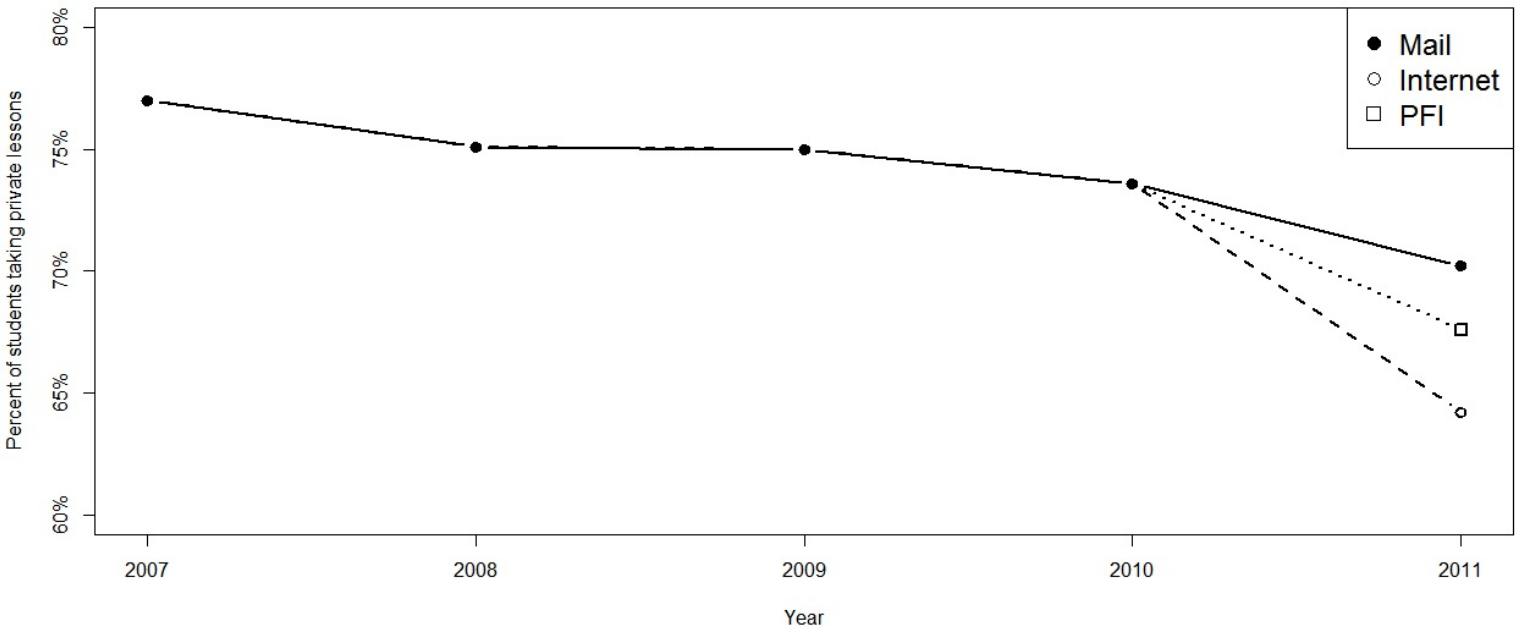
- Mean expense of students taking private education from 2007 to 2011



Application to the Private Education Expenses Survey Data

Result

- Percents of students taking private lessons from 2007 to 2011



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 research topic: extend to combine Big data(measurement error and selection bias) and survey data (accurate and representative).

Thank You !