Introduction to Survey Data Integration

Jae-Kwang Kim

Iowa State University

May 20, 2014
Outline

1. Introduction
2. Survey Integration Examples
3. Basic Theory for Survey Integration
4. NASS application
5. Discussion
1. Introduction

**Question:** What do you mean by “integration” in surveys?
- integrating survey and administrative data through unit record linkage
- improving coherence across data collections by using standard classifications and questions
- rationalizing content between surveys, including census
- greater standardization of methods, tools, IT systems and processes
- combining separate sample surveys into one survey vehicle.

1. Introduction

- **Our focus** is the last part of the several meaning of survey integration. “combining separate sample surveys into one survey vehicle”
- Most sample surveys are run as stand-alone surveys.
- However, survey integration is a new area of research to achieve the three conflicting goals:
  1. Minimize the cost associated with surveys
  2. Maximize the information (or efficiency of the survey estimation)
  3. Minimize the respondent burden
Two approaches for survey integration

- **Macro approach:** Obtain improve estimates by combining all available information
  - General Least Squares (GLS) estimation, GLS weighting
- **Micro approach:** Create a single data (single survey vehicle) that contains all available information
  - Synthetic data imputation
  - Record Linkage
  - Statistical Matching, or data fusion
2. Survey Integration Examples

- How to combine several source of information?
- **Example 1**: US Census of housing and population
  - Short Form: 100% sample (obtain basic demographic information)
  - Long form: about 16% sample (obtain other social and economic information as well as demographic information)
- Classical two-phase sampling problem: Calibration weighting for demographic variable to match known population counts from short form (Deming & Stephan, 1940).
Example 2: Consumer Expenditure Survey (Zieschang, 1990)

- Diary survey: Observe $X$, $Y$
- Interview survey (quarterly): Observe $X$

Two surveys are obtained independently from the same target population. (uses the same sampling frame.)

Two estimates of $X$, $\hat{X}_1$ and $\hat{X}_2$, can be different because of the sampling errors.

How to incorporate the information from the quarterly interview survey to diary survey estimate?
2. Survey Integration Examples

GLS Weighting

- Two parameters with three estimates:
  1. Survey one: Observe $\hat{X}_1$
  2. Survey two: Observe $\hat{X}_2$ and $\hat{Y}_2$

- GLS model

\[
\begin{pmatrix}
\hat{X}_1 \\
\hat{X}_2 \\
\hat{Y}_2
\end{pmatrix}
= \begin{pmatrix}
1 & 0 \\
1 & 0 \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
X \\
Y
\end{pmatrix}
+ \begin{pmatrix}
e_1 \\
e_2 \\
e_3
\end{pmatrix}
\tag{1}
\]

where

\[
V
\begin{pmatrix}
e_1 \\
e_2 \\
e_3
\end{pmatrix}
= \begin{pmatrix}
V_{xx1} & 0 & 0 \\
0 & V_{xx2} & V_{xy2} \\
0 & V_{xy2} & V_{yy2}
\end{pmatrix}
\]
GLS Weighting (Continued)

- GLS model: We may write

\[ Y = Z \theta + e \]

where \( e \sim (0, V) \).

- GLS estimator of \( \theta \):

\[ \hat{\theta}_{GLS} = \left( Z' V^{-1} Z \right)^{-1} Z' V^{-1} Y \]

- Note that the GLS estimator is a linear combination of the three observed information:

\[ \hat{\theta}_{GLS} = \alpha_1 \hat{X}_1 + \alpha_2 \hat{X}_2 + \alpha_3 \hat{Y}_2 \]

In fact, we have

\[ \hat{X}_{GLS} = \frac{V_{xx2}}{V_{xx1} + V_{xx2}} \hat{X}_1 + \frac{V_{xx1}}{V_{xx1} + V_{xx2}} \hat{X}_2 \]

and

\[ \hat{Y}_{GLS} = \hat{Y}_2 + \left( \hat{X}_{GLS} - \hat{X}_2 \right) \frac{V_{xy2}}{V_{xx2}} \]
2. Survey Integration Examples

- **Example 3:** SAIPE (Fay & Herriot, 1979), combines small area level information
  - Current Population Survey (obtain $\hat{Y}_i$, estimated poverty rate for area $i$)
  - Census long form data ($x_{1i}$: poverty rate for the census year for area $i$)
  - Food stamp participation rate ($x_{2i}$)
  - Pseudo poverty rate for children from tax return information & tax nonfiler rate ($x_{3i}$)

- Fay-Herriot model

\[
\hat{Y}_i = Y_i + e_i = (x_i'\beta + u_i) + e_i
\]

$x_i = (1, x_{1i}, x_{2i}, x_{3i})'$: predictors, $e_i$: sampling error, $u_i$: area $i$ random effect (=model error).
3. Basic Theory

- **General Setup**
  - Survey $A$: Measures $\hat{Y}_{i,a}$, subject to sampling error.
  - Survey $B$: Measures $\hat{X}_{i,b}$, subject to sampling error.
  - $E_A(\hat{Y}_{i,a}) \neq E_B(\hat{X}_{i,b})$ due to the structural difference between the surveys

- **Structural difference (or systematic difference)**
  - due to different mode of survey
  - due to time difference
  - due to frame difference

- **Example**: Objective Yield Survey vs Agricultural Yield Survey and December Agricultural Surveys
Two error models (for area $i$)

**Sampling error model**

$$
\begin{align*}
\hat{Y}_{i,a} &= Y_i + a_i \\
\hat{X}_{i,b} &= X_i + b_i
\end{align*}
$$

where $(a_i, b_i)$ represents the sampling error such that

$$
\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{yy,i} & V_{yx,i} \\ V_{xy,i} & V_{xx,i} \end{pmatrix}
$$

**Structural error modeling**

$$X_i = \beta_0 N_i + \beta_1 Y_i + e_i, \quad e_i \sim (0, N_i \sigma^2_e)$$

where $N_i$ is the (known) population size of area $i$. 

3. Basic Theory

- Structural error model describes the relationship between the two survey measurement up to sampling error.
  - $Y$: target measurement item (variable of primary interest)
  - $X$: inaccurate measurement of $Y$ with possible systematic bias.
- If both $X$ and $Y$ measure the same item (with different survey modes), structural error model is essentially a measurement error model. ($\beta_0 = 0, \beta_1 = 1$ means no measurement bias.)
If the parameters in the structural error model are known, \( \hat{Y}_{i,b} \equiv \beta_1^{-1}(\hat{X}_{i,b} - \beta_0 N_i) \) is also an unbiased estimator of \( Y_i \), computed from survey B. Estimator \( \hat{Y}_{i,b} \) using consistent \( (\hat{\beta}_0, \hat{\beta}_1) \) is often called **synthetic estimator**.

How to combine the two estimators?

1. Best Unbiased Prediction approach
2. GLS (or GMM) approach
3. Basic Theory

- **Best Unbiased Prediction approach:**
  1. Specify a model for sampling error: \( \hat{f}(\hat{X}_i, \hat{Y}_i | X_i, Y_i) \)
  2. Specify a parametric model for structural error: \( g(Y_i | X_i; \theta) h(X_i; \alpha) \)
  3. Use Bayes rule to combine the two models and obtain the posterior distribution

\[
f(Y_i | \hat{X}_i, \hat{Y}_i; \theta, \alpha) \propto \int \hat{f}(\hat{X}_i, \hat{Y}_i | X_i, Y_i)g(Y_i | X_i; \theta)h(X_i; \alpha)dX_i.
\]

4. Compute conditional expectation

\[
E \left( Y_i | \hat{X}_i, \hat{Y}_i; \theta, \alpha \right) = \int Y_i f(Y_i | \hat{X}_i, \hat{Y}_i; \theta, \alpha)dY_i
\]
3. Basic Theory

- **Best Unbiased Prediction** approach:
  - Best prediction under correct model specification.
  - Requires strong model assumptions.
  - Computation can be heavy (MCMC)
  - Depends on unknown superpopulation parameter
    1. Empirical Bayes
    2. Hierarchical Bayes
3. Basic Theory

- Alternative approach: **GLS (Generalized least squares)** approach
  - Does not require distributional assumptions. Only require moment assumptions (second moment assumptions)
  - Combine information through GLS, not through Bayes

- Recall : GLS method

\[
y = Z\beta + e, \quad e \sim (0, V)
\]

\[
\Rightarrow \hat{\beta}_{GLS} = (Z'V^{-1}Z)^{-1}Z'V^{-1}y
\]
3. Basic Theory

- **GLS approach** to combine two error models:

\[ y = Z\theta + e, \quad e \sim (0, V) \]

\[ \iff \left( \begin{array}{c} \hat{Y}_{i,a} \\ \beta_{1}^{-1}(\hat{X}_{i,b} - N_{i}\beta_{0}) \end{array} \right) = \left( \begin{array}{c} 1 \\ 1 \end{array} \right) Y_{i} + \left( \begin{array}{c} u_{1i} \\ u_{2i} \end{array} \right) \]

where \( u_{1i} = a_{i} \) and \( u_{2i} = \beta_{1}^{-1}(b_{i} + e_{i}) \). Thus,

\[
\left( \begin{array}{c} u_{1i} \\ u_{2i} \end{array} \right) \sim \left[ \begin{array}{c} 0 \\ 0 \end{array} \right], \left( \begin{array}{cc} V_{yy,i} & \beta_{1}^{-1}V_{xy,i} \\ \beta_{1}^{-1}V_{xy,i} & \beta_{1}^{-2}(V_{xx,i} + N_{i}\sigma_{e}^{2}) \end{array} \right) \right].
\]
3. Basic Theory

- **GLS estimator**: Best linear unbiased estimator of $Y_i$ based on the linear combination of $\hat{Y}_{i,a}$ and $\hat{Y}_{i,b} = \beta_1^{-1}(\hat{X}_{i,b} - N_i\beta_0)$.

- Under the current setup,

$$\hat{Y}_i^* = \omega_i \hat{Y}_{i,a} + (1 - \omega_i) \hat{Y}_{i,b}$$

where

$$\omega_i = \frac{N_i\sigma_e^2 + \beta_1^2 V_{xx,i} - \beta_1 V_{xy,i}}{N_i\sigma_e^2 + \beta_1^2 V_{xx,i} + V_{yy,i} - 2\beta_1 V_{xy,i}}$$

- The GLS estimator is sometimes called **composite estimator**. In practice, we need to use $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}_e^2$. 
3. Basic Theory

- Parameter estimation for the structural model
  1. **Case 1**: Matching measurement $X$ and measurement $Y$ is possible (e.g.: Survey A sample is a subset of survey B sample.)
  2. **Case 2**: Matching is not possible.

- In Case 1, we can easily obtain a consistent estimator of the model parameters from the set where units have both $X$ and $Y$ observed. (Unit level modeling)

- In Case 2, we may use area level model to match $\hat{X}_i$ (from survey B) and $\hat{Y}_i$ (from survey A) for each area.
3. Basic Theory

- To estimate \((\beta_0, \beta_1)\), we express

\[
\hat{X}_{i,b} = N_i \beta_0 + Y_i \beta_1 + e_i + b_i \\
\hat{Y}_{i,a} = Y_i + a_i
\]

- Direct regression of \(\hat{X}_{i,b}\) on \((N_i, \hat{Y}_{i,a})\) does not work because \(\hat{Y}_{i,a}\) has sampling error.

- The area-level model takes the form of measurement error model (Fuller, 1987).

- Parameter estimation can be performed using the measurement error model estimation methods. (Details skipped.)
4. NASS application
Improving Crop Acreage Estimation

- National Agricultural Statistical Service (NASS) under US Department of Agriculture provides acreage estimates for crops.
- Several sources of information for crop acreage
  - June Area Survey (JAS) data
  - Satellite imagery classification data. Cropland Data Layer (CDL)
  - Farm Service Agency (FSA) data
Several sources of information for crop acreage (for area $i$: analysis district)

- $\hat{Y}_i$: estimates from JAS (subject to sampling error)
- $\hat{X}_{1i}$: estimates from CDL
- $\hat{X}_{2i}$: FSA estimates

Even though satellite image data is not subject to sampling error, the CDL estimates are subject to sampling error due to classification from JAS data.

FSA estimates can be modified to reduce the coverage error by modeling propensity scores for program participation. Thus, FSA estimates are subject to some type of sampling error where the sampling mechanism here refers to the program participation.
Figure: The 2009 cropland data layer products. The legend identifies aggregated agricultural and non-agricultural land cover categories by decreasing acreage.
4. NASS application
Improving Crop Acreage Estimation

We can construct separate structural error models

\[ X_{1i} = \beta_{10} N_i + \beta_{11} Y_i + e_{1i} \]
\[ X_{2i} = \beta_{20} N_i + \beta_{21} Y_i + e_{2i} \]

where

\[
\begin{pmatrix}
  e_{1i} \\
  e_{2i} \\
\end{pmatrix}
\sim
\begin{pmatrix}
  0 \\
  0 \\
\end{pmatrix},
\begin{pmatrix}
  N_i \sigma_{e1}^2 & 0 \\
  0 & N_i \sigma_{e2}^2 \\
\end{pmatrix}
\]
GLS method can be applied to separate structural error models to get

\[
\begin{pmatrix}
\hat{Y}_i \\
\hat{Y}_{i,1} \\
\hat{Y}_{i,2}
\end{pmatrix} =
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix} Y_i +
\begin{pmatrix}
u_{1i} \\
u_{2i} \\
u_{3i}
\end{pmatrix},
\]

where \( \hat{Y}_{i,1} = \hat{\beta}_{11}^{-1}(\hat{X}_{1i} - \hat{\beta}_{10} N_i) \) and \( \hat{Y}_{i,2} = \hat{\beta}_{21}^{-1}(\hat{X}_{2i} - \hat{\beta}_{20} N_i) \) are two different synthetic estimators of \( Y_i \).

We may use a simpler covariance matrix to achieve computational simplicity.
4. NASS application

Some Results

Figure: Comparison of the estimates in Indiana State
4. NASS application

Some Results

Table: CV for Corn Acreage (%)

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JAS</td>
<td>GLS</td>
</tr>
<tr>
<td>CA</td>
<td>12.6</td>
<td>8.1</td>
</tr>
<tr>
<td>IN</td>
<td>5.7</td>
<td>2.0</td>
</tr>
<tr>
<td>NE</td>
<td>4.9</td>
<td>1.7</td>
</tr>
<tr>
<td>PN</td>
<td>9.6</td>
<td>3.3</td>
</tr>
</tbody>
</table>
5. Discussion

- Two error models: sampling error model and structural error model.
- GLS method provides a useful tool for combining two models.
- Does not rely on parametric distributional assumptions.
- Requires correct specification of the variance-covariance matrix for optimal estimation.
- Simpler form of covariance matrix can be used for computational simplicity over statistical efficiency.
REFERENCES


