Estimation of Measurement Variance for Unbalanced Data in Surveys

Roshanak Alimohammadi

Department of Mathematics
Alzahra University
Tehran, Iran
r_alimohammadi@alzahra.ac.ir

Abstract

Survey implementation is one of the common methods to data collection. Precision of survey results is of prime importance due to their applications in decision making and planning. To determine precision in surveys, it is necessary to estimate sampling and nonsampling variances. Measurement error is an unavoidable part of nonsampling error and it is considered as a main part of errors. In this paper, a typical model of measurement error is considered and based on the model, estimators of model components are determined for unbalanced data sets. Furthermore, estimation of measurement variance is assessed and the relations for determining measurement variance of some linear statistics are presented under some assumptions.

Mathematics Subject Classification: 62P99

Keywords: Measurement Variance, Measurement Error Model, Precision, Data Quality, Unbalanced Data

1 Introduction

Successful planning and decision making requires statistical information with high quality. precision is a main concept related to data accuracy. To evaluate accuracy of survey results, it is necessary to estimate total survey error, that is both sampling and nonsampling errors. Wolter [17] has mentioned diverse methods of sampling variance estimation in complex surveys. Groves [9] has studied nonsampling errors. Groves, Fowler and colleagues [10] have assessed total survey error.

Measurement error is a main part of nonsampling error. To quantify this source of error, a model of measurement error may be applied. In this purpose
many researches are carried out and there are two mainly types of measurement error models in the literature: census bureau model and ANOVA model. In this article, ANOVA method is considered for modeling of measurement error.

2 Measurement Error Model

Census bureau model introduced by Hansen, Hurvitz and Bershad [11] and extended by Fellegi ([7], [8]). Lessler and Kalsbeek [14] have described this model.

Mahalanobis [15] and Sukhatme and Seth [16] were first statisticians who used ANOVA for measurement error modeling. After them, several researchers such as Kish [13], Biemer and Stokes [5] and Biemer and Forsman [4] have extended their works. Biemer and Trewin [6] have considered a general model.

Kish [13] has considered an ANOVA model in which the response from the \( j \)th unit to the \( i \)th interviewer is expressed as:

\[
y_{ij} = \mu_{ij} + A_i + e_{ij}
\]

Where \( \mu_{ij} \) is the true value of \( ij \)th unit, \( A_i \) is the effect of \( i \)th interviewer on any interview and may be considered as containing other sources of error. Hartley and Rao [12] have proposed the following model:

\[
y_{ps} = \mu_{ps} + A_i + C_c + \delta A_{ps} + \delta C_{ps}
\]

Where the estimated value (\( y_{ps} \)), can be obtained as a function of true value, interviewer error, coder error, interviewer-observation interaction and coder-observation interaction respectively. Bassie and Fabbris [3] have presented a measurement error model including supervisor, interviewer and respondent effects, as follows:

\[
y_{hi} = \mu_{hi} + S_h + A_{hi} + B_{hi} + d_{ij}
\]

Where estimated value is computed as a function of the true value, supervisor error, interviewer error and respondent error (respectively) in a supervised interview-reinterview setting.

Biemer and Trewin [6] have considered two general models for measurement error, one for continuous data and one for binary data. Here the first type is described. Based on their notation, \( U = \{1, 2, \ldots, N\} \) is a label set for the target population containing \( N \) units and \( S = \{1, 2, \ldots, n\} \) denotes the label of sample units so that \( n=mI \). It is assumed that \( S \) is partitioned into \( I \) assignments of \( m=n/I \) units. \( S_i \) is the set of units assigned to the \( i \)th operator and \( d_{ij} \) is the error of \( j \)th unit in \( S_i \) for \( j = 1, \ldots, m \) and \( i = 1, \ldots, I \). For \( j \in S_i \), it is assumed that \( d_{ij} \) is sum of two error terms, \( A_i \) and \( B_{ij} \), where \( A_i \) is an operator error which is assumed to be same for all units in the \( i \)th operator.
Estimation of measurement variance

assignment and $B_{ij}$ is the elementary error due to respondent as well as other sources of error including the operator. Thus, the model for $ij$th observation is:

$$y_{ij} = \mu_{ij} + d_{ij} = \mu_{ij} + A_i + B_{ij}$$

(1)

Where $A_i$ can be considered as fixed or random and $B_{ij}$ as random variables.

Ayhan [2] has proposed the following model for the measurement error components in supervised interview-reinterview surveys:

$$y_{hkirc} = \mu_{hkirc} + K_{hk} + A_{hki} + B_{hkir} + C_{hkirc}$$

Where the estimated value ($y_{hkirc}$) is a function of true value, controller error ($K_{hk}$), interviewer error ($A_{hki}$), respondent error ($B_{hkir}$) and coder error ($C_{hkirc}$). Groves, Fowler and colleagues [10] have presented some recommendations for reducing interviewer variance.

Alimohammadi and Navvabpour [1] proposed a model of measurement error in face to face surveys as follows:

$$y_{ijkl} = \mu_{ijkl} + A_i + B_{ij} + C_{ijk} + D_{ijkl} + R_{ijkl}$$

(2)

Where $y_{ijkl}$ is the observed value and $\mu_{ijkl}$ is the real value of $st$ unit of $l$th interviewer for $k$th controller related to $j$th supervisor in the $i$th state. $A_i$ is the effect of states, $B_{ij}$ is effect of the related supervisor, $C_{ijk}$ indicates controller (and coder) effect, $D_{ijkl}$ as interviewer effect and $R_{ijkl}$ is effect of respondent.

Index $i = \{1, 2, ..., I\}$ is applied for states, $j = \{1, 2, ..., J_i\}$ for supervisor, $k = \{1, 2, ..., K_{ij}\}$ for controller, $l = \{1, 2, ..., L_{ijk}\}$ for interviewer and $s = \{1, 2, ..., S_{ijkl}\}$ is index of respondents.

In model (2), effect of respondents nested in interviewers, interviewer effect in controller effect, controller effect in supervisor and all of the effects nested in state effect. Usually in statistical surveys, effects $A$, $B$ and $C$ are fixed and $D$ and $R$ are random variables. Then model (2) is considered as a nested mixed effect ANOVA model.

3 Estimation of Variance Components

In this paper, estimation of measurement variance for unbalanced data is assessed under some assumptions. In this purpose, variance components of a typical measurement error model is assessed for unbalanced data. To quantify measurement error in surveys, a typical model of measurement error is considered as follows:

$$y_{ij} = \mu_{ij} + \alpha_i + \beta_{ij}$$

(3)
Where $\mu_{ij}$ is real value and $y_{ij}$ is observed value of $ij$th unit, $A_i$ is effect of $i$th interviewer and $\beta_{ij}$ is effect of $j$th respondent related to $i$th interviewer. (This component may be contain the other sources of errors. However, if the other sources of errors are ignorable, respondent error can be estimated correctly. Otherwise, for estimating respondent effect, interviews should be done several times. Often in the surveys, repeating the interviews is not possible.) Model (3) is a random effect ANOVA model.

Index $i = \{1, 2, ..., a\}$ is applied for interviewers, $j = \{1, 2, ..., N_i\}$ is index of respondents related to $i$th interviewer and $N_i$ is workload of $i$th interviewer.

To estimate the model effects by ANOVA estimation method, ANOVA table of model (3) is obtained as follows:

<table>
<thead>
<tr>
<th>Var source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$a - 1$</td>
<td>$\sum_{i=1}^{a} N_i (\bar{y}<em>i - \bar{y}</em>.)^2$</td>
<td>$(N - \sum_{i=1}^{a} \frac{N_i^2}{N}) \sigma_A^2 + \sigma_M^2 + \sigma_B^2$</td>
</tr>
<tr>
<td>B</td>
<td>$(N - a)$</td>
<td>$\sum_{i=1}^{a} \sum_{j=1}^{N_i} (\bar{y}_{ij} - \bar{y}_i)^2$</td>
<td>$\sigma_B^2 + \sigma_M^2$</td>
</tr>
</tbody>
</table>

Where $N$ is the sample size (the number of total respondents) and $a$ is the number of interviewers. By replacing $z = \sigma_B^2 + \sigma_M^2$, it can be shown that ANOVA estimators of variance components $z$ and $\sigma_A^2$ are obtained as follows:

$$\hat{z} = MSB$$

and for $MSA \geq MSB$:

$$\hat{\sigma}_A^2 = \frac{MSA - MSB}{(N - \sum_{i=1}^{a} \frac{N_i^2}{N})/(a - 1)}$$

If $MSD < MSB$, $\hat{\sigma}_A^2$ is not reasonable. Then for such data sets, $\hat{\sigma}_A^2 = 0$ is considered. Therefore, if $MSD < MSB$:

$$\hat{z} = MSB$$

$$\hat{\sigma}_A^2 = 0$$

In this section, variance components of model (3) are obtained for unbalanced data.

In the next section, measurement variance of linear estimators are presented based on the results of this section.

4 Estimation of Measurement Variance for Unbalanced Data

To estimate measurement variance for linear statistics, measurement variance for Mean and Sum estimators are assessed and the relations are obtained based on estimators of variance components in section 3.
By considering model (3) and the ANOVA table, Mean Squares of effect A (MSA) equals to \( MSA = \frac{SSA}{a-1} \), where SSA is Sum of Squares of effect A. \( MSB = \frac{SSB}{N-a} \) is Mean Squares of effect B where SSB is Sum of Squares of effect B.

Multi-stage cluster sampling designs are usually apply in surveys. Therefore, in this section measurement variance of linear estimators for two-stage cluster sampling design is considered (as a useful sampling design in surveys), \( Var_M(\bar{y}) \) can be written as:

\[
Var_M(\bar{y}) = \frac{1}{n} [Var(\sum_{b=1}^{m} \frac{1}{m} y_{ab})] = \frac{1}{n^2} [Var(\sum_{a=1}^{n} \bar{y}_{am})]
\]

Since:

\[
Var(\bar{y}_{am}) = Var(\sum_{b=1}^{m} \frac{1}{m} y_{ab})
\]

\[
= \frac{1}{m^2} (\sum_{b=1}^{m} (Var(y_{ab}) + \sum_{b \neq b'} Cov(y_{ab}, y_{ab'}))
\]

\[
= \frac{1}{m^2} [m(\sigma_A^2 + z) + m(m-1)\sigma_A^2]
\]

\[
= \sigma_A^2 + \frac{z}{m}
\]

Where \( n \) is the number of sampled clusters, \( L \) is the number of interviewers, \( h \) is the number of clusters assigned to each interviewer such that \( n=hL \), \( m \) is the number of sampled units from each of the sampled clusters and \( z = \sigma_B^2 + \sigma_M^2 \). Therefore, the estimator of measurement variance is obtained as follows:

\[
Var_M(\bar{y}) = \frac{1}{n} [\sigma_A^2 (h(1 - \frac{1}{m}) + \frac{1}{m}) + \frac{\hat{z}}{m}]
\]  \hfill (4)

Where \( \hat{\sigma}_A^2 \) and \( \hat{z} \) are the ANOVA estimators of \( \sigma_A^2 \) ans \( z \).

As a similar way, measurement variance of Sum estimator (T) can be obtained as:

\[
Var_M(T) = \sum_n \sum_i y_{ij} = Var(\sum_{b=1}^{m} \sum_{a=1}^{n} \frac{1}{m} y_{ab}) = Var(\sum_{a=1}^{n} \bar{y}_{am})
\]

\[
= (\sum_{a=1}^{n} Var(\bar{y}_{am}) + \sum_{a \neq a'} Cov(\bar{y}_{am}, \bar{y}_{a'm}))
\]

\[
= n(\sigma_A^2 + \frac{z}{m}) + n(h-1)(\frac{m-1}{m})\sigma_A^2
\]

\[
= n[\sigma_A^2 (h(1 - \frac{1}{m}) + \frac{1}{m}) + \frac{z}{m}]
\]

By replacing ANOVA estimator of \( \sigma_A^2 \) and \( z \) in relation (4), estimator of measurement variance for mean (\( \bar{y} \)) can be obtained as:
For $MSA \geq MSB$:
\[
\hat{Var}_M(\bar{y}) = \frac{1}{n}[\hat{\sigma}_A^2(h(1 - \frac{1}{m}) + \frac{1}{m} + \frac{\hat{z}}{m})]
\]
\[
= \frac{1}{n}(\frac{(MSA - MSB)}{(N - \frac{\sum_i n_i^2}{N})/(a - 1)}(h(1 - \frac{1}{m}) + \frac{1}{m} + \frac{MSB}{m})
\]

For $MSD < MSR$:
\[
\hat{Var}_M(\bar{y}) = \frac{MSB}{nm}
\]

Furthermore, measurement variance for sum estimator ($T$) is obtained as follows:

For $MSA \geq MSB$:
\[
\hat{Var}_M(T) = n[\hat{\sigma}_A^2(h(1 - \frac{1}{m}) + \frac{1}{m} + \frac{\hat{z}}{m})]
\]
\[
= n[\frac{(MSA - MSB)}{(N - \frac{\sum_i n_i^2}{N})/(a - 1)}(h(1 - \frac{1}{m}) + \frac{1}{m} + \frac{MSB}{m})
\]

For $MSD < MSB$:
\[
\hat{Var}_M(T) = n\frac{MSB}{m}
\]

The proposed relations in this section may be used to determine the proportion of measurement error in survey precision of the mentioned linear estimators.

5 Conclusion

Precision of survey results is of prime importance due to their applications. Measurement error is an essential part of nonsampling error. In this paper, estimation of measurement variance for unbalanced data sets is investigated and the relations are proposed for mean and sum estimators. By applying these relations, variance due to measurement error can be determined for unbalanced data sets. Therefore, application of the paper results is proposed to determine this part of survey precision.

ACKNOWLEDGEMENTS. This research is supported financially by Alzahra University.
References


Received: January, 2011