BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-83/05

RESPONSE ERRORS IN REPEATED SURVEYS WITH DUPLICATED OBSERVATIONS

by

Tin Chiu Chua
Under the supervision of Wayne A. Fuller
From the Department of Statistics
Iowa State University

Recommended by: Myron J. Katzoff

Report issued: July 27, 1983

Response errors in repeated surveys with duplicated observations

Tin Chiu Chua

Under the supervision of Wayne A. Fuller
From the Department of Statistics
Iowa State University

The analysis of response error models for categorical data that form an r x r contingency table is considered. Individuals are placed in the row and column classes on the basis of two interviews. It is assumed that the errors in the row and in the column classifications are independent. It is also assumed that the error in the classification of an individual depends only on the individual's true class. A parametric model for the probability that an individual belonging to the i-th class is classified in the j-th class is proposed.

Reinterview on one of the dimensions is conducted in order to estimate the classification probabilities. Two kinds of reinterview procedures are performed by the U.S. Bureau of the Census in the Current Population Survey. In the first kind, the reinterviewers are not given the original responses. In the second kind, the original responses are given to the reinterviewers and a reconciliation is made after the responses are collected in the reinterview. The Gauss-Newton procedure for the nonlinear model is used to estimate the parameters of the classification model from data collected in the three interviews.

The determination of the optimal number of replicates to observe for the estimation of the simple errors-in-variables model is

considered. It is assumed that the cost of obtaining an observation is the same for every unit. For a fixed total cost, the optimal ratio of the number of units with duplicated observations to the total number of units is obtained by minimizing the variance of the estimator of the slope in the simple linear errors-in-variables regression model. Extension of replicated designs to three observations per unit is considered under the condition that all the units in the sample are observed twice. Tables of optimal designs are constructed for some specific values of the parameters of the model. The optimal design for the case where the observed values are dichotomous is also considered.

TABLE OF CONTENTS

Response errors in repeated surveys with duplicated observations

by

Tin Chiu Chua

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:

In Charge of Major Work

For the Major Department

For the Graduate College

Iowa State University
Ames, Iowa

1983

I. INTRODUCTION

Data that are collected from individuals by personal interview are known to be subject to response error. Nonsampling errors have long been recognized and discussed. In an expository paper on errors in survey samples, Deming (1944) lists 13 different factors that affect the usefulness of surveys. Four of these factors are related to response errors:

(1) Variability in response;

(2) Bias and variation arising from the interviewer;

(3) Imperfections in the design of the questionnaire; and

(4) Processing errors involved in coding, editing and punching of data.

The variability in a respondent's responses in repeated interviews may be due to a lack of understanding of the questions, difficulty in determining his "true-value" for the question, or to the lack of information required to answer correctly. The interviewer may contribute to variability in responses by giving different interpretations to questions and by failing to understand the subject and purpose of the survey.

In a review paper on the effect of the question on survey responses, Kalton and Schuman (1982) discuss several studies which show that the survey responses are sensitive to the precise wording, format

and placement of the questions asked. The nature of question wording and form effects on response errors remains an important area of study.

It is generally assumed that the true values of characteristics under study exist for each individual. For variables such as age, sex, and total income, the definition of the individual true values does not present major problems. However, for variables such as attitude toward social issues and preference for a certain product, it is more difficult to define the individual true values. Hansen, et al. (1951) suggest three criteria for the definition of the true value for an individual:

(1) It must be uniquely defined;

(2) It must be defined in such a manner that the purposes of the survey are met; and

(3) It should be defined in terms of operations which can be carried through, even though it might be difficult or expensive to perform the operations.

For a situation in which survey response for a given individual can be considered as coming from a population of conceptual responses for that individual, it may be appropriate to define the individual true value as the expected response obtained under certain well-defined survey conditions.

In this dissertation, it is assumed that a random sample of n individuals is taken from a population of N individuals and that all or part of the selected individuals are interviewed twice.

In Chapter III, we consider response errors in classificatory problems where individuals are classified in an r x r contingency table. Particular attention is given to the structure of response models for which the sample marginal proportions are unbiased estimators of the corresponding population proportions. The response errors for the response in the row and column classes are assumed to be independent. The response errors in the two responses from the interview-reinterview process on one of the marginal classes are assumed to be either independent or dependent, depending on the interview procedure.

In Chapters IV and V, we consider the problem of determining the optimum (minimum variance) number of replicated observations and unreplicated observations for the estimation of a simple linear model where both the independent and dependent variables are subject to response errors. In Chapter IV, the response errors are assumed to be normally distributed. The case where the observed and true values can take only the values zero and one is treated in Chapter V.

II. REVIEW OF LITERATURE

Response errors, sometimes called measurement errors, have long been recognized as one of the major problems in surveys. The effect of response errors can be quite severe in statistical data analysis. It has been reported that there are interactions between respondents, interviewers and crew leaders which produce correlated meausurement errors, (e.g., Evaluation and research program of the U.S. censuses of population and housing 1960: effects of interviewers and crew leaders, (1968)). The recording of data for processing can also result in errors in the data. Pearson (1902) studied the measuring variability of human beings by conducting two experiments. From the study, Pearson (1902) found that

(1) The mean errors differed significantly from zero;

(2) For a given measurer, the size of the bias varied throughout the series of trials when the errors were grouped in successive sets of 25.

(3) The errors were not, in general, normally distributed; and

(4) The errors of two apparently independent observers in measuring the same quantity were positively correlated.

Cochran (1968) gave a short description of the experiments conducted by Pearson (1902) in a review paper on measurement errors.

Mahalanobis (1946) reported on the survey work of the Indian Statistical Institute, and in particular described efforts to measure

and control reporting errors. Interpenetrating samples were incorporated in crop surveys in order to measure the overall error and to measure the reliability of the enumerators. Sukhatme and Seth (1952) questioned the use of interpenetrating samples as a regular feature of sample surveys. They argued that

(1) The limitation on the size of the samples rendered replicated samples an ineffective tool for detecting discrepancies in field work; and

(2) The cost of replicated samples was very high.

For the case where the nonsampling errors are likely to be large, Sukhatme and Seth (1952) recommended the use of interpenetrating samples only at the pilot stage for improving the questionnaire and the method of training the interviewers, rather than as an integral part of a large-scale survey. They further noted that if nonsampling errors could not be controlled by improving the questionnaire and training to the level of accuracy with which information is desired to be sought, one would hesitate to conduct a sample survey on a probability basis.

Eckler and Pritzker (1951) reported that the U.S. Bureau of the Census attempts to develop programs for measuring the accuracy of all censuses and surveys which it conducted. The technique involved post-enumeration surveys in which more highly trained enumerators were used for the reinterviewing process. These studies led to improvements in the efficiency of the census and survey designs (Eckler and Hurwitz (1958)).

Hansen, et al. (1951) carefully discussed the concepts of response errors. They defined the individual response error as the difference between a sample response and the true value for the individual. The response error had an expected value (individual response bias) and a random component of variation around that expected value. They presented a response model with the following assumptions:

(1) There is a population of N individuals and a population of K interviewers;

(2) There is a true value for each individual; and

(3) There is zero correlation between the random components of responses for two different individuals with two different interviewers.

Under the response model, Hansen, et al. (1951) considered the estimation of the response variance due to interviewers, using survey data obtained from an interpenetrating sample design in which n individuals are randomly assigned to each of k randomly selected interviewers. For this design, the response variance of the individual respondents could not be estimated.

Sukhatme and Seth (1952) discussed a general response model by expressing it as

$$y_{ijk} = x_i + \alpha_j + \delta_{ij} + \varepsilon_{ijk} , \qquad (2.1)$$

where y_{ijk} denotes the sample response obtained by the j-th enumerator ($j = 1,2,\ldots, m$) from the i-th sample respondent ($i = 1,2,\ldots, \ell$) on the k-th occasion ($k = 0,1,2,\ldots, n_{ij}$); x_i denotes the true value for the i-th respondent who is selected randomly from a finite or an infinite population with mean μ and variance σ^2; α_j denotes the effect of the j-th enumerator in the enumeration of many respondents; δ_{ij} denotes the interaction between the j-th enumerator and the i-th respondent and ε_{ijk} denotes the random deviation associated with y_{ijk} that is not accounted for by interviewer and interaction effects. Analysis-of-variance type estimators for (linear combinations of) the variance components in the model were presented for different types of sampling designs: (1) a unit is observed once only, (2) a unit is observed p times by the same enumerator, (3) a unit is observed once by each of p enumerators, and (4) some of the units are observed once and some are observed twice. They also gave separate consideration to the cases where the interviewers were fixed and where they were randomly selected from a larger population of enumerators.

Hanson and Marks (1958) used the method of the analysis of variance to estimate interviewer effects in the enumerator variance study of the 1950 census of population conducted by the U.S. Bureau of the Census. The study was based on the data obtained by 984 interviewers covering 1,778 enumeration districts. They found that the significant interviewer effects were mostly due to (1) a tendency for the interviewer to omit or alter the question involved or to assume the

answer; (2) a relatively high degree of ambiguity, subjectivity, or complexity in the question; (3) a tendency to alter respondent replies because of additional questioning.

Eckler and Hurwitz (1958) reported additional empirical investigations into interviewer effects on the 1950 census of population. The study involved about 700 enumerators in 125 strata with an average population of about 6,500 each. The effect of interviewer variability was measured by comparing the between-enumerator and within-enumerator mean squares. An approximate F-test indicated that the between-enumerator variability was statistically significant on nearly all of the items tested. The study also showed that the interviewer variability was relatively large for a small area, but small for an area that was the responsibility of many interviewers. While these results indicated that there was a possible way of reducing the effect of interviewer variability, Eckler and Hurwitz (1958) warned that attempts to reduce response variability may lead to an increase in biases. For example, resorting to self-enumeration in order to eliminate the effect of enumerator variability may result in many respondents misinterpreting the question of the questionnaire and giving results that are biased.

Hansen, Hurwitz and Bershad (1961) presented a summary of the conceptual ideas and response model formulations that have evolved in the U.S. Bureau of the Census. They presented their response model in the context of estimation of the proportion of individuals that belong to a given class of a finite population. The model has been discussed

and applied in several publications including Pritzker and Hanson (1962), Hansen, Hurwitz and Pritzker (1964), Bailar (1968), the U.S. Bureau of the Census (1968, 1972), and Bailar and Dalenius (1969).

Hansen, Hurwitz and Bershad (1961) assumed that a survey was conceptually repeatable under the same general conditions and that the responses from sample individuals were described by some (unknown) probability distribution. An observation on the j-th unit in the survey is designated by x_{jt}, where x_{jt} has the value 1 if the j-th unit is assigned to the particular class under consideration on the t-th trial, and has the value zero otherwise. An estimate of the population mean is

$$p_t = \frac{1}{n} \sum_{j=1}^{n} x_{jt} , \tag{2.2}$$

where n is the number of units in the sample.

The variance of p_t is

$$\mathrm{Var}(p_t) = E(p_t - \bar{P})^2 + 2E(p_t - \bar{P})(\bar{P} - P) + E(\bar{P} - P)^2 , \tag{2.3}$$

where $P = E(p_t)$, $P_j = E(x_{jt}|j)$ and $\bar{P} = \frac{1}{n} \sum_{j=1}^{n} P_j$.

The first term in (2.3) is defined as the response variance which can be expressed as

$$\sigma^2_{\bar{d}_t} = E\{(p_t - \bar{P})^2\} = E\{(\frac{1}{n}\sum_{j=1}^{n} d_{jt})^2\} = E\{(\bar{d}_t)^2\} ,$$

(2.4)

where $d_{jt} = x_{jt} - P_j$ is the response deviation.

The third term in (2.3) is defined to be the sampling variance of p_t and the second term in (2.3) is twice the covariance of $\bar{d}_t$ and $\bar{P}$. The second term is considered to be trivial in Hansen, Hurwitz and Bershad's (1961) discussion.

The response variance $\sigma^2_{\bar{d}_t}$ can be expressed as

$$\sigma^2_{\bar{d}_t} = \frac{1}{n}\sigma_d^2[1 + \rho(n - 1)] , \qquad (2.5)$$

where $\sigma_d^2 = E(d_{jt}^2) = \frac{1}{N}\sum_{j=1}^{N} P_j(1 - P_j)$ is the simple response variance and $\rho = E(d_{jt}d_{kt})/\sigma_d^2$ for $j \neq k$ is the intraclass correlation among the response deviations in a trial.

Hansen, Hurwitz and Bershad (1961) found that the impact of even a very small intraclass correlation was substantial when the sample size n was quite large. This can be seen from an examination of (2.5).

Two methods were suggested by Hansen, Hurwitz and Bershad (1961) for estimating the response variance. The replication method repeats the survey procedure on the same sample. The method of interpenetrating samples divides the sample randomly into several subsamples with each interviewer assigned to one of the subsamples.

Hansen, Hurwitz and Pritzker (1964) defined the index of inconsistency as the ratio of the simple responses variance to the total variance of individual responses; that is,

$$I = \sigma_d^2/\sigma_{p_t}^2 , \qquad (2.6)$$

where $\sigma_{p_t}^2 = \mathrm{Var}(p_t)$. For a binomial random variable, the total variance $\sigma_{p_t}^2$ is $P(1 - P)$, where P is the expectation of the sample mean.

The response model defined by Fellegi (1964) was similar to that of Hansen, Hurwitz and Bershad (1961). His sampling design, however, involved both interpenetration and replication. He represented the assignments for the two surveys by $\{(S_{i(1)}, S_{i(2)}) , i = 1,2,\ldots, k\}$, where $S_{i(1)}$ and $S_{i(2)}$ denote the interview assignments for the i-th enumerator in the original and reinterview surveys, respectively. For any given interviewer, $S_{i(1)}$ and $S_{i(2)}$ are not the same.

In evaluating the reinterview procedures, Bailar (1968) followed the response model developed by Hansen, Hurwitz and Bershad (1961) to study the effect of the time lag between the census or survey and reinterview and the effect of the reinterviewers having access to the original responses. By comparing estimates of the simple response variance and estimates of the bias for several characteristics from three samples of the 1960 Census enumerated population, Bailar (1968) concluded that the best procedure was one in which the reinterview was

relatively close in time to the original interview and one in which the reinterview did not have access to the original responses.

Bailar and Dalenius (1969) presented the statistical theory and methods for measuring the contribution of response variability to the overall error of a survey. The method of replication and the method of interpenetration in the sample dimension are considered. In the trial dimension, Bailar and Dalenius (1969) considered cases where the same enumerator was used in all the trials or different enumerators were used in different trials. Different sampling schemes were discussed for estimating the response variance and the correlated component. The choice of a sampling scheme was decided by the following factors:

(1) The variance components that are to be estimated;

(2) The cost of a survey; and

(3) The change of the general conditions of a survey due to the time lage between trials.

Bailar (1976) reported that a study of the components of error might lead to methods of improving the accuracy and reliability of survey data. Suppose that one of the purposes of a survey is to estimate a mean, $\bar{X}$, and the data are to be collected by k interviewers, each with a random assignment of n sample units. Simple random sampling is used. By ignoring the finite population correction factors, the mean square error of the sample mean, $\bar{x}$, can be expressed as

$$MSE(\bar{x}) = \frac{\sigma_s^2}{kn} + \frac{\sigma_R^2}{kn}[1 + (n - 1)\rho_R] + \frac{2(n - 1)}{kn}\sigma_{RS} + B^2 \quad ,(2.7)$$

where $(kn)^{-1}\sigma_s^2$ is the sampling variance; $(kn)^{-1}\sigma_R^2$ is the simple response variance; $(kn)^{-1}\rho_R\sigma_R^2(n - 1)$ is the variability caused by the correlation between response deviation of elements in the sample; $2(kn)^{-1}(n - 1)\sigma_{RS}$ is the covariance of response and sampling deviations of different units; and B is the bias of $\bar{x}$.

Equation (2.7) shows that the correlated component of response variance decreases directly as the number of interviewers increases, but not as the number of sampling units within an interviewer's assignment increases. In this way, it is different from the sampling variance. Thus, the correlated component of response variance may be larger than the sampling variance. Bailar (1976) reported that a 1950 study of enumerator variance showed that for areas of 6,500 persons, this component of total variance for a complete census by direct enumeration was at about the same level as a sampling variance for an estimate based on self-enumeration for a 25 percent sample of the population. The results were one reason why the Census Bureau turned to the use of self-enumeration techniques in the 1960 census.

Bryson (1965) studied the effect of misclassification on the bias of the sample proportion in the estimate of the population proportion when the item in a sample is from a binomial population. The upper and lower bounds for the bias were derived based on assumptions regarding magnitudes of the probability of misclassification when each of the two

interviewers independently classified the items in a single sample. Let the minimum values of p_{12} and p_{22} be denoted by $\underline{p}_{12}$ and $\underline{p}_{22}$, respectively, where p_{12} is the probability that an item is classified in class A' by the first interviewer given that it is in A' and p_{22} is the probability that an item is classified in class A' by the second interviewer given that it is in class A'. Bryson (1965) obtained the following inequality

$$\overline{\%\ \text{Bias}} < \frac{E(\frac{x+y}{2}) + \frac{E(x)E(y)}{\underline{p}_{12}\underline{p}_{22}E(x+y+z)}}{E(w) - \frac{E(x)E(y)}{\underline{p}_{12}\underline{p}_{22}E(x+y+z)}} \times 100 , \qquad (2.8)$$

where $\overline{\%\ \text{Bias}}$ is the upper bound of the ratio of the bias of the sample proportion in class A to the population proportion in class A; w, x, y and z are the proportions of the sample that are classified in class A by the first interviewer and in class A by the second interviewer, in class A' by the first interviewer and in class A by the second interviewer, in class A by the first interviewer and in class A' by the second interviewer, and in class A' by both interviewers, respectively. The inequality for $\underline{\%\ \text{Bias}}$, the lower bound of % Bias, is

$$\underline{\%\ \text{Bias}} > -\frac{E\left(\frac{x+y}{2}\right) + \frac{E(x)E(y)}{p_{11}p_{21}E(w+x+y)}}{E(w+x+y) + \frac{E(x)E(y)}{p_{11}p_{21}E(w+x+y)}} \times 100 , \qquad (2.9)$$

where p_{11} and p_{21} are the probabilities that an item is classified in class A by the first interviewer and the second interviewer, respectively, given that it is in class A and p_{11} and p_{21} are the minimum value that p_{11} and p_{21} can take, respectively. When the sample size is sufficiently large, the expected values can be replaced by the observed values. Krishnaswami and Nath (1968) extended the results to the multinomial population.

The methods of analysis of variance have been used by several authors to estimate the variance component associated with enumerators. Examples are Eckler and Hurwitz (1958), Hanson and Marks (1958), Kish (1962), Stock and Hochstim (1951). Battese, Fuller and Hickman (1976) considered a simple components-of-variance model involving enumerator effects, sampling deviations and respondent-response errors. Battese, Fuller and Hickman (1976) assumed that a simple random sample of $rm(m-1)$ respondents was chosen from the population of interest and m enumerators were randomly selected from a large pool of available enumerators. The sample respondents were randomly divided into $m(m-1)$ groups, each of r respondents. The i-th enumerator interviewed $(m-1)$ respondent groups and reinterviewed another $(m-1)$ respondent groups that were first interviewed by the

j-th enumerator for $j = 1,2,\ldots, m$, $j \neq i$. This interpenetrating and replicated survey design was assumed to be applied to several strata of the population of interest.

Battese, Fuller and Hickman (1976) expressed the model as

$$Y_{ik1} = y_k + \beta_i + \varepsilon_{ik1}, \quad k = 1,2,\ldots, r$$

$$Y_{jk2} = y_k + \beta_j + \varepsilon_{jk2}, \quad k = 1,2,\ldots, r, \qquad (2.10)$$

where Y_{ik1} denotes the response of the k-th respondent interviewed by the i-th enumerator at time 1 and Y_{jk2} is the response of the k-th respondent interviewed by the j-th enumerator at time 2; y_k denotes the true value for the k-th respondent; β_i denotes the random effect of the i-th enumerator; ε_{ik1} and ε_{jk2} denote the respondent-response errors that are associated with the interview and reinterview responses, respectively. They also assumed that β_i and ε_{ikt}, $t = 1,2$, are independently distributed with zero means and variances σ^2_β and $\sigma^2_{\varepsilon_k}$, respectively; that β_i and ε_{ikt} are uncorrelated with the true values; and the true value, y_k, is equal to the sum of a stratum mean, μ, and a "sampling deviation" e_k. The sampling deviations for all individuals in the population are assumed to have zero mean and variance σ^2_e. The response errors, ε_{ikt} and the sampling deviation, e_k are assumed to have finite fourth moments.

Using a least-squares regression procedure, Battese, Fuller and Hickman (1976) obtained the estimators for the variances of enumerator effects, for the variance of the sampling deviation and for the average of the respondents response variances.

Hartley and Rao (1978) assumed that

(1) the survey is of a stratified multistage design in which the last stage units are drawn with equal probabilities;

(2) the errors are additive;

(3) all correlations between the errors contributed by a particular error source are generated through an additive model; and

(4) there is no systematic bias from any of the error sources.

They expressed the model in the form

$$y_{ps} = \eta_{ps} + b_i + c_c + \delta b_{ps} + \delta c_{ps} , \qquad (2.11)$$

where the index s labels the s-th elementary unit; the index p is a composite label indexing the last but one stage unit within the next higher stage unit ... within a primary unit within a stratum; y_{ps} is the recorded observation for the elementary unit labeled (p, s) ; η_{ps} is the true content for elementary unit labeled (p, s) ; b_i is the error contributed by the i-th interviewer common to all units interviewed by the i-th interviewer; c_c is the error contributed by the c-th coder common to all units coded by the c-th order; δb_{ps} is

the elementary interviewer error for the content item of unit (p, s) and δc_{ps} is the elementary coder error for the content item of unit (p, s). They assumed that b_i and c_c are random samples from an infinite population of interviewer and coder errors with means zero and variances σ_b^2 and σ_c^2, respectively. Also, δb_{ps} and δc_{ps} are assumed to have means zero and variances $\sigma_{\delta b}^2$ and $\sigma_{\delta c}^2$, respectively. The b_i and c_c are assumed to be independent of one another and independent of the η_{ps}, δb_{ps} and δc_{ps}. No restriction is applied on η_{ps}, δb_{ps} and δc_{ps}.

Using the simple mixed model ANOVA techniques, Hartley and Rao (1978) provided a method of estimating the overall variance of a linear estimator of the form $c'(p)\bar{y}$, where $\bar{y}$ is the vector of primary-sample means $\bar{y}_p$ and the coefficient vector $c(p)$ depends on the set of selected primaries p.

Bross (1954) discussed the effect of misclassification on testing the hypothesis that the proportions of two independent populations were equal. Under the assumption that the same classification system was used in both samples, Bross (1954) found that the size of the ordinary chi-square test was not affected by ignoring misclassification, but the power of the test was drastically reduced.

Mote and Anderson (1965) considered two simple response models in an investigation of the effect of misclassification on the size and power of chi-square goodness-of-fit tests for categorical data. The first model assumed equal probabilities of misclassification into one

of r available categories. The second response model assumed that there were only misclassifications in classes adjoining the true classes to which individuals belong. Mote and Anderson (1965) showed that, with these response models, hypothesis tests concerning the class proportions that ignored classification errors had greater size and smaller power than tests that were modified to account for classification errors.

Assakul and Proctor (1967) considered two cases of the effect of misclassification on the test of independence in a two-way contingency table. When errors of classification in the row direction were independent of those in the column variable, Assakul and Proctor (1967) found that the usual chi-square test had the announced level of significance, but the power of the test was smaller. When the errors for the marginals were not independent and under the assumption that the misclassification probabilities were known, Assakul and Proctor (1967) proposed a test criterion.

Koch (1969) studied the effects of nonsampling errors on measures of association in a 2 x 2 contingency table under the model due to the U.S. Bureau of the Census. The sample estimate for a measure of association was expressed in the form of a Taylor series approximation involving cell probabilities. Then, the response model was applied in a term by term fashion. The relative effects of sampling errors and response errors on the variability of the estimated measure of association could be interpreted in terms of a sampling variance component and a response variance component.

Fleiss (1981) discusses the effect of classification errors on the estimation of population proportions. A clinical trial example is given. Methods of controlling and measuring the classification errors are also presented.

The study of Mote and Anderson (1965) is extended by Korn (1981) to contingency tables of dimension greater than two. Korn (1981) studied the effect of the classification errors on the analysis of hierarchical log-linear models. It is assumed that

(1) For each dimension of the table, the conditional probability that an individual is observed with error at a particular level of that dimension, given its true level of that dimension, does not depend on the true levels of that individual in the other dimensions of the table.

(2) Given its true levels in all the dimensions of the table, the conditional probability that an observation is misclassified into a certain level of a certain dimension is independent of whether that observation was misclassified in the other dimensions of the table.

In an $I \times J \times K$ contingency table, let π_{ijk} be the probability a randomly chosen individual from a large population would be classified into cell (ijk) of the table if observed with no classification error. Let τ_{ijk} be the probability an individual is classified into cell (ijk) with classification error. Then

$$\tau_{ijk} = \sum_{i'j'k'} q_{1(ii')} q_{2(jj')} q_{3(kk')} \pi_{i'j'k'} , \qquad (2.12)$$

where $q_{\ell(ii')}$ is the conditional probability that an individual is observed in level i of dimension ℓ given that its true level is i'. The fully saturated model for an $I \times J \times K$ table is specified by

$$\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \qquad (2.13)$$

subject to the usual ANOVA-like constraints. Hierarchical log-linear models postulate certain u terms in (2.13) to be identically zero with the condition that the lower order relatives of every u term present in the model are also present in the model. A model is said to be preserved by classification error in dimension ℓ if when π satisfies the model, τ does also.

Korn (1981) shows that a hierarchical log-linear model is preserved by classification error in dimension ℓ of the table if and only if the minimal set contains exactly one u term having an ℓ as a subscript where the minimal set of u terms for a hierarchical log-linear model is defined to be the set of u terms such that the model is specified by all the lower order relatives of this set.

Korn (1982) provides an expression for the approximate upper bound of the asymptotic relative efficiency of tests between nested log-linear models using misclassified data versus those using data with no classification errors. This efficiency depends on the probabilities of

data being misclassified into the wrong classes of the contingency table. It is shown that the loss of efficiency due to misclassification can be substantial.

Giesbrecht (1957) considered the classification of individuals into the four groups defined by the presence or absence of two attributes. The presence (or absence) of the first attribute is denoted by A (or $\bar{A}$), and the presence (or absence) of the second attribute is denoted by B (or $\bar{B}$). The four classes involved are denoted by AB, $\bar{A}B$, $A\bar{B}$ and $\overline{AB}$.

For this four-class situation, Giesbrecht (1967) defined ten conditional probabilites from which the probabilities of classification are obtained for each of the four columns. By use of the abbreviation "ac" for actual classification and "tc" for true classification, the conditional probabilities that were defined are

$$\beta_1 = \Pr(\text{ac is B} \mid \text{tc is B})$$

$$\beta_0 = \Pr(\text{ac is } \bar{B} \mid \text{tc is } \bar{B})$$

$$\alpha_{11} = \Pr(\text{ac is AB} \mid \text{tc is AB and ac is B})$$

$$\alpha_{01} = \text{PR}(\text{ac is } \bar{A}B \mid \text{tc is } \bar{A}B \text{ and ac is B})$$

$$\alpha_{10} = \Pr(\text{ac is AB} \mid \text{tc is } A\bar{B} \text{ and ac is B})$$

$$\alpha_{00} = \Pr(\text{ac is } \bar{A}\bar{B} \mid \text{tc is } \overline{AB} \text{ and ac is } B)$$

$$\nu_{11} = \Pr(\text{ac is } A\bar{B} \mid \text{tc is } AB \text{ and ac is } \bar{B})$$

$$\nu_{01} = \Pr(\text{ac is } \overline{AB} \mid \text{tc is } \bar{A}B \text{ and ac is } \bar{B})$$

$$\nu_{10} = \Pr(\text{ac is } A\bar{B} \mid \text{tc is } A\bar{B} \text{ and ac is } \bar{B})$$

$$\nu_{00} = \Pr(\text{ac is } \overline{AB} \mid \text{tc is } \overline{AB} \text{ and ac is } \bar{B}) \; . \qquad (2.14)$$

The ten conditional probabilities defined by Giesbrecht (1967) do not represent the most general response model for the two-attribute case. To obtain the probabilities of assigned classifications for each of the true classes, the four probabilities $\Pr(\text{ac is } B \mid \text{tc is } AB)$, $\Pr(\text{ac is } B \mid \text{tc is } \bar{A}B)$, $\Pr(\text{ac is } \bar{B} \mid \text{tc is } A\bar{B})$ and $\Pr(\text{ac is } \bar{B} \mid \text{tc is } \overline{AB})$, need to be defined. Giesbrecht implicity assumes that

$$\beta_1 = \Pr(\text{ac is } B \mid \text{tc is } AB) = \Pr(\text{ac is } B \mid \text{tc is } \bar{A}B)$$

and

$$\beta_0 = \Pr(\text{ac is } \bar{B} \mid \text{tc is } A\bar{B}) = \Pr(\text{ac is } \bar{B} \mid \text{tc is } \overline{AB}) \; . \qquad (2.15)$$

These assumptions reduce the number of independent classification probabilities from twelve to ten.

However, even with Giesbrecht's reduction of parameters from fifteen to thirteen, the response model cannot be estimated from an experiment with independent classification of sample individuals at two trials of a survey.

Bershad (1967) studied the effect of response errors on "gross change" tables under a simple response errors model. The assumptions of his model are:

(1) At any point in time, each individual in the population belongs to one of the two groups, U or $\bar{U}$;

(2) Individuals in a sample are classified into the two classes such that different classifications are (i) independent of one another; and (ii) dependent only on the true status of the individual at the time of classification;

(3) The sample proportion of group U is an unbiased estimator for the true proportion of group U at that time; and

(4) The proportion of group U in the population is the same in the two months considered.

Under these assumptions, Bershad (1967) showed that the expected proportion classified in group U in the first month but in group $\bar{U}$ in the second month, a_{12}, was not equal to the true proportion, A_{12}. The relationship between the true proportion, A_{12}, and the expected proportion a_{12}, is given by

$$A_{12} = [a_{12} - P(1 - P)I]/(1 - I) , \qquad (2.16)$$

where P is the proportion in group U in a given month; and I is the index of inconsistency.

Koop (1974) considered a linear estimator of the form

$$T(s) = \sum_{i \epsilon s} \beta(i, s) x_{it} \qquad (2.17)$$

for estimating the population total subject to response errors where s is a selected sample and x_{it} is the response of the i-th unit at trial t. When x_{it} is free from error, it is well-known that the estimator $T(s)$ will be unbiased for all x if

$$\sum_{\{s \,:\, i \epsilon s\}} \beta(i, s)p(s) = 1 \quad \text{for} \quad i = 1,2,\ldots, N . \qquad (2.18)$$

Koop (1974) showed that a linear estimator with coefficients, $\beta(i, s)$, satisfying (2.18) and having the least mean square error did not exist except for the uni-cluster design. He also showed that the estimators of the variance of linear estimators given by standard theory were always negatively biased.

Battese and Fuller (1974) obtained estimates of the response probabilities from categorical data by assuming an unbiased response model. A response model is said to be unbiased if the expected value of the sample proportion is equal to the population proportion. Battese and Fuller (1974) suggested a model for the response probability β_{ij},

where β_{ij} is the probability that a randomly selected individual belonging to the j-th class is classified in the i-th class. The Battese-Fuller model is

$$\beta_{ij} = 1 - \alpha + \alpha P_i \quad , \quad i = j$$

$$= \alpha P_i \quad , \quad i \neq j \quad , \tag{2.19}$$

where P_i is the population proportion of the i-th class. In this model, the probability of incorrect response depends upon the true probabilities and upon the parameter α.

With each individual classified twice and assuming that the first and the second classification are independent, Battese and Fuller (1974) show that the expectation of p_{ij} is a nonlinear function of α and of the P_i's, where p_{ij} is the proportion of the sample which is classified in class i at the first trial and in class j at the second trial. Using the Gauss-Newton method of nonlinear estimation, Battese and Fuller (1974) obtained estimators of α and of the P_i's and the asymptotic properties of the estimators.

III. A RESPONSE MODEL FOR CATEGORICAL DATA CLASSIFIED IN A TWO-WAY TABLE

A. Introduction

It is assumed that each individual in a sampled population belongs to one of a set of r^2 classes. Let P_{ij} be the proportion in the i-th row class and j-th column class of the population. Let $P_{i.}$ and $P_{.j}$ be the marginal proportions for the i-th row class and j-th column class, respectively. Thus,

$$P_{i.} = \sum_{j=1}^{r} P_{ij}$$

and

$$P_{.j} = \sum_{i=1}^{r} P_{ij} . \qquad (3.1)$$

Assume that a sample of size n is selected and is interviewed twice. On the basis of the two interviews, the n individuals are classified into one of the r^2 classes. In the first interview, called 'Trial-1,' individuals are placed in the r row classes and in the second interview, called 'Trial-2,' individuals are classified into the r column classes. The sample classification and the true classification are not necessarily the same.

It is assumed that the probabilities of classification depend on the true classes to which the individuals belong and are characterized

by the classification probabilities $\gamma_{k(i)}$ and $\kappa_{\ell(j)}$, $i,j,k,\ell = 1,\ldots, r$, where $\gamma_{k(i)}$ is the probability that an individual belonging to the i-th row class is classified into the k-th row class and $\kappa_{\ell(j)}$ is the probability that an individual belonging to the j-th column class is classified in the ℓ-th column class. Because all individuals are placed in one of the classes, it follows that $\sum_{k=1}^{r} \gamma_{k(i)} = 1$ and $\sum_{\ell=1}^{r} \kappa_{\ell(j)} = 1$ for all $i,j = 1,2,\ldots, r$. It is also assumed that the classifications on the two trials are independent.

If the classification probabilities $\gamma_{k(i)}$ and $\kappa_{\ell(j)}$ are known, then an unbiased estimator for the P_{ij} can be obtained from the two trials survey. Let p_{ij} be the sample proportion of i-th row class and j-th column class. We have

$$E(p_{ij}) = \sum_{\ell=1}^{r} \sum_{k=1}^{r} \gamma_{i(k)} \kappa_{j(\ell)} P_{k\ell} \qquad (3.2)$$

Let

$$\underset{\sim}{p}' = (p_{11}, p_{21},\ldots, p_{r1},\ldots, p_{1r}, p_{2r},\ldots, p_{rr})$$

and

$$\underset{\sim}{P}' = (P_{11}, P_{21},\ldots, P_{r1},\ldots, P_{1r}, P_{2r},\ldots, P_{rr}) \ . \qquad (3.3)$$

Let $\underset{\sim}{\Gamma}$ and $\underset{\sim}{K}$ be two r by r matrices such that their (i,j)-th elements are $\gamma_{i(j)}$ and $\kappa_{i(j)}$, respectively. Then Equation (3.2) can be expressed as

$$E(\underset{\sim}{p}) = (\underset{\sim}{K} \blacksquare \underset{\sim}{\Gamma})\underset{\sim}{P} , \tag{3.4}$$

where $\blacksquare$ is the Kronecker product. If the inverse of $\underset{\sim}{K} \blacksquare \underset{\sim}{\Gamma}$ exists, an unbiased estimator for $\underset{\sim}{P}$ is

$$\underset{\sim}{P} = (\underset{\sim}{K} \blacksquare \underset{\sim}{\Gamma})^{-1} \underset{\sim}{p}$$

$$= (\underset{\sim}{K}^{-1} \blacksquare \underset{\sim}{\Gamma}^{-1})\underset{\sim}{p} . \tag{3.5}$$

In most cases, the classification probabilities are unknown. Thus, a reinterview procedure is incorporated into the survey procedure to study the classification errors. Individuals in the sample are classified in the r column classes by a reinterviewer. No original interviewer is used to reinterview his/her own interview cases. The reinterview is called 'Trial-3.'

Two kinds of reinterview processes are conducted by the U.S. Census Bureau in the reinterview program of the Current Population Survey. In the first, the reinterview is conducted with no reference to the original responses. In this case, the classification is characterized by the classification probabilities $\nu_{\ell(j)}$, where $\nu_{\ell(j)}$,

$\ell,j = 1,2,\ldots, r$, is the probability that an individual belonging to the j-th column class is placed in the ℓ-th column class in the reinterview. In this case, the classification in 'Trial-3' is assumed to be independent of the previous two trials and the data collected from the three trials are called unreconciled data.

In the second type of reinterview process, reinterviewers are given the original responses and are instructed to consult them after a first reinterview response has been given. Reconciliation is done on a separate form containing the original responses. In this case, the original interview classification and the reconciled reinterview classification are not independent. The data collected from the three trials survey are called reconciled data. A suggested model for the probability that an individual is classified in the t-th column class by the reconciled reinterview, given that the individual is in the j-th true column class and is classified in the ℓ-th column class on the original interview is

$$\omega_{t(j\ell)} = \phi + (1 - \phi)\tau_{t(j)} \quad , \quad \ell = t$$

$$= (1 - \phi)\tau_{t(j)} \quad , \quad \ell \neq t, \quad j,\ell,t = 1,2,\ldots, r \ ,$$

(3.6)

where $0 < \phi < 1$. That is, for ϕ of the time the response in the reinterview is the same as that reported on the original response and for the remaining $(1 - \phi)$ of the time the response in the reinterview follows the classification probabilities $\tau_{t(j)}$. Thus, ϕ is a measure of the persistence from the first interview to the second.

Let p_{ijk} be the proportion of the sample which is classified in the i-th row class at Trial-1, in the j-th column class at Trial-2 and in the k-th column class at Trial-3. Then for unreconciled data,

$$P_{ijk} \equiv E(p_{ijk}) = \sum_{\ell=1}^{r} \sum_{m=1}^{r} P_{\ell m} \gamma_{i(\ell)} \kappa_{j(m)} \nu_{k(m)} \qquad (3.7)$$

and for reconciled data

$$P_{ijk} \equiv E(p_{ijk}) = \sum_{\ell=1}^{r} \sum_{m=1}^{r} P_{\ell m} \gamma_{i(\ell)} \kappa_{j(m)} [\phi\ \delta_{jk} + (1 - \phi)\tau_{k(m)}] , \qquad (3.8)$$

where δ_{jk} is Kronecker's delta.

The general classification model contains $4r(r-1) + 1$ independent classification probabilities and $r^2 - 1$ independent population proportions. We develop a classification model in which the classification probabilities are expressed as functions of a reduced number of independent parameters.

B. The Classification Model

Battese and Fuller (1974) consider a classification model for classifying individuals to a one-way table. They assume that the sample response is a function of the population parameters P_i, $i = 1,2,\ldots, r$. The classification probabililties, β_{ij}, $i,j = 1,2,\ldots, r$ suggested by Battese and Fuller are

$$\begin{aligned} \beta_{ij} &= 1 - \alpha + \alpha P_i , \quad i = j \\ &= \alpha P_i , \qquad i \neq j , \end{aligned} \tag{3.9}$$

where β_{ij} is the probability that an individual belonging to the j-th class is classified in the i-th class and α is a constant in the interval $[0, 1]$. For this classification model, the sample proportion for any given class unbiasedly estimates the true proportion belonging to the class. We propose a classification model which is an extension of the Battese-Fuller classification model.

Assume that the marginal population proportions $P_{i.}$ and $P_{.j}$, $i,j = 1,2,\ldots, r$, are positive. Let the probability that an individual belonging to the i-th class is classified in the j-th class, $i \neq j$, be proportional to the conditional population proportion of the j-th class given that the element belongs either to the i-th or j-th class. Let the constant of proportionality be α_{ij}. Then the proposed classification probabilities $\gamma_{k(i)}$ and $\kappa_{\ell(j)}$, $i,j,k,\ell = 1,2,\ldots, r$ are

$$\gamma_{k(i)} = [1 - \sum_{t=1}^{r} \alpha_{ti} P_{t.} (P_{t.} + P_{k.})^{-1}] \delta_{ki}$$

$$+ \alpha_{ki} P_{k.} (P_{k.} + P_{i.})^{-1} , \quad i,k = 1,2,\ldots, r$$

(3.10)

and

$$\kappa_{\ell(j)} = [1 - \sum_{t=1}^{r} \alpha_{tj} P_{.t} (P_{.t} + P_{.\ell})^{-1}] \delta_{\ell j}$$

$$+ \alpha_{\ell j} P_{.\ell} (P_{.\ell} + P_{.j})^{-1} , \quad \ell,j = 1,2,\ldots, r ,$$

(3.11)

where δ_{ij} is Kronecker's delta, $\alpha_{ii} = 0$, $\alpha_{ij} = \alpha_{ji}$, $i \neq j$ and α_{ij}, $i,j = 1,2,\ldots, r$ are constants in the interval $[0,1]$.

The classification probabilities defined in (3.10) and (3.11) are such that the row and column marginal sample proportions obtained at Trial-1 and Trial-2 are unbiased for the row and column marginal population proportions, respectively. That is,

$$E(p_{i..}) = \sum_{\ell=1}^{r} P_{\ell.} \gamma_{i(\ell)}$$

$$= \sum_{\ell=1}^{r} P_{\ell.} \{[1 - \sum_{t=1}^{r} \alpha_{t\ell} P_{t.} (P_{t.} + P_{i.})^{-1}] \delta_{i\ell}$$

$$+ \alpha_{i\ell} P_{i.} (P_{i.} + P_{\ell.})^{-1}\}$$

$$= P_{i.} - \sum_{t=1}^{r} \alpha_{t\ell} P_{i.} P_{t.} (P_{t.} + P_{i.})^{-1}$$

$$+ \sum_{\ell=1}^{r} \alpha_{i\ell} P_{i.} P_{\ell.} (P_{i.} + P_{\ell.})^{-1}$$

$$= P_{i.} \tag{3.12}$$

and similarly

$$E(p_{.j.}) = P_{.j} \, . \tag{3.13}$$

When $r = 2$, Equation (3.10) and (3.11) are the classification probabilities defined in the Battese-Fuller model.

It has been observed that the sample proportions obtained from the reinterview are not the same as the sample proportions obtained from the original interview. In order to preserve the form for the response probabilities defined in Equation (3.11), we replace the $P_{.j}$ by different parameters in the classification probabilities of the reinterview. Thus, for the unreconciled data the $\nu_{\ell(j)}$, $\ell, j = 1,2,\ldots, r$, are written as

$$\nu_{\ell(j)} = [1 - \sum_{t=1}^{r} \alpha_{tj} U_t (U_t + U_j)^{-1}]\delta_{\ell j} + \alpha_{\ell j} U_\ell (U_\ell + U_j)^{-1}$$

$$\ell,j = 1,2,\ldots, r , \qquad (3.14)$$

and for the reconciled data the $\tau_{\ell(j)}$, $\ell,j = 1,2,\ldots, r$, is expressed as

$$\tau_{\ell(j)} = [1 - \sum_{t=1}^{r} \alpha_{tj} R_t (R_t + R_j)^{-1}]\delta_{\ell j} + \alpha_{\ell j} R_\ell (R_\ell + R_j)^{-1}$$

$$\ell,j = 1,2,\ldots, r . \qquad (3.15)$$

Two submodels can be considered. In one, the Us and the Rs satisfy $\Sigma_{j=1}^{r} U_j = 1$ and $\Sigma_{j=1}^{r} R_j = 1$, while in the other model the Us and Rs are unrestricted.

From Equation (3.14), the expectation of the sample proportion in the j-th class obtained from the reinterview procedure without reconciliation is

$$\sum_{\ell=1}^{r} P_{.\ell} \nu_{j(\ell)}$$

$$= \sum_{\ell=1}^{r} P_{.\ell} \{[1 - \sum_{t=1}^{r} \alpha_{t\ell} U_t (U_t + U_\ell)^{-1}]\delta_{\ell j}$$

$$+ \alpha_{j\ell} U_j (U_j + U_\ell)^{-1}\}$$

$$= P_{.j} - \sum_{t=1}^{r} \alpha_{tj} (U_t + U_j)^{-1} U_t P_{.j}$$

$$+ \sum_{\ell=1}^{r} \alpha_{j\ell}(U_j + U_\ell)^{-1} U_j P_{.\ell}$$

$$= P_{.j} - \sum_{t=1}^{r} \alpha_{t\ell}(U_j + U_t)^{-1}(U_j P_{.t} - U_t P_{.j}) \ . \qquad (3.16)$$

Also from Equations (3.6) and (3.15) the expectation of the sample proportion in the j-th class obtained from the reinterview procedure with reconciliation is

$$P_{.j} - (1 - \beta) \sum_{t=1}^{r} \alpha_{t\ell}(R_j + R_t)^{-1}(R_j P_{.t} - R_t P_{.j}) \ . \qquad (3.17)$$

Thus, the column marginal sample proportions obtained from the two kinds of reinterview procedures are not unbiased for the column marginal population proportions unless $P_{.j}P_{.t}^{-1} = R_j R_t^{-1}$ and $P_{.j}P_{.t}^{-1} = U_j U_t^{-1}$.

By substituting Equations (3.10), (3.11), and (3.14) into Equation (3.7) for the unreconciled data and by substituting Equations (3.10), (3.11) and (3.15) into Equation (3.8) for the reconciled data, the expectations of the sample proportions p_{ijk} for the three trial survey can be expressed as a nonlinear function of α_{ij} , $i < j = 1,2,\ldots, r$; $P_{i.}$; $P_{.j}$; U_j; and P_{ij}, $i, j = 1,2,\ldots, r-1$; for the unreconciled data and of ϕ; α_{ij}, $i < j = 1,2,\ldots, r$; $P_{i.}$; $P_{.j}$; R_j ; and P_{ij} , $i,j = 1,2,\ldots, r-1$ for the reconciled data. Thus, the Gauss-Newton procedure can be used to obtain estimates of the parameters.

Let

$$\underset{\sim}{Y} = (P_{111}, P_{112}, \ldots, P_{11r}, P_{121}, P_{122}, \ldots, P_{12r}, \ldots, P_{rr1}, \ldots, P_{rr,r-1})' \tag{3.18}$$

be the vector of observed proportions, and let $\underset{\sim}{\theta}$ be the vector of parameters, where the parameters are P_{ij}, $P_{i.}$, $P_{.j}$, $i,j = 1,2,\ldots, r-1$ and α_{ij}, $i < j = 1,2,\ldots, r$. Then

$$\underset{\sim}{Y} = \underset{\sim}{P}(\underset{\sim}{\theta}) + \underset{\sim}{e} , \tag{3.19}$$

where $\underset{\sim}{P}(\underset{\sim}{\theta})$ denotes the vector of expected values of the sample proportions in $\underset{\sim}{Y}$ expressed as functions of the vector $\underset{\sim}{\theta}$; and $\underset{\sim}{e}$ denotes the vector of deviations of the observed proportions from the expected proportions. Let $\underset{\sim}{V}$ be the covariance matrix of $\underset{\sim}{e}$. Then

$$\underset{\sim}{V} = n^{-1}\{\text{Diag}[\underset{\sim}{P}(\underset{\sim}{\theta})] - \underset{\sim}{P}(\underset{\sim}{\theta})[\underset{\sim}{P}(\underset{\sim}{\theta})]'\} . \tag{3.20}$$

Let $\tilde{\underset{\sim}{\theta}}$ be an initial estimate of $\underset{\sim}{\theta}$. Then the one-step Gauss-Newton estimator for $\underset{\sim}{\theta}$, denoted by $\hat{\underset{\sim}{\theta}}$, is

$$\hat{\underset{\sim}{\theta}} = \tilde{\underset{\sim}{\theta}} + \hat{\underset{\sim}{d}} , \tag{3.21}$$

where

$$\hat{\underset{\sim}{d}} = [\underset{\sim}{F}'(\tilde{\underset{\sim}{\theta}})\tilde{\underset{\sim}{V}}^{-1}\ \underset{\sim}{F}(\tilde{\underset{\sim}{\theta}})]^{-1}\ \underset{\sim}{F}'(\tilde{\underset{\sim}{\theta}})\tilde{\underset{\sim}{V}}^{-1}[\underset{\sim}{Y} - \underset{\sim}{P}(\tilde{\underset{\sim}{\theta}})]\ , \tag{3.22}$$

$\underset{\sim}{F}(\tilde{\underset{\sim}{\theta}})$ denotes the matrix of partial derivatives of $\underset{\sim}{P}(\underset{\sim}{\theta})$ with respect to $\underset{\sim}{\theta}$ evaluated at $\tilde{\underset{\sim}{\theta}}$, and

$$\tilde{\underset{\sim}{V}} = n^{-1}\{\mathrm{Diag}[\underset{\sim}{P}(\tilde{\underset{\sim}{\theta}})] - \underset{\sim}{P}(\tilde{\underset{\sim}{\theta}})[\underset{\sim}{P}(\tilde{\underset{\sim}{\theta}})]'\}\ . \tag{3.23}$$

Assume that the initial estimator, $\tilde{\underset{\sim}{\theta}}$, satisfies the condition

$$\tilde{\underset{\sim}{\theta}} - \underset{\sim}{\theta} = O_p(n^{-1/2}) \tag{3.24}$$

and the matrix $\underset{\sim}{F}'(\underset{\sim}{\theta}^0)\underset{\sim}{V}^{-1}\ \underset{\sim}{F}(\underset{\sim}{\theta}^0)$ is nonsingular for every $\underset{\sim}{\theta}^0$ in an open subset of B of the parameter space where the true parameter $\underset{\sim}{\theta}$ belongs to B. Then, it can be shown that (see, for example, Fuller (1976, Chapter 5))

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \xrightarrow{L} N(\underset{\sim}{0},\ [\underset{\sim}{F}'(\underset{\sim}{\theta})\underset{\sim}{V}^{-1}\ \underset{\sim}{F}(\underset{\sim}{\theta})]^{-1}) \tag{3.25}$$

and

$$S^2 = [\underset{\sim}{Y} - \underset{\sim}{P}(\hat{\underset{\sim}{\theta}})]'\hat{\underset{\sim}{V}}^{-1}[\underset{\sim}{Y} - \underset{\sim}{P}(\hat{\underset{\sim}{\theta}})] \tag{3.26}$$

converges in law to χ^2, where χ^2 is distributed as a chi-square random variable with $2^{-1}(r-1)(2r^2 - r - 2)$ degree of freedom.

The statistical package SAS (1982) provides a quite efficient program for nonlinear estimation. In practice, several iterations are performed by the program until the reduction of the residual sum of squares for two consecutive iterations is less than a specified constant.

C. Example

In the monthly CPS sample conducted by the U.S. Bureau of the Census, information on the employment status of individuals is collected. In a given month, each individual is classified into one of the following categories: Employed, Unemployed and Not in the Labor Force (NILF). As a part of the quality control procedures, about 1 of 18 units in the monthly CPS sample is reinterviewed. The original interviewers do not know which household will be reinterviewed by senior interviewers and supervisors during the reinterview. No original interviewer is used to interview his/her own interview cases.

In the reinterview process, a reconciliation is done for 80 percent of the reinterview sample. Reinterviewers are given the original responses and instructed to consult them only after the reinterview responses have been given. Reconciliation is done on a separate form containing the original responses. For the other 20 percent of the sample, no reconciliation is made.

The survey responses in January and two interviews in February of 1979 with reconciliation and no reconciliation in the reinterview are given in Table 1 and Table 2, respectively. The size of the sample is 3,198.

Table 1. Reported employment status in January, February and February reinterviews, where reconciliation is made in the reinterview process

Employed in January at Trial-1

February Trial-2 class	February Trial-3 class			
	Employed	Unemployed	NILF	Total
Employed	1,428	4	12	1,444
Unemployed	2	19	2	23
NILF	6	1	43	50
Total	1,436	24	57	1,517

Unemployed in January at Trial-1

February Trial-2 class	February Trial-3 class			
	Employed	Unemployed	NILF	Total
Employed	22	2	0	24
Unemployed	3	34	2	39
NILF	1	2	15	18
Total	26	38	17	81

NILF in January at Trial-1

February Trial-2 class	February Trial-3 class			
	Employed	Unemployed	NILF	Total
Employed	39	1	7	47
Unemployed	0	21	5	26
NILF	9	16	1,003	1,028
Total	48	38	1,015	1,101

Table 2. Reported employment status in January, February and February reinterviews, where no reconciliation is made in the reinterview process

Employed in January at Trial-1

February Trial-2 class	February Trial-3 class			
	Employed	Unemployed	NILF	Total
Employed	248	2	6	256
Unemployed	0	3	0	3
NILF	2	0	8	10
Total	250	5	14	269

Unemployed in January at Trial-1

February Trial-2 class	February Trial-3 class			
	Employed	Unemployed	NILF	Total
Employed	6	0	0	6
Unemployed	0	8	0	8
NILF	0	2	1	3
Total	6	10	1	17

NILF in January at Trial-1

February Trial-2 class	February Trial-3 class			
	Employed	Unemployed	NILF	Total
Employed	8	0	0	8
Unemployed	0	4	1	5
NILF	8	1	191	200
Total	16	5	192	213

Additional data on the reinterview process from the second quarter 1978 to the fourth quarter 1980 are also available. The responses to the original interview and reinterview during that period with no reconciliation and with reconciliation in the reinterview are given in Table 3 and Table 4, respectively.

Let the three categories Employed, Unemployed and NILF be indexed by 1, 2, and 3, respectively. It is hoped that the α_{ij} parameters of the model proposed in Section B will remain relatively constant over time. Then, estimates of the α_{ij}'s can be obtained from data collected during the period beginning with the second quarter of 1978 and ending with the fourth quarter of 1980. It is assumed that no individual was reinterviewed more than once during that period of time. This is a policy of the Census Bureau.

The classification probabilities suggested in Equations (3.11), (3.14), and (3.15) are used for the original interview and reinterview of the grouped data and also for the original interview and reinterview of the unreconciled data. The α's are assumed to be the same in the classification probabilities for both data sets. For the reinterview on the reconciled data, different α_{13} and α_{23} are used for the two interviews. We also assume that $\Sigma_{i=1}^{3} U_i = 1$ and $\Sigma_{i=1}^{3} R_i = 1$.

Let p_{ij} be the sample proportion in the i-th class on the original interview and in the j-th class on the reinterview. Let

$$\underline{p} = (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32})' .$$

Table 3. Employment Status, Original Interview by Reinterview with No Reconciliation in the Reinterview, 2nd Quarter 1978 Through 4th Quarter 1980

Original Interview	Reinterview			
	Employed	Unemployed	NILF	Total
Employed	15,619	123	485	16,227
Unemployed	114	770	195	1,079
NILF	416	275	10,307	10,998
Total	16,149	1,168	10,987	28,304

Table 4. Employment Status, Original Interview by Reinterview with Reconciliation in the Reinterview, 2nd Quarter 1978 Through 4th Quarter 1980

Original Interview	Reinterview			
	Employed	Unemployed	NILF	Total
Employed	77,535	112	264	77,911
Unemployed	155	4,913	140	5,208
NILF	864	592	50,858	52,314
Total	78,554	5,617	51,262	135,433

The covariance matrix, $\underset{\sim}{V}$, of p is obtained under the assumption that the sample observations are distributed as multinomial random variables. Thus,

$$\hat{\underset{\sim}{V}} = n^{-1}[\mathrm{Diag}(\underset{\sim}{p}) - \underset{\sim}{p}\,\underset{\sim}{p}'] \ .$$

This is a gross approximation because the sample is selected according to a multistage sampling scheme. Also, within every selected household every member is interviewed. Thus, there exists the cluster effect. Another effect that could not be identified from the available data includes interviewer effects. It is hoped that these effects are small enough so that the multinomial approximation will be adequate for the computation of estimates. With this estimator of $\underset{\sim}{V}$, the computational procedure for the nonlinear model is simplified a great deal. The estimates obtained using the Gauss-Newton procedure for the nonlinear model are

$$\hat{\alpha}_{12} = \underset{(0.0053)}{0.0564}\ , \qquad \hat{\alpha}_{13} = \underset{(0.0013)}{0.0344}\ ,$$

$$\hat{\alpha}_{23} = \underset{(0.0096)}{0.1192}\ , \qquad \hat{P}_{.1} = \underset{(0.0020)}{0.5749}\ ,$$

$$\hat{P}_{.2} = \underset{(0.0017)}{0.0384}\ , \qquad \hat{U}_1 = \underset{(0.0382)}{0.5315}\ ,$$

$$\hat{U}_2 = \underset{(0.0092)}{0.0570}\ , \qquad \hat{R}_1 = \underset{(0.0151)}{0.9729}\ ,$$

$$\hat{R}_2 = \underset{(0.0148)}{0.0271}\ , \qquad \hat{\phi} = \underset{(0.0137)}{0.7289}\ ,$$

$$\hat{\alpha}^{*}_{13} = \underset{(0.0038)}{0.0412}\ , \qquad \hat{\alpha}^{*}_{23} = \underset{(0.0036)}{0.0326}\ ,$$

where $\hat{\alpha}^*_{13}$ and $\hat{\alpha}^*_{23}$ are the estimates of the parameters α_{13} and α_{23} that appear in the reinterview classification probabilities $\tau_{\ell(j)}$ of the reconciled data. The residual sum of squares is 8.34 with 4 degrees of freedom. The 5 percent point of the chi-square distribution with 4 degrees of freedom is 9.49. Thus, the fitted model is consistent with the observed data. The standard errors of these estimates are calculated under the multinomial assumption. Because of the clustered nature of the sample, it is expected that the standard errors are biased downward.

To analyze the data obtained in January and February 1979, we combine that data with the grouped 1978-80 data. Before doing so, a careful look at the data set reveals that the marginal proportions of the reinterview in February on the reconciled data are not consistent with the corresponding marginal proportions of the grouped 1978-80 data. Thus, only the parameters of α_{13}, α_{12}, α_{23}, U_1 and U_2 are assumed to be the same for the grouped data as for the 1979 data. In constructing estimated standard errors, it is assumed that the grouped data are independent of the sample data collected in January and February 1979.

Let p_{ijk} be the sample proportion of the i-th class of January, j-th class of February and k-th class of February reinterview. Let $\underset{\sim}{p}$ be the column vector of P_{ijk}'s . Due to the fact that there are some zeroes in $\underset{\sim}{p}$, we propose an approximate estimate of the covariance matrix, $\underset{\sim}{V}$, of $\underset{\sim}{p}$. Let

$$\underset{\sim}{Z} = (n + 27)^{-1}(n\ \underset{\sim}{p} + 1) \ .$$

Then, an estimate of $\underset{\sim}{V}$ is

$$\tilde{\underset{\sim}{V}} = n^{-1}[\mathrm{Diag}(\underset{\sim}{Z}) - \underset{\sim}{Z}\ \underset{\sim}{Z}'] \ .$$

With this $\tilde{\underset{\sim}{V}}$, the Gauss-Newton estimates for the parameters are

$$\hat{\alpha}_{12} = \underset{(0.0049)}{0.0558}\ , \qquad \hat{\alpha}_{13} = \underset{(0.0012)}{0.0334}\ ,$$

$$\hat{\alpha}_{23} = \underset{(0.0087)}{0.1161}\ , \qquad \hat{U}_1 = \underset{(0.0348)}{0.5267}\ ,$$

$$\hat{U}_2 = \underset{(0.0082)}{0.0572}\ , \qquad \hat{R}_1 = \underset{(0.0901)}{0.4917}\ ,$$

$$\hat{R}_2 = \underset{(0.0312)}{0.0853}\ , \qquad \hat{\phi} = \underset{(0.0703)}{0.6381}\ ,$$

$$\hat{\alpha}^*_{13} = \underset{(0.0135)}{0.0462}\ , \qquad \hat{\alpha}^*_{23} = \underset{(0.0916)}{0.2405}$$

with $\hat{P}_{ij}$, $\hat{P}_{i.}$ and $\hat{P}_{.j}$, $i,j = 1,2,3$ shown in Table 5. The sum of squares of the residuals for the nonlinear model is 33.53 with 39 degrees of freedom. The usual chi-square value can be calculated by the following equation

$$\chi^2 = \sum_{\ell=1}^{r}\sum_{j=1}^{r}\sum_{k=1}^{r} (n\ P_{ijk}(\hat{\theta}) - n\ p_{ijk})^2 (n\ P_{ijk}(\hat{\theta}))^{-1} \ ,$$

where $n\ p_{ijk}$ is the observed value of the i-th class in January, j-th class in February and k-th class in February-reinterview and $n\ P_{ijk}(\hat{\underset{\sim}{\theta}})$

is the expected value of the i-th class in January, j-th class in February, and k-th class in February-reinterview. With these estimates of the parameters, the chi-square value is 39.49. The 5 percent value of a chi-square distribution with 39 degrees of freedom is 53.56. Thus, the model fitted is consistent with the observed survey response of the employment status in January and February of 1979.

The usual maximum likelihood estimates of P_{ij}, i,j = 1,2,3 , based on the original interviews conducted in January and February 1979 are shown in Table 6. Table 6 is constructed under the assumption that no classification error exists. The size of the sample is 3,198. By comparing the figures in Table 5 and Table 6, one sees that the estimates of the diagonal elements P_{ii} adjusted for the classification error are larger than the maximum likelihood estimates constructed under the assumption of no response error. The estimates of the off diagonal elements, P_{ij} , adjusted for the classification errors are, in general, smaller than the simple proportions. The biggest differences are for the proportions changing classes between NILF and employed from January to February. The differences are about six times the standard deviations of the simple proportions, where the standard deviations are obtained under the multinomial assumption. The two estimated movements between unemployed and NILF are also reduced substantially, while the estimated movements between employed and unemployed are only slightly smaller than the original sample proportions. One expects the Gauss-Newton estimates of the row and column marginal probabilities to be

Table 5. Gauss-Newton Estimates of Probabilities

	February			
January	Employed	Unemployed	NILF	Total
Employed	0.5499 (0.0081)	0.0066 (0.0018)	0.0042 (0.0023)	0.5607 (0.0081)
Unemployed	0.0081 (0.0018)	0.0200 (0.0028)	0.0010 (0.0015)	0.0291 (0.0029)
NILF	0.0019 (0.0022)	0.0053 (0.0019)	0.4030 (0.0080)	0.4102 (0.0080)
Total	0.5599 (0.0081)	0.0319 (0.0030)	0.4082 (0.0080)	1.0000

Table 6. The Maximum Likelihood Estimates of Probabilities, Assuming No Classification Error

	February			
January	Employed	Unemployed	NILF	Total
Employed	0.5316 (0.0088)	0.0081 (0.0016)	0.0188 (0.0024)	0.5585 (0.0084)
Unemployed	0.0094 (0.0017)	0.0147 (0.0021)	0.0066 (0.0014)	0.0307 (0.0031)
NILF	0.0172 (0.0023)	0.0097 (0.0017)	0.3839 (0.0085)	0.4108 (0.0087)
Total	0.5582 (0.0089)	0.0325 (0.0031)	0.4093 (0.0087)	1.0000

equal to the simple proportions of the row and column marginal probabilities due to the fact that the classification probabilities satisfy the marginal unbiased property. Our estimates came out slightly different because the Gauss-Newton estimates are obtained from the reconciled and unreconciled data sets and the simple proportions are calculated by using the first interviews in January and February of the combined data set. The estimate of α_{23} is the largest of the estimates and indicates that mistakes in classification between unemployed and NILF have the highest probability.

The classification probabilities for January and February are shown in Table 7. The probabilities are constructed using the January and February marginals from Table 6. From these two sets of classification probabilities, one can also obtain estimates of P_{ij} by using the Equation (3.5). That is,

$$\underset{\sim}{P} = (\underset{\sim}{k}^{-1} \otimes \underset{\sim}{\Gamma}^{-1})\underset{\sim}{p}$$

where $\underset{\sim}{\Gamma}$ and $\underset{\sim}{k}$ are the matrices of classification probabilities of January and February, respectively, and $\underset{\sim}{P}$ and $\underset{\sim}{p}$ are defined in Equation (3.3).

Table 7. Estimated Classification Probabilities for January and February 1979

Reported class		True class		
		Employed	Unemployed	NILF
January	Employed	0.9829	0.0529	0.0192
	Unemployed	0.0029	0.8391	0.0081
	NILF	0.0142	0.1080	0.9807
February	Employed	0.9828	0.0527	0.0193
	Unemployed	0.0031	0.8397	0.0085
	NILF	0.0141	0.1076	0.9722

IV. AN ERRORS-IN-VARIABLES MODEL WITH REPLICATED OBSERVATIONS ON SOME UNITS

A. Introduction

Consider the following errors-in-variables model

$$y_t = \beta_0 + \beta_1 x_t + q_t \,,$$

$$Y_t = y_t + w_t \,, \qquad \text{and}$$

$$X_t = x_t + u_t \,, \quad t = 1,2,\ldots, n \,, \tag{4.1}$$

here

$$\begin{pmatrix} x_t \\ q_t \\ w_t \\ u_t \end{pmatrix} \sim NI\left(\begin{pmatrix} \mu_x \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & 0 & 0 & 0 \\ 0 & \sigma_{qq} & 0 & 0 \\ 0 & 0 & \sigma_{ww} & \sigma_{wu} \\ 0 & 0 & \sigma_{wu} & \sigma_{uu} \end{pmatrix} \right) ,$$

ıd x_t and y_t are the true values of the variables of interest which ınnot be measured exactly. Instead, Y_t and X_t are observed. Under ıis setup, the random variable q_t is the error in the equation and ıe random varibles w_t and u_t are the measurement errors of y_t ıd x_t, respectively.

Let (S_{uu}, S_{ww}, S_{uw}) be unbiased estimators of $(\sigma_{uu}, \sigma_{ww}, \sigma_{uw})$, where the matrix

$$\underset{\sim}{S} = \begin{pmatrix} S_{uu} & S_{uw} \\ S_{wu} & S_{ww} \end{pmatrix}$$

is distributed as a Wishart distribution with d degrees of freedom. Let (S_{uu}, S_{ww}, S_{uw}) and (X_t, Y_t) be independent for $t = 1,2,\ldots, n$. Fuller (1980) obtained an estimator, $\hat{\beta}_1$, of β_1 and the limiting distribution of $\hat{\beta}_1$. The estimator $\hat{\beta}_1$ is

$$\hat{\beta}_1 = (m_{XX} - S_{uu})^{-1}(m_{XY} - S_{uw}) , \tag{4.2}$$

where $m_{XX} = n^{-1} \Sigma_{t=1}^{n}(X_t - \bar{X})^2$, $m_{XY} = n^{-1} \Sigma_{t=1}^{n}(X_t - \bar{X})(Y_t - \bar{Y})$, $\bar{X} = n^{-1} \Sigma_{t=1}^{n} X_t$, $\bar{Y} = n^{-1} \Sigma_{t=1}^{n} Y_t$. The limiting distribution of $n^{1/2}(\hat{\beta}_1 - \beta_1)$ is $N[0, V(n^{1/2}\hat{\beta}_1)]$, where

$$V(n^{1/2}\hat{\beta}_1) = \sigma_{xx}^{-2}[\sigma_{xx}\sigma_{vv} + \sigma_{uu}\sigma_{vv} + \sigma_{uv}^2 + n(\sigma_{uu}\sigma_{rr} + \sigma_{ur}^2)] , \tag{4.3}$$

$$\sigma_{vv} = \sigma_{rr} + \sigma_{qq} ,$$

$$\sigma_{rr} = \sigma_{ww} + \beta_1^2 \sigma_{uu} - 2\beta_1\sigma_{uw} ,$$

$$\sigma_{uv} = \sigma_{ur} = \sigma_{uw} - \beta_1 \sigma_{uu} ,$$

and $\eta = nd^{-1}$ is a fixed number.

Suppose that among the n units, d units are observed twice. Without loss of generality, let them be the first d units. Then

$$Y_{ti} = y_t + w_{ti}$$

and

$$X_{ti} = x_t + u_{ti} , \quad i = 1,2, \quad t = 1,2,\ldots, d ,$$

where (w_{ti}, u_{ti}) , $i = 1,2$, $t = 1,2,\ldots, d$, (w_t, u_t) , $t = d+1,\ldots, n$ are independent bivariate normal vectors with mean zero and variance-covariance matrix

$$\begin{pmatrix} \sigma_{ww} & \sigma_{wu} \\ \sigma_{wu} & \sigma_{uu} \end{pmatrix} .$$

Let

$$S_{ww} = (2d)^{-1} \sum_{t=1}^{d} (Y_{t1} - Y_{t2})^2$$

$$S_{uu} = (2d)^{-1} \sum_{t=1}^{d} (X_{t1} - X_{t2})^2$$

and

$$S_{uw} = (2d)^{-1} \sum_{t=1}^{d} (X_{t1} - X_{t2})(Y_{t1} - Y_{t2}) . \qquad (4.4)$$

Then S_{ww}, S_{uu} and S_{uw} are unbiased estimators of σ_{ww}, σ_{uu} and σ_{uw}, respectively.

Let $(\bar{Y}_{t.}, \bar{X}_{t.})$ be the mean of (Y_{t1}, X_{t1}) and (Y_{t2}, X_{t2}) for $t = 1,2,\ldots, d$. From (4.1), the model becomes

$$y_t = \beta_0 + \beta_1 x_t + q_t$$

$$\bar{Y}_{t.} = y_t + \bar{w}_{t.}$$

$$\bar{X}_{t.} = x_t + \bar{u}_{t.} , \quad t = 1,2,\ldots, d , \qquad (4.5)$$

where

$$\bar{w}_{t.} = 2^{-1}(w_{t1} + w_{t2})$$

$$\bar{u}_{t.} = 2^{-1}(u_{t1} + u_{t2})$$

and

$$\begin{pmatrix} x_t \\ q_t \\ \bar{w}_{t.} \\ \bar{u}_{t.} \end{pmatrix} \sim NI\left(\begin{pmatrix} \mu_x \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & 0 & 0 & 0 \\ 0 & \sigma_{qq} & 0 & 0 \\ 0 & 0 & \tfrac{1}{2}\sigma_{ww} & \tfrac{1}{2}\sigma_{wu} \\ 0 & 0 & \tfrac{1}{2}\sigma_{wu} & \tfrac{1}{2}\sigma_{uu} \end{pmatrix} \right) .$$

Under the normality assumption, $(\bar{Y}_{t.}, \bar{X}_{t.})$ and (S_{ww}, S_{uu}, S_{wu}) are independent for $t = 1,2,\ldots, d$. Thus, from (4.2) and (4.3), the estimator for β_1 based on the first d units is

$$\hat{\beta}_{1,d} = (m_{\bar{X}\,\bar{X}} - \tfrac{1}{2} S_{uu})^{-1}(m_{\bar{X}\,\bar{Y}} - \tfrac{1}{2} S_{wu}) \qquad (4.6)$$

with asymptotic variance

$$V(\hat{\beta}_{1,d}) = \sigma_{xx}^{-2}\{d^{-1}[\sigma_{xx}(\sigma_{qq} + \tfrac{1}{2}\sigma_{rr}) + \tfrac{1}{2}\sigma_{uu}(\sigma_{qq} + \tfrac{1}{2}\sigma_{rr}) + \tfrac{1}{4}\sigma_{ur}^2]$$

$$+ \tfrac{1}{4} d^{-1}(\sigma_{uu}\sigma_{rr} + \sigma_{ur}^2)\} , \qquad (4.7)$$

where

$$m_{\bar{X}\,\bar{X}} = d^{-1} \sum_{t=1}^{d} (\bar{X}_{t.} - \bar{X}_{..})^2 ,$$

$$m_{\bar{X}\,\bar{Y}} = d^{-1} \sum_{t=1}^{d} (\bar{X}_{t.} - \bar{X}_{..})(\bar{Y}_{t.} - \bar{Y}_{..}) ,$$

$$\bar{X}_{..} = d^{-1} \sum_{t=1}^{d} \bar{X}_{t.} ,$$

$$\bar{Y}_{..} = d^{-1} \sum_{t=1}^{d} \bar{Y}_{t.} ,$$

$$\sigma_{rr} = \sigma_{ww} - 2\beta_1\sigma_{wu} + \beta_1^2 \sigma_{uu} ,$$

$$\sigma_{ur} = \sigma_{wu} - \beta_1\sigma_{uu} .$$

Similarly, since (Y_t, X_t) and (S_{ww}, S_{uu}, S_{wu}) are independent for $t = d+1,\ldots, n$, the estimator for β_1 based on the last $n-d$ units is

$$\hat{\beta}_{1,n-d} = (m_{XX} - S_{uu})^{-1}(m_{XY} - S_{wu}) \qquad (4.8)$$

with asymptotic variance

$$V(\hat{\beta}_{1,n-d}) = \sigma_{xx}^{-2}\{(n-d)^{-1}[\sigma_{xx}(\sigma_{qq} + \sigma_{rr}) + \sigma_{uu}(\sigma_{qq} + \sigma_{rr}) + \sigma_{ur}^2]$$

$$+ d^{-1}(\sigma_{uu}\sigma_{rr} + \sigma_{ur}^2)\} , \qquad (4.9)$$

where

$$m_{XX} = (n-d)^{-1} \sum_{t=d+1}^{n} (X_t - \bar{X})^2 ,$$

$$m_{YY} = (n-d)^{-1} \sum_{t=d+1}^{n} (X_t - \bar{X})(Y_t - \bar{Y}) ,$$

$$\bar{X} = (n-d)^{-1} \sum_{t=d+1}^{n} X_t ,$$

$$\bar{Y} = (n-d)^{-1} \sum_{t=d+1}^{n} Y_t \ .$$

Also from the result of Appendix A, the asymptotic covariance of $\hat{\beta}_{1,d}$ and $\hat{\beta}_{1,n-d}$ is

$$\mathrm{Cov}(\hat{\beta}_{1,d}, \hat{\beta}_{1,n-d}) = \sigma_{xx}^{-2}[(2d)^{-1}(\sigma_{uu}\sigma_{rr} + \sigma_{ur}^2)] \ . \quad (4.10)$$

The covariance is positive because both estimators use the estimated covariances (S_{uu}, S_{uw}, S_{ww}) .

Let

$$V_{11} = V(\hat{\beta}_{1,d}) \ ,$$

$$V_{22} = V(\hat{\beta}_{1,n-d})$$

and

$$V_{12} = \mathrm{cov}(\hat{\beta}_{1,d}, \hat{\beta}_{1,n-d}) \ . \quad (4.11)$$

To find the optimal linear combination of the two estimators, let

$$\hat{\beta}_{1,p} = p\,\hat{\beta}_{1,d} + (1-p)\hat{\beta}_{1,n-d} \ ,$$

where p is to be determined. From the results of Appendix B, the p that minimizes the variance of $\hat{\beta}_{1,p}$ is

$$p^* = (V_{11} - 2V_{12} + V_{22})^{-1}(V_{22} - V_{12}) , \tag{4.12}$$

and the variance of $p^*\hat{\beta}_{1,d} + (1-p^*)\hat{\beta}_{1,n-d}$ is

$$(V_{11} + V_{22} - 2V_{12})^{-1}(V_{11}V_{22} - V_{12}^2) . \tag{4.13}$$

B. Determination of Number of Duplicate Measurements Units

Assume that the cost of obtaining one observation is c units. Then the total cost, T, for the survey is

$$T = c(n+d) , \tag{4.14}$$

where d is the number of the units that have duplicate measurements. We assume that it is not practical to observe a unit more than twice. Let

$$\eta = nd^{-1} . \tag{4.15}$$

Given the total cost T, the value η that minimizes the variance (4.13) is obtained as follows. From (4.14) and (4.15), (4.13) becomes

$$\frac{c(\eta+1)}{T\,\sigma_{xx}^2}\left\{\frac{[A+(\eta-1)D][B+\frac{1}{4}D]-\frac{1}{4}(\eta-1)D^2}{A+(B+\frac{1}{4}D)(\eta-1)}\right\} \equiv f(\eta)\ ,$$

(4.16)

where

$$A = \sigma_{xx}(\sigma_{qq}+\sigma_{rr})+\sigma_{uu}(\sigma_{qq}+\sigma_{rr})+\sigma_{ur}^2\ ,$$

$$B = \sigma_{xx}(\sigma_{qq}+\tfrac{1}{2}\sigma_{rr})+\tfrac{1}{2}\sigma_{uu}(\sigma_{qq}+\tfrac{1}{2}\sigma_{rr})+\tfrac{1}{4}\sigma_{ur}^2\ ,$$

$$D = \sigma_{uu}\sigma_{rr}+\sigma_{ur}^2\ . \qquad (4.17)$$

Differentiating $f(\eta)$ with respect to η and setting the derivatives equal to zero, we have

$$\frac{\partial f(\eta)}{\partial \eta} = \frac{c}{\sigma_{xx}^2 T}\left\{\frac{BD(B+\frac{D}{4})(\eta-1)^2+2ABD(\eta-1)+A^2(B+\frac{D}{4})+2ABD-2A(B+\frac{D}{4})^2}{[A+(B+\frac{D}{4})(\eta-1)]^2}\right\}$$

$$= 0\ . \qquad (4.18)$$

Solving the equation (4.18) for η, we get

$$\eta-1 = [BD(B+\tfrac{D}{4})]^{-1}\{-ABD \pm [ABD(B-\tfrac{D}{4})^2(2B + \tfrac{1}{2}D - A)]^{1/2}\} .$$

(4.19)

The second derivative of $f(\eta)$ with respect to η is

$$\frac{\partial^2 f(\eta)}{\partial \eta^2} = \frac{c}{\sigma_{xx}^2 T}\left\{\frac{2A(B - \frac{D}{4})^2(2(B + \frac{D}{4})-A)}{[A+(B + \frac{D}{4})(\eta-1)]^3}\right\} \qquad (4.20)$$

which is positive when

$$\eta-1 = [BD(B + \tfrac{D}{4})]^{-1}\{-ABD + [ABD(B - \tfrac{1}{4}D)^2(2B + \tfrac{1}{2}D-A)]^{1/2}\} .$$

(4.21)

Thus, $f(\eta)$ is a minimum when (4.21) holds.

From (4.15), $\eta > 1$, since $d < n$. It follows that

$$\eta = 1 \quad \text{if} \quad [BD(B + \tfrac{D}{4})]^{-1}\{-ABD+[ABD(B - \tfrac{1}{4}D)^2(2B + \tfrac{1}{2}D-A)]^{1/2}\} < 0$$

$$= 1+[BD(B + \tfrac{D}{4})]^{-1}\{-ABD+[ABD(B - \tfrac{1}{4}D)^2(2B + \tfrac{1}{2}D-A)]^{1/2}\} \text{ otherwise.}$$

(4.22)

But

$$[BD(B + \tfrac{D}{4})]^{-1}\{-ABD+[ABD(B - 1/4\, D)^2(2B + 1/2\, D-A)]^{1/2}\} < 0$$

$$<=> -ABD[B + \tfrac{D}{4}][AB + \tfrac{AD}{4} + B\,D - 2B^2 - \tfrac{D^2}{8}] < 0$$

$$<=> AB + \tfrac{AD}{4} + BD - 2B^2 - \tfrac{D^2}{8} > 0$$

$$<=> A(B + \tfrac{D}{4}) - 2(B - \tfrac{D}{4})^2 > 0$$

$$<=> 1/2\, D^2 + [3\sigma_{xx}\sigma_{qq} + 2\,\sigma_{xx}\sigma_{rr} + 2\,\sigma_{uu}\sigma_{qq}]\tfrac{D}{2}$$

$$- [\sigma_{xx}(\sigma_{qq} + 1/2\,\sigma_{rr}) + 1/2\,\sigma_{uu}\sigma_{qq}]\sigma_{xx}\sigma_{qq} > 0 \quad .(4.23)$$

Let

$$E = \sigma_{xx}(\sigma_{qq} + 1/2\,\sigma_{rr}) + 1/2\,\sigma_{uu}\sigma_{qq} \quad . \qquad (4.24)$$

Then, Equation (4.23) can be written as

$$[D + (2E - \tfrac{\sigma_{xx}\sigma_{qq}}{2}) - (4E^2 + \tfrac{\sigma_{xx}^2\sigma_{qq}^2}{4})^{1/2}][D+(2E - \tfrac{\sigma_{xx}\sigma_{qq}}{2})$$

$$+ (4E + \tfrac{\sigma_{xx}^2\sigma_{qq}^2}{2})^{1/2}] > 0 \quad . \qquad (4.25)$$

The inequality (4.25) holds if and only if

$$D > -(2E - \frac{\sigma_{xx}\sigma_{qq}}{2}) + (4E^2 + \frac{\sigma_{xx}^2\sigma_{qq}^2}{2})^{1/2} \qquad (4.26)$$

since $D = \sigma_{uu}\sigma_{rr} + \sigma_{ur}^2$ is positive.

Hence, we conclude that

$$\eta = 1 \quad \text{if} \quad D > -(2E - \frac{\sigma_{xx}\sigma_{qq}}{2}) + (4E^2 + \frac{\sigma_{xx}^2\sigma_{qq}^2}{4})^{1/2}$$

$$= 1+[BD(B+\frac{D}{4})]^{-1}\{-ABD+[ABD(B - 1/4\, D)^2(2B + 1/2\, D-A)]^{1/2}\}, \text{ otherwise,}$$

(4.27)

where A, B, D and E are defined in (4.17) and (4.24).

Tables for the optimal η corresponding to certain values of β_1, σ_{ww}, σ_{uu}, σ_{qq} and σ_{xx} are tabulated and are shown in Table 8. With known σ_{ww}, σ_{uu} and σ_{wu}, (w, u) can always be transformed into two independent random variables with equal variances. Therefore, without loss of generality, the tables are for $\sigma_{ww} = \sigma_{uu}$ and $\sigma_{wu} = 0$. From the tables, it can be seen that η is decreasing with respect to $\sigma_{xx}^{-1}\sigma_{uu}$, which shows that when measurement errors are large, more units with replicated observations are needed in order to obtain a better estimate for the errors.

Table 8. The optimal values of $\eta = nd^{-1}$ for given β, $\sigma_{xx}^{-1}\sigma_{uu}$, $\sigma_{xx}^{-1}\sigma_{qq}$.

$\beta_1 = 0.00$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.05	0.10	0.20	0.30	0.40	0.50	0.60
0.05	4.79	2.14	1.00	1.00	1.00	1.00	1.00
0.10	6.72	3.25	1.28	1.00	1.00	1.00	1.00
0.30	11.36	5.72	2.72	1.61	1.00	1.00	1.00
0.50	14.54	7.36	3.62	2.27	1.54	1.07	1.00
0.70	17.14	8.67	4.32	2.77	1.94	1.41	1.04
0.90	19.38	9.81	4.91	3.19	2.28	1.70	1.29

$\beta_1 = 0.20$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.05	0.10	0.20	0.30	0.40	0.50	0.60
0.05	4.60	2.03	1.00	1.00	1.00	1.00	1.00
0.10	6.47	3.11	1.20	1.00	1.00	1.00	1.00
0.30	10.94	5.50	2.60	1.52	1.00	1.00	1.00
0.50	14.00	7.08	3.47	2.16	1.45	1.00	1.00
0.70	16.50	8.35	4.14	2.65	1.85	1.33	1.00
0.90	18.66	9.44	4.72	3.06	2.18	1.61	1.21

Table 8. (continued)

$\beta_1 = 0.60$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.05	0.10	0.20	0.30	0.40	0.50	0.60
0.05	3.57	1.44	1.00	1.00	1.00	1.00	1.00
0.10	5.11	2.35	1.00	1.00	1.00	1.00	1.00
0.30	8.70	4.33	1.94	1.03	1.00	1.00	1.00
0.50	11.13	5.60	2.66	1.57	1.00	1.00	1.00
0.70	13.11	6.61	3.22	1.99	1.31	1.00	1.00
0.90	14.82	7.48	3.69	2.33	1.59	1.11	1.00

$\beta_1 = 1.00$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.05	0.10	0.20	0.30	0.40	0.50	0.60
0.05	2.58	1.00	1.00	1.00	1.00	1.00	1.00
0.10	3.82	1.61	1.00	1.00	1.00	1.00	1.00
0.30	6.62	3.21	1.29	1.00	1.00	1.00	1.00
0.50	8.47	4.21	1.88	1.00	1.00	1.00	1.00
0.70	9.97	4.99	2.33	1.33	1.00	1.00	1.00
0.90	11.26	5.66	2.71	1.61	1.01	1.00	1.00

Table 8 (continued)

$\beta_1 = 1.40$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.05	0.10	0.20	0.30	0.40	0.50	0.60
0.05	1.86	1.00	1.00	1.00	1.00	1.00	1.00
0.10	2.90	1.06	1.00	1.00	1.00	1.00	1.00
0.30	5.18	2.41	1.00	1.00	1.00	1.00	1.00
0.50	6.65	3.23	1.31	1.00	1.00	1.00	1.00
0.70	7.82	3.87	1.69	1.00	1.00	1.00	1.00
0.90	8.84	4.40	2.00	1.08	1.00	1.00	1.00

$\beta_1 = 1.80$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.05	0.10	0.20	0.30	0.40	0.50	0.60
0.05	1.34	1.00	1.00	1.00	1.00	1.00	1.00
0.10	2.24	1.00	1.00	1.00	1.00	1.00	1.00
0.30	4.18	1.84	1.00	1.00	1.00	1.00	1.00
0.50	5.41	2.54	1.00	1.00	1.00	1.00	1.00
0.70	6.37	3.08	1.22	1.00	1.00	1.00	1.00
0.90	7.20	3.53	1.49	1.00	1.00	1.00	1.00

To illustrate the use of the tables, we assume that one has enough money for 1,000 observations. For design purposes, it is assumed that $\beta_1 = 1$, the variance of the measurement error in X and Y is 10 percent of the variance of x and the variance of the error in the equation is 50 percent of the variance of x. Then, the optimal value of η is 4.21. This means that

$$d = (4.21)^{-1} n$$

$$1000 = n + (4.21)^{-1} n$$

and, hence,

$$n = 808$$

$$d = 192 .$$

The optimal design is to select 808 individuals and to make duplicate measurements on 192 of those individuals.

C. Extension of Duplicate Measurements to Triple Measurements

Given an errors-in-variables model (4.1) and the simple cost function (4.14), the value of $\eta = d^{-1} n$ that minimizes the variance of $\hat{\beta}_1$ was obtained in Section B, where n is the total number of units selected in a sample and d is the number of the sampling units that

are observed twice. Table 8 contains the optimal value of n corresponding to specific values of β_1, σ_{ww}, σ_{uu}, σ_{qq} and σ_{xx}.

For certain values of β_1, σ_{ww}, σ_{uu}, σ_{qq} and σ_{xx}, the optimal value of n is equal to one. That is, all the sampling units should be observed twice. We now determine if triple observations should be obtained for certain parameter configurations. The cost function described in (4.14), where the cost of obtaining one observation is c, will be used. The result is developed in a general case, where the number of units with $k+1$ observations is determined given that all the units are observed k times, for $k = 2,3,4,\ldots$.

Assume that at least k observations are obtained for each of the n sampling units. Let d be the number of units for which $k+1$ observations are obtained. Without loss of generality, let them be the first d units. Then

$$Y_{ti} = y_t + w_{ti}$$

and

$$X_{ti} = x_t + u_{ti}$$

for $i = 1,2,\ldots, k+1$ if $t = 1,2,\ldots, d$ and $i = 1,2,\ldots, k$ if $t = d+1,\ldots, n$, where (w_{ti}, u_{ti}), $i = 1,2,\ldots, k+1$, $t = 1,2,\ldots, d$, (w_{ti}, u_{ti}), $i = 1,2,\ldots, k$, $t = d+1,\ldots, n$ are

independent bivariate normal vectors with mean zero and variance covariance matrix

$$\begin{pmatrix} \sigma_{ww} & \sigma_{wu} \\ \sigma_{wu} & \sigma_{uu} \end{pmatrix} .$$

Let

$$S_{ww} = [n(k-1)+d]^{-1}\Big[\sum_{t=1}^{d}\sum_{i=1}^{k+1}(Y_{ti}-\bar{Y}_{t.})^2 + \sum_{t=d+1}^{n}\sum_{i=1}^{k}(Y_{ti}-\bar{Y}_{t.})^2\Big] ,$$

$$S_{uu} = [n(k-1)+d]^{-1}\Big[\sum_{t=1}^{d}\sum_{i=1}^{k+1}(X_{ti}-\bar{X}_{t.})^2 + \sum_{t=d+1}^{n}\sum_{i=1}^{k}(X_{ti}-\bar{X}_{t.})^2\Big] ,$$

and

$$S_{uw} = [n(k-1)+d]^{-1}\Big[\sum_{t=1}^{d}\sum_{i=1}^{k+1}(X_{ti}-\bar{X}_{t.})(Y_{ti}-\bar{Y}_{t.}) + \sum_{t=d+1}^{n}\sum_{i=1}^{k}(X_{ti}-\bar{X}_{t.})(Y_{ti}-\bar{Y}_{t.})\Big] , \quad (4.28)$$

where

$$\bar{Y}_{t.} = (k+1)^{-1}\sum_{i=1}^{k+1} Y_{ti} ,$$

$$\bar{X}_{t.} = (k+1)^{-1}\sum_{i=1}^{k+1} X_{ti} , \quad t = 1,2,\ldots, d ,$$

and

$$\bar{Y}_{t.} = k^{-1} \sum_{i=1}^{k} Y_{ti} ,$$

$$\bar{X}_{t.} = k^{-1} \sum_{i=1}^{k} X_{ti} , \quad t = d+1, \ldots, n .$$

Then S_{ww}, S_{uu} and S_{uw} are unbiased estimators of σ_{ww}, σ_{uu} and σ_{uw}, respectively.

For the first d units, the model can be rewritten as

$$y_t = \beta_0 + \beta_1 x_t + q_t ,$$

$$\bar{Y}_{t.} = y_t + \bar{w}_{t.} ,$$

$$\bar{X}_{t.} = x_t + \bar{u}_{t.} , \quad t = 1,2, \ldots, d , \tag{4.29}$$

where

$$\bar{w}_{t.} = (k+1)^{-1} \sum_{t=1}^{k+1} w_{ti} ,$$

$$\bar{u}_{t.} = (k+1)^{-1} \sum_{i=1}^{k+1} u_{ti}$$

and

$$\begin{pmatrix} x_t \\ q_t \\ \bar{w}_{t.} \\ \bar{u}_{t.} \end{pmatrix} \sim NI\left(\begin{pmatrix} \mu_x \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & 0 & 0 & 0 \\ 0 & \sigma_{qq} & 0 & 0 \\ 0 & 0 & (k+1)^{-1}\sigma_{ww} & (k+1)^{-1}\sigma_{wu} \\ 0 & 0 & (k+1)^{-1}\sigma_{wu} & (k+1)^{-1}\sigma_{uu} \end{pmatrix} \right) .$$

Under the normality assumption, $(\bar{Y}_{t.}, \bar{X}_{t.})$ and (S_{ww}, S_{uu}, S_{wu}) are independent for $t = 1,2,\ldots, d$. Thus, from (4.2) and (4.3), the estimator for β_1 based on the first d units is

$$\hat{\beta}_{1,d} = (m_{\bar{X}\bar{X},d} - (k+1)^{-1} S_{uu})^{-1}(m_{\bar{X}\bar{Y},d} - (k+1)^{-1} S_{wu}) \tag{4.30}$$

with asymptotic variance

$$\begin{aligned} V(\hat{\beta}_{1,d}) = \{d^{-1}\{\sigma_{xx}[\sigma_{qq} + (k+1)^{-1}\sigma_{rr}] \\ + (k+1)^{-1}\sigma_{uu}[\sigma_{qq} + (k+1)^{-1}\sigma_{rr}] \\ + (k+1)^{-2}\sigma_{ur}^2\} + \{[n(k+1)+d](k+1)^2\}^{-1}(\sigma_{uu}\sigma_{rr} \\ + \sigma_{ur}^2)\}\sigma_{xx}^{-2} , \end{aligned} \tag{4.31}$$

where

$$m_{\bar{X}\,\bar{X},d} = d^{-1} \sum_{t=1}^{d} (\bar{X}_{t.} - \bar{X}_{..})^2 ,$$

$$m_{\bar{X}\,\bar{Y},d} = d^{-1} \sum_{t=1}^{d} (\bar{X}_{t.} - \bar{X}_{..})(\bar{Y}_{t.} - \bar{Y}_{..}) ,$$

$$\bar{X}_{..} = d^{-1} \sum_{t=1}^{d} \bar{X}_{t.} ,$$

$$\bar{Y}_{..} = d^{-1} \sum_{t=1}^{d} \bar{X}_{t.} ,$$

$$\sigma_{rr} = \sigma_{ww} + \beta_1^2 \sigma_{uu} - 2\beta_1\sigma_{wu} ,$$

and

$$\sigma_{ur} = \sigma_{wu} - \beta_1\sigma_{uu} .$$

Similarly, the estimator for β_1 based on the last $n-d$ units is

$$\hat{\beta}_{1,n-d} = (m_{\bar{X}\,\bar{X},n-d} - k^{-1} S_{uu}^{-1}) \; (m_{\bar{X}\,\bar{Y},n-d} - k^{-1} S_{wu}) \quad (4.32)$$

with asymptotic variance

$$V(\hat{\beta}_{1,n-d}) = \{(n-d)^{-1}[\sigma_{xx}(\sigma_{qq} + k^{-1} \sigma_{rr}) + k^{-1} \sigma_{uu}(\sigma_{qq} + k^{-1} \sigma_{rr})$$

$$+ k^{-2}\sigma_{ur}^2] + \{[n(k-1) + d]k^2\}^{-1}(\sigma_{uu}\sigma_{rr} + \sigma_{ur}^2)\}\sigma_{xx}^{-2} ,$$

(4.33)

where

$$m_{\bar{X}\,\bar{X},n-d} = (n-d)^{-1} \sum_{t=d+1}^{n} (\bar{X}_{t.} - \bar{X}_{..})^2 ,$$

$$m_{\bar{X}\,\bar{Y},n-d} = (n-d)^{-1} \sum_{t=d+1}^{n} (\bar{X}_{t.} - \bar{X}_{..})(\bar{Y}_{t.} - \bar{Y}_{..}) ,$$

$$\bar{X}_{..} = (n-d)^{-1} \sum_{t=d+1}^{n} \bar{X}_{t.} ,$$

and

$$\bar{Y}_{..} = (n-d)^{-1} \sum_{t=d+1}^{n} \bar{Y}_{t.} .$$

Also, the asymptotic covariance of $\hat{\beta}_{1,d}$ and $\hat{\beta}_{1,n-d}$ is

$$\mathrm{Cov}(\hat{\beta}_{1,d}, \hat{\beta}_{1,n-d}) = \{k(k+1)[n(k-1) + d]\}^{-1}(\sigma_{uu}\sigma_{rr}$$

$$+ \sigma_{ur}^2)\sigma_{xx}^{-2} . \qquad (4.34)$$

Let

$$A = \sigma_{xx}(\sigma_{qq} + k^{-1}\sigma_{rr}) + k^{-1}\sigma_{uu}(\sigma_{qq} + k^{-1}\sigma_{rr}) + k^{-2}\sigma_{ur}^2 ,$$

$$B = \sigma_{xx}[\sigma_{qq} + (k+1)^{-1}\sigma_{rr}] + (k+1)^{-1}\sigma_{uu}[\sigma_{qq} + (k+1)^{-1}\sigma_{rr}] + (k+1)^{-2}\sigma_{ur}^2$$

and

$$D = \sigma_{uu}\sigma_{rr} + \sigma_{ur}^2 . \qquad (4.35)$$

Then, Equations (4.31), (4.33) and (4.34) can be expressed as

$$V_{11} = V(\hat{\beta}_{1,d}) = \{d^{-1}B + (k+1)^{-2}[n(k-1) + d]^{-1}D\}\sigma_{xx}^{-2} ,$$

$$V_{22} = V(\hat{\beta}_{1,n-d}) = \{(n-d)^{-1}A + k^{-2}[n(k-1) + d]^{-1}D\}\sigma_{xx}^{-2}$$

and

$$V_{12} = Cov(\hat{\beta}_{1,d}, \hat{\beta}_{1,n-d}) = k^{-1}(k+1)^{-1}[n(k-1) + d]^{-1}D\,\sigma_{xx}^{-2} . \qquad (4.36)$$

From (4.13), the variance of $p^* \hat{\beta}_{1,d} + (1-p^*)\hat{\beta}_{1,n-d}$ is

$$(V_{11} + V_{22} - 2V_{12})^{-1}(V_{11}V_{22} - V_{12}^2) , \qquad (4.37)$$

where $p^* = (V_{11} - 2V_{12} + V_{22})^{-1}(V_{22} - V_{12})$ is the optimal value of p that minimizes the variance of

$$\hat{\beta}_{1,p} = p\,\hat{\beta}_{1,d} + (1-p)\hat{\beta}_{1,n-d} .$$

If the cost of obtaining an observation is c units, then the total cost, T, for the survey is

$$T = c(kn + d) . \qquad (4.38)$$

Let

$$\eta = n\,d^{-1} . \qquad (4.39)$$

Given the total cost T, the value η that minimizes the variance (4.37) is obtained as follows. From (4.38) and (4.39), (4.37) becomes

$$\left(\frac{c}{T\sigma_{xx}^2}\right)[k(\eta-1)+k+1]\{AB+k^{-2}[(\eta-1)(k-1)+k]^{-1}(\eta-1)BD$$

$$+ (k+1)^{-2}[(\eta-1)(k-1)+k]^{-1}AD\}\{A+(\eta-1)B$$

$$+ [k(k+1)]^{-2}[(\eta-1)(k-1)+k]^{-1}(\eta-1)D\}^{-1}$$

$$\equiv f(\eta) \ . \tag{4.40}$$

Further simplification shows that $f(\eta)$ can be expressed as

$$\begin{aligned} f(\eta) = \left(\frac{c}{T\sigma_{xx}^2}\right)\{&[k(k-1)AB + k^{-1}\,BD](\eta-1)^2 \\ &+ [(2k-1)AB + k^{-2}(k+1)BD \\ &+ (k+1)^{-2}\,kAD](\eta-1) + k(k+1)AB \\ &+ (k+1)^{-1}\,AD\}\{(k-1)B(\eta-1)^2 + [(k-1)A + kB \\ &+ k^{-2}(k+1)^{-2}\,D](\eta-1) + kA\}^{-1} \ . \end{aligned} \tag{4.41}$$

Let

$$a_1 = k(k-1)AB + k^{-1}\,BD \ ,$$

$$b_1 = (2k^2 - 1)AB + k^{-2}(k+1)BD + (k+1)^{-2}\,kAD \ ,$$

$$c_1 = k(k+1)AB + (k+1)^{-1}\,AD \ ,$$

$$a_2 = (k-1)B \ ,$$

$$b_2 = (k-1)A + kB + k^{-2}(k+1)^{-2} D ,$$

and

$$c_2 = k A . \tag{4.42}$$

Then, Equation (4.41) becomes

$$f(\eta) = \left(\frac{c}{T\sigma_{xx}^2}\right) \frac{a_1(\eta-1)^2 + b_1(\eta-1) + c_1}{a_2(\eta-1)^2 + b_2(\eta-1) + c_2} . \tag{4.43}$$

Differentiating $f(\eta)$ with respect to η and setting the derivative equal to zero, we have

$$\frac{\partial f(\eta)}{\partial \eta} = \left(\frac{c}{T\sigma_{xx}^2}\right) \frac{(a_1b_2 - a_2b_1)(\eta-1)^2 + 2(a_1c_2 - a_2c_1)(\eta-1) + b_1c_2 - b_2c_1}{[a_2(\eta-1)^2 + b_2(\eta-1) + c_2]^2}$$

$$\equiv 0 . \tag{4.44}$$

Solving the equation (4.44) for η, we get

$$\eta-1 = (a_1b_2 - a_2b_1)^{-1}\{-(a_1c_2 - a_2c_1) \pm [(a_1c_2 - a_2c_1)^2$$

$$- (a_1b_2 - a_2b_1)(b_1c_2 - b_2c_1)]^{1/2}\} . \tag{4.45}$$

Further checking on the second derivative of $f(\eta)$ with respect to η shows that $f(\eta)$ is a minimum at

$$\eta-1 = (a_1b_2 - a_2b_1)^{-1}\{-(a_1c_2 - a_2c_1) + [(a_1c_2 - a_2c_1)^2 - (a_1b_2 - a_2b_1)(b_1c_2 - b_2c_1)]^{1/2}\} . \quad (4.46)$$

Since $d < n$, it follows that $\eta > 1$. Let

$$Q = (a_1c_2 - a_2c_1)^2 - (a_1b_2 - a_2b_1)(b_1c_2 - b_2c_1) . \quad (4.47)$$

If $Q < 0$, then two cases are to be considered.

Case I. If $a_1b_2 - a_2b_1 > 0$, then $\frac{\partial f(\eta)}{\partial \eta} > 0$ for all η which implies that $f(\eta)$ is a monotone increasing function. Thus, for $\eta > 1$, $f(\eta)$ is a minimum at $\eta = 1$.

Case II. If $a_1b_2 - a_2b_1 < 0$, then $\frac{\partial f(\eta)}{\partial \eta} < 0$ for all η which implies that $f(\eta)$ is a monotone decreasing function. Thus, for $\eta > 1$, $f(\eta)$ is a minimum at $\eta = \infty$.

For the case where $Q > 0$, if

$$(a_1b_2 - a_2b_1)^{-1}\{-(a_1c_2 - a_2c_1) + Q^{1/2}\} > 0 ,$$

then

$$\eta = 1 + (a_1 b_2 - a_2 b_1)^{-1}\{-(a_1 c_2 - a_2 c_1) + Q^{1/2}\} \ ;$$

if

$$(a_1 b_2 - a_2 b_1)^{-1}\{-(a_1 c_2 - a_2 c_1) + Q^{1/2}\} < 0 \ ,$$

then $\eta = 1$ when $a_2^{-1} a_1 < c_2^{-1} c_1$ and $\eta = \infty$ when $a_2^{-1} a_1 > c_2^{-1} c_1$.

To find out if triple observations are needed for some units when it is known that replicated observations are obtained on all the units, let $k = 2$. Tables for the optimal η corresponding to certain values of β_1, σ_{ww}, σ_{uu}, σ_{qq} and σ_{xx} are tabulated and are shown in Table 9. Without loss of generality, the tables are for $\sigma_{ww} = \sigma_{uu}$ and $\sigma_{wu} = 0$. From the tables, the optimal value of η is decreasing with respect to $\sigma_{uu}\sigma_{xx}^{-1}$. When $\eta = \infty$, that is, $d = 0$, only two observations are to be taken on all the sampling units.

Table 9. The optimal values of $\eta = nd^{-1}$ for given values of β_1, $\sigma_{xx}^{-1}\sigma_{uu}$ and $\sigma_{xx}^{-1}\sigma_{qq}$ under the assumption that at least two observations are taken on all sampling units.

$\beta_1 = 0.00$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.10	0.20	0.30	0.40	0.50	0.60	0.70
0.05		∞	4.40	1.00	1.00	1.00	1.00
0.10			∞	10.39	1.25	1.00	1.00
0.30				∞	∞	∞	8.06
0.50						∞	∞
0.70							∞
0.90							∞

$\beta_1 = 0.20$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.10	0.20	0.30	0.40	0.50	0.60	0.70
0.05		∞	3.02	1.00	1.00	1.00	1.00
0.10			∞	5.52	1.00	1.00	1.00
0.30				∞	∞	∞	4.75
0.50					∞	∞	∞
0.70						∞	∞
0.90							∞

Table 9 (continued)

$\beta_1 = 0.60$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.10	0.20	0.30	0.40	0.50	0.60	0.70
0.05		∞	1.00	1.00	1.00	1.00	1.00
0.10		∞	15.96	1.00	1.00	1.00	1.00
0.30				∞	∞	2.35	1.00
0.50				∞	∞	∞	6.24
0.70					∞	∞	∞
0.90						∞	∞

$\beta_1 = 1.00$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.10	0.20	0.30	0.40	0.50	0.60	0.70
0.05	∞	1.41	1.00	1.00	1.00	1.00	1.00
0.10		∞	1.00	1.00	1.00	1.00	1.00
0.30			∞	9.80	1.26	1.00	1.00
0.50			∞	∞	23.66	2.00	1.00
0.70				∞	∞	14.21	2.20
0.90					∞	∞	7.79

Table 9 (continued)

$\beta_1 = 1.40$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.10	0.20	0.30	0.40	0.50	0.60	0.70
0.05	∞	1.00	1.00	1.00	1.00	1.00	1.00
0.10		2.86	1.00	1.00	1.00	1.00	1.00
0.30		∞	44.57	1.09	1.00	1.00	1.00
0.50			∞	11.45	1.32	1.00	1.00
0.70			∞	∞	4.81	1.22	1.00
0.90				∞	∞	2.81	1.03

$\beta_1 = 1.80$

$\sigma_{xx}^{-1}\sigma_{qq}$	$\sigma_{xx}^{-1}\sigma_{uu}$ 0.10	0.20	0.30	0.40	0.50	0.60	0.70
0.05	54.82	1.00	1.00	1.00	1.00	1.00	1.00
0.10	∞	1.00	1.00	1.00	1.00	1.00	1.00
0.30		∞	1.76	1.00	1.00	1.00	1.00
0.50		∞	∞	1.43	1.00	1.00	1.00
0.70			∞	5.71	1.04	1.00	1.00
0.90			∞	∞	2.29	1.00	1.00

V. ON THE DETERMINATION OF THE NUMBER OF REPLICATED OBSERVATIONS FOR AN ERRORS-IN-VARIABLES MODEL WITH BINOMIAL OBSERVATIONS

A. Introduction

In the previous chapter, we considered a simple errors-in-variables model with the assumption that the response errors are normally distributed. The model of the previous chapter was

$$y_t = \beta_0 + \beta_1 x_t + q_t ,$$

$$Y_t = y_t + w_t ,$$

$$X_t = x_t + u_t , \quad t = 1,2,\ldots, n , \tag{5.1}$$

where

$$\begin{pmatrix} x_t \\ q_t \\ w_t \\ u_t \end{pmatrix} \sim NI\left(\begin{pmatrix} \mu_x \\ 0 \\ 0 \\ 0 \end{pmatrix} , \begin{pmatrix} \sigma_{xx} & 0 & 0 & 0 \\ 0 & \sigma_{qq} & 0 & 0 \\ 0 & 0 & \sigma_{ww} & \sigma_{wu} \\ 0 & 0 & \sigma_{wu} & \sigma_{uu} \end{pmatrix} \right)$$

and x_t and y_t are the true values of the variables of interest which cannot be measured exactly.

If the true values x and y are restricted to the two values, zero and one, and if the observed values X^* and Y^* are also

restricted to be zero or one, then any measurement errors must be correlated with the true value x.

Let u_t^* and w_t^* be the difference between the observed value X_t^* and the true value x_t and the difference between the observed value Y_t^* and the true value of y_t, respectively. Then

$$u_t^* = X_t^* - x_t$$

and

$$w_t^* = Y_t^* - y_t \,. \tag{5.2}$$

Let $u_t^*|(x_t, y_t)$ and $w_t^*|(x_t, y_t)$ denote the conditional random variables of u_t^* and w_t^* given (x_t, y_t), respectively. Assume that $u_t^*|(x_t, y_t)$ and $w_t^*|(x_t\ y_t)$ are independent and $u_t^*|(x_t, y_t)$ and $w_t^*|(x_t, y_t)$ are distributed as $u_t^*|x_t$ and $w_t^*|y_t$, respectively. That is, the response error u_t^* depends on the true value of x_t only and the response error w_t^* depends on the true value of y_t only.

Let $\gamma_{j(i)}$ be the probability that $X_t^* = j$ given that $x_t = i$ and $K_{j(i)}$ be the probability that $Y_t^* = j$ given that $y_t = i$, for $i,j = 0,1$. Let the fraction of the population whose true value of x is i and whose true value of y is j be P_{ij}, $i,j = 0,1$. Let the fraction of the population whose true value of x is one be $P_{1.}$ and let the population fraction of the observed X-values that are one be

P_X . Also, let the population fraction of true y-values that have the value one be $P_{.1}$ and let the population fraction of observed Y-values that have the value one be P_Y . Then

$$P_{1.} = P_{11} + P_{10}$$

$$P_{.1} = P_{01} + P_{11}$$

$$P_X = P_{1.}\gamma_{1(1)} + P_{0.}\gamma_{1(0)}$$

and

$$P_Y = P_{.1}K_{1(1)} + P_{.0}K_{1(0)} , \tag{5.3}$$

where

$$P_{0.} = 1 - P_{1.} \quad \text{and} \quad P_{.0} = 1 - P_{.1} .$$

Let

$$X_t = a_x(X_t^* - b_x) ,$$

$$Y_t = a_y(Y_t^* - b_y) , \tag{5.4}$$

where $a_x = (\gamma_{1(1)} - \gamma_{1(0)})^{-1}$, $b_x = \gamma_{1(0)}$, $a_y = (K_{1(1)} - K_{1(0)})^{-1}$, and $b_y = K_{1(0)}$. Then, by expressing

$$u_t = X_t - x_t ,$$

$$w_t = Y_t - y_t ,$$

we have

$$E(X_t|x_t) = x_t ,$$

$$E(X_t) = P_{1.} ,$$

$$E(Y_t|y_t) = y_t ,$$

$$E(Y_t) = P_{.1} ,$$

and

$$Cov(x_t, u_t) = Cov(x_t, w_t) = 0 . \qquad (5.5)$$

Also, assuming $E(q_t|x_t) = 0$, we have

$$Cov(u_t, q_t) = 0$$

and

$$\beta_0 = P_{0\cdot}^{-1} P_{01}$$

$$\beta_0 + \beta_1 = P_{1\cdot}^{-1} P_{11} \cdot \qquad (5.6)$$

B. The Variances of Estimators of β_1

From Equation (5.4) and (5.5), the response model (5.1) is written as

$$y_t = \beta_0 + \beta_1 x_t + q_t$$

$$Y_t = y_t + w_t$$

$$X_t = x_t + u_t , \qquad t = 1,2,3,\ldots, n , \qquad (5.7)$$

where (x_t, q_t, w_t, u_t) are independent for $t = 1,2,\ldots, n$, $u_t|(x_t, y_t)$ and $w_t|(x_t, y_t)$ are independent with

$$E(u_t) = E(w_t) = 0 ,$$

$$Cov(x_t, u_t) = Cov(x_t, w_t) = Cov(u_t, q_t) = 0 ,$$

and

$$\beta_1 = (P_{11} - P_{1\cdot}P_{\cdot 1})[P_{1\cdot}(1 - P_{1\cdot})]^{-1} .$$

Suppose that among the n units, d units are observed twice. Without loss of generality, let them be the first d units. Then

$$Y_{ti} = y_t + w_{ti}$$

and

$$X_{ti} = x_t + u_{ti} , \quad i = 1,2, \quad t = 1,2,\ldots, d , \qquad (5.8)$$

where (w_{ti}, u_{ti}) , $i = 1,2$, $t = 1,2,\ldots, d$, (w_t, u_t) , $t = d+1,\ldots, n$ are independent.

From the previous chapter, we consider estimators of β_1 based on the first d units and the last $n-d$ units. The two estimators are

$$\hat{\beta}_{1,d} = (m_{\bar{X}\bar{X}} - \tfrac{1}{2} S_{uu})^{-1}(m_{\bar{X}\bar{Y}} - \tfrac{1}{2} S_{wu}) ,$$

and

$$\hat{\beta}_{1,n-d} = (m_{XX} - S_{uu})^{-1}(m_{XY} - S_{wu}) , \qquad (5.9)$$

where

$$S_{uu} = (2d)^{-1} \sum_{t=1}^{d} (X_{t1} - X_{t2})^2 ,$$

$$S_{uw} = (2d)^{-1} \sum_{t=1}^{d} (X_{t1} - X_{t2})(Y_{t1} - Y_{t2}) ,$$

$$m_{XX} = (n - d)^{-1} \sum_{t=d+1}^{n} (X_t - \bar{X})^2 ,$$

$$m_{XY} = (n - d)^{-1} \sum_{t=d+1}^{n} (X_t - \bar{X})(Y_t - \bar{Y}) ,$$

$$\bar{X} = (n - d)^{-1} \sum_{t=d+1}^{n} X_t ,$$

$$\bar{Y} = (n - d)^{-1} \sum_{t=d+1}^{n} Y_t ,$$

$$m_{\bar{X}\bar{X}} = d^{-1} \sum_{t=1}^{d} (\bar{X}_{t.} - \bar{X}_{..})^2 ,$$

$$m_{\bar{X}\bar{Y}} = d^{-1} \sum_{t=1}^{d} (\bar{X}_{t.} - \bar{X}_{..})(\bar{Y}_{t.} - \bar{Y}_{..}) ,$$

$$\bar{X}_{t.} = 2^{-1}(X_{t1} + X_{t2}) , \quad t = 1,2,\ldots, d$$

$$\bar{Y}_{t.} = 2^{-1}(Y_{t1} + Y_{t2}) , \quad t = 1,2,\ldots, d$$

$$\bar{X}_{..} = d^{-1} \sum_{t=1}^{d} \bar{X}_{t.} ,$$

$$\bar{Y}_{..} = d^{-1} \sum_{t=1}^{d} \bar{Y}_{t.} ,$$

and (X_t, Y_t) is defined in (5.4). The fourth moments exist for the random variables u_{ti}, w_{ti}, $i = 1,2$, x_t, $t = 1,2,\ldots, d$, u_t, w_t, x_t, $t = d+1,\ldots, n$. Thus, following Fuller (1980), the asymptotic variances of $\hat{\beta}_{1,d}$ and $\hat{\beta}_{1,n-d}$ and the asymptotic covariance of $\hat{\beta}_{1,d}$ and $\hat{\beta}_{1,n-d}$ are

$$V(\hat{\beta}_{1,d}) \doteq V_{11} = [\frac{1}{d} B + \frac{1}{4d} D - \frac{1}{2d} F][(P_{1.}(1 - P_{1.})]^{-2} ,$$

$$V(\hat{\beta}_{1,n-d}) \doteq V_{22} = [\frac{1}{n-d} A + \frac{1}{d} D][P_{1.}(1 - P_{1.})]^{-2} ,$$

$$\mathrm{Cov}(\hat{\beta}_{n-d}, \hat{\beta}_{d}) \doteq V_{12} = [\frac{1}{2d} D - \frac{1}{d} F][P_{1.}(1 - P_{1.})]^{-2} , \quad (5.10)$$

where

$$A = V\{[x_t - E(x_t)]q_t + q_t u_t + [x_t - E(x_t)]r_t + u_t r_t - \sigma_{ur}\}$$

$$B = V\{[x_t - E(x_t)]q_t + q_t \bar{u}_t + [x_t - E(x_t)]\bar{r}_t + \bar{u}_t \bar{r}_t - \frac{1}{2}\sigma_{ur}\}$$

$$D = V\{\frac{1}{2}(u_{t1} - u_{t2})(r_{t1} - r_{t2}) - \sigma_{ur}\}$$

$$F = \mathrm{Cov}\{[x_t - E(x_t)]q_t + q_t \bar{u}_t + [x_t - E(x_t)]\bar{r}_t + \bar{u}_t \bar{r}_t - \tfrac{1}{2}\sigma_{ur} ,$$

$$\frac{1}{2}(u_{t1} - u_{t2})(r_{t1} - r_{t2}) - \sigma_{ur}\} ,$$

$$r_t = w_t - \beta_1 u_t ,$$

$$\bar{u}_t = \frac{1}{2} (u_{t1} + u_{t2}) ,$$

$$\bar{r}_t = \frac{1}{2} (r_{t1} + r_{t2}) ,$$

and

$$\sigma_{ur} = \mathrm{Cov}(u_t, r_t) .$$

Further simplication of A, B, D , and F shows that

$$A = E\{[x_t - E(x_t)]^2 q_t^2 + q_t^2 u_t^2 + [x_t - E(x_t)]^2 r_t^2 + u_t^2 r_t^2 - \sigma_{ur}^2$$

$$+ 2[x_t - E(x_t)]r_t^2 u_t\} ,$$

$$B = E\{[x_t - E(x_t)]^2 q_t^2 + \frac{1}{2} q_t^2 u_t^2 + \frac{1}{2} [x_t - E(x_t)]^2 r_t^2 + \bar{u}_t^2 \bar{r}_t^2$$

$$- \frac{1}{4} \sigma_{ur}^2 + \frac{1}{2} [x_t - E(x_t)]r_t^2 u_t\} ,$$

$$D = E(\frac{1}{2} u_t^2 r_t^2 + u_{t1} r_{t1} u_{t2} r_{t2} + \frac{1}{2} u_{t1}^2 r_{t2}^2 - \sigma_{ur}^2) ,$$

$$F = E\{\tfrac{1}{2}\,[x_t - E(x_t)]u_t r_t^2 + \tfrac{1}{4}\,u_t^2 r_t^2 - \tfrac{1}{4}\,u_t^2\,r^2 - \tfrac{1}{4}\,u_{t1}^2 u_{t2}^2 - \tfrac{1}{2}\,\sigma_{ur}^2\}\ ,$$

(5.11)

where

$$E\{[x_t - E(x_t)]^2\, q_t^2\} = P_{1.}(1 - P_{1.})[\beta_0(1 - \beta_0) + \beta_1(1 - \beta_1)(1 - P_{1.})$$

$$- 2\,\beta_1\beta_0(1 - P_{1.})]\ ,$$

$$E(u_t^2\, q_t^2) = a_x^2\, b_x(1 - b_x)[P_{.1}(1 - P_{.1}) + \beta_1^2\, P_{1.}(1 - P_{1.})$$

$$- 2\,\beta_1(P_{11} - P_{1.}P_{.1})]$$

$$+ [a_x(1 - 2\gamma_{1(1)}) + 1]P_{11}P_{10}P_{1.}^{-1}\ ,$$

$$E\{[x_t - E(x_t)]^2\, r_t^2\} = [a_y^2\, b_y(1 - b_y) + \beta_1^2\, a_x^2\, b_x(1 - b_x)]P_{1.}(1 - P_{1.})$$

$$+ \beta_1^2[a_x(1 - 2\gamma_{1(1)}) + 1](1 - P_{1.})^2\, P_{1.}$$

$$+ [a_y(1 - 2K_{1(1)}) + 1][(1 - 2P_{1.})P_{11} + P_{1.}^2 P_{.1}]\ ,$$

$$E(u_t^2\, r_t^2) = a_y^2 b_y(1 - b_y)a_x^2 b_x(1 - b_x) + a_x^4 b_x(1 - 4b_x + 6b_x^2 - 3b_x^3)$$

$$+ [a_y(1 - 2K_{1(1)}) + 1]\{a_x^2 b_x(1 - b_x)P_{.1}$$

$$+ [a_x(1 - 2\gamma_{1(1)}) + 1]P_{11}\}$$

$$+ \{a_y^2 b_y(1 - b_y)[a_x(1 - 2\gamma_{1(1)}) + 1]$$

$$+ \beta_1^2 a_x^4(1 - 4b_x + 6b_x^2 - 4b_x^3)(\gamma_{1(1)} - \gamma_{1(0)})$$

$$- \beta_1^2[4a_x^3(1 - 3b_x + 3b_x^2)$$

$$- 6a_x^2(1 - 2b_x) + 4a_x]\gamma_{1(1)}$$

$$+ \beta_1^2[4a_x^3 b_x^3 + 6a_x^2 b_x^2 + 4a_x b_x + 1]\}P_{1.} ,$$

$$E\{[x_t - E(x_t)]u_t r_t^2\} = [a_x^3(1 - 3b_x + 3b_x^2)(\gamma_{1(1)} - \gamma_{1(0)})$$

$$- 3a_x^2(1 - 2b_x)\gamma_{1(1)}$$

$$+ 3a_x\gamma_{1(1)} - 3a_x^2 b_x^2 - 3a_x b_x - 1]P_{1.}(1 - P_{1.}) ,$$

$$\sigma_{ur} = - \beta_1\{a_x^2 b_x(1 - b_x) + [a_x(1 - 2\gamma_{1(1)}) + 1]P_{1.}\} ,$$

$$E(\bar{u}_t^2 \bar{r}_t^2) = \frac{1}{8} E(u_t^2 r_t^2 + u_{t1}^2 r_{t2}^2 + 2u_{1t} r_{1t} u_{2t} r_{2t}) ,$$

$$E(u_{t1}^2 r_{t2}^2) = a_x^2 b_x(1 - b_x)[a_y^2 b_y(1 - b_y) + \beta_1^2 a_x^2 b_x(1 - b_x)]$$

$$+ a_x^2 b_x(1 - b_x)[a_y(1 - 2K_{1(1)}) + 1]P_{.1}$$

$$+ a_x^2 b_x(1 - b_x)\beta_1^2[a_x(1 - 2\gamma_{1(1)} + 1]P_{1.}$$

$$+ [a_x(1 - 2\gamma_{1(1)}) + 1]\{\{a_y^2 b_y(1 - b_y) + \beta_1^2 a_x^2 b_x(1 - b_x)$$

$$+ \beta_1^2[a_x(1 - 2\gamma_{1(1)}) + 1]\}P_{1.}$$

$$+ [a_y(1 - 2K_{1(1)}) + 1]P_{11}\} ,$$

$$E(u_{t1}r_{t1}u_{t2}r_{t2}) = \beta_1^2 E(u_{t1}^2 u_{t2}^2) ,$$

and

$$E(u_{t1}^2 u_{t2}^2) = [a_x^2 b_x(1 - b_x)]^2 + [a_x(1 - 2\gamma_{1(1)}) + 1]\{2a_x^2 b_x(1 - b_x)$$

$$+ [a_x(1 - 2\gamma_{1(1)}) + 1]\}P_{1.} .$$

From Appendix B, the θ that minimizes the variance of $\hat{\beta}_{1,\theta}$ is given by

$$\theta^* = (V_{11} - 2V_{12} + V_{22})^{-1}(V_{22} - V_{12}) , \qquad (5.12)$$

where $\hat{\beta}_{1,\theta} = \theta\hat{\beta}_{1,d} + (1 - \theta)\hat{\beta}_{1,n-d}$. The variance of $\hat{\beta}_{1,\theta^*}$ is

$$(V_{11} + V_{22} - 2V_{12})^{-1}(V_{11}V_{22} - V_{12}^2)$$

$$= \{d\, P_{1.}^2 (1 - P_{1.})^2 [A + (n - 1)(B + \tfrac{1}{4} D + \tfrac{3}{2} F)]\}^{-1} \{[A$$

$$+ (n - 1)D][B + \tfrac{1}{4} D - \tfrac{1}{2} F] - (n - 1)(\tfrac{1}{2} D - F)^2\} ,$$

(5.13)

where

$$n = nd^{-1} . \tag{5.14}$$

C. Determination of Number of Replicated Measurement Units

Assume that the cost of obtaining an observation is c per unit. Then the total cost, denoted by T, for a survey of n units in which d are observed twice is

$$T = c(n + d) , \tag{5.15}$$

where d is the number of the units that have replicate measurements. Suppose that the total cost T for the survey is fixed. The value n, where $n = nd^{-1}$, that minimizes the variance (5.13) subject to the cost function (5.15) is obtained as follows. From (5.14) and (5.15), (5.13) becomes

$$\frac{c(n+1)}{T\,P_{1.}^2(1-P_{1.})^2}\left\{\frac{[A+(n-1)D][B+\frac{1}{4}D-\frac{1}{2}F]-(n-1)(\frac{1}{2}D-F)^2}{A+(n-1)(B+\frac{1}{4}D+\frac{3}{2}F)}\right\}$$

$$\equiv f(n)\ . \qquad (5.16)$$

Differentiating $f(n)$ with respect to n and setting the derivative equal to zero, we have

$$\frac{\partial f(n)}{\partial n} = \frac{c}{P_{1.}^2(1-P_{1.})^2T}[A+(n-1)(B+\tfrac{1}{4}D+\tfrac{3}{2}F)]^{-2}\{[DB(B+\tfrac{1}{4}D)$$

$$+(2B+\tfrac{1}{8}D)DF+(\tfrac{1}{2}D-B)F^2-\tfrac{3}{2}F^3](n-1)^2$$

$$+[2ABD+ADF-2AF^2](n-1)+A^2(B+\tfrac{1}{4}D)$$

$$+2ABD-2A(B+\tfrac{1}{4}D)^2+[2A(\tfrac{1}{4}D-B)$$

$$-\tfrac{1}{2}A^2]F-\tfrac{1}{2}AF^2=0\ . \qquad (5.17)$$

Let

$$A^* = DB(B+\tfrac{1}{4}D)+(2B+\tfrac{1}{8}D)DF+(\tfrac{1}{2}D-B)F^2-\tfrac{3}{2}F^3$$

$$B^* = 2ABD+ADF-2AF^2\ ,$$

$$C^* = A^2(B + \tfrac{1}{4} D) + 2ABD - 2A(B + \tfrac{1}{4} D)^2 + [2A(\tfrac{1}{4} D - B) - \tfrac{1}{2} A^2]F$$

$$- \tfrac{1}{2} AF^2 \ . \qquad (5.18)$$

Solving the Equation (5.17) for η , we get

$$\eta - 1 = (2A^*)^{-1}[- B^* \pm (B^{*2} - 4A^*C^*)^{1/2}] \ . \qquad (5.19)$$

Two cases are to be considered. If $B^{*2} - 4A^* C^* > 0$, then the η in Equation (5.19) is a real number. The second derivative of $f(\eta)$ with respect to η is

$$\frac{\partial^2 f(\eta)}{\partial \eta^2} = \frac{c}{P_{1.}^2 (1-P_{1.}^2)^2 \ T} \left\{ \frac{2A^*(\eta - 1) + B^*}{[A + (\eta-1)(B + 1/4 \ D + 3/2 \ F)]^2} \right.$$

$$\left. - \frac{2[A^*(\eta - 1)^2 + B^*(\eta - 1) + C^*](B + 1/4 \ D + 3/2 \ F)}{[A + (\eta - 1)(B + 1/4 \ D + 3/2 \ F)]^3} \right\}$$

(5.20)

which is positive when

$$\eta - 1 = (2A^*)^{-1}[- B^* + (B^{*2} - 4A^* \ C^*)^{1/2}] \ . \qquad (5.21)$$

But $\eta > 1$, it follows that

$$\eta = 1 \quad \text{if} \quad (2A^*)^{-1}[- B^* + (B^{*2} - 4A^*C^*)^{1/2}] < 0$$

$$= 1 + (2A)^{-1}[- B^* + (B^{*2} - 4A^*C^*)^{1/2}] \quad \text{otherwise} .$$

(5.22)

When $B^{*2} - 4A^*C^* < 0$ and $A^* > 0$, it can be shown that $\frac{\partial f(\eta)}{\partial \eta}$ is greater than zero for all $\eta > 1$, which implies that $f(\eta)$ is a monotone increasing function. Thus, for $\eta > 1$, $f(\eta^*)$ is a minimum when $\eta^* = 1$.

When $B^{*2} - 4A^*C^* < 0$ and $A^* < 0$, then $\frac{\partial f(\eta)}{\partial \eta}$ is less than zero for all $\eta > 1$, which implies that $f(\eta)$ is a monotone decreasing function. This situation would not occur because at least one unit with replicate observations is needed in order to estimate the variances of the errors.

Table 10 contains the optimal value of η where the response probabilities are given by the unbiased response model proposed by Battese and Fuller (1974). The Battese-Fuller model is

$$\gamma_{i(j)} = 1 - \alpha + \alpha P_{i.} \qquad i=j$$

$$= \alpha P_{i.} \qquad i \neq j ,$$

$$K_{i(j)} = 1 - \alpha + \alpha P_{.i} \qquad i=j$$

$$= \alpha P_{.i} \qquad i \neq j \ , \tag{5.23}$$

where α is the parameter of the model.

The tables show that when α increases, the optimal value of η decreases. Intuitively, it says that if the probability of making a correct classification is small, more units have to be observed twice. For example, assume that one has enough money for 1,000 observations. For design purposes, it is assumed that $P_{11} = 0.50$, $P_{1.} = 0.80$ and $P_{.1} = 0.60$. If $\alpha = 0.05$, then by using Equation (5.23), $\gamma_{1(1)} = 0.99$ and $\kappa_{1(1)} = 0.98$. The optimal value of η is 8.37, and thus, $n = 893$ and $d = 107$. If $\alpha = 0.15$, then $\gamma_{1(1)} = 0.97$ and $\kappa_{1(1)} = 0.94$, which are smaller than the classification probabilities obtained for $\alpha = 0.05$. For $\alpha = 0.15$, the optimal value of η is 2.43 and the values of n and d are 708 and 292, respectively. This shows that more units have to be observed twice when the true classification probabilities are small.

We compare the optimal value of η obtained under the actual distribution of x, u, w, and q with the optimal value of η obtained under the assumption that these random variables are normally distributed with the mean vector and covariance matrix defined in Equation (5.1). Under the normality assumption the optimal value of η is calculated using the Equation (4.27) derived in Chapter IV. For

Table 10. The optimal values of n for selected values of α, $P_{1.}$, $P_{.1}$ and P_{11}. (Notation: A * means that the three values $P_{1.}$, $P_{.1}$ and P_{11} are incompatible.)

$\alpha = 0.05$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	0.90	0.80
0.90	2.45	1.00
0.80	3.48	1.00

$\alpha = 0.10$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	0.90	0.80
0.90	1.00	1.00
0.80	2.04	1.00

$\alpha = 0.15$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	0.90	0.80
0.90	1.00	1.00
0.80	1.30	1.00

$\alpha = 0.20$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	1.00	1.00
0.90	1.00	1.00
0.80	1.00	1.00

$\alpha = 0.05$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	2.63	1.00
0.60	4.94	2.98	1.00

$\alpha = 0.10$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	1.22	1.00
0.60	2.83	1.68	1.00

Table 10 (continued)

$\alpha = 0.15$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	1.00	1.00
0.60	1.79	1.00	1.00

$\alpha = 0.20$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	1.00	1.00
0.60	1.12	1.00	1.00

$\alpha = 0.05$, $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	*	8.37
0.60	4.75	8.84	2.67

$\alpha = 0.10$, $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	*	4.09
0.60	2.53	4.25	1.45

$\alpha = 0.15$, $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	*	2.43
0.60	1.47	2.51	1.00

$\alpha = 0.20$ $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.00
0.80	*	*	1.48
0.60	1.00	1.53	1.00

different values of $P_{1.}$, $P_{.1}$, P_{11} and α, the optimal value of n under the normality assumption is tabulated in Table 11. By comparing Table 10 with Table 11, one sees that the normal approximation to the Bernoulli distribution does not perform well in this case.

Table 11. The optimal values of η for selected values of α, $P_{1.}$, $P_{.1}$ and P_{11} under the normality assumptions on the response errors. (Notation: A * means that the three values $P_{1.}$, $P_{.1}$ and P_{11} are incompatible.)

$\alpha = 0.05$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	0.90	0.80
0.90	9.40	5.20
0.80	5.20	1.00

$\alpha = 0.10$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	0.90	0.80
0.90	4.32	2.21
0.80	2.21	1.00

$\alpha = 0.15$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	0.90	0.80
0.90	2.50	1.06
0.80	1.06	1.00

$\alpha = 0.20$, $P_{11} = 0.80$

$P_{1.}$	$P_{.1}$	
	0.90	0.80
0.90	1.51	1.00
0.90	1.00	1.00

$\alpha = 0.05$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	7.58
0.80	*	8.75	5.73
0.60	7.58	5.73	1.00

$\alpha = 0.10$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	3.42
0.80	*	4.00	2.49
0.60	3.42	2.49	1.00

Table 11 (continued)

$\alpha = 0.15$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.90
0.80	*	2.29	1.26
0.60	1.90	1.26	1.00

$\alpha = 0.20$, $P_{11} = 0.60$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.06
0.80	*	1.35	1.00
0.60	1.06	1.00	1.00

$\alpha = 0.05$, $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	8.61
0.80	*	*	9 43
0.60	8.61	9.43	6.01

$\alpha = 0.10$, $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	3.93
0.80	*	*	4.33
0.60	3.93	4.33	2.63

$\alpha = 0.15$, $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	2.24
0.80	*	*	2.51
0.60	2.24	2.51	1.36

$\alpha = 0.20$, $P_{11} = 0.50$

$P_{1.}$	$P_{.1}$		
	0.90	0.80	0.60
0.90	*	*	1.31
0.80	*	*	1.51
0.60	1.31	1.51	1.00

VI. BIBLIOGRAPHY

Assakul, K. and C. H. Proctor. 1967. Testing independence in two-way contingency tables with data subject to misclassification. Psychometrika 32:67-76.

Bailar, B. A. 1968. Recent research in reinterview procedures. Journal of the American Statistical Association 63:41-63.

Bailar, B. A. 1976. Some sources of error and their effect on census statistics. Demography 13:273-286.

Bailar, B. A. and T. Dalenius. 1969. Estimating the response variance components of the U.S. Burea of the Census survey model. Sankhya, Series B, 31:341-360.

Battese, G. E. and W. A. Fuller. 1974. An unbiased response model for analysis of categorical data. Unpublished report to the U.S. Bureau of the Census.

Battese, G. E., W. A. Fuller and R. D. Hickman. 1976. Estimation of response variances from interview-reinterview surveys. Journal of the Indian Society of Agricultural Statistics 28(2):1-14.

Bershad, M. A. 1967. Gross changes in the presence of response errors. Unpublished memorandum U.S. Bureau of the Census, Washington, D.C.

Bross, I. 1954. Misclassification in 2 x 2 tables. Biometrics 10:478-486.

Bryson, M. R. 1965. Errors of classification in a binomial population. Journal of the American Statistical Association 60:217-224.

Cochran, W. G. 1968. Errors of measurement in statistics. Technometrics 10:637-666.

Deming, W. E. 1944. On errors in surveys. American Sociological Review 9:359-369.

Eckler, A. R. and W. N. Hurwitz. 1958. Response variance and biases in censuses and surveys. Bulletin de Institut International de Statistique 36(2):12-35.

Eckler, A. R. and L. Pritzker. 1951. Measuring the accuracy of enumerative surveys. Bulletin de Institut International de Statistique 33(4):7-24.

Fellegi, I. P. 1964. Response variance and its estimation. Journal of the American Statistical Association 59:1016-1041.

Fleiss, J. L. 1981. Statistical methods for rates and proportions. 2nd ed. John Wiley & Sons, New York.

Fuller, W. A. 1976. Introduction to statistical time series. John Wiley & Sons, New York.

Fuller, W. A. 1980. Properties of some estimators for the errors-in-variables model. The Annals of Statistics 8:407-422.

Giesbrecht, F. G. 1967. Classification errors and measures of association in contingency tables. Proceedings of the Social Statistical Section of ASA 1967:271-276.

Hansen, M. H., W. N. Hurwitz and M. A. Bershad. 1961. Measurement errors in censuses and surveys. Bulletin de Institut International de Statistique 38(2):359-374.

Hansen, M. H., W. N. Hurwitz and L. Pritzker. 1964. The estimation and interpretation of gross differences and the simple response variance. Pages 111-136 in C. R. Rao ed. Contributions to statistics. Pergamon Press, Oxford.

Hansen, M. H., W. N. Hurwitz, E. S. Mark and W. P. Mauldin. 1951. Response errors in surveys. Journal of the American Statistical Association 46:147-190.

Hanson, R. H. and E. S. Marks. 1958. Influence of the interviewer on the accuracy of survey results. Journal of the American Statistical Association 53:635-655.

Hartley, H. O. and J. N. K. Rao. 1978. Estimation of nonsampling variance components in sample surveys. Pages 35-43 in Nambrodiri, N. K. ed. Survey sampling and measurement. Academic Press, New York.

Kalton, G. and H. Schuman. 1982. The effect of the question on survey response: a review. Journal of the Royal Statistical Society, Series A, 145:42-73.

Kish, L. 1962. Studies of interviewer variance for attitudinal variables. Journal of the American Statistical Association 57:92-115.

Koch, G. G. 1969. The effect of nonsampling errors on measures of association in 2 x 2 contingency tables. Journal of the American Statistical Association 64:852-863.

Koch, G. G. 1973. An alternative approach to multivariate response error models for sample survey data with aplication to estimators involving subclass means. Journal of the American Statistical Association 68:906-913.

Koop, J. C. 1974. Note for a unified theory of estimation for sample surveys taking into account response errors. Metrika 21:19-39.

Korn, E. L. 1981. Hierarchical log-linear models not preserved by classification error. Journal of the American Statistical Association 76:110-113.

Korn, E. L. 1982. The asymptotic efficiency of tests of using misclassified data in contingency tables. Biometrics 38:445-450.

Krishnaswani, P. and R. Nath. 1968. Bias in multinomial classification. Journal of the American Statistical Association 63:298-303.

Mahalanobis, P. C. 1946. Recent experiments in statistical sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society 109:325-378.

Mote, V. L. and R. L. Anderson. 1965. An investigation of the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data. Biometrika 52(1):95-109.

Nathan, G. 1973. Response errors of estimators based on different samples. Sankhya Series A 35:205-220.

Pearson, K. 1902. On the mathematical theory of errors of judgment. Philosophical Transactions of the Royal Society of London A 198:235-299.

Pritzker, L. and R. Hanson. 1962. Measurement errors in the 1960 census of population. Proceeding of the Social Sciences Section of the American Statistical Association 1962:80-90

Rao, C. R. 1973. Linear statistical inference and its applications. John Wiley & Sons, New York.

Statistical Analysis System. 1982. SAS user's guide: statistics. SAS Institute Inc., Cary, North Carolina.

Stock, J. S. and J. R. Hochstim. 1951. A method of measuring interviewer variability. Public Opinion Quarterly 15:322-234.

Sukhatme, P. V. and G. R. Seth. 1952. Non-sampling errors in surveys. Journal of the Indian Society of Agricultural Statistics 4:5-41.

U.S. Bureau of the Census. 1968. Evaluation and research program of the U.S. censuses of population and housing, 1960: Effects of interviewers and crew leaders. Series ER 60 No. 7, Washington, D.C.

U.S. Bureau of the Census. 1972. Evaluation and research program of the U.S. census of population and housing, 1960: Effects of different reinterview techniques on estimates of simple response variances. Series ER 60 No. 7, Washington, D.C.

U.S. Bureau of the Census. 1974. Evaluation and research program of the U.S. censuses population and housing, 1970: Accuracy of data for selected population characteristics as measured by reinterviews. PHC(E)-9, Washington, D.C.

U.S. Bureau of the Census. 1975. Evaluation and research program of the U.S. censuses population and housing, 1970: Accuracy of data for selected population characteristics as measured by the 1970 CPS-Census match. PHC(E)-11, Washington, D.C.

VII. ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to Dr. W. A. Fuller for introducing me to the area of classification errors. Through his assistance, guidance and patience this dissertation became possible.

I greatly appreciate the financial support given me by the Statistical Laboratory during the course of my graduate program at Iowa State University. I had such a wonderful time in the Survey Section that it is difficult for me to leave.

Thanks go to Jane Stowe for her efficient and excellent typing of this manuscript.

Finally, I would like to thank my parents and other members of the family for their continuous support and encouragement during my many years of graduate study.

VIII. APPENDIX A

In Chapter IV, the asymptotic covariance of $\hat{\beta}_{1,d}$ and $\hat{\beta}_{1,n-d}$ is expressed as

$$\mathrm{Cov}(\hat{\beta}_{1,d}, \hat{\beta}_{1,n-d}) = \sigma_{xx}^{-2}[(2d)^{-1}(\sigma_{uu}\sigma_{rr} + \sigma_{ur}^2)] \ . \qquad (8.1)$$

We derive the above expression using the notation of Chapter IV. Recall that

$$\hat{\beta}_{1,d} = (m_{\bar{X}\,\bar{X}} - \tfrac{1}{2} S_{uu})^{-1}(m_{\bar{X}\,\bar{Y}} - \tfrac{1}{2} S_{wu})$$

and

$$\hat{\beta}_{1,n-d} = (m_{XX} - S_{uu})^{-1}(m_{XY} - S_{wu}) \ .$$

Because the sample moments are converging to the population moments, we can expand $\hat{\beta}_{1,d}$ and $\hat{\beta}_{1,n-d}$ in Taylor's series about the population values to obtain

$$\hat{\beta}_{1,d} - \beta_1 = \sigma_{xx}^{-1}(m_{\bar{X}\,\bar{v}} - \tfrac{1}{2} S_{ur}) + O_p(n^{-1})$$

and

$$\hat{\beta}_{1,n-d} - \beta_1 = \sigma_{xx}^{-1}(m_{Xv} - S_{ur}) + O_p(n^{-1}) \ ,$$

where

$$\bar{v}_{t.} = q_t + \bar{w}_{t.} - \beta_1 \bar{u}_{t.} , \quad t = 1,2,\ldots, d ,$$

$$v_t = q_t + w_t - \beta_1 u_t , \quad t = d+1,\ldots, n ,$$

and

$$r_t = w_t - \beta_1 u_t , \quad t = d+1,\ldots, n .$$

The $\underline{m}_{\underline{X}\,\bar{v}}$ is obtained from the first d sampling units and M_{Xv} is obtained from the last $n - d$ sampling units. Thus, $\underline{m}_{\underline{X}\,\bar{v}}$, M_{Xv} and S_{ur} are independent. The asymptotic covariance of $\hat{\beta}_{1,d}$ and $\hat{\beta}_{1,n-d}$ is

$$\mathrm{Cov}[\sigma_{xx}^{-1}(\underline{m}_{\underline{X}\,\bar{v}} - \tfrac{1}{2} S_{ur}), \sigma_{xx}^{-1}(m_{Xv} - S_{ur})]$$

$$= \sigma_{xx}^{-2} \mathrm{Cov}(\tfrac{1}{2} S_{ur}, S_{ur})$$

$$= (2\sigma_{xx}^2)^{-1} \mathrm{Var}(S_{ur})$$

$$= \sigma_{xx}^{-2}[(2d)^{-1}(\sigma_{uu}\sigma_{rr} + \sigma_{ur}^2)] .$$

IX. APPENDIX B

Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be two estimators of β with variances V_{11} and V_{22}, respectively. Let V_{12} be the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$. Consider an estimator $\hat{\beta}_p$ which is a linear combination of $\hat{\beta}_1$ and $\hat{\beta}_2$. That is,

$$\hat{\beta}_p = p\,\hat{\beta}_1 + (1 - p)\hat{\beta}_2 , \tag{9.1}$$

where p is any constant. The variance of $\hat{\beta}_p$ is

$$p^2 V_{11} + (1 - p)^2 V_{22} + 2p(1 - p)V_{12} . \tag{9.2}$$

Let p^* denote the value of p that gives the smallest variance of $\hat{\beta}_p$. Then p^* can be obtained as follows. By equating the first derivative of the variance of $\hat{\beta}_p$ that is taken with respect to p to zero, we have

$$2p\,V_{11} - 2(1 - p)V_{22} + 2(1 - 2p)V_{12} = 0 . \tag{9.3}$$

Then

$$p^* = (V_{11} + V_{12} - 2V_{12})^{-1}(V_{22} - V_{12}) . \tag{9.4}$$

The second derivative of the variance of $\hat{\beta}_p$ taken with respect to p is

$$2(V_{11} + V_{22} - 2V_{12}) \ . \tag{9.5}$$

Equation (9.5) is greater than or equal to zero with equality if and only if the difference of $\hat{\beta}_1$ and $\hat{\beta}_2$ is a constant. Thus, $\hat{\beta}_{p^*}$, where p^* is defined in Equation (9.4) has the smallest variance among all the estimators with the form defined in Equation (9.1). The variance of $\hat{\beta}_{p^*}$ is

$$\begin{aligned} & p^{*2} V_{11} + (1 - p^*)^2 V_{22} + 2p^*(1 - p^*)V_{12} \\ & = (V_{11} + V_{22} - 2V_{12})^{-2}[(V_{22} - V_{12})^2 V_{11} \\ & \qquad + (V_{11} - V_{12})^2 V_{22} \\ & \qquad + 2(V_{22} - V_{12})(V_{11} - V_{12})V_{12}] \\ & = (V_{22} + V_{11} - 2V_{12})^{-2}[V_{22}V_{11}(V_{22} + V_{11} - 2V_{12}) \\ & \qquad - V_{12}^2(V_{22} + V_{11} - 2V_{12})] \\ & = (V_{22} + V_{11} - 2V_{12})^{-1}(V_{22}V_{11} - V_{12}^2) \ . \end{aligned} \tag{9.6}$$