NASA Contractor Report 177933

NASA-CR-177933
19850024666

STATISTICAL METHODS FOR EFFICIENT DESIGN OF COMMUNITY SURVEYS OF RESPONSE TO NOISE: RANDOM COEFFICIENTS REGRESSION MODELS

Thomas J. Tomberlin

THE BIONETICS CORPORATION
Hampton, Virginia

THOMAS J. TOMBERLIN, CONSULTANT (SUBCONTRACTOR)
Montreal, Quebec, Canada

Contract NAS1-16978
July 1985

National Aeronautics and Space Administration

Langley Research Center
Hampton, Virginia 23665

TABLE OF CONTENTS

1. Introduction

Many social surveys have as their main purpose, the analysis of relationships between variables. In particular, studies of public reactions to aircraft noise generally have as a principal goal, the estimation of regression parameters for a model predicting annoyance as a function of various measures of noise exposure. For example, in studying the trade-off between noise levels and numbers of events, the following two-variable regression model is commonly employed:

$$(1.1) \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

where, y_i is a measure of annoyance associated with the ith individual in the population, x_{1i} and x_{2i} are measures of the noise level and number of events, respectively, to which the ith individual is exposed, and the e_i are independent random disturbance terms.

For simple random sample designs, the model given in (1.1) is quite adequate. Discussion of design and analysis issues for such models appears in many standard texts on multiple regression. See for example, Draper and Smith (1981) or Neter, Wasserman and Kutner (1985).

More commonly, samples for social surveys are drawn using complex sampling schemes, usually stratified, multi-stage cluster sample designs. For example, studies of residents' responses to noise most often consist of interviews with samples of individuals drawn from a number of different compact study areas, usually neighborhoods. In order to design such studies, it is necessary to determine the numbers of individuals and numbers of study areas to include in order to achieve specified research objectives. The statistical techniques developed in this report provide a basis for these sample design decisions.

Optimal design and estimation for means and totals using these designs is well understood. (Cochran, 1963) On the other hand, no such consensus exists for design and estimation for regression parameters using such samples. One approach to this problem is described by Kalton (1983) in an earlier NASA Contractor Report. His methodology is briefly described in Section 2, below. Kalton employed regression models which incorporate nested random intercepts associated with various stages of a multi-stage cluster sample design. For cases where there is little or no variability in

predictor variables within clusters, this approach provides useful results. However, when such variability does exist, it can lead to results which seem counter-intuitive.

In Section 3, we build on the models proposed by Kalton. Three regression models are presented for which the regression parameters themselves are considered random, with components of variability corresponding to the stages of a multi-stage cluster sample design. These models differ in the assumptions regarding variability of model parameters. For each, sampling variances and covariances are derived for estimates of linear regression parameters and ratios of parameters. Optimal allocation of sample resources across the stages of the design is derived for each situation. These allocations depend, in part, on estimates of variability at the various levels of the sample designs. Variance component estimates for this purpose are derived. These estimates could be obtained from existing data, contributing to the efficient design of future surveys. In Section 4, we apply some of these techniques in a simple example. Finally, the results of this research are summarized in Section 5.

2. Models with Fixed Regression Slopes.

The task of designing complex sample surveys for estimating regression parameters has been addressed elsewhere. Specifically, Kalton (1983) considered essentially the same problem in an earlier NASA Contractor Report. The simplest sampling situation considered by Kalton is that of a two-stage design. In the first stage, primary sampling units (PSU's) are selected. In the second, individuals are sampled within the selected PSU's. For example, for a survey around a single airport, PSU's might correspond to Census Tracts, and individuals within these PSU's might correspond to households within Census Tracts. The first multiple regression model for such a design considered by Kalton is a classical nested random effects model:

$$(2.1) \quad y_{ij} = B_{0i} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + e_{ij}$$

where, B_{0i} and e_{ij} are random effects corresponding to PSU's and individuals within PSU's, repectively. (See equation (11) in Kalton.)

Note that the slope parameters in equation (2.1), β_1 and β_2, are constant. They do not vary from cluster to cluster. While this assumption is standard for such random effects analysis of covariance models, the design implications are somewhat counter-intuitive. As was noted by Kalton, assuming a standard cost model, if the within-cluster variability of

the x-variables is the same as that of the associated population variability, a sample of a single cluster can be the implied optimal design for estimating slope parameters or functions of slope parameters. This seems counter-intuitive because one suspects that the structure of the relationship (the slope parameters) is probably not constant across clusters. Under such circumstances, drawing a sample of a single cluster would not be considered a safe alternative. In the following sections, we consider models which incorporate this structural variability.

3. Random Coefficients Regression Models.

The principal difference between the models considered in this section and those employed by Kalton is that here, the slopes are allowed to vary from cluster to cluster. They are assumed to be random, with components of variability associated with the various stages in a multi-stage cluster sample design. Below, three model-design combinations are presented. In the first two of these, estimates of regression slopes and of variance components are functions of individual slope estimates associated with observations belonging to the same penultimate stage sampling units.

Model I is for a two-stage design as described in Section 2 above. For such designs, penultimate stage sampling units are PSU's. Model III incorporates structural varibility at two levels for a three stage design. Such a sampling scheme might be used for a national or regional survey with cities or counties serving as PSU's, Census Tracts selected within PSU's as secondary sampling units (SSU's), and finally households within SSU's.

For some samples, an estimation strategy based on regression parameter estimates obtained within penultimate stage sampling units is feasible, however for many others, it is not. For example, in order to obtain slope estimates within a cluster, there must be variability in the predictor variables in that cluster. In the present case, that means variability in the noise exposure measurements within penultimate clusters. For many available studies, there is little or no variability in these predictors at the penultimate cluster level and the procedures described for Models I and II are not feasible.

Model III is also based on a three-stage design. It contains random effects which incorporate variability in regression slope parameters only at the first stage. Estimates depend on individual slope estimates associated with data from PSU's. Such a procedure could be applied to data from most available studies. It should be noted however that should substantial variability in regression slopes exist at the SSU level, conclusions obtained from analyses based on Model III should be limited to designs with similar

final stage characteristics. That is, in designing future studies, only designs which have PSU's of size comparable to those associated with PSU's in the studies used to obtain estimates of variance components should be considered. Any variability in slope parameters associated with SSU's in these studies will be included in the estimate of the variance component associated with PSU's. However, the effect on the estimated component will change depending on the number of SSU's per PSU.

3.1 Model I: Two-stage design, parameters random at the first stage.

3.1.1 The model and the design.

First, we consider a two-stage sample design, together with a model which incorporates variability in regression parameters at the first stage:

Design: Select n PSU's, and within the ith PSU, select n_i individuals for a total sample size of $\sum_i n_i = n$.

(3.1) $y_{ij} = B_{0i} + B_{1i}x_{1ij} + B_{2i}x_{2ij} + e_{ij}$, where

$B_{mi} = \beta_m + a_{mi}$; m=0,1,2, and where

$E(a_{mi}) = 0$; m=0,1,2

$E(a_{mi}a_{m'i}) = \sigma_{amm'}$; m,m'=0,1,2

$E(e_{ij}) = 0$; j=1,2,...,n_i ; i=1,2,...,n

$E(e_{ij}^2) = \sigma_e^2$; j=1,2,...,n_i ; i=1,2,...,n

$E(e_{ij}e_{i'j'}) = 0$; i≠i' or j≠j'

$E(e_{ij}a_{mj'}) = 0$; m=0,1,2 ; j,j'=1,...,n_i ; i=1,...,n

Here, for simplicity, the independent variables x_{1ij} and x_{2ij} are corrected for within PSU means. That is,

(3.2) $x_{mij} = x'_{mij} - \bar{x}'_{mi\cdot}$; m=1,2 ; j=1,...,n_i ; i=1,...,n

where x'_{mij} is the raw, uncorrected measurement, and where $\bar{x}'_{mi\cdot} = \sum_j x'_{mij}/n_i$, is the mean of the uncorrected measurements for the ith PSU.

3.1.2 Estimating regression parameters, associated sampling variances, and optimal design.

Let b_{mi} be the usual least squares estimate for B_{mi}, m=0,1,2 , calculated within PSU's. That is,

$$(3.3) \qquad b_{0i} = \sum_j y_{ij}/n_i = \bar{y}_{i\cdot} ,$$

$$b_{1i} = \frac{(\sum_j x_{2ij}^2)(\sum_j x_{1ij}y_{ij}) - (\sum_j x_{1ij}x_{2ij})(\sum_j x_{2ij}y_{ij})}{(\sum_j x_{1ij}^2)(\sum_j x_{2ij}^2) - (\sum_j x_{1ij}x_{2ij})^2}, \text{ and}$$

$$b_{2i} = \frac{(\sum_j x_{1ij}^2)(\sum_j x_{2ij}y_{ij}) - (\sum_j x_{1ij}x_{2ij})(\sum_j x_{1ij}y_{ij})}{(\sum_j x_{1ij}^2)(\sum_j x_{2ij}^2) - (\sum_j x_{1ij}x_{2ij})^2}.$$

Estimates of the β_m , m=0,1,2 are obtained by averaging the PSU estimates, i.e.,

$$(3.4) \qquad \hat{\beta}_m = \sum_i b_{mi}/n \; ; \; m=0,1,2.$$

Then, since the within PSU least squares estimates, the b_{mi}, are unbiased for the associated B_{mi} , the overall estimates, $\hat{\beta}_m$ are unbiased for the parameters β_m . That is,

$$(3.5) \qquad E(\hat{\beta}_m) = E_i [E_{j|i} (\sum_i b_{mi}/n)]$$

$$= E_i [\sum_i B_{mi}/n]$$

$$= \sum_i \beta_m/n = \beta_m .$$

Here, the notation, $E_{j|i}$ represents the conditional expectation taken over all samples of individuals (the j's) for a fixed set of PSU's (the i's).

The sampling variances of these estimates are determined as follows,

(3.6) $$\mathrm{Var}(\hat{\beta}_0) = E_i\,[\mathrm{Var}_{j|i}\,(\hat{\beta}_0)] + \mathrm{Var}_i\,[E_{j|i}\,(\hat{\beta}_0)]$$

$$= E_i\,(\sigma_e^2/n_i) + \mathrm{Var}_i\,(\sum_i B_{0i}/n)$$

$$= [\sigma_e^2\,[\sum_i (1/n_i)]\,/\,n] + (\sigma_{a00}/n)\,.$$

The conditional variance notation, $\mathrm{Var}_{j|i}$ is defined in a manner analogous to that of the conditional expectation described above for equation (3.5). If there is a constant PSU size, $n_i = \sum_i n_i/n = \bar{n}_.$, then (3.6) simplifies as follows,

(3.7) $$\mathrm{Var}(\hat{\beta}_0) = (\sigma_e^2/n\bar{n}_.) + (\sigma_{a00}/n)$$

$$= (\sigma_e^2/n_.) + (\sigma_{a00}/n)\,,$$

where, $n_. = \sum_i n_i$ is the total sample size.

For $\hat{\beta}_1$, the sampling variance is developed as follows,

(3.8) $$\mathrm{Var}(\hat{\beta}_1) = E_i\,[\mathrm{Var}_{j|i}\,(\hat{\beta}_1)] + \mathrm{Var}_i\,[E_{j|i}\,(\hat{\beta}_1)]$$

$$= E_i\,[(\sigma_e^2/n^2)\sum_i \sigma_{i22}/n_i(\sigma_{i11}\sigma_{i22} - \sigma_{i12}^2)] + \mathrm{Var}_i(\sum_i B_{1i}/n)$$

$$= [(\sigma_e^2/n^2)\sum_i \Delta_{i22}/n_i\,] + \sigma_{a11}/n\,,\ \text{where}$$

$$\Delta_{i22} = \sigma_{i22}/(\sigma_{i11}\sigma_{i22} - \sigma_{i12}^2)\,,\ \text{and where}$$

σ_{i11} = variance of x_{1ij} within the ith PSU,

σ_{i22} = variance of x_{2ij} within the ith PSU, and

σ_{i12} = covariance of x_{1ij} and x_{2ij} within the ith PSU.

Again, if $n_i = \bar{n}_.$, this simplifies to

(3.9) $\mathrm{Var}(\hat{\beta}_1) = \sigma_e^2 \bar{\Delta}_{22}/\bar{n}. + \sigma_{a11}/n$, where

$\bar{\Delta}_{22} = \Sigma_i \Delta_{i22}/n$.

Similarly, for $\hat{\beta}_2$,

(3.10) $\mathrm{Var}(\hat{\beta}_2) = [(\sigma_e^2/n^2) \sum_i \Delta_{i11}/n_i] + \sigma_{a22}/n$, where

$\Delta_{i11} = \sigma_{i11}/(\sigma_{i11}\sigma_{i22} - \sigma_{i12}^2)$, and if $n_i = \bar{n}.$,

$\mathrm{Var}(\hat{\beta}_2) = \sigma_e^2 \bar{\Delta}_{11}/\bar{n}. + \sigma_{22}/n$, where

$\bar{\Delta}_{11} = \sum_i \Delta_{i11}/n$.

Under the simple cost model assumed by Kalton, $C = C_0 + nC_a + \bar{n}.C_b$, where C_0 is the fixed cost of the survey, C_a is the average cost of including a cluster in the sample, and C_b is the average cost of including an individual in the sample, the optimum cluster size for estimating β_1 is given by,

(3.11) $\bar{n}.(\mathrm{opt}) = [\sigma_e^2 \bar{\Delta}_{22}/\sigma_{a11}]^{1/2} [C_a/C_b]^{1/2}$

3.1.3 Estimating ratios of regression parameters, associated sampling variances and optimal design.

Finally, for designing a sample to estimate ratios of regression coefficients, one requires the sampling covariances of the estimates, $\hat{\beta}_1$ and $\hat{\beta}_2$,

(3.12) $\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2) = [(-\sigma_e^2/n^2) \sum_i \Delta_{i12}/n_i] + \sigma_{a12}/n$, where

$\Delta_{i12} = \sigma_{i12}/(\sigma_{i11}\sigma_{i22} - \sigma_{i12}^2)$, and if $n_i = \bar{n}.$,

$\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2) = -\sigma_e^2 \bar{\Delta}_{12}/\bar{n}. + \sigma_{a12}/n$, where

$\bar{\Delta}_{12} = \Sigma_i \Delta_{i12}/n$.

For estimating the ratio of the two regression coefficients, $R = (\beta_2/\beta_1)$, we propose to use $\hat{R} = (\hat{\beta}_2/\hat{\beta}_1)$. Then, we can use the Taylor

expansion method to obtain an approximation of the variance of this ratio estimate,

$$(3.13) \quad \mathrm{Var}(\hat{R}) \approx \beta_1^{-2} [\mathrm{Var}(\hat{\beta}_2) + R^2 \mathrm{Var}(\hat{\beta}_1) - 2R \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)] .$$

Again, if there is a constant cluster size of $n_i = \bar{n}.$, this simplifies to

$$\begin{aligned} (3.14) \quad \mathrm{Var}(\hat{R}) &\approx \beta_1^{-2} [(\sigma_e^2 \bar{\Delta}_{11}/n. + \sigma_{a22}/n) \\ &+ R^2 (\sigma_e^2 \bar{\Delta}_{22}/n + \sigma_{a11}/n) \\ &- 2R (-\sigma_e^2 \bar{\Delta}_{12}/n. + \sigma_{a12}/n)] \\ &= \beta_1^{-2} [[\sigma_e^2 (\bar{\Delta}_{11} + R^2 \bar{\Delta}_{22} + 2R \bar{\Delta}_{12}) / n.] \\ &+ [(\sigma_{a22} + R^2 \sigma_{a11} - 2R \sigma_{a12}) / n]] . \end{aligned}$$

Under the cost model described above, the optimum PSU size for estimating the ratio is given by,

$$\begin{aligned} (3.15) \quad \bar{n}.(\mathrm{opt}) &= \{\sigma_e^2 (\bar{\Delta}_{11} + R^2 \bar{\Delta}_{22} + 2R \bar{\Delta}_{12}) \\ &/ (\sigma_{a22} + R^2 \sigma_{a11} - 2R \sigma_{a12})\}^{1/2} \{C_a/C_b\}^{1/2} . \end{aligned}$$

3.1.4 Estimating variance components.

Thus, in order to determine the optimum PSU size for estimating regression coefficients and ratios of regression coefficients, one requires the following information, the average cost parameters, C_a and C_b, the design characteristics, the variance and covariance components of the random parameters, and an approximation for the true ratio R. The design characteristics describe the within PSU distribution of the x variables in terms of $\bar{\Delta}_{11}$, $\bar{\Delta}_{22}$, and $\bar{\Delta}_{12}$. These in turn depend on σ_{i11}, σ_{i22}, and σ_{i12} calculated within clusters. The variance components of the random parameters, σ_{a11}, σ_{a22}, σ_{a12}, and σ_e^2 can be estimated from previous surveys using the methods described below.

The residual variance, σ_e^2 , is estimated in the usual manner as,

$$(3.16)\qquad \hat{\sigma}_e^2 = \frac{\sum_i \sum_j (y_{ij} - \hat{y}_{ij})^2}{\sum_i (n_i - 3)} = \frac{\sum_i \sum_j (y_{ij} - \hat{y}_{ij})^2}{n_\cdot - 3n}$$

The remaining components are estimated as follows:

$$(3.17)\qquad \hat{\sigma}_{a11} = \sum_i (b_{1i} - \hat{\beta}_1)^2/(n-1) - \hat{\sigma}_e^2 \left(\sum_i \Delta_{i22}/n_i\right)/n$$

or for constant cluster size, $n_i = \bar{n}_\cdot$,

$$\hat{\sigma}_{a11} = \sum_i (b_{1i} - \hat{\beta}_1)^2/(n-1) - \hat{\sigma}_e^2 \, \bar{\Delta}_{22}/\bar{n}_\cdot .$$

$$(3.18)\qquad \hat{\sigma}_{a22} = \sum_i (b_{2i} - \hat{\beta}_2)^2/(n-1) - \hat{\sigma}_e^2 \left(\sum_i \Delta_{i11}/n_i\right)/n$$

or for constant cluster size, $n_i = \bar{n}_\cdot$,

$$\hat{\sigma}_{a22} = \sum_i (b_{2i} - \hat{\beta}_2)^2/(n-1) - \hat{\sigma}_e^2 \, \bar{\Delta}_{11}/\bar{n}_\cdot .$$

$$(3.19)\qquad \hat{\sigma}_{a12} = \sum_i (b_{1i} - \hat{\beta}_1)(b_{2i} - \hat{\beta}_2)/(n-1) + \hat{\sigma}_e^2 \left(\sum_i \Delta_{i12}/n_i\right)/n$$

or for constant cluster size, $n_i = \bar{n}_\cdot$,

$$\hat{\sigma}_{a12} = \sum_i (b_{1i} - \hat{\beta}_1)(b_{2i} - \hat{\beta}_2)/(n-1) - \hat{\sigma}_e^2 \, \bar{\Delta}_{12}/\bar{n}_\cdot .$$

3.2 Model II: Three-Stage Design, Parameters Random at the First Two Stages.

3.2.1 The model and the design.

Now, we consider a three-stage sample design, together with a model which incorporates variability in regression parameters at the first two stages:

Design: Select n PSU's, within the ith PSU, select n_i SSU's and within the ijth SSU, select n_{ij} individuals.

(3.20) $y_{ijk} = B_{0ij} + B_{1ij}x_{1ijk} + B_{2ij}x_{2ijk} + e_{ijk}$, where

$B_{mij} = \beta_m + a_{mi} + c_{mij}$; m=0,1,2 ; j=1,...,n_i ; i=1,...,n .

$E(e_{ijk}) = E(a_{mi}) = E(c_{mij}) = 0$; m=0,1,2 ;
k=1,...,n_{ij} ; j=1,...,n_i ; i=1,...,n .

$E(a_{mi}a_{m'i}) = \sigma_{amm'}$; m,m'=0,1,2 ; i=1,...,n .

$E(c_{mij}c_{m'ij}) = \sigma_{cmm'}$; m,m'=0,1,2 ; j=1,...,n_i ; i=1,...,n ,

$E(a_{mi}a_{m'i'}) = 0$; m,m'=0,1,2 ; i≠i' .

$E(c_{mij}c_{m'i'j'}) = 0$; m,m'=0,1,2 ; i≠i or j≠j' .

$E(a_{mi}c_{m'i'j}) = 0$; m,m'=0,1,2 ;

$E(a_{mi}e_{i'jk}) = E(c_{mij}e_{i'j'k}) = 0$; m=0,1,2 .

The last four lines of (3.20) imply that the e_{ijk} , a_{mi} , and c_{mij} terms are independent of each other.

3.2.2 Estimation, sampling variances, and optimal design.

Here again, regressions are carried out within the penultimate stage sampling units, that is within SSU's. Let b_{mij} be the usual, within SSU least squares estimate of B_{mij}, and let

(3.21) $\bar{b}_{mi\cdot} = \sum_j b_{mij}/n_{ij}$,

$\hat{\beta}_m = \sum_i \bar{b}_{mi\cdot}/n$, and finally

$\hat{R} = \hat{\beta}_2/\hat{\beta}_1$.

Using an argument analagous to that used in section 3.1, it is easy to see that the regression parameter estimates, $\hat{\beta}_m$, are unbiased for the parameters β_m . Sampling variances and covariances are also derived in a similar manner,

(3.22) $\mathrm{Var}(b_{1ij}) = \mathrm{Var}_{ij}\, E_{klij}(b_{1ij}) + E_{ij}\, \mathrm{Var}_{klij}(b_{1ij})$

$= \mathrm{Var}_{ij}(a_{1i} + c_{1ij}) + E_{ij}\,[\sigma_e^2\, \sigma_{ij22} / n_{ij}(\sigma_{ij11}\sigma_{ij22} - \sigma_{ij12}^2)]$

$= \sigma_{a11} + \sigma_{c11} + \sigma_e^2\, \Delta_{ij22} / n_{ij}$. Further,

(3.23) $\mathrm{Cov}(b_{1ij}, b_{1ij'}) = \mathrm{Cov}_{ij,ij'}\, E_{klij,ij'}(b_{1ij}, b_{1ij'})$

$+ E_{ij,ij'}\, \mathrm{Cov}_{klij'ij'}(b_{1ij}, b_{1ij'})$

$= \mathrm{Cov}_{ij,ij'}(a_{1i} + c_{1ij}\,,\, a_{1i} + c_{1ij'}) + 0 = \sigma_{a11}$. Therefore,

(3.24) $\mathrm{Var}(\hat{\beta}_1) = n^{-2} \sum_i n_i^{-2} \sum_j (\sigma_{a11} + \sigma_{c11} + \sigma_e^2\, \Delta_{ij22} / n_{ij})$

$+ n^{-2} \sum_i n_i^{-2} \sum_j \sum_{j \neq j'} \sigma_{a11}$.

$= (\sigma_{a11} + \sigma_{c11})\, n^{-2} (\sum_i n_i^{-1}) + (\sigma_e^2/n^2)\, [\sum_i n_i^{-2} (\sum_j \Delta_{ij22}/n_{ij})]$

$+ \sigma_{a11}\, n^{-2} \sum_i (n_i - 1)/n_i$.

$= (\sigma_{a11}/n) + (\sigma_{c11}/n^2) \sum_i n_i^{-1} + (\sigma_e^2/n^2)\, [\sum_i n_i^{-2} (\sum_j \Delta_{ij22}/n_{ij})]$.

In equations (3.22-24), $\sigma_{ijmm'}$ and $\Delta_{ijmm'}$ are defined to be the variances, covariances and functions of these, for predictor variables x_{1ij} and x_{2ij}, calculated within the ijth SSU, analogous to the definitions of $\sigma_{imm'}$ and $\Delta_{imm'}$ given in equation (3.8).

Now, if we have a constant PSU and SSU size, $n_i = \bar{n}.$ and $n_{ij} = \bar{n}_i. = \bar{n}..$, and if the design characteristics Δ_{ij22} are constant over SSU's at $\overline{\Delta}_{22}$, then we have,

(3.24) $\mathrm{Var}(\hat{\beta}_1) = (\sigma_{a11}/n) + (\sigma_{c11}/n\bar{n}.) + (\sigma_e^2 \overline{\Delta}_{22}/n\bar{n}.\bar{n}..)$

$= (\sigma_{a11}/n) + (\sigma_{c11}/n.) + (\sigma_e^2 \overline{\Delta}_{22}/n..)$.

Under similar assumptions, the $\mathrm{Var}(\hat{\beta}_2)$ and $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ are seen to be,

(3.25) $\mathrm{Var}(\hat{\beta}_2) = (\sigma_{a22}/n) + (\sigma_{c22}/n_\cdot) + (\sigma_e^2\overline{\Delta}_{11}/n_{\cdot\cdot})$, and

(3.26) $\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2) = (\sigma_{a12}/n) + (\sigma_{c12}/n_\cdot) - (\sigma_e^2\overline{\Delta}_{12}/n_{\cdot\cdot})$.

Therefore, for the estimate of the ratio between the two coefficients, $\hat{R} = \hat{\beta}_2/\hat{\beta}_1$,

$$(3.27)\quad \mathrm{Var}(\hat{R}) \approx \beta_1^{-2}\,[\mathrm{Var}(\hat{\beta}_2) + R^2\,\mathrm{Var}(\hat{\beta}_1) - 2R\,\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)]$$

$$= \beta_1^{-2}\,[[(\sigma_{a22}/n) + (\sigma_{c22}/n_\cdot) + \sigma_e^2\overline{\Delta}_{11}/n_{\cdot\cdot})]$$

$$+ R^2\,[(\sigma_{a11}/n) + (\sigma_{c11}/n_\cdot) + (\sigma_e^2\overline{\Delta}_{22}/n_{\cdot\cdot})]$$

$$- 2R\,[(\sigma_{a12}/n) + (\sigma_{c12}/n_\cdot) - (\sigma_e^2\overline{\Delta}_{12}/n_{\cdot\cdot})]]$$

$$= \beta_1^{-2}\,[(\sigma_{a22} + R^2\,\sigma_{a11} - 2R\,\sigma_{a12})/n$$

$$+ (\sigma_{c22} + R^2\,\sigma_{c11} - 2R\,\sigma_{c12})/n_\cdot$$

$$+ \sigma_e^2\,(\overline{\Delta}_{11} + R^2\,\overline{\Delta}_{22} + 2R\,\overline{\Delta}_{12})/n_{\cdot\cdot}\,]\,.$$

Using these results, combined with a simple cost model for three-stage cluster sampling we can arrive at optimum sampling unit sizes in a manner similar to that used for Model I. Let,

(3.28) $C = C_0 + C_1 n + C_2 n_\cdot + C_3 n_{\cdot\cdot}$,

where, C_0 is the constant overhead cost of the survey, C_1 is the average cost of including a PSU in the sample, C_2 is the average cost of including an SSU in the sample, and C_3 is the average cost of including an individual in the sample. Then, using the Cauchy-Schwartz inequality, we have the following condition for optimum allocation for estimating the ratio R:

$$(3.29)\quad (\sigma_{a22} + R^2\,\sigma_{a11} - 2R\,\sigma_{a12})^{1/2} / (n\,C_1^{1/2})$$

$$= (\sigma_{c22} + R^2\,\sigma_{c11} - 2R\,\sigma_{c12})^{1/2} / (n_\cdot\,C_2^{1/2})$$

$$= \sigma_e(\overline{\Delta}_{11} + R^2\,\overline{\Delta}_{22} + 2R\,\overline{\Delta}_{12})^{1/2} / (n_{\cdot\cdot}\,C_3^{1/2})$$

This relation translates into the following optimum sampling unit sizes:

$$(3.30)\quad \bar{n}.(opt) = [(\sigma_{c22} + R^2\,\sigma_{c11} - 2R\,\sigma_{c12})C_1 \;/\; (\sigma_{a22} + R^2\,\sigma_{a11} - 2R\,\sigma_{a12})C_2]^{1/2}$$

$$(3.31)\quad \bar{n}..(opt) = \sigma_e\,[(\bar{\Delta}_{11} + R^2\,\bar{\Delta}_{22} + 2R\,\bar{\Delta}_{12})C_2 \;/\; (\sigma_{c22} + R^2\,\sigma_{c11} - 2R\,\sigma_{c12})C_3]^{1/2}\,.$$

3.2.3 Estimating variance components.

In order to determine necessary sample sizes and optimum allocation, one needs some idea of cost parameters, design characteristics summarized by $\bar{\Delta}_{11}$, $\bar{\Delta}_{22}$, and $\bar{\Delta}_{12}$, an approximation of the ratio R and its denominator β_1, and the variance components σ_{a11}, σ_{a22}, σ_{a12}, σ_{c11}, σ_{c22}, σ_{c12}, and σ_e^2. These may be obtained from previous survey data using the following estimates:

$$(3.32)\quad \hat{\sigma}_e^2 = \sum_i\sum_j (y_{ijk}-\hat{y}_{ijk})^2/(n..-3n.)$$

$$(3.33)\quad \hat{\sigma}_{c11} = [\sum_i\sum_j (b_{1ij}-\bar{b}_{1i\cdot})^2 - \hat{\sigma}_e^2 \sum_i[(n_i-1)/n_i] \sum_j \Delta_{ij22}/n_{ij}]/(n.-n),$$

which if $n_i=\bar{n}.$ and $n_{ij}=\bar{n}..$, further simplifies to,

$$\sum_i\sum_j (b_{1ij}-\bar{b}_{1i\cdot})^2/(n.-n) - \hat{\sigma}_e^2\,\bar{\Delta}_{22}/\bar{n}..$$

$$(3.34)\quad \hat{\sigma}_{c22} = [\sum_i\sum_j (b_{2ij}-\bar{b}_{2i\cdot})^2 - \hat{\sigma}_e^2 \sum_i[(n_i-1)/n_i] \sum_j \Delta_{ij11}/n_{ij}]/(n.-n),$$

which if $n_i=\bar{n}.$ and $n_{ij}=\bar{n}..$, further simplifies to,

$$\sum_i\sum_j (b_{2ij}-\bar{b}_{2i\cdot})^2/(n.-n) - \hat{\sigma}_e^2\,\bar{\Delta}_{11}/\bar{n}..$$

$$(3.35)\quad \hat{\sigma}_{c12} = [\sum_i\sum_j (b_{1ij}-\bar{b}_{1i\cdot})(b_{2ij}-\bar{b}_{2i\cdot}) + \hat{\sigma}_e^2 \sum_i[(n_i-1)/n_i] \sum_j \Delta_{ij12}/n_{ij}]/(n.-n),$$

which if $n_i=\bar{n}.$ and $n_{ij}=\bar{n}..$, further simplifies to,

$$\sum_i\sum_j (b_{1ij}-\bar{b}_{1i\cdot})(b_{2ij}-\bar{b}_{2i\cdot})/(n.-n) + \hat{\sigma}_e^2\,\bar{\Delta}_{12}/\bar{n}..$$

$$(3.36)\quad \hat{\sigma}_{a11} = [\sum_i(\bar{b}_{1i}-\hat{\beta}_1)^2 - \hat{\sigma}_{c11}[(n-1)/n]\sum_i n_i^{-1} - \hat{\sigma}_e^2[(n-1)/n]\sum_i\sum_j \Delta_{ij22}/n_i.n_{ij}] / (n-1),$$

which if $n_i=\bar{n}.$ and $n_{ij}=\bar{n}..$, further simplifies to,

$$\sum_i(\bar{b}_{1i}-\hat{\beta}_1)^2/(n-1) - \hat{\sigma}_{c11}/\bar{n}. - \hat{\sigma}_e^2\,\bar{\Delta}_{22}/\bar{n}..\ ,$$

$$(3.37)\quad \hat{\sigma}_{a22} = [\sum_i(\bar{b}_{2i}-\hat{\beta}_2)^2 - \hat{\sigma}_{c22}[(n-1)/n]\sum_i n_i^{-1} - \hat{\sigma}_e^2[(n-1)/n]\sum_i\sum_j \Delta_{ij11}/n_i.n_{ij}] / (n-1),$$

which if $n_i=\bar{n}.$ and $n_{ij}=\bar{n}..$, further simplifies to,

$$\sum_i(\bar{b}_{2i}-\hat{\beta}_2)^2/(n-1) - \hat{\sigma}_{c22}/\bar{n}. - \hat{\sigma}_e^2\,\bar{\Delta}_{11}/\bar{n}..\ ,\ \text{and finally,}$$

$$(3.38)\quad \hat{\sigma}_{a12} = [\sum_i(\bar{b}_{1i}-\hat{\beta}_1)(\bar{b}_{2i}-\hat{\beta}_2) - \hat{\sigma}_{c12}[(n-1)/n]\sum_i n_i^{-1} + \hat{\sigma}_e^2[(n-1)/n]\sum_i\sum_j \Delta_{ij12}/n_i.n_{ij}] / (n-1),$$

which if $n_i=\bar{n}.$ and $n_{ij}=\bar{n}..$, further simplifies to,

$$\sum_i(\bar{b}_{1i}-\hat{\beta}_1)(\bar{b}_{2i}-\hat{\beta}_2)/(n-1) - \hat{\sigma}_{c12}/\bar{n}. + \hat{\sigma}_e^2\,\bar{\Delta}_{11}/\bar{n}..\ .$$

3.3 Model III: Three-stage design, slope parameters random at the first two stages, error terms nested, individuals within SSU's within PSU's.

3.3.1 The model and the design.

As stated earlier, it is not always possible to estimate regression parameters at the SSU level as required with Model II. For this reason, we introduce a model-design combination based on the same sampling scheme as that for Model II, but with a model which allows for variability of slope parameters among first stage units only.

Design: Select n PSU's, within the ith PSU, select n_i SSU's and within the ijth SSU, select n_{ij} individuals.

(3.39) $y_{ijk} = B_{0ij} + B_{1i}x_{1ijk} + B_{2i}x_{2ijk} + e_{ijk}$, where

$B_{0ij} = \beta_0 + a_{0i} + c_{0ij}$; $j=1,...,n_i$; $i=1,...,n$.

$B_{mi} = \beta_m + a_{mi}$; $m=1,2$; $i=1,...,n$

$E(e_{ijk}) = E(a_{mi}) = E(c_{mij}) = 0$; $m=0,1,2$;
$k=1,...,n_{ij}$; $j=1,...,n_i$; $i=1,...,n$.

$E(c_{0ij}^2) = \sigma_{c00}$,

$E(a_{mi}a_{m'i}) = \sigma_{amm'}$; $m,m'=0,1,2$; $i=1,...,n$.

$E(a_{mi}a_{m'i'}) = 0$; $m,m'=0,1,2$; $i \neq i'$.

$E(c_{0ij}c_{0i'j'}) = 0$; $i \neq i$ or $j \neq j'$.

$E(a_{mi}c_{0i'j}) = 0$; $m=0,1,2$;

$E(a_{mi}e_{i'jk}) = E(c_{mij}e_{i'j'k}) = 0$; $m=0,1,2$.

3.3.2 The estimates, associated sampling variances, and optimal design.

Consider the following estimates for the slope parameter β_{1i} associated with the ith PSU:

(3.40)
$$b_{1i} = \frac{(\sum_j\sum_k x_{2ijk}^2)(\sum_j\sum_k x_{1ijk}y_{ijk}) - (\sum_j\sum_k x_{1ijk}x_{2ijk})(\sum_j\sum_k x_{2ijk}y_{ijk})}{(\sum_j\sum_k x_{1ijk}^2)(\sum_j\sum_k x_{2ijk}^2) - (\sum_j\sum_k x_{1ijk}x_{2ijk})^2}$$

$$= \frac{(\sum_j\sum_k x_{2ijk}^2)(\sum_j\sum_k x_{1ijk}y_{ijk}) - (\sum_j\sum_k x_{1ijk}x_{2ijk})(\sum_j\sum_k x_{2ijk}y_{ijk})}{d_i}$$

where, $d_i = (\sum_j\sum_k x_{1ijk}^2)(\sum_j\sum_k x_{2ijk}^2) - (\sum_j\sum_k x_{1ijk}x_{2ijk})^2$

$= n_i.^2(\sigma_{i11}\sigma_{i22} - \sigma_{i12}^2)$.

Here, the x_{mijk} are mean corrected within PSU's as for Model I. Then for the estimate of the overall mean slope, β_1, consider

$$(3.41) \qquad \hat{\beta}_1 = \sum_i b_{1i}/n \ .$$

It can be shown, using the following logic, that $\hat{\beta}_1$, as defined in equations (3.40) and (3.41) above, is unbiased for the mean slope β_1. First, notice that the observations y_{ijk} can be expressed as follows:

$$(3.42) \qquad y_{ijk} = \beta_0 + a_{0i} + c_{0ij} + (\beta_1 + a_{1i})x_{1ijk} + (\beta_2 + a_{2i})x_{2ijk} + e_{ijk} \ .$$

Now, the conditional expectation of b_{1i} given that the ith PSU is in the sample is given by,

$$(3.43) \qquad E_{jkli}(b_{1i}) = E_{jkli}\, d_i^{-1} \sum_j \sum_k [(\sum_j \sum_k x_{2ijk}^2)x_{1ijk} - (\sum_j \sum_k x_{1ijk}x_{2ijk})x_{2ijk}]y_{ijk}$$

$$= d_i^{-1}\, E_{jkli}\, [(\sum_j \sum_k x_{2ijk}^2)\,[\sum_j \sum_k x_{1ijk}c_{0ij} + (\sum_j \sum_k x_{1ijk}^2)(\beta_1 + a_{1i})$$

$$+ (\sum_j \sum_k x_{1ijk}x_{2ijk})(\beta_2 + a_{2i})]$$

$$- (\sum_j \sum_k x_{1ijk}x_{2ijk})[\sum_j \sum_k x_{2ijk}c_{0ij} + \sum_j \sum_k x_{1ijk}x_{2ijk}(\beta_1 + a_{1i})$$

$$+ \sum_j \sum_k x_{2ijk}^2(\beta_2 + a_{2i})]] \ .$$

$$= d_i^{-1}\ E_{jkli}\, [\sum_j [\sum_k\ (\sum_i \sum_k x_{2ijk}^2)x_{1ijk} - (\sum_j \sum_k x_{1ijk}x_{2ijk})x_{2ijk}]c_{0ij}$$

$$+ [(\sum_j \sum_k x_{1ijk}^2)(\sum_j \sum_k x_{2ijk}^2) - \sum_j \sum_k x_{1ijk}x_{2ijk})^2]\,(b_{1i} + a_{1i})$$

$$= \beta_1 + a_{1i} \ .$$

Thus, $E(\hat{\beta}_1) = E_i\, E_{jkli}\, (\hat{\beta}_1) = \beta_1$. That is, $\hat{\beta}_1$ is unbiased for β_1. Similarly, it can be shown that, $\hat{\beta}_2$, defined in an analogous fashion is unbiased for β_2.

Using a similar treatment, sampling variances and covariances can be derived for these estimates. For $\hat{\beta}_1$ we have,

$$(3.44) \qquad Var(\hat{\beta}_1) = E_i\, Var_{jkli}\, (\hat{\beta}_1) + Var_i\, E_{jkli}\, (\hat{\beta}_1) \ . \quad \text{Now,}$$

(3.45) $\mathrm{Var}_i\, E_{jkli}\,(\hat{\beta}_1) = \mathrm{Var}_i\{\sum_i a_{1i}/n\} = \sigma_{a11}/n$, and

(3.46) $\mathrm{Var}_{jkli}\,(\hat{\beta}_1) = \sum_i \mathrm{Var}_{jkli}\,(b_{1i}) / n^2$. Now,

(3.47) $\mathrm{Var}_{jkli}\,(b_{1i}) = d_i^{-2}\,\mathrm{Var}_{jkli} \sum_j\sum_k A_{1ijk}y_{ijk}$, where

(3.48) $A_{1ijk} = (\sum_j\sum_k x_{2ijk}^2)x_{1ijk} - (\sum_j\sum_k x_{1ijk}x_{2ijk})x_{2ijk}$

$= n_i.\sigma_{i22}x_{1ijk} - n_i.\sigma_{i12}x_{2ijk}$. So,

(3.49) $\mathrm{Var}_{jkli}\,(b_{1i}) = d_i^{-2}\,\{\sum_j\sum_k A_{1ijk}^2\,\sigma_e^2 + \sum_j(\sum_k A_{1ijk})^2\,\sigma_{c00}\}$. Now,

(3.50) $\sum_j\sum_k A_{1ijk}^2 = n_i.^2\,[\sigma_{i22}\sum_j\sum_k x_{1ijk}^2 + \sigma_{i12}\sum_j\sum_k x_{2ijk}^2$

$- 2\sigma_{i22}\sigma_{i12}\sum_j\sum_k x_{1ijk}x_{2ijk}\,]$.

$= n_i.^3\,(\sigma_{i22}\sigma_{i11} + \sigma_{i12}^2\sigma_{i22} - 2\sigma_{i22}\sigma_{i12}^2)$

$= n_i.^3\,\sigma_{i22}\,(\sigma_{i11}\sigma_{i22} - \sigma_{i12}^2) = n_i.\,\sigma_{i22}\,d_i$, and

(3.51) $\sum_j(\sum_k A_{1ijk})^2 = n_i.^2\,\sum_j(\sum_k \sigma_{i22}x_{1ijk} - \sigma_{i12}x_{2ijk})^2$

$= n_i.^2\,\sum_j n_{ij}^2\,(\sigma_{i22}^2\bar{x}_{1ij}.^2 + \sigma_{i12}^2\bar{x}_{2ij}.^2$

$- 2\sigma_{i22}\sigma_{i12}\bar{x}_{1ij}.\bar{x}_{2ij}.)$.

Now, if we have a constant SSU size, $n_{ij}=\bar{n}_i.=n_i./n_i$, this simplifies to,

(3.52) $n_i.^2\,\bar{n}_i.\,(\sigma_{i22}\,\Omega_{i11}\,\sigma_{i11} + \sigma_{i12}^2\,\Omega_{i22}\,\sigma_{i22} - 2\,\sigma_{i22}\,\sigma_{i12}\,\Omega_{i12})$,

where, for example,

(3.53) $\Omega_{i11} = (\sum_j n_{ij}\bar{x}_{1ij}.^2)/(\sum_j\sum_k x_{ijk}^2)$

is the proportion of within PSU variance "explained" by SSU's. Now if $\Omega_{i11}=\Omega_{i22}=\Omega_{i12}=\Omega_i$, this further simplifies to,

$$(3.54)\qquad \sum_j(\sum_k A_{1ijk})^2 = n_{i\cdot}^2\,\bar{n}_{i\cdot}\,\Omega_i\,\sigma_{i22}\,(\sigma_{i22}\sigma_{i11}-\sigma_{i12}^2)$$

$$= (n_{i\cdot}/n_{\cdot})\,\Omega_i\,\sigma_{i22}\,d_i\,.$$

So, assuming that the SSU's are of constant size within the ith PSU, i.e. that $n_{ij}=\bar{n}_{i\cdot}=n_{i\cdot}/n_i$, and that the proportions of within ith PSU variances and covariance of the two predictor variables explained by SSU's are all equal to Ω_i, we have,

$$(3.55)\qquad Var_{jkli}\,b_{1i} = d_i^{-2}\{n_{i\cdot}\sigma_{i22}d_i\sigma_e^2 + \sigma_{c11}\bar{n}_{i\cdot}\Omega_i\sigma_{i22}d_i\}$$

$$= \{n_{i\cdot}\sigma_{i22}(\sigma_e^2 + \Omega_i\sigma_{c00}/n_i)\}/d_i\ .$$

Now, if we further assume constant number of individuals per PSU, and a constant number of SSU's per PSU, as well as constant within PSU design characteristics,

$$(3.56)\qquad n_{i\cdot} = \sum_i n_{i\cdot}/n = n_{\cdot\cdot}/n = \bar{n}_{\cdot\cdot}\ ,$$

$$n_i = \sum_i n_i/n = n_{\cdot}/n = \bar{n}_{\cdot}\ ,\ \text{and}$$

$\Omega_i = \Omega$ and $d_i = d$, we have,

$$(3.57)\qquad Var\,(\hat{\beta}_1) = n^{-2}\sum_i[n_{i\cdot}\,\sigma_{i22}\,(\sigma_e^2 + \Omega\sigma_{c00}/n_i)\,/\,d_i] + \sigma_{a11}/n$$

$$= n^{-2}\sum_i\,[n_{\cdot\cdot}/n]\,[\sigma_{22}\,(\sigma_e^2+\Omega\sigma_{c00}n/n_{\cdot})\,/\,d] + \sigma_{a11}/n$$

$$= (\Delta_{22}/n_{\cdot\cdot})\,(\sigma_e^2 + \Omega\sigma_{c00}/\bar{n}_{\cdot}) + \sigma_{a11}/n$$

$$= (\sigma_{a11}/n) + \Delta_{22}\,[(\Omega\sigma_{c00}/n_{\cdot}) + \sigma_e^2/n_{\cdot\cdot}]\,.$$

Similarly for $\hat{\beta}_2$ and the covariance, we have

$$(3.58)\qquad Var\,(\hat{\beta}_2) = (\sigma_{a22}/n) + \Delta_{11}\,[(\Omega\sigma_{c00}/n_{\cdot}) + \sigma_e^2/n_{\cdot\cdot}]\ ,\ \text{and}$$

$$(3.59)\qquad Cov\,(\hat{\beta}_1,\hat{\beta}_2) = (\sigma_{a12}/n) - \Delta_{12}\,[(\Omega\sigma_{c00}/n_{\cdot}) + \sigma_e^2/n_{\cdot\cdot}]\,.$$

Thus, if we estimate the ratio of the expected values of the two slope parameters by the ratio of the individual parameter estimates, $\hat{R} = \hat{\beta}_2/\hat{\beta}_1$, we have,

$$(3.60)\quad \mathrm{Var}(\hat{R}) = \beta_1^{-2}\{(\sigma_{a22}/n) + \Delta_{11}[(\Omega\sigma_{c00}/n.) + \sigma_e^2/n..] + R^2[(\sigma_{a11}/n) + \Delta_{22}[(\Omega\sigma_{c00}/n.) + \sigma_e^2/n..]] - 2R[(\sigma_{a12}/n) - \Delta_{12}[(\Omega\sigma_{c00}/n.) + \sigma_e^2/n..]]\}$$

$$= \beta_1^{-2}\{(\sigma_{a22}+R^2\sigma_{a11}-2R\sigma_{a12})/n + (\Delta_{11}+R^2\Delta_{22}+2R\Delta_{12})(\Omega\sigma_{c00}/n. + \sigma_e^2/n..)\}.$$

Using the cost model described in Section 3.2.2, equation (3.28), for Model II, the Cauchy-Schwartz inequality yields the following optimum unit sizes for estimating R:

$$(3.61)\quad \bar{n}.(\mathrm{opt}) = \{(\Delta_{11}+R^2\Delta_{22}+2R\Delta_{12})\Omega\sigma_{c00}C_1 / (\sigma_{a22}+R^2\sigma_{a11}-2R\sigma_{a12})C_2\}^{1/2}, \text{ and}$$

$$(3.62)\quad \bar{n}..(\mathrm{opt}) = (\sigma_e^2/\sigma_{c00}\Omega)^{1/2}(C_2/C_3)^{1/2}.$$

3.3.3 Estimating variance components.

Again, estimates of variances and covariances of the random parameters are required for determing sample allocations. These can be obtained as follows:

$$(3.63)\quad \hat{\sigma}_e^2 = \frac{\sum_j\sum_k\sum_k(y_{ijk}-\hat{y}_{ijk})^2}{\sum_i\sum_j(n_{ij}-3)},$$

$$(3.64)\quad \hat{\sigma}_{c00} = \frac{\sum_i\sum_j(\bar{y}_{ij}.-\bar{y}_{i}..)^2}{\sum_i(n_i-1)} - \frac{\hat{\sigma}_e^2}{n_{ij}},$$

$$(3.65)\qquad \hat{\sigma}_{a11} = \frac{\sum_i (b_{1i}-\hat{\beta}_1)^2}{n-1} - \Delta_{22}\hat{\sigma}_e^2 - \frac{\Omega\Delta_{22}\hat{\sigma}_{c00}}{\bar{n}.},$$

$$(3.66)\qquad \hat{\sigma}_{a22} = \frac{\sum_i (b_{2i}-\hat{\beta}_2)^2}{n-1} - \Delta_{11}\hat{\sigma}_e^2 - \frac{\Omega\Delta_{11}\hat{\sigma}_{c00}}{\bar{n}.}, \text{ and}$$

$$(3.67)\qquad \hat{\sigma}_{a12} = \frac{\sum_i (b_{1i}-\hat{\beta}_1)(b_{2i}-\hat{\beta}_2)}{n-1} + \Delta_{12}\hat{\sigma}_e^2 + \frac{\Omega\Delta_{12}\hat{\sigma}_{c00}}{\bar{n}.}.$$

4. An Example.

Fields and Walker (1982) describe a study of the effects of railway noise on residents based on a social survey of 1453 respondents in 75 areas in Great Britain. Among other things, estimates of 24 hour L_{eq} dB(A) and the difference between daytime 15 hour L_{eq} and nightime L_{eq} were obtained for each individual surveyed using physical noise measurements. Several measurements of annoyance were obtained using an interviewer administered questionnaire. Here, we make use of the four category verbal scale of annoyance obtained from Question 17 of that study: "Does the noise of the trains bother or annoy you?" Possible responses ranged from (1) "Not at all," to (4) "Very much."

For purposes of illustration, we regress this annoyance variable (y_{ij}) on the two independent variables, 24 hour L_{eq} (x_{1ij}) and the daytime-nightime L_{eq} difference (x_{2ij}). An ordinary least squares regression yields the following parameter estimates:

$$(4.1)\qquad b_0 = 0.0327, \quad se(b_0) = 0.132$$

$$b_1 = 0.0287, \quad se(b1) = 0.00229$$

$$b_2 = -0.00110, \quad se(b_2) = 0.00500$$

$$\hat{\sigma}_e = 0.911$$

The standard errors in (4.1) are those obtained from the least squares analysis assuming simple random sampling, and are provided only as rough indications of sampling variability.

We now analyse these data assuming Model I as described in equation (3.1). In doing so, regression coefficients are estimated within PSU's, and then averaged over PSU's. As stated in Section 3, this requires an acceptable joint distribution of independent variables within PSU's. In particular, there must be reasonable variation in both x_{1ij} and x_{2ij} and the two independent variables must not be too highly correlated within PSU's.

The acceptability of these within PSU design characteristics can be determined by inspecting the design measures Δ_{i11} and Δ_{i22} as described in equation (3.8) above. After eliminating those PSU's with insufficient variability in independent variables, we are left with 44 PSU's for Model I analysis.

Least squares estimates of the regression slopes were calculated within each of the 44 PSU's, and overall estimates of the two slopes were obtained as the averages of these PSU estimates as described in equation (3.4). By treating each PSU as an independent replicate, estimates of standard errors of these estimates can be obtained based on the variability of the individual PSU estimates. The results of this analysis are given below:

(4.2) $\hat{\beta}_1 = 0.0527 \quad se(\hat{\beta}_1) = 0.0169$

$\hat{\beta}_2 = 0.689 \quad se(\hat{\beta}_2) = 0.725$

$\hat{R} = 13.07 \quad se(\hat{R}) = 14.57$

Variance components estimates are obtained by using equations (3.16-19) as,

(4.3) $\hat{\sigma}_e = 0.659$

$\hat{\sigma}_{a11} = 0.00819$

$\hat{\sigma}_{a22} = 4.331$

$\hat{\sigma}_{a12} = 0.0331$

Using these estimates in the formulas developed in Section 3 for optimum sample design, equations (3.11) and (3.15) we get the following estimated optimum cluster sizes for estimating the three parameters, β_1, β_2 and R

(4.4) $\bar{n}.(opt)$ for estimating $\beta_1 = 4.82(C_a/C_b)^{1/2}$

$\bar{n}.(opt)$ for estimating $\beta_2 = 13.8(C_a/C_b)^{1/2}$

$\bar{n}.(opt)$ for estimating $R = 13.8(C_a/C_b)^{1/2}$

For example, for estimating the ratio R, optimum cluster sizes for various ratios of C_a/C_b are given below:

C_a/C_b	5	10	25	50
$\bar{n}.(opt)$	31	44	69	98

These calculations are only meant to be illustrative, and several qualifications concerning these estimates should be made. First, it should be noted that the estimate of β_2, and consequently that of the ratio R, is not very precise. It follows that the corresponding variance components and hence the optimum cluster sizes are not very precisely determined either. In addition, the actual design used in the Fields and Walker study does not correspond exactly to that described in Model I. A three-stage design rather than a two-stage design was employed. As a result, these calculations are relevant only for designs with PSU sizes similar to those observed in the study used here.

5. Summary and Conclusions

Interview studies of residents' response to noise are often based on two-stage sample designs. For these designs, samples of individuals are drawn within samples of compact study areas. In a typical survey, such a compact study area could consist of a neighborhood or a set of adjacent households. If the variability of the noise exposure variables within these compact study areas is not large, then the techniques described by Kalton (1983) can be used to determine optimal cluster design. On the other hand, if there is substantial within area variation in noise exposure, then the possibility of variability in the structural relationships (the "true" regression coefficients) over clusters should be considered. In such situations, the methods described for Model I in Section 3 can be used to assist in sample design.

Other noise studies have been based on a more complex design, a three-stage design. In such a design, samples of individuals are drawn from samples of compact study areas, which in turn are drawn from a sample of larger areas. For example, for a multi-airport study, these larger areas might correspond to cities. If there is substantial variability in noise exposure within compact study areas, the techniques based on Model III described in Section 3 can be used to assist in sample design. On the other hand, if there is substantial variability in noise exposure within compact study areas, then the methods described for Model II should be employed. These methods allow for the possibility of variability in the "true" regression coefficients among compact areas.

The statistical techniques described in this report can be used to provide assistance in designing noise surveys. It should also be noted that these techniques are more generally useful in a broad range of sample survey applications. Indeed, the conclusions regarding multi-stage sample design are applicable for any two-variable linear regression model of the form given in equation (1.1).

REFERENCES

Cochran, William G. (1977), Sampling Techniques, Third Ed., New York: Wiley.

Draper, N. R. and H. Smith (1981), Applied Regression Analysis, New York: Wiley.

Fields, J. M. and J. G. Walker (1982), "The Response to Railway Noise in Residential Areas in Great Britain", Journal of Sound and Vibration, Vol. 85, No. 2,177-255.

Kalton, Graham (1983), "Estimating Regression Coefficients from Clustered Samples: Sampling Errors and Optimum Sample Allocation", NASA Contractor Report 166117.

Neter, John, William Wasserman, and Michael Kutner (1985), Applied Linear Statistical Models, Second Edition, Homewood, IL: Irwin.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-177933		

4. Title and Subtitle	5. Report Date
Statistical Methods for Efficient Design of Community Surveys of Response to Noise: Random Coefficients Regression Models	July 1985
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Thomas J. Tomberlin	
	10. Work Unit No.

9. Performing Organization Name and Address		11. Contract or Grant No.
Bionetics Corporation 20 Research Drive Hampton, VA 23666	Subcontractor: Thomas J. Tomberlin 3615 Durocher Street Montreal Quebec CANADA H2X1E9	NAS1-16978

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Aeronautics and Space Administration Washington, DC 20546	Contractor Report
	14. Sponsoring Agency Code 505-35-13-53

15. Supplementary Notes

Langley Technical Monitor: Clemans A. Powell

16. Abstract

Research studies of residents' responses to noise consist of interviews with samples of individuals who are drawn from a number of different compact study areas (usually neighborhoods). In order to design such studies, it is necessary to determine the numbers of individuals and numbers of study areas which must be included to achieve the research objectives. The statistical techniques developed in this report provide a basis for those sample design decisions. These techniques are suitable for a wide range of sample survey applications. A sample may consist of a random sample of residents selected from a sample of compact study areas, or in a more complex design, of a sample of residents selected from a sample of compact study areas which has in turn been selected from a sample of larger areas (e.g. cities). The techniques may be applied to estimates of the effects on annoyance of noise level, numbers of noise events, the time-of-day of the events, ambient noise levels, or other factors. Methods are provided for determining, in advance, how accurately these effects can be estimated for different sample sizes and study designs. Using a simple cost function, they also provide for optimum allocation of the sample across the stages of the design for estimating these effects. These techniques are developed via a regression model in which the regression coefficients are assumed to be random, with components of variance associated with the various stages of a multi-stage sample design.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Cluster Samples Multi-stage Designs Optimal Design Random Effects Noise Annoyance	Unclassified Unlimited Subject Category - 71

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	25	A02

End of Document