nasa CR-166,117

NASA Contractor Report 166117

ESTIMATING REGRESSION COEFFICIENTS FROM CLUSTERED SAMPLES: SAMPLING ERRORS AND OPTIMUM SAMPLE ALLOCATION

Graham Kalton

UNIVERSITY OF MICHIGAN
Survey Research Center
Ann Arbor, Michigan 48106

Contract NAS1-16107
May 1983

NASA
National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23665

TABLE OF CONTENTS

#
N83-25496

1. Introduction

A number of surveys have been conducted around airports to study the relationship between the level of exposure to aircraft noise experienced by people living in the area and their annoyance with it. A two-stage sample is commonly adopted, selecting a sample of clusters at the first-stage and then a sample of individuals within selected clusters at the second stage. Most airports have maps of noise contours which are often used for stratification at the first stage; generally a disproportionate stratified sample of clusters is drawn, oversampling those in the high noise areas. Often all individuals in a selected cluster are assumed to experience the same noise exposure, which may therefore be measured by a single set of physical measurements in each sampled cluster.

In the simplest case, the regression coefficient for the simple regression of annoyance (y) on noise level (x) is the quantity of interest. Frequently annoyance is regressed on several noise-related independent variables, in which case the ratio of regression coefficients is often of interest (as with the noise and number index NNI). The issues addressed in this report are (1) the method of calculating standard errors for the estimated regression coefficients and for the ratio of estimated regression coefficients with a clustered two- or three-stage sample design and (2) the optimum way of allocating the sample across the stages of the sample design.

2. Regression Model

One approach to the specification of the regression is to take the regression coefficient in the population sampled as the quantity of interest. This population regression coefficient is

$$B = \sum^{N}(X_i - \bar{X})(Y_i - \bar{Y})/\sum(X_i - \bar{X})^2$$

for the population of size N. Under this approach, B may be estimated by

$$b = \sum^{n} w_i(x_i - \bar{x})(y_i - \bar{y})/\sum^{n} w_i(x_i - \bar{x})^2$$

where $\bar{x} = \Sigma w_i x_i/\Sigma w_i$, $\bar{y} = \Sigma w_i y_i/\Sigma w_i$ and w_i are weights inversely proportional to individuals' selection probabilities. Then the standard error of b may be estimated by techniques such as balanced repeated replication or jackknife repeated replication (Kish and Frankel, 1970, 1974). These techniques can take full account of the disproportionate stratification and clustering in the sample design.

The attraction of treating the quantity of interest as a parameter of the finite population (B) is the avoidance of the model assumptions required for standard regression analysis. However, the consequence of not making such assumptions is that the sample estimator b estimates B only for the specific population sampled, and cannot be readily applied to other populations. For the problem under study, the aim is to estimate a more general parameter, applicable to a wide range of populations (i.e. populations around a range of existing and proposed airports). For this reason, some regression model seems essential.

The assumptions made with the standard linear regression model $y_i = \beta_o + \beta x_i + e_i$ are that $E(e_i) = 0$, $V(e_i) = \sigma^2$ and $\text{Cov}\ (e_i, e_k) = 0$ for $i \neq k$. Under these assumptions β may be estimated by

$$b = \overset{n}{\Sigma}(x_i - \bar{x})(y_i - \bar{y})/\overset{n}{\Sigma}(x_i - \bar{x})^2 \qquad (1)$$

with $\bar{x} = \Sigma x_i/n$ and $\bar{y} = \Sigma y_i/n$. The variance of b is

$$V(b) = \sigma^2/\overset{n}{\Sigma}(x_i - \bar{x})^2 \qquad (2)$$

With this model the x's are considered fixed by the design. The choice of x-values affects the magnitude of V(b), but the above formulae apply whatever values of x are chosen. From the sampling perspective, the x's are mainly determined by the disproportionate stratification, and the formulae automatically reflect this aspect of the sample design. To the extent that the sampled x's are not fixed by the design, the formulae may be treated as conditional on the x's obtained.

While the standard regression model readily accommodates the effect of disproportionate stratification by x, it does not suitably reflect the clustering in the sample design. The clusters used in sample designs almost always exhibit some degree of homogeneity with respect to the variables under study, and this homogeneity has also been found to occur with regression residuals. The consequence of this homogeneity is that the assumption $\mathrm{Cov}(e_i,e_k) = 0$ does not hold for individuals i and k in the same cluster. To handle this feature, the model may be extended to

$$y_{ij} = \beta_o + \beta x_i + \alpha_i + e_{ij}$$

where the subscripts (i,j) refer to individual j in cluster i, and α_i is the cluster effect of cluster i. The α_i are random effects with $E(\alpha_i) = 0$. Under the further assumption $E(\alpha_i|x) = 0$, or $\mathrm{Cov}(\alpha_i,x) = 0$, b in (1) remains unbiased for β, but equation (2) no longer holds for the variance of b. It should be noted that estimators of β that are more efficient than b are available for this model; however, for simplicity, we will consider only the simple estimator b.

3. Variance of b

With the double subscript notation the estimated regression coefficient b in (1) may be expressed as

$$b = \Sigma\Sigma(x_i - \bar{x})(y_{ij} - \bar{y})/\Sigma n_i(x_i - \bar{x})^2$$

$$= \Sigma\Sigma(x_i - \bar{x})y_{ij}/\Sigma n_i(x_i - \bar{x})^2$$

$$= \Sigma n_i(x_i - \bar{x})\bar{y}_i/\Sigma n_i(x_i - \bar{x})^2$$

where there are n_i sampled individuals in cluster i and $\bar{y}_i = \Sigma y_{ij}/n_i$.

Conditional on the x_i's, the variance of b is then

$$V(b) = \Sigma n_i^2(x_i - \bar{x})^2 V(\bar{y}_i)/[\Sigma n_i(x_i - \bar{x})^2]^2$$

Under the model of $y_{ij} = \beta_o + \beta x_i + \alpha_i + e_{ij}$ with $V(e_{ij}) = \sigma_e^2$, and $V(\alpha_i) = \sigma_\alpha^2$,

$$V(\bar{y}_i) = V(\alpha_i + \bar{e}_i) = \sigma_\alpha^2 + (\sigma_e^2/n_i)$$

Thus

$$V(b) = \Sigma n_i^2(x_i - \bar{x})^2(\sigma_\alpha^2 + \sigma_e^2/n_i)/[\Sigma n_i(x_i - \bar{x})^2]^2$$

$$= \{\sigma_\alpha^2\Sigma n_i^2(x_i - \bar{x})^2/[\Sigma n_i(x_i - \bar{x})^2]^2\} + \{\sigma_e^2/\Sigma n_i(x_i - \bar{x})^2\} \quad (3)$$

In the special case when the same subsample size is taken from each cluster, $n_i = \bar{n}$, V(b) reduces to

$$V(b) = [\sigma_\alpha^2 + (\sigma_e^2/\bar{n})]/\overset{a}{\Sigma}(x_i - \bar{x})^2 \quad (4)$$

$$= [(\sigma_\alpha^2/a) + (\sigma_e^2/n)]/\sigma_x^2 \quad (5)$$

where σ_x^2 is defined as $\overset{a}{\Sigma}(x_i - \bar{x})^2/a$ and a is the number of sampled clusters.

Defining the intra-class correlation coefficient for the clusters as the proportion of the variance of the y_{ij} conditional on the x_i that is accounted for by the cluster effect, i.e. $\rho = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_e^2)$, V(b) may be alternatively expressed as

$$V(b) = (\sigma^2/n)[1 + (\bar{n} - 1)\rho]/\sigma_x^2 \qquad (6)$$

where $\sigma^2 = \sigma_\alpha^2 + \sigma_e^2$.

An estimator of V(b) may be obtained by substituting estimates $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_e^2$ for σ_α^2 and σ_e^2 in (3) or (4). The quantity σ_e^2 may be estimated by the residual mean square from a one-way analysis of variance of the y-values by clusters, that is by

$$\hat{\sigma}_e^2 = \Sigma\Sigma(y_{ij} - \bar{y}_i)^2/(n - a)$$

where n is the total sample size and a is the number of sampled clusters. Then σ_α^2 may be estimated by

$$\hat{\sigma}_\alpha^2 = [\Sigma\Sigma(y_{ij} - b_o - bx_i)^2 - (n - 2)\hat{\sigma}_e^2]/[\lambda(a - 2)]$$

where $\Sigma\Sigma(y_{ij} - b_o - bx_i)^2$ is the residual sum of squares from the regression of y on x, b_o is the sample estimate of the intercept β_o, and $\lambda = (n^2 - \Sigma n_i^2)/n(a - 2)$ (for λ, see Anderson and Bancroft, 1952, Section 25.2; Snedecor and Cochran, 1980, Section 13.7).

4. Optimum subsample size, $\bar{n}$

In this section, we consider the optimum allocation of the sample between the first and second stages of the sample. We assume that the same subsample size $\bar{n}$ is taken from each selected cluster; the results obtained can also be applied as an approximation to situations where the subsample size varies to a small extent between clusters, in which case $\bar{n}$ represents the average subsample size. We assume a simple cost model of the form $C = aC_a + nc$, where C_a is the cost of including a cluster in the sample, c is the cost of including an individual, and $n = a\bar{n}$ is the total sample size.

For given σ_x^2, the optimum choice of $\bar{n}$ that minimizes $V(b)$ for fixed total cost C may then be readily obtained from the Cauchy-Schwartz inequality as follows. Write $V(b) = \Sigma x_i^2$, where $x_1 = \sigma_\alpha/\sqrt{a}$ and $x_2 = \sigma_e/\sqrt{n}$, and $C = \Sigma y_i^2$, where $y_1 = \sqrt{aC_a}$ and $y_2 = \sqrt{nc}$. Then the product $V(b).C$ is minimized when

$$(x_1/y_1) = (x_2/y_2)$$

i.e. when

$$\sigma_\alpha/a\sqrt{C_a} = \sigma_e/n\sqrt{c}$$

or

$$\bar{n}_{opt} = (\sigma_e/\sigma_\alpha)(C_a/c)^{1/2} \qquad (7)$$

This result can be equivalently expressed in terms of the cluster intra-class correlation as

$$\bar{n}_{opt} = [(1 - \rho)/\rho]^{1/2}[C_a/c]^{1/2} \qquad (8)$$

5. Example

A study of traffic noise was carried out with a sample of n = 2933 cases in a = 53 clusters (Langdon, 1976). Of the 2933 cases, 2881 provided responses which are analyzed here. The average number of respondents per cluster is thus $\bar{n}$ = 54.358; the cluster sizes varied markedly, from the lowest of 20 respondents to the highest of 109 respondents. The dependent variable for the regression is the answer to the question "How do you feel about traffic noise here?" (the end points of the scale are labelled "definitely satisfactory" and "definitely unsatisfactory") and the independent variable is the noise level (24 hour Leq dB(A)). The regression coefficient is b = 0.07971.

The following sums of squares (SS) and degrees of freedom (d.f.) were obtained for the regression of annoyance on noise level:

Source	d.f.	SS
Regression	1	293.1871
Residuals	2879	10200.1250
Total	2880	10493.3121

The analysis of variance of the annoyance scores by clusters yielded the following results:

Source	d.f.	SS
Clusters	52	1525.0925
Residuals	2828	8968.2196
Total	2880	10493.3121

From these results the following analysis of variance table for the regression residuals is constructed:

Residuals	d.f.	SS	MS	E(MS)
Between clusters after regression	51*	1231.9054	24.15501	$\lambda\sigma_\alpha^2 + \sigma_e^2$
Within clusters	2828	8968.2196	3.17122	σ_e^2
Total regression residuals	2879	10200.1250		

*Note that one degree of freedom is used for the regression.

The residual variance σ_e^2 is estimated by the within clusters residual mean square, i.e. $\hat{\sigma}_e^2 = 3.17122$. The expected value of the between clusters after regression residual mean square is $\lambda\sigma_\alpha^2 + \sigma_e^2$, where $\lambda = (n^2 - \Sigma n_i^2)/nd$ and d is the degrees of freedom for the between clusters after regression residuals. An approximate value for λ is the average sample size per cluster, $\bar{n} = 54.358$. With n = 2881, $\Sigma n_i^2 = 174,571$ and d = 51, the exact value of λ is 55.302. Using this exact value,

$$\hat{\sigma}_\alpha^2 = (24.15501 - 3.17122)/55.302$$

$$= 0.37944$$

and $$\hat{\rho} = \hat{\sigma}_\alpha^2/(\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2) = 0.1069 \text{ or } 10.7\%.$$

With $\hat{\rho} = 0.1069$, from (8)

$$\bar{n}_{opt} = 2.891[C_a/c]^{1/2}$$

Values of $\bar{n}_{opt}$ for various ratios of C_a/c are given below:

C_a/c	5	10	20	30	40	50
$\bar{n}_{opt}$	6	9	13	16	18	20

A variance estimate for b is obtained by substituting sample estimates in (3). Using

$$\Sigma n_i^2(x_i - \bar{x})^2 = \Sigma n_i^2 x_i^2 - 2\bar{x}\Sigma n_i^2 x_i + \bar{x}^2 \Sigma n_i^2,$$
$$\Sigma n_i(x_i - \bar{x})^2 = (n - 1)\sigma_x^2$$

where $\bar{x} = 70.5917$, $\sigma_x = 4.0026$ is the standard deviation of x, $\Sigma n_i^2 x_i^2 = 869{,}779{,}520.5$, $\Sigma n_i^2 x_i = 12{,}303{,}941.25$ and $\Sigma n_i^2 = 174{,}571$, the following results are obtained:

$$\Sigma n_i^2(x_i - \bar{x})^2 = 2587392.84$$
$$\Sigma n_i(x_i - \bar{x})^2 = 46139.92346.$$

Substituting these values and $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_e^2$ from above in (3) gives

$$v(b) = (46.11601 + 6.87305) \times 10^{-5}$$
$$= 5.2989 \times 10^{-4}.$$

The estimate of the variance of b from the standard regression analysis is 7.6788×10^{-5}, so that ignoring the cluster design underestimates the variance by a factor of 6.90. This factor corresponds approximately to the multiplier $[1 + (\bar{n} - 1)\hat{\rho}] = 6.70$ in (6).

Note that an approximate variance estimate for b is obtained by assuming $n_i = \bar{n}$ and using equation (5). Then $v^*(b) \doteq 5.1558 \times 10^{-4}$. This value is fairly close to that obtained above, even in this case where the n_i are subject to substantial variation. This approximate variance estimate is 6.70 times as large as the estimate of the variance of b from standard regression analysis: this factor is the multiplier $[1 + (\bar{n} - 1)\hat{\rho}] = 6.70$.

6. Extension to regression with two independent variables

We turn now to a linear regression of y on two independent variables x and z, both of which are constant within clusters:

$$y_i = \beta_o + \beta_x x_i + \beta_z z_i + e_i$$

Under the standard assumptions that $E(e_i) \doteq 0$, $V(e_i) = \sigma_r^2$ and $Cov(e_i, e_k) = 0$ for $i \neq k$, β_x and β_z may be estimated by

$$b_x = \{\Sigma(z_i - \bar{z})^2\Sigma(x_i - \bar{x})(y_i - \bar{y})$$
$$- \Sigma(x_i - \bar{x})(z_i - \bar{z})\Sigma(z_i - \bar{z})(y_i - \bar{y})\}/\Delta \quad (9)$$

$$b_z = \{\Sigma(x_i - \bar{x})^2\Sigma(z_i - \bar{z})(y_i - \bar{y})$$
$$- \Sigma(x_i - \bar{x})(z_i - \bar{z})\Sigma(x_i - \bar{x})(y_i - \bar{y})\}/\Delta \quad (10)$$

where $\Delta = \Sigma(x_i - \bar{x})^2\Sigma(z_i - \bar{z})^2 - [\Sigma(x_i - \bar{x})(z_i - \bar{z})]^2$. Under this model the x's and z's are considered fixed by the design. The choice of combinations of x and z values affects the precision of the estimators.

Consider the estimators b_x and b_z under the model

$$y_{ij} = \beta_o + \beta_x x_i + \beta_z z_i + \alpha_i + e_{ij} \quad (11)$$

where α_i is the cluster effect of cluster i, which is assumed to be a random effect with $E(\alpha_i) = 0$. Under the further assumptions that the α_i are uncorrelated with the x's and the z's, b_x and b_z remain unbiased for β_x and β_z. Since b_x and b_z are of the same form, simply with x and z interchanged, it will suffice to obtain the variance of one of them, say b_x. Using the double subscript notation and letting the sum of squares of the z's be

$$\Sigma\Sigma(z_i - \bar{z})^2 = \Sigma n_i(z_i - \bar{z})^2 = S_{zz}$$

and the sum of cross-products of the z's and the x's be $\Sigma\Sigma(x_i - \bar{x})(z_i - \bar{z}) = \Sigma n_i(x_i - \bar{x})(z_i - \bar{z}) = S_{xz}$, b_x may be expressed as

$$\begin{aligned} b_x &= [S_{zz}\Sigma\Sigma(x_i - \bar{x})y_{ij} - S_{xz}\Sigma\Sigma(z_i - \bar{z})y_{ij}]/\Delta \\ &= \Sigma\Sigma[(S_{zz}(x_i - \bar{x}) - S_{xz}(z_i - \bar{z})]y_{ij}/\Delta \\ &= \Sigma n_i[S_{zz}(x_i - \bar{x}) - S_{xz}(z_i - \bar{z})]\bar{y}_i/\Delta \\ &= \Sigma n_i C_i \bar{y}_i/\Delta \end{aligned}$$

where $C_i = S_{zz}(x_i - \bar{x}) - S_{xz}(z_i - \bar{z})$.

Conditional on the x's and z's, the variance of b_x is then

$$V(b_x) = \Sigma n_i^2 C_i^2 V(\bar{y}_i)/\Delta^2$$

Under the model given by (11) with $V(e_{ij}) = \sigma_e^2$ and $V(\alpha_i) = \sigma_\alpha^2$,

$$V(\bar{y}_i) = V(\alpha_i + \bar{e}_i) = \sigma_\alpha^2 + (\sigma_e^2/n_i)$$

Thus

$$V(b_x) = [\sigma_\alpha^2 \Sigma n_i^2 C_i^2 + \sigma_e^2 \Sigma n_i C_i^2]/\Delta^2 \qquad (12)$$

In the special case when $n_i = \bar{n}$, $V(b_x)$ reduces to

$$V(b_x) = \bar{n}^2 \Sigma C_i^2 [\sigma_\alpha^2 + (\sigma_e^2/\bar{n})]/\Delta^2 \qquad (13)$$

Defining the intra-class correlation coefficient for the clusters as $\rho = \sigma_\alpha^2/\sigma^2$, where $\sigma^2 = \sigma_\alpha^2 + \sigma_e^2$, $V(b_x)$ may be expressed as

$$V(b_x) = \bar{n}\Sigma C_i^2 \sigma^2 [1 + (\bar{n} - 1)\rho]/\Delta^2. \qquad (14)$$

In order to obtain the optimum subsample size, $\bar{n}$, it is useful to express ΣC_i^2 and Δ^2 in terms of the variances of x and z and the covariance between x and z, which are defined as $\sigma_x^2 = \Sigma(x_i - \bar{x})^2/a$, $\sigma_z^2 = \Sigma(z_i - \bar{z})^2/a$ and $\sigma_{xz} = \Sigma(x_i - \bar{x})(z_i - \bar{z})/a$. Using this notation, $S_{zz} = \bar{n}a\sigma_z^2$, $S_{xz} = \bar{n}a\sigma_{xz}$,

$$\begin{aligned}\Sigma C_i^2 &= \bar{n}^2a^2\Sigma[\sigma_z^2(x_i - \bar{x}) - \sigma_{xz}(z_i - \bar{z})]^2\\ &= \bar{n}^2a^3(\sigma_z^4\sigma_x^2 - 2\sigma_z^2\sigma_{xz}^2 + \sigma_{xz}^2\sigma_z^2)\\ &= \bar{n}^2a^3\sigma_z^2(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)\end{aligned}$$

and $$\Delta^2 = \bar{n}^4a^4(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)^2$$

Substituting these values in (13) gives

$$V(b_x) = A[(\sigma_\alpha^2/a) + (\sigma_e^2/n)] \qquad (15)$$

where $$A = \sigma_z^2/(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2).$$

The form of $V(b_x)$ in (15) is now the same as that for the regression with the single independent variable in (5), with A replacing σ_x^2. It therefore follows that the optimum value of $\bar{n}$ is given by equations (6) or (7), namely

$$\bar{n}_{opt} = (\sigma_e/\sigma_\alpha)(C_a/c)^{1/2} = [(1 - \rho)/\rho]^{1/2}[C_a/c]^{1/2}. \qquad (16)$$

Note, however, that σ_e^2 is now the residual variance from the multiple regression. This residual variance can in general be expected to be smaller than that for the simple regression, and hence the value of $\bar{n}_{opt}$ will also be smaller. Since the formula for $\bar{n}_{opt}$ depends only on ρ and the cost ratio, it is the same for both regression coefficients, i.e. the optimum allocation is given by (16) whether β_x or β_z is being estimated.

In order to estimate the variance of b_x, first note that $\Sigma n_i C_i^2 = \Sigma n_i(z_i - \bar{z})^2\Delta$, so that $V(b_x)$ in (12) may be written as

$$V(b_x) = (\sigma_\alpha^2 \Sigma n_i^2 C_i^2/\Delta^2) + (\sigma_e^2 \Sigma n_i(z_i - \bar{z})^2/\Delta) \qquad (17)$$

where $\Sigma n_i^2 C_i^2$ and Δ may be computed as

$$\begin{aligned}\Sigma n_i^2 C_i^2 = {} & [\Sigma n_i(z_i - \bar{z})^2]^2 \Sigma n_i^2(x_i - \bar{x})^2 \\ & + [\Sigma n_i(x_i - \bar{x})(z_i - \bar{z})]^2 \Sigma n_i^2(z_i - \bar{z})^2 \\ & - 2\Sigma n_i(z_i - \bar{z})^2 \Sigma n_i(x_i - \bar{x})(z_i - \bar{z}) \Sigma n_i^2(x_i - \bar{x})(z_i - \bar{z})\end{aligned}$$

$$\Delta = \Sigma n_i(x_i - \bar{x})^2 \Sigma n_i(z_i - \bar{z})^2 - [\Sigma n_i(x_i - \bar{x})(z_i - \bar{z})]^2$$

The variance of b_x can then be estimated by substituting sample estimates of σ_α^2 and σ_e^2 in (17). As with the simple regression (p. 6), σ_e^2 may be estimated by

$$\hat{\sigma}_e^2 = \Sigma\Sigma(y_{ij} - \bar{y}_i)^2/(n - a)$$

Then, noting that the regression sum of squares now has two degrees of freedom, σ_α^2 may be estimated by

$$\hat{\sigma}_\alpha^2 = [\Sigma\Sigma(y_{ij} - b_o - b_x x_i - b_z z_i)^2 - (n - 3)\hat{\sigma}_e^2]/\lambda(a - 3).$$

(With a multiple regression with K independent variables, σ_α^2 may be estimated by

$$\hat{\sigma}_\alpha^2 = [\Sigma\Sigma(y_{ij} - \hat{y}_{ij})^2 - (n - K - 1)\hat{\sigma}_e^2]/\lambda(a - K - 1)$$

where $\hat{y}_{ij} = b_o + \Sigma b_k x_{ki}$ are the predicted values from the regression.)

7. Ratio of two regression coefficients

With aircraft noise surveys one common analysis is to run a regression of respondents' annoyance with aircraft noise (y) on the levels of noise (x) and numbers of the noise events (z) to which they are exposed. The level of noise and number of noise events may be combined into a noise and number index (NNI). For this purpose the ratio of the regression coefficients, $t = b_z/b_x$, is needed. This section demonstrates that the optimum choice of $\bar{n}$ (assumed constant for all clusters) for estimating t is the same as that given in equations 8 and 16. The results in this section are derived using two slightly different applications of the Taylor's series expansion method for obtaining large-sample approximations to the variances of complex statistics.

First Application

Using the notation of the previous section, b_x may be expressed as $\Sigma n_i C_i \bar{y}_i/\Delta$, and b_z may be similarly expressed as $\Sigma n_i D_i \bar{y}_i/\Delta$. Thus

$$t = \Sigma n_i D_i \bar{y}_i/\Sigma n_i C_i \bar{y}_i \qquad (18)$$

Treating t as a function of the random variables $\bar{y}_i$, the approximate variance of t for large samples may be obtained from the Taylor's series expansion method. From this method the approximate variance of t is equal to that of its linear substitute, t^*, where

$$t^* = \Sigma(\delta t/\delta\bar{y}_i)\bar{y}_i$$

and $(\delta t/\delta\bar{y}_i)$ is evaluated at $\bar{y}_i = E(\bar{y}_i) = \bar{Y}_i$, say.

Now

$$V(t^*) = \Sigma(\delta t/\delta\bar{y}_i)^2 V(\bar{y}_i) = \Sigma(\delta t/\delta\bar{y}_i)^2[\sigma_\alpha^2 + (\sigma_e^2/n_i)]$$

Thus, in general, under the model given in equation (1),

$$V(t) \simeq \Sigma(\delta t/\delta\bar{y}_i)^2[\sigma_\alpha^2 + (\sigma_e^2/n_i)] \qquad (19)$$

Assuming a constant subsample size, $n_i = \bar{n}$,

$$V(t) \simeq \Sigma(\delta t/\delta\bar{y}_i)^2 a(\sigma^2/n)[1 + (\bar{n} - 1)\rho] \qquad (20)$$

$$= K(\sigma^2/n)[1 + (\bar{n} - 1)\rho] \qquad (21)$$

where $K = a\Sigma(\delta t/\delta\bar{y}_i)^2$. This equation is of the same form as (6) with σ_x^2 replaced by $1/K$. Thus, providing K is not a function of a or n, the optimum value of $\bar{n}$ for estimating t is the same as that for estimating b, i.e. the value given by equation (8).

The following derivation demonstrates that K does not depend on a or n. First, from (18) with $n_i = \bar{n}$ it follows that

$$\delta t/\delta\bar{y}_i = \{(\Sigma C_i\bar{Y}_i)D_i - (\Sigma D_i\bar{Y}_i)C_i\}/(\Sigma C_i Y_i)^2 \qquad (22)$$

so that K is

$$a\{(\Sigma C_i\bar{Y}_i)^2\Sigma D_i^2+(\Sigma D_i\bar{Y}_i)^2\Sigma C_i^2-2(\Sigma C_i\bar{Y}_i)(\Sigma D_i\bar{Y}_i)(\Sigma C_iD_i)\}/(\Sigma C_i\bar{Y}_i)^4$$

where $\Sigma D_i\bar{Y}_i = \bar{n}a^2(\sigma_x^2\sigma_{Yz} - \sigma_{xz}\sigma_{Yx})$

$\Sigma C_i\bar{Y}_i = \bar{n}a^2(\sigma_z^2\sigma_{Yx} - \sigma_{xz}\sigma_{Yz})$.

$$\Sigma C_i^2 = \bar{n}^2 a^3 \sigma_z^2(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)$$

$$\Sigma D_i^2 = \bar{n}^2 a^3 \sigma_x^2(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)$$

$$\Sigma C_i D_i = - \bar{n}^2 a^3 \sigma_{xz}(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)$$

and $$\sigma_{Yz} = \Sigma(\bar{Y}_i - \bar{Y})(z_i - \bar{z})/a \doteq \Sigma\bar{Y}_i(z_i - \bar{z})/a$$

$$\sigma_{Yx} = \Sigma(\bar{Y}_i - \bar{Y})(x_i - \bar{x})/a = \Sigma\bar{Y}_i(x_i - \bar{x})/a$$

All the terms in the numerator of K have a common factor of $\bar{n}^4 a^8$ and the denominator has this same common factor. On cancellation of this factor, K is seen to be a function only of σ_x^2, σ_z^2, σ_{xz}, σ_{Yx}, σ_{Yz}. Thus K does not depend on $\bar{n}$ or a.

Given values of σ_x^2, σ_z^2, σ_{xz} and estimates of σ_{Yx}, σ_{Yz}, σ_α^2 and σ_e^2, an estimate of V(t) can be obtained by substituting these values and estimates in equation (19) using $(\delta t/\delta\bar{y}_i)$ from (22).

Second Application

An alternative approach for obtaining V(t) is to start with the Taylor expansion of the ratio $t = b_z/b_x$. Thus

$$V(t) \doteq \beta_x^{-2}[V(b_z) + \tau^2 V(b_x) - 2\tau C(b_x,b_z)] \tag{23}$$

where $\tau = \beta_z/\beta_x$ and $C(b_x,b_z)$ is the covariance of b_x and b_z. From (15)

$$V(b_x) = \sigma_z^2[(\sigma_\alpha^2/a) + (\sigma_e^2/n)]/(\sigma_x^2\sigma_z^2 - \sigma_{xz})^2 \tag{24}$$

and $$V(b_z) = \sigma_x^2[(\sigma_\alpha^2/a) + (\sigma_e^2/n)]/(\sigma_x^2\sigma_z^2 - \sigma_{xz})^2 \tag{25}$$

The covariance term is obtained from

$$C(b_x,b_z) = \Sigma(\delta b_x/\delta\bar{y}_i)(\delta b_z/\delta\bar{y}_i)V(\bar{y}_i),$$

with other terms in the summation being zero since $C(\bar{y}_i,\bar{y}_j) = 0$ for $i \neq j$. Expressions for b_x and b_z are

$$b_x = [\sigma_z^2\Sigma(x_i - \bar{x})\bar{y}_i - \sigma_{xz}\Sigma(z_i - \bar{z})\bar{y}_i]/a(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)$$

$$b_z = [\sigma_x^2\Sigma(z_i - \bar{z})\bar{y}_i - \sigma_{xz}\Sigma(x_i - \bar{x})\bar{y}_i]/a(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)$$

Thus $C(b_x,b_z)$ is

$$\frac{[\sigma_\alpha^2+(\sigma_e^2/\bar{n})]\Sigma[\sigma_z^2(x_i - \bar{x})-\sigma_{xz}(z_i-\bar{z})][\sigma_x^2(z_i-\bar{z})-\sigma_{xz}(x_i-\bar{x})]}{a^2(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)^2}$$

$$= -\sigma_{xz}[(\sigma_\alpha^2/a) + (\sigma_e^2/n)]/(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2) \qquad (26)$$

Substituting (24), (25) and (26) in (23) gives

$$V(t) = \frac{[(\sigma_\alpha^2/a) + (\sigma_e^2/n)][\sigma_x^2 + \tau^2\sigma_z^2 + 2\tau\sigma_{xz}]}{\beta_x^2(\sigma_x^2\sigma_z^2 - \sigma_{xz}^2)}$$

With $\rho_{xz} = \sigma_{xz}/\sigma_x\sigma_z$,

$$V(t) = \frac{[(\sigma_\alpha^2/a)+(\sigma_e^2/n)][(1/\sigma_z^2)+(\tau^2/\sigma_x^2)+(2\tau\rho_{xz}/\sigma_x\sigma_z)]}{\beta_x^2(1 - \rho_{xz}^2)} \qquad (27)$$

A variance estimate $\hat{V}(t)$ is obtained by substituting sample estimates $\hat{\sigma}_\alpha^2$, $\hat{\sigma}_e^2$, t and b_x for the respective unknown parameters in (27).

The accuracy of the approximate variance of the ratio $t = b_z/b_x$ obtained by the Taylor expansion method depends on the coefficient of variation of the denominator of the ratio, i.e. $CV(b_x) = \sqrt{V(b_x)}/\beta_x$. A $CV(b_x)$ of less than 0.2 and preferably less than 0.1 is required if the Taylor expansion method is to produce a satisfactory approximation of V(t). (It should be noted that a low $CV(b_x)$ also ensures that the bias of t is negligible). A check should be made that the estimated coefficient of variation $cv(b_x) = \sqrt{v(b_x)}/b_x$ is less than 0.2; if this condition is not satisfied, the Taylor expansion variance estimate should not be used. In any case, if this condition is not satisified, the utility of the index t should be critically examined.

For the first application of the Taylor expansion method, the equivalent condition is that the coefficient of variation of the denominator, i.e. $CV(\Sigma n_i C_i \bar{y}_i)$, should be small, less than 0.2 and preferably less than 0.1.

8. The case of variable x in clusters

The previous sections have assumed that x_i is a constant value within a cluster. We now consider the case where x takes different values within a cluster, individual j in cluster i having a value x_{ij}. The regression coefficient β is assumed to be the same within each cluster. In this case the treatment of the simple regression discussed in section 3 is modified as follows.

The simple regression coefficient is now given by

$$\begin{aligned} b &= \Sigma\Sigma(x_{ij} - \bar{x})y_{ij}/\Sigma\Sigma(x_{ij} - \bar{x})^2 \\ &= \Sigma\Sigma\xi_{ij}y_{ij}/\Sigma\Sigma\xi_{ij}^2 \end{aligned} \qquad (28)$$

with $\xi_{ij} = (x_{ij} - \bar{x})$. The variance of b is then

$$V(b) = \{\Sigma_i\Sigma_j\xi_{ij}^2 V(y_{ij}) + \Sigma_i\Sigma\Sigma_{j\neq k}\xi_{ij}\xi_{ik}C(y_{ij},y_{ik})\}/(\Sigma\Sigma\xi_{ij}^2)^2$$

where $C(y_{ij}, y_{ik})$ is the covariance of y_{ij} and y_{ik}.

Now $\qquad V(y_{ij}) = \sigma_\alpha^2 + \sigma_e^2$

and $\qquad \begin{aligned} C(y_{ij}, y_{ik}) &= E[y_{ij} - E(y_{ij})][y_{ik} - E(y_{ik})] \\ &= E[(\alpha_i + e_{ij})(\alpha_i + e_{ik})] \\ &= \sigma_\alpha^2 \end{aligned}$

Thus

$$\begin{aligned} V(b) &= \{\sigma_\alpha^2[\Sigma\Sigma\xi_{ij}^2 + \Sigma\Sigma\Sigma\xi_{ij}\xi_{ik}] + \sigma_e^2(\Sigma\Sigma\xi_{ij}^2)\}/(\Sigma\Sigma\xi_{ij}^2)^2 \\ &= \{\sigma_\alpha^2[\Sigma_i(\Sigma_j\xi_{ij})^2] + \sigma_e^2(\Sigma\Sigma\xi_{ij}^2)\}/(\Sigma\Sigma\xi_{ij}^2)^2 \\ &= \{\sigma_\alpha^2\Sigma n_i^2(\bar{x}_i - \bar{x})^2/[\Sigma\Sigma(x_{ij} - \bar{x})^2]^2\} + \{\sigma_e^2/\Sigma\Sigma(x_{ij} - \bar{x})^2\} \end{aligned} \quad (29)$$

This formula is the generalization of (3): substituting $x_{ij} = \bar{x}_i = x_i$ in (29) yields (3).

In order to examine the optimum subsample size, consider the case with $n_i = \bar{n}$. Denoting the proportion of the variance in x explained by the clusters as

$$\eta^2 = \bar{n}\Sigma(\bar{x}_i - \bar{x})^2/\Sigma\Sigma(x_{ij} - \bar{x})^2, \qquad (30)$$

the variance of b is given by

$$V(b) = [(\sigma_\alpha^2\eta^2/a) + (\sigma_e^2/n)]/\sigma_x^2 \qquad (31)$$

or

$$V(b) = (\sigma^2/n)[1 + (\bar{n}\eta^2 - 1)\rho]/\sigma_x^2 \qquad (32)$$

These formulae are the same as equations (5) and (6) except that σ_α^2 is replaced by $\sigma_\alpha^2\eta^2$ in (5) and $\bar{n}$ is replaced by $\bar{n}\eta^2$ in (6). Thus by redefining x in Section 4 to be $\sigma_\alpha\eta/\sqrt{a}$, the optimum value of $\bar{n}$ is obtained directly as

$$\bar{n}_{opt} = (\sigma_e/\eta\sigma_\alpha)(C_a/c)^{1/2} \qquad (33)$$

or equivalently as

$$\bar{n}_{opt} = [(1 - \rho)/\rho]^{1/2}[C_a/c]^{1/2}(1/\eta) \qquad (34)$$

Note that if $\eta = 1$, i.e. $x_{ij} = \bar{x}_i$ for all j, then $\bar{n}_{opt}$ reduces to that obtained in Section 4. If $\eta = 0$, i.e. the cluster means for x are all the same so that the variability in x is all within the clusters, $\bar{n}_{opt} = n$, with only one cluster being sampled.

9. A three-stage design

In this section we consider a three stage sample design. At the first stage a primary sampling units (PSU's) are selected; next n_i second stage units (SSU's) are selected within PSU i; and finally n_{ij} elements are selected in second stage unit j in PSU i. The regression model with a single independent x-variable extends to

$$y_{ijk} = \beta_o + \beta x_{ij} + \alpha_i + \delta_{ij} + e_{ijk} \qquad (35)$$

where α_i is the cluster effect of PSU i and δ_{ij} is the cluster effect of SSU ij. The α_i and δ_{ij} are random effects with $E(\alpha_i) = E(\delta_{ij}) = 0$, $E(\alpha_i^2) = \sigma_\alpha^2$, $E(\delta_{ij}^2) = \sigma_\delta^2$, and

$E(\alpha_i|x) = E(\delta_{ij}|x) = 0$. The x-variable is assumed to be constant within a SSU, and the regression coefficient is assumed to be the same within each PSU.

The simple regression coefficient may be expressed as

$$b = \frac{\Sigma\Sigma\Sigma(x_{ij} - \bar{x})y_{ijk}}{\Sigma\Sigma\Sigma(x_{ij} - \bar{x})^2} = \frac{\Sigma\Sigma n_{ij}(x_{ij} - \bar{x})\bar{y}_{ij}}{\Sigma\Sigma n_{ij}(x_{ij} - \bar{x})^2} \tag{36}$$
$$= \Sigma\Sigma n_{ij}\xi_{ij}\bar{y}_{ij}/\Sigma\Sigma n_{ij}\xi_{ij}^2$$

where $\xi_{ij} = (x_{ij} - \bar{x})$. Then

$$V(b) = \frac{\Sigma\Sigma n_{ij}^2\xi_{ij}^2 V(\bar{y}_{ij})}{(\Sigma\Sigma n_{ij}\xi_{ij}^2)^2} + \frac{\Sigma\Sigma\Sigma n_{ij}\xi_{ij}n_{ik}\xi_{ik}C(\bar{y}_{ij}, \bar{y}_{ik})}{(\Sigma\Sigma n_{ij}\xi_{ij}^2)^2} \tag{37}$$

Now $\bar{y}_{ij} = \beta_o + \beta x_{ij} + \alpha_i + \delta_{ij} + \bar{e}_{ij}$, so that

$$V(\bar{y}_{ij}) = \sigma_\alpha^2 + \sigma_\delta^2 + (\sigma_e^2/n_{ij}) \tag{38}$$

and $\text{Cov}(\bar{y}_{ij}, \bar{y}_{ik}) = E(\alpha_i + \delta_{ij} + \bar{e}_{ij})(\alpha_i + \delta_{ik} + \bar{e}_{ik})$

$$= \sigma_\alpha^2 \text{ for } j \neq k. \tag{39}$$

Thus

$$V(b) = \frac{\sigma_\alpha^2(\Sigma\Sigma n_{ij}^2\xi_{ij}^2 + \Sigma\Sigma\Sigma n_{ij}\xi_{ij}n_{ik}\xi_{ik})}{(\Sigma\Sigma n_{ij}\xi_{ij}^2)^2}$$
$$+ \frac{\sigma_\delta^2\Sigma\Sigma\ n_{ij}^2\xi_{ij}^2}{(\Sigma\Sigma n_{ij}\xi_{ij}^2)^2} + \frac{\sigma_e^2\Sigma\Sigma n_{ij}\xi_{ij}^2}{(\Sigma\Sigma n_{ij}\xi_{ij}^2)^2}$$

$$= \frac{\sigma_\alpha^2\Sigma(\Sigma n_{ij}\xi_{ij})^2}{(\Sigma\Sigma n_{ij}\xi_{ij}^2)^2} + \frac{\sigma_\delta^2\Sigma\Sigma n_{ij}^2\xi_{ij}^2}{(\Sigma\Sigma n_{ij}\xi_{ij}^2)^2} + \frac{\sigma_e^2}{\Sigma\Sigma n_{ij}\xi_{ij}^2} \tag{40}$$

Consider now the case where the same number of SSU's is taken from each PSU, $n_i = d$, and the same number of elements is taken from each SSU, $n_{ij} = \bar{n}$. The V(b) in (40) reduces to

$$V(b) = \frac{\sigma_\alpha^2 \Sigma(\Sigma\xi_{ij})^2}{(\Sigma\Sigma\xi_{ij}^2)^2} + \frac{\sigma_\delta^2}{\Sigma\Sigma\xi_{ij}^2} + \frac{\sigma_e^2}{\bar{n}\Sigma\Sigma\xi_{ij}^2} \qquad (41)$$

Denoting the variance of x as

$$\sigma_x^2 = \Sigma\Sigma(x_{ij} - \bar{x})^2/ad \qquad (42)$$

and the proportion of σ_x^2 explained by the PSU's as

$$\eta^2 = \frac{d\Sigma(\bar{x}_i - \bar{x})^2}{\Sigma\Sigma(x_{ij} - \bar{x})^2} = \frac{\Sigma(\Sigma\xi_{ij})^2}{d\Sigma\Sigma\xi_{ij}^2}, \qquad (43)$$

V(b) may be expressed as

$$V(b) = \left[\frac{\sigma_\alpha^2\eta^2}{a} + \frac{\sigma_\delta^2}{ad} + \frac{\sigma_e^2}{ad\bar{n}}\right]/\sigma_x^2 \qquad (44)$$

To determine the optimum values of $\bar{n}$ and c, consider the simple cost model $C = ac_a + adc_d + \bar{n}adc$, where c_a is cost of including a PSU, c_d is the cost of including a SSU and c the cost of including an element in the sample.

For given σ_x^2, the optimum choice of $\bar{n}$ and d that minimize V(b) for fixed total cost C can be obtained from the Cauchy-Schwartz inequality as follows. Write $V(b) = \Sigma V_i/n_i = \Sigma u_i^2$ where $V_1 = \sigma_\alpha^2\eta^2/\sigma_x^2$, $V_2 = \sigma_\delta^2/\sigma_x^2$, $V_3 = \sigma_e^2/\sigma_x^2$, $n_1 = a$, $n_2 = ad$, $n_3 = ad\bar{n}$, and write

$C = \Sigma c_i n_i = \Sigma w_i^2$, where $c_1 = c_a$, $c_2 = c_d$ and $c_3 = c$. Then the product VC is minimized when u_i/w_i = constant. This condition requires first that $(u_2/w_2) = (u_3/w_3)$, i.e. that

$$\frac{(\sigma_\delta/\sigma_x\sqrt{ad})}{\sqrt{c_d}\ \sqrt{a}} = \frac{(\sigma_e/\sigma_x\sqrt{(nad)})}{\sqrt{c}\ \sqrt{(nad)}}$$

so that $$\bar{n}_{opt} = (c_d/c)^{1/2}(\sigma_e/\sigma_\delta). \qquad (45)$$

The condition also requires that $(u_1/w_1) = (u_2/w_2)$, i.e. that

$$\frac{(\sigma_\alpha\eta/\sigma_x\sqrt{a})}{\sqrt{c_a}\ \sqrt{a}} = \frac{(\sigma_\delta/\sigma_x\sqrt{ad})}{\sqrt{c_d}\ \sqrt{ad}}$$

so that $$d_{opt} = (c_a/c_d)^{1/2}(\sigma_\delta/\eta\sigma_\alpha) \qquad (46)$$

The optimum values of $\bar{n}_{opt}$ and d_{opt} given by (45) and (46) may then be combined with the constraint of the total cost C to determine the value of a.

References

Anderson, R.L. and T.A. Bancroft (1952). Statistical Theory in Research. McGraw-Hill, New York.

Langdon, F.J. (1976). Noise nuisance caused by road traffic in residential areas. Parts I and II. Journal of Sound and Vibration, 47, no. 2, 243-282.

Kish, L. and M.R. Frankel (1970). Balanced repeated replication for standard errors. Journal of the American Statistical Association, 65, 1071-1094.

Kish, L. and M.R. Frankel (1974). Inference from complex samples. Journal of the Royal Statistical Society, B, 36, 1-37.

Snedecor, G.W. and W.G. Cochran (1980). Statistical Methods. 7th ed. Iowa State University Press, Ames, Iowa.

1. Report No
NASA CR-166117

2 Government Accession No

3 Recipient's Catalog No

4 Title and Subtitle
Estimating Regression Coefficients from Clustered Samples: Sampling Errors and Optimum Sample Allocation

5 Report Date
May 1983

6 Performing Organization Code

7 Author(s)
Graham Kalton

8 Performing Organization Report No

10 Work Unit No

9 Performing Organization Name and Address
Survey Research Center
Institute for Social Research
University of Michigan
Ann Arbor, Michigan 48106

11 Contract or Grant No
NAS1-16107

13 Type of Report and Period Covered
Contractor Report

12 Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Washington, DC 20546

14 Sponsoring Agency Code

15 Supplementary Notes
Langley Technical Monitor: Sherman A. Clevenson
Final Report

16 Abstract

A number of surveys have been conducted to study the relationship between the level of aircraft or traffic noise exposure experienced by people living in a particular area and their annoyance with it. These surveys generally employ a clustered sample design which affects the precision of the survey estimates. Regression analysis of annoyance on noise measures and other variables is often an important component of the survey analysis. This report provides formulae for estimating the standard errors of regression coefficients and ratio of regression coefficients that are applicable with a two- or three-stage clustered sample design. Using a simple cost function, it also determines the optimum allocation of the sample across the stages of the sample design for the estimation of a regression coefficient.

17. Key Words (Suggested by Author(s))
Regression Coefficients
Cluster Sample
Optimal Subsample Size
Noise Surveys

18 Distribution Statement
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	26	A03

* For sale by the National Technical Information Service, Springfield Virginia 22161

End of Document