A REGRESSION MODEL FOR THE DATA FROM TWO STAGE CLUSTER SAMPLES

Jai W. Choi, National Center for Health Statistics

INTRODUCTION

Large sample survey data such as those of the National Health Interview Survey were often collected by groups or clusters rather than by individual units to minimize the cost. Thus, such data consisted of the first-stage clusters, primary sampling units (PSU's), each including a number of second-stage clusters (segments), and each segment including a number of third-stage clusters (households). The final cluster (household) included a number of final units (persons).

More than one member of households are often interviewed. The information from the same household members is correlated for certain variables, for instance, race and hereditary diseases. The members in the segment are also correlated for such variables as household income in some neighborhoods. Since a PSU includes many heterogeneous people over wider areas, there would be little correlation among them for most variables.

This problem of correlation arises not only in nested survey data, but also in many other situations. When repeated measurements are taken by one investigator or from the same subject, the results are often correlated. The siblings, the animals in the same litter, or plants in the same plot may be correlated.

The estimate of regression studies for such data should differ from those estimates for simple random sample. The aim of this paper is to present an estimator of the regression coefficient from two-stage clustered samples, when the regression coefficient for the first-stage cluster varies from cluster to cluster and when the members in the respective cluster are correlated. The model can be formulated as follows.

Denote, u, for each unit of a finite population, U, containing N identifiable first-stage clusters, each first-stage cluster including B_i second-stage clusters, each second-stage cluster including M_{ij} units.

U includes N exclusive and exhaustive first-stage clusters, $U_1, \ldots, U_i, \ldots, U_N$, and U_i includes B_i second-stage clusters, $U_{i1}, \ldots, U_{ij}, \ldots, U_{iB_i}$,

with $U_{ij} = (u_{ij1}, \ldots, u_{ijM_{ij}})$.

The random variable Y_{ijk}, and the observation, x_{ijk}, on X, associated with the unit u_{ijk}, are assumed to have the linear relationship,

$(i = 1,\ldots,N;\ j = 1,\ldots,B_i;\ k = 1,\ldots,M_{ij})$

$$Y_{ijk} = \beta_i x_{ijk} + e_{ijk} \qquad (1.1)$$

for some vector $\boldsymbol{\beta}' = (\beta_1,\ldots,\beta_N)$, when the random deviations, e_{ijk}, satisfy

$$E(e_{ijk} / x_{ijk}) = 0 \text{ for all i, j and k,}$$

and $E(e_{ijk}\, e_{i'j'k'} / x_{ijk}\, x_{i'j'k'}) =$

$$\begin{cases} \sigma_i^2 ; & (i,j,k)=(i',j',k') \\ \rho_2\sigma_i^2 ; & (i,j)=(i'j'),\ k\neq k' \\ \rho_1\sigma_i^2 ; & i=i',\ j\neq j' \\ 0 ; & \text{otherwise} \end{cases} \qquad (1.2)$$

where ρ_2 is the common pairwise intracluster correlation in the second-stage cluster, while ρ_1 is the common pairwise intracluster correlation in the first-stage cluster, excluding those pairs when both members of the pair belong to the same second-stage cluster.

Pfeffermann and Nathan (1981) used a model:

$E(e_{ij} / x_{ij}) = 0$ and

$$E(e_{ij}\, e_{i'j'} / x_{ij}\, x_{i'j'}) = \begin{cases} \sigma_i^2 \text{ if } (i,j)=(i',j') \\ 0, \quad \text{otherwise.} \end{cases}$$

This model for a one-stage cluster sample is extended to the data from a two-stage cluster sample, when the correlations among the members in the cluster cannot be ignored.

The regression models for individual first-stage clusters may assume different vectors of coefficients. But the second-stage clusters in the first-stage cluster may assume the same vector.

Following Pfeffermann and Nathan (1981), we assume that the unknown coefficients, β_i's in the model (1.1) are uncorrelated random variables rather than fixed with the same expectation and variance, that is,

$$\beta_i = \beta + \nu_i \quad (i = 1, \ldots, N) \qquad (1.3)$$

where $E(\nu_i) = 0$, and $E(\nu_i\, \nu_{i'}) = \begin{cases} \delta^2;\ i = i' \\ 0;\ \text{otherwise} \end{cases}$

where β_i is the coefficient in the equation (1.1).

The aim of this paper is to estimate the linear combination,

$$\beta_W = \sum_i^N w_i \beta_i,$$

of the regression coefficients with known weights, w_i, on the basis of a sample, S_i, which includes units from only part of the clusters, U_i.

In a regression model, $Y_i = x_i\beta + e_i$, with $E(Y_i/x_i) = x_i\beta$ and $Var(Y_i/x_i) = \sigma_i^2 I$, it has been shown that the ordinary least squares estimator is the best linear unbiased estimate of the regression coefficient β and

$$\hat{\beta} = (x'x)^{-1} x'y = S_{12} / S_2^2 \qquad (1.3)$$

where $S_{12} = \sum_i x_i y_i$ and $S_2^2 = \sum_i x_i^2$.

In recent statistical literature, the regression model for data from non-simple random samples has been widely discussed. DeMets and Halperin (1977) estimated the maximum likelihood estimate of β, conditioning on a third variable, x_3 say, considered as a design variable, and the trivariate normal distribution of (x_1, x_2, x_3) is assumed.

Adjustment for the sample selection takes place through the design variable x_3, that is not the sample probabilities, nor get into the model as an independent variable, but x_3 provides the information on sample selection.

$$\hat{\beta}_{12} = \frac{S_{12} + \dfrac{S_{13} S_{23}}{S_3^2} \left(\dfrac{\hat{\sigma}_3^2}{S_3^2} - 1 \right)}{S_2^2 + \dfrac{S_{23}^2}{S_3^2} \left(\dfrac{\hat{\sigma}_3^2}{S_3^2} - 1 \right)} \qquad (1.4)$$

which reduces to the unconditional ordinary least square estimate, S_{12}/ S_2^2, when the sample value S_3^2 approaches the population variance $\hat{\sigma}_3^2$ for a large n.

Holt et al. (1980), Nathan and Holt (1980), and later Pfeffermann and Holmes (1985) further discussed such conditional methods. It is complicated to have the design features reflected through conditional treatment mainly because it is difficult to derive the conditional distributions except for some simple cases.

A design variable x_3 say, may be included as an independent variable in the regression model. The regression coefficients can be estimated by the standard method of least squares.

Under a more general model $E(y \mid x) = x\beta$ and $Var(y \mid x) = V$, where V is the diagonal matrix with σ_i^2 for $i = i'$ and $= 0$ otherwise, the weighted least squares estimator (WLSE) is given by

$$\hat{\beta} = (x' V^{-1} x)^{-1} x' V^{-1} y \qquad (1.6)$$

where the covariance matrix V could be constructed to include other conditions suitable to the data. For instance, Royall (1986) used

$$Var(Y_i) = \sigma^2 x_i,$$

where the variance is proportional to the size of response x_i.

The weighting could be done directly from the sample information. Holt et al. (1980) used the diagonal matrix D with inclusion probabilities on the diagonal, and obtained the probability weighted estimate of β, replacing V with D in (1.6) as

$$\hat{\beta} = (x'D^{-1}x)^{-1} x'D^{-1} y \qquad (1.7)$$

DuMouchel and Duncan (1983) repeatedly estimated OLSE of β_i, i = 1, ..., L for L strata and the weighted average over all strata is obtained by

$$\hat{\beta} = \textstyle\sum_i w_i \hat{\beta}_i \,/\, \sum_i w_i . \qquad (1.8)$$

where w_i's are the weights of sample strata.

DuMouchel and Duncan (1983), Konijn (1962) and Porter (1973) employed the design-based probabilities or sample sizes. Their estimation is similar to the total for finite population under design based approaches. Since they assumed no prior conditions on the coefficients, their estimators did not have optimal properties.

Pfeffermann and Nathan (1981), Sedransk (1977), Lindley and Smith (1972) and Scott and Smith (1969) formalized the models that defined the relationship between the coefficients.

Section 2 shows the weighted estimator of the regression coefficient and its variance. The weighted empirical estimator is shown in Section 3, substituting the unknown parameters with their empirical estimates.

2. ESTIMATION OF β_w BY EXTENDED LEAST SQUARES

Denote the sample by

$$S = \{ (i,j,k): i \epsilon S^*, j \epsilon S_i, k \epsilon S_{ij} \},$$

where S^* is a sample of n first-stage clusters, S_i is a sample of b_i second-stage clusters in the i-th first-stage cluster, and S_{ij} is a sample of m_{ij} persons in the j-th second-stage cluster which belongs to the i-th first-stage cluster. The sampling design is not specified here except for the fact that it is a multistage probability sample on the set of all possible samples. But for each first-stage cluster sample S_i, we should have a sufficiently large number of units for the estimation of β_i.

We follow Pfeffermann and Nathan (1981), and Duncan and Horn (1973), combining the definitions shown in (1.1)-(1.3), and write a single model for the two-stage clustered data as:

$$\mathbf{Y^o} = \mathbf{X^o} \boldsymbol{\beta^o} + \mathbf{e^o} \qquad (2.1)$$

where

$$\mathbf{X^o} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{0}_{m_1} & & \\ \vdots & & \vdots & \mathbf{0}_{m,N-n} & \mathbf{0}_m \\ \mathbf{0}_{m_n} & \cdots & \mathbf{x}_n & & \\ & & -\mathbf{I}_{N,N} & & \mathbf{1}_N \end{pmatrix}, \quad \boldsymbol{\beta^o} = \begin{pmatrix} \boldsymbol{\beta} \\ \beta \end{pmatrix}, \quad \mathbf{e^o} = \begin{pmatrix} \mathbf{e} \\ \boldsymbol{\nu} \end{pmatrix},$$

and $\mathbf{Y_o}' = (\mathbf{Y}_1', \ldots, \mathbf{Y}_n', \mathbf{0}_N')$, $\mathbf{Y}_i$ denotes the $\mathbf{Y}$ values for the i-th sample cluster S_i, that is,

$$\mathbf{Y}_i = (Y_{i11}, \ldots, Y_{ijk}, \ldots, Y_{ib_i m_{ib_i}}),$$

$\mathbf{x}_i$ denotes the X values for S_i, that is,

$$\mathbf{x}_i = (x_{i11}, \ldots\ldots, x_{ib_i m_{ib_i}}),$$

$\mathbf{0}_{m,N-n}$ is a zero matrix of order m x (N-n), $m = \sum_{i\epsilon S}\sum_{j\epsilon S_i} m_{ij}$; $\mathbf{I}_k$ and $\mathbf{1}_k$ denote an identity matrix and a units vector of order k.

$\mathbf{0}_k'$ denotes a zero vector of order k;

$\mathbf{e}' = (\mathbf{e}_1', \ldots, \mathbf{e}_n')$, where $\mathbf{e}_i'$ denotes the vector of random errors, corrresponding to the sampled units in S_i as shown in (1.2), and $\boldsymbol{\nu}' = (\nu_1, \ldots, \nu_N)$.

Assume $E(e_{ijk},\nu_i) = 0$ for all i, j, and k.

From the model (1.2) and this assumption, we have $\mathbf{E(e^\circ) = 0}_{m+N}$ and $\mathbf{E(e^\circ,e^{\circ\prime})} = \mathrm{diag}(\sigma_i^2\,\mathbf{V}_{m_1},\ldots,\sigma_i^2\,\mathbf{V}_{m_n},\delta^2\,\mathbf{1}_N') = \mathbf{V}$,

where
$$\mathbf{V}_{m_i} = \sigma_i^2\begin{pmatrix} 1 & \rho_2 & \cdot\cdot & \rho_1 & \rho_1 & \cdot\cdot \\ \rho_2 & 1 & \cdot\cdot & \rho_1 & \rho_1 & \cdot\cdot \\ \vdots & \vdots & & \vdots & \vdots & \\ \rho_1 & \rho_1 & \cdot\cdot & 1 & \rho_2 & \cdot\cdot \\ \rho_1 & \rho_1 & \cdot\cdot & \rho_2 & 1 & \cdot\cdot \\ & \cdots & & & \cdots & \end{pmatrix},$$
and $m_i = \sum_{j\in S_i} m_{ij}$.

The error structures depend on common intracluster correlations, cluster sizes, and vaiance σ_i^2.

The extended least squares estimator of $\boldsymbol{\beta}^\circ$:

$$\hat{\boldsymbol{\beta}}^\circ(\mathbf{a}) = (\mathbf{X}^{\circ\prime}\mathbf{V}^{-1}\mathbf{X}^\circ)^{-1}\,\mathbf{X}^{\circ\prime}\mathbf{V}^{-1}\mathbf{Y}^\circ \qquad (2.2)$$

It is tedious to obtain the closed form of this general solution. We show the closed form of (2.2) for j = 1,2 (i = 1, ..., N; k = 1, ..., m_{ij}):

$$\hat\beta_i(a) = \hat\beta_i\,\lambda_i + (1-\lambda_i)\,\hat\beta(a), \qquad (2.3)$$

where $\hat\beta_i$ is the estimator reflecting the two-stage correlations and sampling as:

$$\hat\beta_i = \begin{cases} \dfrac{C_{1i}}{C_{2i}} & i \in S^* \\ 0; & \text{otherwise,} \end{cases} \qquad (2.4)$$

C_{1i} and C_{2i} are given by

$$C_{1i} = \sum_j\sum_k x_{ijk}y_{ijk} + \sum_j^2 g_{ij(j')}(\sum_k x_{ijk}\sum_k y_{ijk}) \qquad (2.5)$$
$$+ \sum_{j\neq j'}^2 c_{ij(j')}(\sum_k x_{ijk}\sum_{k'} y_{ij'k'}),$$

$$C_{2i} = \sum_j\sum_k x_{ijk}^2 + \sum_j^2 g_{ij(j')}(\sum_k x_{ijk})^2 \qquad (2.6)$$
$$+ \sum_{j\neq j'}^2 c_{ij(j')}(\sum_k x_{ijk}\sum_{k'} x_{ij'k'}).$$

C_{2i} are the known numbers, and $g_{ij(j')}$ and $c_{ij(j')}$ ($c_{i1} = c_{i2}$) depended on the ρ_1, ρ_2 and the m_{ij}'s are given by

$$g_{ij(j')} = \frac{-\left(\rho_2\left(1+\rho_2(m_{ij'}-1)\right) - m_{ij'}\rho_1\right)}{\left(1+\rho_2(m_{ij'}-1)\right)\left(1+\rho_2(m_{ij}-1)\right) - m_{ij}m_{ij'}\rho_1^2}, \qquad (2.7)$$

$$c_{ij(j')} = \frac{-\rho_1(1-\rho_2)}{\left(1+\rho_2(m_{ij}-1)\right)\left(1+\rho_2(m_{ij'}-1)\right) - m_{ij}m_{ij'}\rho_1^2},$$

(note $c_{i1} = c_{i2}$ for j = 1,2) (2.8)

$$\hat\beta(a) = \frac{\sum_i \hat\beta_i\,\lambda_i}{\sum_i \lambda_i}, \text{ and } \lambda_i = \begin{cases} \dfrac{\delta^2}{\delta^2 + \dfrac{\sigma_i^2(1-\rho_2)}{C_{2i}}}; & i \in S^* \\ 0; & \text{otherwise.} \end{cases}$$

$\hat\beta_i(a)$ is the weighted average of $\hat\beta_i$ and of the estimator of common expectation, $\hat\beta(a)$, with weights λ_i and $1-\lambda_i$, respectively, while λ_i is the weight approaching one as the ratio $\sigma_i^2(1-\rho_2)/C_{2i} \to 0$.

If $\rho_2 = 1$, $\lambda_i = 1$ and for $\rho_2 = 0$, λ_i is the the ratio, the more weight is given to the $\hat\beta_i$ in the formula (2.3).

If $\rho_1 = 0$ or $\rho_2 = 1$, $c_{ij} = 0$, the third terms in C_{1i} and C_{2i} drop out, and only the first terms remain when $\rho_1 = \rho_2 = 0$.

When the units in the cluster are uncorrelated, $\rho_1 = \rho_2 = 0$, above results are the same as those given by Pfeffermann and Nathan (1981, equation 2.5).

If the vector $(w_1, w_2, \ldots, w_N)$ is denoted for the weights of the N individual clusters, following Pfeffermann and Nathan (1981), the weighted extended least squares estimator $\hat\beta$ (a) is defined as

$$\hat\beta_W(a) = \sum_{i=1}^N w_i\,\hat\beta_i(a). \qquad (2.9)$$

where w_i's are some cluster weights. For the simplicity of the argument, we use $\sum_i w_i = 1$.

When the ratios $t_i = \sigma_i^2(1-\rho_2)/C_{2i} \to 0$ (i=1,..,n) and δ^2 is bounded (or $\delta^2 \to \infty$ and the ratios t_i are bounded),

$$\lim_{t_i\to 0} \sum_i^N w_i\hat\beta_i(a) = \sum_{i=1}^n w_i\hat\beta_i + \sum_{i=n+1}^N w_i\left(\frac{\sum_i^n\hat\beta_i}{n}\right) \qquad (2.10)$$

When $\delta^2 \to 0$, and the ratios are bounded, the β_i's tend to be equal. The expected value and variance of $\hat\beta_i$ given by

$E(\hat\beta_i) = \beta$, hence $E(\hat\beta_W(a)) = \beta$, and

$$\mathrm{Var}(\hat\beta_i) = \delta^2 + \frac{\sigma_i^2\;(A_{1i} + \rho_2 A_{2i} + \rho_1 A_{3i})}{C_{2i}^2} \qquad (2.11)$$

and similarly for $\mathrm{Var}(\hat\beta_W(a))$, where C_{2i} is given in (2.4) and, $x_{ij} = \sum_k^{m_{ij}} x_{ijk}$,

$$A_{1i} = \sum_j\sum_k x_{ijk}^2 + m_{i1}(g_{i1}x_{i1} + c_{i1}x_{i2})^2$$
$$+ m_{i2}(g_{i2}x_{i2} + c_{i2}x_{i1})^2$$

$$+ 2g_{i1}x_{i1}^2 + 2g_{i2}x_{i2}^2 + 4c_{i1}x_{i1}x_{i2}, \quad (2.12)$$

(for j = 1, 2),

$$A_{2i} = \sum_j\sum_{k\neq k'} x_{ijk}x_{ijk'}$$
$$+ m_{i1}(m_{i1} - 1)(g_{i1}x_{i1} + c_{i2}x_{i2})^2$$
$$+ m_{i2}(m_{i2} - 1)(g_{i2}x_{i2} + c_{i1}x_{i1})^2$$
$$+ 2\sum_j g_{ij}x_{ij}^2(m_{ij} - 1)$$
$$+ 2c_{i1}x_{i1}x_{i2}(m_{i1} + m_{i2} - 2), \quad (2.13)$$

$$A_{3i} = 2x_1 x_2 + 2g_{i1}x_{i1}m_{i1} \quad g_{12}x_{i2}m_{i2}$$
$$+ 2c_{i1}^2 x_{i1}x_{i2} m_{i1}m_{i2}$$
$$+ \sum_{j\neq j'}^2 x_{i1}(x_{i2}g_{i2}m_{i2}) + 2c_{i1}(m_{i2}x_{i1}^2 + m_{i1}x_{i2}^2)$$
$$+ 2c_{i1}m_{i1}m_{i2}(g_{i1}x_{i1}^2 + g_{i2}x_{i2}^2). \quad (2.14)$$

3. ESTIMATION OF PARAMETERS

The unknown parameters in the formuala (2.7) may be estimated from the sample information. The proposed estimators of σ_i^2, δ^2, ρ_1, and ρ_2 are as follows: when e_{ijk} are independently distributed, the unbiased and consistent estimator of σ_i^2 (Pfeffermann and Nathan, 1981) is

$$\hat{\sigma}_i^2 = \frac{1}{m_i - 1}\sum_{j\varepsilon S_i}\sum_{k\varepsilon S_{ij}}(Y_{ijk} - \hat{\beta}_i x_{ijk})^2; \quad (3.1)$$

Since e_{ijk} are not independent, but correlated in two-stages, it is a biased estimator.

Following Pfeffermann and Nathan (1981), we may use the squared deviation of $\hat{\beta}_i$ about the average $\hat{\hat{\beta}}$, weighted by $\hat{\lambda}_i$, for the estimation of δ^2 by

$$\delta^2 = \frac{1}{n - 1}\sum_{i\varepsilon S}\hat{\lambda}_i(\hat{\beta}_i - \hat{\beta})^2 \quad (3.2)$$

where $\hat{\hat{\beta}} = \frac{\sum_t \hat{\lambda}_t \hat{\beta}_t}{\sum_t \hat{\lambda}_t}$ and $\hat{\lambda}_i = \frac{\delta^2}{\delta^2 + \frac{\hat{\sigma}_i^2}{C_{2i}}}$,

C_{2i} is the known number given in (2.4), (3.2) is the usual ordinary least square estimator of the expansion model $\beta_i = \beta + e_i$ with $E(e_i) = 0$, $E(e_i\, e_j) = \delta^2/\lambda_i$; i = j and = 0; otherwise.

Dividing both sides of (3.2) by δ^2, we can obtain the estimator $\hat{\delta}^2$ from the solution of this result. There exists a unique solution. We may take the positive and nontrivial solution and zero otherwise.

The usual definition of the intracluster correlation may be used for the continuous variables as

$$\rho_1 = \frac{\sum_i\sum_{j\neq j'}\sum_{k,k'}\frac{(y_{ijk} - x_{ijk}\hat{\beta}_i)(y_{ij'k'} - x_{ij'k'}\hat{\beta}_i)}{H - G}}{(\mathrm{Var}(y_{ijk})\,\mathrm{Var}(y_{ij'k'}))^{\frac{1}{2}}} \quad (3.3)$$

$$\rho_2 = \frac{\sum_i\sum_j\sum_{k\neq k'}\frac{(y_{ijk} - x_{ijk}\hat{\beta}_i)(y_{ijk'} - x_{ijk'}\hat{\beta}_i)}{G}}{(\mathrm{Var}(y_{ijk})\,\mathrm{Var}(y_{ij\,k'}))^{\frac{1}{2}}} \quad (3.4)$$

and $\mathrm{Var}(y_{ijk}) = \sum_i\sum_j\sum_k\frac{(y_{ijk} - x_{ijk}\hat{\beta}_i)^2}{m - 1}$.

where $\hat{\beta}_i$ depented on the ρ_1 and ρ_2, and G, H, m, and m_i are:

$$m_i = \sum_j^{b_i} m_{ij};\ m = \sum_i^n m_i;\ G = \sum_i^n\sum_j^{b_i} m_{ij}(m_{ij} - 1);$$
$$H = \sum_i^n m_i(m_i - 1).$$

The open solution for ρ_1, fixing ρ_2, may be obtained by iteration. Define the function

$$\hat{\rho}_{1,i+1} = f(\hat{\rho}_{1,i}, \rho_2)$$

for the iteration.

The initial value, $\hat{\rho}_{1,0}$, given ρ_2, may be obtained form the sample information in order to start the iteration. The final value $\hat{\rho}_1$ can be obtained when two consecutive estimators differ less than 0.005, and simliary for $\hat{\rho}_2$.

Using the estimators $\hat{\sigma}_i^2$, $\hat{\delta}^2$, $\hat{\rho}_2$, and $\hat{\rho}_1$ in $\hat{\beta}_w(a)$ and denoting the empirical estimator of β_i by $\hat{\beta}_i(e)$ the empirical estimator of β_w is defined to be

$$\hat{\beta}_w(e) = \sum_i^N w_i\hat{\beta}_i(e) \quad (3.5)$$

The behavior of this estimator is to be investigated by simulation. The results from the previous sections can be used to generate the data of the varying intracluster corrlations in the model (1.2), which gives the mean and variance:

$E(Y_{ijk}) = X_{ijk}\beta_i$, and

$$Cov(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} \delta^2 x_{ijk} x_{ijk} + \sigma_i^2; \\ \delta^2 x_{ijk} x_{ij'k'} + \sigma_i^2 \rho_1; \; j \neq j' \\ \delta^2 x_{ijk} x_{ijk'} + \sigma_i^2 \rho_2; \; k \neq k' \\ 0; \text{ otherwise.} \end{cases}$$

ν_i and e_{ijk} in the model (1.2) may be considered as normal $N(0,\delta^2)$ and $N(0, \sigma_i^2)$, respectively.

When all the final sample units are independent, the result is equivalent to that presented by Pfeffermann and Nathan (1981).

The four sets of intracluster correlations, shown below, may be used to generate the four sets of data for regression models.

	1st stage ρ_1	2nd stage ρ_2
1	0.500	1.000
2	0.250	0.500
3	0.125	0.250
4	0.000	0.000

The estimation of regression coefficients from these four sets of correlation can be found and the results may be available soon.

REFERENCES

Choi, J. W., and Casady, J. R. (1982). "χ^2 - Testing of Categorical Data from Nested Design Using the Correction Factor Estimated from Analysis of Variance Components," American Statistical Association 1982 Proceedings of the Section on Survey Research Methods, 531-536.

DeMets, D., and Halperin, M. (1977). "Estimation of a Sample Regression Coefficient in Samples Arising from a Sub-Sampling Procedure," Biometrics, 33, 47-56.

DuMouchel, W. H., and Duncan, G. J. (1983). " Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," Journal of the American Statistical Association, 76, 535-543.

Duncan, D. B., and Horn, S. D. (1972). "Linear Dynamic Recursive Estimation From the Viewpoint of Regression Analysis," Journal of the American Statistical Association, 67, 815-822.

Holt, D., and Smith, T.M.F., and Winter, P. D. (1980). "Regression Analysis of Data from Complex Surveys," Journal of the Royal Statistical Society, Ser. A, 143, 476-487.

Kempthorne, W. J., and Koch, G. G. (1983). "A General Approach for the Analysis of Attribute Data from a Two-stage Nested Design: One and Two Treatments per Cluster," Contribution to Statistics: Essays in Honour of Norman L. Johnson, (P.K. Sen, Editor), North-Holland Publishing Co., 259-281.

Konijn, H. (1962). "Regression Analysis in Sample Surveys," Journal of the American Statistical Association, 57, 590-605.

Landis, J. R., and Koch, G. G. (1977). "A One-way Components of Variance Model for Categorical Data," Biometrics, 33, 671-679.

Lindley, D. V., and Smith, A. F. M. (1972). "Bayes Estimates for the Linear Model," Journal of the Royal Statistical Society, Ser. B, 34, 1-18.

Nathan, G., and Holt, D. (1980). "The effect of Survey Design on Regression Analysis," Journal of the Royal Statistical Society, Ser. B, 47, 377-386.

Pfeffermann, D. and Holmes D. J. (1985). "Robustness Cosiderations in the Choice of a Method of Inference for Regression Analysis of Survey Data," Journal of the Royal Statistical Society, Ser. A, 148, 268-278.

Porter, R. M. (1973). "On the Use of Survey Sample Weights in the Linear Model," Annals of Economic and Social Measurement, 2, 141-158.

Royall, R. M. (1983). "The Prediction Approach to Robust Variance Estimation in Two-stage cluster Sampling," Paper 553, Department of Biostatistics, Johns Hopkins University.

Sedransk, J. (1977). "Sample Problems in the Estimation of the Money Supply," Journal of the American Statistical Association, 72, 516-522.

Scott, A., and Smith, T. M. F. (1969). "Estimation in Multistage Surveys," Journal of the American Statistical Association, 64, 830-840.

Zeger, S. L., and Liang, K. Y. (1985). "Longitudinal Data Analysis for Discrete and Continuous Outcomes," Paper 568, Department of Biostatistics, Johns Hopkins University.

ACKNOWLEDGEMENT
Author appreciates the careful review and helpful comments from Dr. Van Parsons, NCHS, and Mr. Samuel Marcus.