THE ANALYSIS OF SURVEY DATA USING STOCHASTIC REGRESSION COEFFICIENTS WITH APPLICATION TO NHANES DATA

Danny Pfeffermann, Hebrew University and Lisa LaVange, Research Triangle Institute

1. INTRODUCTION

This paper discusses the use of stochastic regression coefficients (SRC) models for analyzing survey data. These models permit different regression coefficients in different small groups of the population, an important advantage when analyzing large heterogeneous populations. The model groups can often be expected to overlap with the design clusters which by their nature are homogeneous with respect to economic and socio-demographic characteristics. The variation of the group regression coefficients can then be modelled as a function of known characteristics of the design groups. This has an important impact on the inference process and allows the problem of design selection bias to be dealt with more easily. The special features of the design require some alterations and modification to classical inference procedures but also offer some additional inference methods which employ the known selection probabilities.

The analysis of survey data using SRC models is currently being investigated at the Research Triangle Institute (RTI) under an ongoing research project with the National Institute of Child Health and Human Development (NICHD). The data base used for the analysis consists of data from the National Health and Nutrition Examination Survey II (NHANES II), conducted between 1976-1980, complemented with 1980 U.S. Census data.

The sampling design used for the NHANES II is discussed in the next section. This design is typical of many other designs used in large scale surveys and motivates some of the analysis procedures discussed in subsequent sections. Section 3 deals with the estimation of the regression coefficients and their expectations, and Section 4 deals with the estimation of the unknown variances. Empirical results of the NHANES analysis are presented in Section 5 to illustrate some of the procedures discussed in previous sections.

2. THE NHANES DESIGN

A detailed description of the NHANES is given in McDowell et al. (1981). The NHANES consists of a stratified multistage probability cluster sample of households in the U.S. At the first stage, primary sampling units (PSUs) that coincide with counties or groups of contiguous counties and selected for the National Health Interview Survey were stratified into 64 "superstrata" based on size, income and racial distribution. One PSU was selected from each stratum with probability proportional to size (PPS). In order to oversample persons with low incomes, Enumeration Districts (EDs) within the selected PSUs were sorted into poverty and non-poverty strata. EDs were then selected separately within each stratum with PPS. Households within EDs were clustered into segments of eight adjacent addresses, and a systematic sample of segments was selected across all the EDs with no more than one segment per ED. At the final stage, persons were selected one per household, roughly, with young and old age groups being oversampled.

3. MODEL AND ESTIMATORS

In what follows we refer to the population groups with different vectors of coefficients as EDs because these are the groups used in the empirical study. Let Y_{ij} be the value of the dependent variable for unit j of ED i and $\underset{\sim}{X}_{ij}$ the corresponding values of (k+1) independent variables so that $x_{ij1} \equiv 1$. The vector values of Y and the design matrix X observed for ED i will be denoted by $\underset{\sim}{Y}_i$ and $\underset{\sim}{X}_i$. The orders of $\underset{\sim}{Y}_i$ and X_i are $(m_i \times 1)$ and $(m_i \times (k+1))$. It is not assumed that $m_i > k+1$ and often $m_i = 1$.

3.1 The Model

$$\underset{\sim}{Y}_i = X_i \underset{\sim}{\beta}_i + \underset{\sim}{\varepsilon}_i \; ; \; E\left(\underset{\sim}{\varepsilon}_i\right) = \underset{\sim}{0} \; , \; E\left(\underset{\sim}{\varepsilon}_i \underset{\sim}{\varepsilon}_i^T\right) = \sigma^2 I \quad (3.1)$$

$$\underset{\sim}{\beta}_i = A_i \underset{\sim}{\gamma} + \underset{\sim}{\eta}_i \; ; \; E\left(\underset{\sim}{\eta}_i\right) = \underset{\sim}{0} \; , \; E\left(\underset{\sim}{\eta}_i \underset{\sim}{\eta}_i^T\right) = \Delta \quad (3.2)$$

where $\Delta = \mathrm{diag}\left(\delta_0^2, \delta_1^2, \ldots \delta_k^2\right)$. It is assumed also that $E\left(\underset{\sim}{\varepsilon}_i \underset{\sim}{\eta}_i^T\right) = 0$ and that residuals pertaining to different EDs are independent. The matrix A_i is of order $[(k+1) \times (k+1)(p+1)]$ and has the form $A_i = I \otimes \underset{\sim}{a}_i$ where $\otimes$ denotes the Kronecker product and $\underset{\sim}{a}_i$ represents the ED characteristics with $a_{i1} \equiv 1$. (In the NHANES example, typical ED characteristics are income and education levels, race composition, etc. See Section 5 for the exact definitions used in the analysis).

The vector $\gamma^T = \left(\gamma_0^T, \gamma_1^T \ldots \gamma_k^T\right)$ is of order $\left(k+1\right)\left(p+1\right)$ so that different vectors of coefficients γ_j are assumed for different components β_{ij} of the vectors $\underset{\sim}{\beta}_i$. In a special case $A_i = I$ implying that the coefficients β_{ij} , i=1,2,... can be considered as random drawings from some common distribution with mean γ_j and variance δ_j^2.

Inserting (3.2) into (3.1) yields the following model for the sample observations:

$$\left.\begin{aligned} &\underset{\sim}{Y}_i = X_i A_i \gamma + X_i \underset{\sim}{\eta}_i + \underset{\sim}{\varepsilon}_i = X_i^* \gamma + \mu_i \\ &E\left(\mu_i\right)=0; \; E\left(\mu_i \mu_i^T\right)=\sigma^2 I + X_i \Delta X_i^T; \; E\left(\mu_i \mu_k^T\right)= 0 \; i \neq k \end{aligned}\right\} \quad (3.3)$$

Notice that the l-th row of the matrix $X_i^*, \left(l=1,\ldots,m_i\right)$ is

$$\underset{\sim}{x}_{il}^{*T} = \left[\underset{\sim}{a}_i^T, x_{ill}\,\underset{\sim}{a}_i^T, \ldots, x_{ilk}\,\underset{\sim}{a}_i^T\right] \quad (3.4)$$

Thus, the model defined by (3.3) can be considered as an extension of the model defined by (3.1) obtained by adding the regressors included in A and the first-order (multiplicative) interactions between the regressors in X_i and A. Furthermore, in many applications the regressors in A are of similar nature to those in X_i so that to some extent (3.3) represents a polynomial regression equation. The model for a sample of n EDs can be written compactly as

$$\underset{\sim}{Y} = X^* \gamma + \mu \; ; \; E(\mu) = 0 \; , \; E(\mu \mu^T) = V \quad (3.5)$$

where $\underset{\sim}{Y}^T = (\underset{\sim}{Y}_1^T \ldots \underset{\sim}{Y}_n^T)$, $X^{*T} = [X_1^{*T} \ldots X_n^{*T}]$, $\mu^T = (\mu_1^T \ldots \mu_n^T)$ and V is block diagonal with $V_{ii} = E(\mu_i \mu_i^T)$.

3.2 Estimation of γ

The matrix X^* would usually be of full rank $P^* = (P+1)(k+1)$. For known variances σ^2 and δ_j^2, the best linear unbiased (BLUE) generalized least squares (GLS) estimator of γ under the model is

$$\hat{\gamma}_{GLS} = (X^{*T} V^{-1} X^*)^{-1} X^{*T} V^{-1} \underset{\sim}{Y} \quad (3.6)$$

In practical applications the variances are unknown and have to be estimated from the sample. Inserting these sample estimators into (3.6) yields the empirical estimators $\hat{\gamma}_E$. Assuming that the variance estimators are consistent, the asymptotic distribution of $\hat{\gamma}_E$ is under general conditions the same as that of $\hat{\gamma}_{GLS}$, Anderson (1973).

The OLS estimator of γ is

$$\hat{\gamma}_{OLS} = (X^{*T} X^*)^{-1} X^{*T} \underset{\sim}{Y} \quad (3.7)$$

Under the model $\hat{\gamma}_{OLS}$ is unbiased and under mild conditions it is consistent.

The performance of $\hat{\gamma}_{BLU}$ and $\hat{\gamma}_{OLS}$ depends on the noninformativeness of the design. A design is noninformative if the selection of the sample is independent of the model regression residuals. Under an informative design, the model holding for the sample data is different from the model holding in the population, and in such cases estimators like $\hat{\gamma}_{GLS}$ and $\hat{\gamma}_{OLS}$ can be severely biased. This may happen in the NHANES case, for example, if not all of the design characteristics determining the selection of EDs are included in the vectors $\underset{\sim}{a}_i$, so that $f(\beta_i | \underset{\sim}{a}_i, i \epsilon S_E) \neq f(\beta_i | \underset{\sim}{a}_i)$ where S_E denotes the sample of EDs. See e.g., Holt Smith and Winter (1980) for discussion and references on the notion of informative designs.

A simple way to deal with the informativeness of the design is to weight every observation vector $(Y_{ij}, \underset{\sim}{x}_{ij})$ by the inverse of its selection probability. The weighted least squares (WLS) estimator can be written as

$$\hat{\gamma}_{WLS} = (X^{*T} W X^*)^{-1} X^{*T} W \underset{\sim}{Y} \quad (3.8)$$

where W is a diagonal matrix with the sampling weights on the main diagonal. The estimator $\hat{\gamma}_{WLS}$ is again unbiased and consistent under the model. However, its main property is that it is approximately design ("P") unbiased and consistent for the Census vector γ_{CEN} which in turn is model ("ξ") unbiased and consistent for γ. The census vector is the estimator that would have been obtained in case of a census. For a definition of consistency in finite population sampling, see e.g., Isaki and Fuller (1982).

It follows from the discussion above that $\hat{\gamma}_{WLS}$ is approximately unbiased and consistent for γ with respect to the $P\xi$ distribution. The variance of $\hat{\gamma}_{WLS}$ can be decomposed as

$$\begin{aligned} Var_{p\xi}(\hat{\gamma}_{WLS}) &= E_\xi[Var_p(\hat{\gamma}_{WLS})] + \\ &+ Var_\xi[E_p(\hat{\gamma}_{WLS})] = E_\xi[Var_p(\hat{\gamma}_{WLS})] + O(M^{-1}) \end{aligned} \quad (3.9)$$

where M denotes the population size. Eq. (3.9) implies that in practical situations where the sampling fractions m/M are very small, estimators of $Var_{p\xi}(\hat{\gamma}_{WLS})$ can be obtained by estimating $Var_p(\hat{\gamma}_{WLS})$. Such estimators are calculated for general multistage probability sampling designs by the commonly used software packages for regression analysis of survey data such as SURREGR (Holt 1977), OSIRIS (1979) and SUPERCARP (Hidiroglou et. al, 1980).

3.3 Prediction of β_i When Variances are Known

The use of SRC models requires the specification of the target vectors of coefficients. In general, these will be of the form $\beta_C = \sum_{i=1}^{N} C_i \beta_i$ with $\sum_{i=1}^{N} C_i = 1$ where "N" denotes the number of population EDs. Simple examples are (i) $C_i = 1/N$ (ii) $C_i = M_i/M$ where M_i is the size of ED i and (iii) $C_i = \frac{1}{T}$ for $i \epsilon U_T$, $C_i = 0$ otherwise, where U_T is a subgroup (domain) of T EDs defined by the ED characteristics. For known variances σ^2 and δ_j^2, j=0...k, the BLUE predictor of β_C under the model is

$$\hat{\beta}_C = \sum_{i=1}^{N} C_i \hat{\beta}_i \text{ where for } i \epsilon S_E,$$

$$\hat{\beta}_i = A_i \hat{\gamma}_{GLS} + \Delta X_i^T V_{ii}^{-1} (\underset{\sim}{Y}_i - X_i^* \hat{\gamma}_{GLS}) \quad (3.10)$$

and $\hat{\beta}_i = A_i \hat{\gamma}_{GLS}$ otherwise (Pfeffermann, 1984). The distributional properties of $\hat{\beta}_i$ are with respect to the joint ξ-distribution of the β_i's and the Y_{ij}'s given the selected sample. Notice from (3.10) that the optimal estimator of β_i for $i \notin S_E$ is the same as the optimal estimator of its expectation. For the sampled EDs these estimators are corrected by taking into account the deviations of the observations $\underset{\sim}{Y}_i$ for their expected means. In cases where the target vector β_C is a population average, these correction factors would usually have a minor effect on the estimator. Thus, an alternative predictor of β_C in such situations is

$\tilde{\beta}_C = \sum_{i=1}^{N} C_i A_i \hat{\gamma}$ where $\hat{\gamma}$ can be any one of the estimators discussed in the previous section.

The estimators considered so far assume that the weighted averages $C_A = \sum_{i=1}^{N} C_i A_i$ are known. When this is not the case, they can be estimated using the sampling weights, i.e.,

$$\hat{C}_A = \sum_{i \epsilon S_E} C_i A_i W_i \Big/ \sum_{i \epsilon S_E} W_i. \qquad (3.11)$$

The estimator $\hat{C}_A$ is an example of the use of the design in an essentially model based analysis.

4. ESTIMATION OF THE UNKNOWN VARIANCES

4.1 Maximum Likelihood Estimators

Estimation of the unknown variances σ^2 and δ_j^2, $j=0,\ldots,k$ is needed for three main reasons: (i) for calculating the empirical estimator $\hat{\gamma}_E$ and the correction factors defined by equation (3.10), (ii) for testing hypotheses regarding the variation of the vectors of coefficients, and (iii) for constructing confidence intervals for the unknown coefficients. Variance estimation in SRC models, or more generally, in variance component models has been a major area of research in recent years. Methods with known properties and feasible computational algorithms have been developed and they are widely discussed in the literature (see e.g., Searle, 1971, Harville, 1977, and the more recent article by Henderson, 1984).

In this section we discuss possible ways of deriving maximum likelihood estimators (m.l.e.) for the unknown variances. In particular we emphasize that m.l.e. can be obtained by repeated use of regression software routines without the need for further software development. This has some additional advantages which are outlined following the description of the computations.

The results presented below are borrowed from results on variance estimation in the 'classical' variance components model for which the V-C matrix of the residuals is linear in the unknown parameters. The general model defined by (3.1) and (3.2) falls under that category as can be seen by rewriting equations (3.3) in the form

$$\underset{\sim}{Y} = X^{*}\gamma + U_0\xi_0 + U_1\xi_1 + \ldots + U_k\xi_k + \xi \qquad (4.1)$$

where U_j is a block diagonal matrix of order $(m \times n)$ with the i-th block being the j-th column $X_{i(j)}$ of the design matrix X_i, and $\xi_j = (\eta_{1j}, \eta_{2j} \cdots \eta_{nj})$. The vector ξ_j consists of the deviations $\left(\beta_{1j} - \underset{\sim}{a}_1^T \gamma_j\right)$ of the regression coefficients in the various groups so that $\mathrm{Var}(\xi_j) = \delta_j^2 I_n$ and $E(\xi_i \xi_j^T) = 0$ for $i \neq j$. It follows that

$$\mathrm{Var}(\underset{\sim}{Y}) = \sum_{j=0}^{k} \delta_j^2 U_j U_j^T + \sigma^2 I_m = V \qquad (4.2)$$

The model defined by (4.1) and (4.2) is the classical mixed model of variance components. Assuming that the error terms ξ_j and ξ have a normal distribution, the likelihood equations are (Anderson, 1973),

$$\gamma = \left(X^{*T} V^{-1} X^{*}\right)^{-1} X^{*T} V^{-1} \underset{\sim}{Y} \qquad (4.3)$$

$$B\,\underset{\sim}{\delta}^{*} = \underset{\sim}{c} \qquad (4.4)$$

where

$$B = (B_{pq}),\ B_{pq} = \mathrm{trace}\left[V^{-1}\left(U_p U_p^T\right) V^{-1}\left(U_q U_q^T\right)\right]$$

$$\underset{\sim}{c} = (c_p),\ (c_p) = \left[\left(\underset{\sim}{Y} - X^{*}\gamma\right)^T V^{-1} U_p U_p^T V^{-1}\left(\underset{\sim}{Y} - X^{*}\gamma\right)\right]$$

$$\text{and } \underset{\sim}{\delta}^{*T} = \left(\delta_0^T \ldots \delta_k^T, \sigma^2\right).$$

The common statistical software packages such as SAS and BMDP do not include procedures for solving the likelihood equations (4.4) for the general case where the elements of the matrices X^{*} and U_j are different from zeroes and ones. A SAS procedure which puts out a SAS data set containing the coefficients for the system (4.4) for given (prior) values of the elements of V has been written at North Carolina State University (Giesbrecht, 1985). Iterating on this system of equations with newly obtained estimators for $\underset{\sim}{\delta}^{*}$ used as prior values for defining V in the next iteration yields m.l.e. of the unknown variances provided the system converges and no negative values are encountered.

An alternative procedure to obtain m.l.e. is to transfer the problem into a problem of generalized regression analysis. This is done by fitting a linear model to the squares and cross products of the residuals

$\mu_{ij} = \left(Y_{ij} - \underset{\sim}{x}_{ij}^{*T}\gamma\right)$, i.e.,

$$\left.\begin{aligned} E\left(\mu_{ij}^2\right) &= \delta_0^2 + x_{ij1}^2\delta_1^2 + \ldots + x_{ijk}^2\delta_k^2 + \sigma^2 = \\ &= \left(z_{ij}^T, 1\right)\underset{\sim}{\xi}^* \\ E\left(\mu_{ij}\mu_{il}\right) &= \delta_0^2 + x_{ij1}x_{il1}\delta_1^2 + \ldots \\ &\quad + \left(x_{ijk}\, x_{ilk}\right)\delta_k^2 = \left(z_{ij1}^T, 0\right)\underset{\sim}{\xi}^* \quad j\neq 1 \end{aligned}\right\} \quad (4.5)$$

$$E\left(\mu_{ij}\,\mu_{i*j*}\right) = 0 \quad \text{for } i\neq i*$$

$$\left.\begin{aligned} \mathrm{Var}\left(\mu_{ij}^2\right) &= 2\left(\underset{\sim}{x}_{ij}^T \Delta \underset{\sim}{x}_{ij}\right)^2 \\ \mathrm{Cov}\left(\mu_{ij}\mu_{il},\ \mu_{ip}\mu_{iq}\right) &= \left(\underset{\sim}{x}_{ij}^T\Delta\underset{\sim}{x}_{ip}\right)\left(\underset{\sim}{x}_{il}^T\Delta\underset{\sim}{x}_{iq}\right) \\ &\quad + \left(\underset{\sim}{x}_{ij}^T\Delta\underset{\sim}{x}_{iq}\right)\left(\underset{\sim}{x}_{il}^T\Delta\underset{\sim}{x}_{ip}\right) \end{aligned}\right\} \quad (4.6)$$

$$\mathrm{Cov}\left(\mu_{ij}\mu_{il}, \mu_{i*j*}\mu_{i*l*}\right) = 0 \quad \text{for } i\neq i*$$

Equations (4.5) follow directly from (3.3). Equations (4.6) translate equations (19) of Anderson (1973) to the present model. Let $\tilde{\mu}_i$ be the vector of squares and cross products of the residuals μ_{ij} corresponding to ED i and arranged in some convenient order, and define $\tilde{\mu}^T = \left(\tilde{\mu}_1^T \ldots \tilde{\mu}_n^T\right)$. Let Z_i be the matrix consisting of the rows $\left(z_{ij}^T, 1\right)$ and $\left(z_{ij1}^T, 0\right)$ arranged in the same order as the elements of $\tilde{\mu}_i$ and define $Z^T = \left(Z_1^T \ldots Z_n^T\right)$. It follows from (4.5) and (4.6) that $\tilde{\mu}_i = Z_i\underset{\sim}{\xi}^* + \underset{\sim}{r}_i$, $E\left(\underset{\sim}{r}_i\right) = \underset{\sim}{0}$, $E\left(\underset{\sim}{r}_i\underset{\sim}{r}_i^T\right) = \Lambda_i$ where Λ_i is defined by (4.6). The linear model holding for $\tilde{\mu}$ is

$$\tilde{\mu} = Z\underset{\sim}{\xi}^* + \underset{\sim}{r};\ E\left(\underset{\sim}{r}\right) = \underset{\sim}{0},\ E\left(\underset{\sim}{r}\underset{\sim}{r}^T\right) = \Lambda = \mathrm{diag}\left[\Lambda_1 \ldots \Lambda_n\right] \quad (4.7)$$

It follows from (4.7) that for a given, 'known' V-C matrix Λ, the BLUE estimators of the unknown variances are obtained as

$$\tilde{\underset{\sim}{\xi}}^* = \left(Z^T \Lambda^{-1} Z\right)^{-1} Z^T \Lambda^{-1} \tilde{\mu} \quad (4.8)$$

Notice from (4.5) that as long as some of the ED's contain more than one observation, δ_0^2 and σ^2 can be estimated separately.

Anderson (1973) has shown that m.l.e. of $\underset{\sim}{\xi}^*$ can be obtained by iterating between (4.3) and (4.8) with μ replaced by $\left(\underset{\sim}{Y} - X^*\hat{\underset{\sim}{\gamma}}\right)$. (See also Brown and Burgess, 1984 for further discussion and examples.) Thus, an alternative to solving the system of equations in (4.4) is to solve equation (4.8) with $\hat{\mu} = \left(\underset{\sim}{Y} - X^*\hat{\underset{\sim}{\gamma}}\right)$.

It is important to emphasize that the cross-products of residuals μ_{ij} and μ_{i*j*} pertaining to different ED's i and i* are not considered in the model (4.7). This is so because for $i\neq i*$

$$E(\mu_{ij}\mu_{i*j*}) = \mathrm{Cov}(\mu_{ij}\mu_{i*j*}, \mu_{pq}\mu_{pq*}) = 0 \quad (4.9)$$

Thus, the set of cross-products $\{\mu_{ij}\mu_{i*j*}, i\neq i*\}$ have zero expectations and they are uncorrelated with the squares and cross-products of residuals included in the vector $\tilde{\mu}$. Their inclusion in the model (4.7) has therefore no impact on the estimation of $\underset{\sim}{\xi}^*$.

The immediate implication of (4.9) is that for situations in which there are many ED's with only a few observations in each, employing the model (4.7) and iterating between (4.3) and (4.8) is a simple operation. Notice in this respect that

$$V^{-1} = \mathrm{diag}\left[V_1^{-1} \ldots V_n^{-1}\right],\ V_i^{-1} = \left(\sigma^2 I_{mi} + X_i \Delta X_i^T\right)^{-1}$$
$$= \frac{1}{\sigma^2}\left[I - X_i\left(X_i^T X_i + \sigma^2\Delta^{-1}\right)^{-1}X_i^T\right] \quad (4.10)$$

so that the computation of $\hat{\gamma}$ and $\hat{\underset{\sim}{\xi}}^*$ involves only the inversion of matrices of orders (k+1), (k+2) and (k+1)(p+1).

4.2 Estimation of Variances of Variance Estimators Under Informative and Noninformative Designs

Estimators for the model based variances of the variance estimators are readily obtained from (4.8) and are given by

$$\hat{\mathrm{Var}}\left(\hat{\underset{\sim}{\xi}}^*\right) = \left(Z^T \hat{\Lambda}^{-1} Z\right)^{-1} \quad (4.11)$$

where $\hat{\Lambda}$ denotes the estimator of Λ obtained in the last iteration of the procedure described above.

Design based estimators of the vatiances and covariances of $\hat{\underset{\sim}{\xi}}^*$ can be obtained using a survey regression software package such as SURREGR (Holt, 1977). The use of such estimators becomes essential when the design is informative, in which case $\hat{\underset{\sim}{\xi}}^*$ has to be replaced by the probability weighted estimator in each stage of the analysis.

Another important advantage of the regression method for deriving m.l.e. is that it permits a simple way of constraining the estimates to nonnegative values. An unconstrained solution of the likelihood equations can yield negative variance estimators. Instead of solving (4.8), one can solve the minimization problem

$$\min_{\underset{\sim}{\xi}^* \geq 0} \left[\left(\hat{\tilde{\mu}} - Z\underset{\sim}{\xi}^*\right)^T \Lambda^{-1} \left(\hat{\tilde{\mu}} - Z\underset{\sim}{\xi}^*\right)\right] \quad (4.12)$$

By iterating between (4.3) and (4.12) with the most recently obtained estimators of $\underset{\sim}{\xi}^*$ inserted

in V and Λ to obtain new estimators, the solution of (4.12) reduces to a problem of minimizing a quadratic function subject to linear inequality constraints for which there are computer programs available. Brown and Burgess (1984) discuss the use of constrained variance estimators and give empirical results which illustrate important features of the approach.

5. EMPIRICAL STUDY

5.1 Statement of Problem

The method of fitting stochastic regression coefficients models was applied to NHANES II data to assess the relationship between blood lead and blood pressure. Several articles appearing recently in the literature have dealt with this problem using data from both the NHANES I and NHANES II surveys. Pirkle, et al. (1985) found a significant relationship between blood lead and both systolic and diastolic blood pressure in white males aged 40-59 years while controlling for other dietary and anthropometic blood pressure correlates. Harlan, et al. (1985) also found blood lead to be a significant predictor of both systolic and diastolic blood pressure in men aged 12 to 74 years, but not in women. Shaper and Pocock (1985) used data from the British Regional Heart Study and in contrast to the other studies found no significant relationship between these two variables in middle-aged men.

Because the main goal of the empirical study was to investigate methodological issues as opposed to deriving the "best" explanatory model, it was decided to begin with previously published model equations, thereby taking advantage of the extensive variable selection procedures used to derive these models. Analysis was thus carried out on the subgroup consisting of all males aged 12-74 years using Harlan's model as a basis.

5.2 Model Definitions and Results

In order to fit SRC models to NHANES data, it was decided to define population subgroups according to EDs, (the design clusters selected at the second stage of the NHANES sample) for two reasons: (i) EDs are generally thought to be homogeneous clusters of households, and (ii) design information and selection probabilities, were available for each ED. This information had been merged from the 1970 U.S. Census tapes onto the NHANES II sample of household segments.

As explained above, the model for diastolic blood pressure derived by Harlan, et al. (1985) for males aged 12-74 years was used as a basis for fitting stochastic regression coefficients models to the NHANES data. This model was refit to the sample after deleting observations for which the ED affiliation was missing (n = 2676 versus n = 2818 in Harlan's analysis). The model results are presented in column 5 of Table 1. The coefficients were estimated by probability weighted least squares using RTI's survey regression software package SURREGR (Holt, 1977).

SRC models as defined by equations (3.1) and (3.2) were fit to the data. The vectors of ED specific regression coefficients were modelled as functions of the following ED characteristics: medium family income, proportion of non-whites, proportion of persons aged 25+ with less than 9 years of education, urban/rural living areas, and a poverty stratum indicator. The domain of analysis consisted of 1,892 EDs.

The significant γ coefficients of the models identified for the regression coefficients (equation 3.2) are presented in columns 2-4 of Table 1. As can be seen, a significant portion of the variation among the ED specific regression coefficients for age, age^2, race and particularly blood lead is explained by two ED characteristics; the proportion of non-whites and the proportion of persons with education of less than 9 years of school.

The last column of the table was calculated as

$$\hat{\beta}_C = \left[\sum_{i=1}^{n} M_i \hat{\beta}_i/\Pi_i\right] / \left[\sum_{i=1}^{n} M_i/\Pi_i\right] \text{ where } \{M_i\}$$

and $\{\Pi_i\}$ are the ED sizes and selection probabilities respectively. The entries of this column estimate the average regression relationship in the finite population and are found to be very similar to the coefficients obtained by Harlan's model presented in Table 1.

Summary

The results obtained at this stage suggest that whereas a unique regression equation can be assumed to assess the average effect of the regressors included in the original model, such a model may be inappropriate for studying the relationships in subdomains of the population. Obviously, a more extensive analysis is needed to support these early suggestions. We have started programming the estimation of the unknown variances using the procedures described in Section 4. At the same time, we are trying to locate some additional variables to be incorporated in the regression equation explaining the variation of the regression coefficients among the population EDs.

REFERENCES

Anderson, T.W. (1973). "Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure." The Annals of Statistics, Vol. 1, 135-141.

Brown, K.G. and Burgess, M.A. (1984). "On Maximum Likelihood and Restricted Maximum Likelihood Approaches to Estimation of Variance Components." J. Statist. Comput. Simul., Vol. 4, 1-18.

Giesbrecht, F. (1985). "MIXMOD, A SAS Procedure for Analyzing Mixed Models. Mimeograph Series No. 1659, North Carolina State University, Raleigh, NC.

Harlan, W.R., Landis, J.R., Schmouder,R.L., Goldstein, N.G. and Harlan, L.C. (1985). "Blood Lead and Blood Pressure, JAMA, 253, 530-534.

Harville, D.A. (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." J. Amer. Statist. Ass., 72, 320-340.

Henderson, C.R. (1984). "ANOVA, MIVQUE, REML and ML Algorithms for Estimation of Variances and Covariances." Proceedings of the 50th Anniversary Conference, Iowa State Statistical Laboratory, H. David and H.T. David, Editors.

Hidiroglou, M., Fuller, W.A. and Hickman, R.D. (1980). "SUPER CARP." Statistical Laboratory, Iowa State University, Ames, Iowa.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). "Regression Analysis of Data from Complex Surveys." J.R. Statist. Soc. A, 143, 474-487.

Holt, M.M. (1977). SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data." Research Triangle Institute, Research Triangle Park, NC.

Isaki, C.T. and Fuller, W.A. (1982). "Survey Design Under the Regression Superpopulation Model." J. Amer. Statist. Ass., 77, 89-96.

McDowell, A., Engel, A., Massey, J.T. and Maurer, K. (1981). "Plan and Operation of the Second National Health and Nutrition Examination Survey 1976-80." DHHS Publication No. (PHS) 81-1317, U.S. Department of Health and Human Services, National Center for Health Statistics, Hyattsville, MD.

OSIRIS (1984). Survey Research Center Computer Support Group, The Institute for Social Research, The University of Michigan, Ann Arbor, MI.

Pfeffermann, D. (1984). "On Extension of the Gauss-Markov Theorem to the Case of Stochastic Regression Coefficients." J.R. Statist. Soc. B, 46, 139-148.

Pirkle, J.L., Schwartz, J., Landis, J.R. and Harlan, W.R. (1985). "The Relationship Between Blood Lead Levels and Blood Pressure and Its Cardiovascular Risk Implications." American Journal of Epidemiology, 121, 246-258.

Searle, S.R. (1971). "Topics in Variance Component Estimation." Biometrics 27, 1-76.

Shaper, A.G. and Pocock, S.J. (1985). "Blood Lead and Blood Pressure." British Medical Journal, 291, 1147-1149.

TABLE 1. MODELS IDENTIFIED FOR DISTINCT COEFFICIENTS, ESTIMATES OF REGRESS COEFFICIENTS

	Models Identified[1]			Est. of Coefficients	
		ED Characteristics			
β_j Coeff.	Constant	Prop. Non. W	Prop. Edu. < 9	Unique Reg Line[2]	Weighted Avg. of ED Coeff.[3]
			Coefficients		
Intercept	43.486		-17.871	37.407	38.492
Age	0.449	0.987		0.524	0.539
Age^2	- 0.004	-0.010		- 0.005	- 0.005
Body mass	0.960			0.949	0.960
Race			8.332	- 0.812	2.328
Race x Age	-	-	-	0.091	-
Lead (ln)		-7.241	6.298	1.378	1.095
Hemoglobin	1.207			1.258	1.207
Serum zinc (ln)	- 3.263			- 3.223	- 3.263

[1]Table shows significant coefficients at the 0.07 level in the equaion

$$E(\beta_{ik}) = \gamma_{0k} + a_{i1}\ \gamma_{1k} + \ldots + a_{ip}\ \gamma_{pk}$$

[2]Model published by Harlan et al. (1985).

[3]Coefficients estimated as $\hat{\underset{\sim}{\beta}}_i = (\sum_1^n M_i\ A_i\ \hat{\underset{\sim}{\gamma}}_{WLS}/\Pi_i)\ /\sum_1^n\ M_i/\Pi_i$