ORS Working Paper Series
Number 62

COVARIANCE ESTIMATES FOR REGRESSION PARAMETERS FROM COMPLEX SAMPLE DESIGNS:

Application of the Weighted Maximum Likelihood Estimator to Linear and Logistic Regression Analysis in Which Observations Might Not be Independent

BARRY V. BYE*
JANICE M. DYKACZ*
SALVATORE J. GALLICCHIO*

Office of Research and Statistics
Social Security Administration

September 1994

The authors' names appear in alphabetical order.

*Office of Research and Statistics, Social Security Administration,
Suite 209, 4301 Connecticut Avenue, NW., Washington, DC 20008

ABSTRACT

Statistical methods of variance estimation are presented in this paper for the analysis of survey data involving complex sample designs. With certain complex sample designs, estimation of the covariance matrices in linear and logistic regression is not straightforward. The design may be complex because of disproportionate sampling of strata, necessitating the use of weights, or because the observations are not independent, or possibly both. Examples are given from projects at the Social Security Administration, and computer programs written in SAS (Statistical Analysis System) are provided.

INTRODUCTION

The purpose of this Working Paper is to present statistical methods and provide Statistical Analysis System (SAS) computer programs to compute covariance matrices for linear and logistic regression coefficients estimated from observations from complex sample designs. These designs involve either stratified disproportionate sampling-- resulting in the use of case weights in estimation processes--or observations which are not independent, or both. As of this writing, there are several projects underway at the Social Security Administration (SSA) that may require regression analysis from complex samples.

The Office of Program Integrity Reviews (OPIR) has recently undertaken a project designed to monitor the quality of disability decisions made by Administrative Law Judges (ALJs) at the Hearing level. The sample of ALJ decisions for the ongoing analysis is stratified by decisional outcome (allow, deny) and by region and race, and samples of equal size are drawn from each of the 40 strata (10 regions x 2 race groups x 2 decisions), thus yielding sample sizes that are not proportionate to the corresponding population totals in the strata. As part of the analysis, OPIR plans to estimate logistic regression equations with the ALJ decision as the outcome variable. Because the sample has been stratified on the dependent variable (that is, the stratification is endogenous), the estimates of regression coefficients that do not take the sample design into account may be biased, even in large samples (Manski and McFadden (1981) and Manski and Lerman (1977)). For OPIR's sample design, biased estimates are almost assured because the sample is stratified on both dependent and independent variables. (The associations between dependent and independent variables are artifacts of the sample design when the population is stratified on both dependent and independent variables and disproportionate sampling is used.) In OPIR's sample, it may also be the case that observations on which the estimates of the coefficients are based are not independent. Dependencies may exist between observations from the same ALJ or the same hearing office.

A second project that requires special calculations for covariances of estimated regression coefficients involves the analysis of reported earnings of Supplemental Security Income (SSI) recipients. The purpose of this analysis is to identify the characteristics of SSI blind or disabled recipients who make work attempts after receipt of SSI benefits. The data available are cross-section time-series data for a 1-percent sample of SSI recipients. The outcome measures are derived from annual earnings data obtained from the Continuous Work History Sample (Smith, 1989). The planned analysis calls for the use of logistic regression, with a binary dependent variable indicating whether reported earnings are greater than 0 in a given year, and linear regression, with the logarithm of the earnings as the dependent variable for those person-year observations with earnings greater than 0. The regressors are individual characteristics taken from the Supplemental Security Record and data files containing information pertaining to the application for disability benefits. The annual observations for each individual will in general not be independent even when conditioned on regressors. Although the maximum of the classical likelihood function provides consistent estimates of the regression coefficients (Liang and Zeger, 1986), special computations are needed for the covariance matrix of regression coefficients (Bye and Riley, 1989).

A third project involving regression estimates from complex samples is the analysis of return to work of disabled beneficiaries from SSA's New Beneficiary Followup (NBF) survey that was conducted in 1992. Part of the NBF disabled sample came from the survivors of the original New Beneficiary Survey (NBS) panel that was constituted in 1982. The original panel was augmented by a second sample of beneficiaries who were part of the original NBS population, not selected originally, and had earnings posted to their SSA earnings records after entitlement. For those analyses where return to work is a dependent variable, the sampling seems to be endogenous and could result in biased estimates of regression coefficients if the sample design is not taken into account via the case weights. Dependence between observations is probably not significant (if it exists, it is attributable only to interviewer effects), but covariance estimators will be needed that take the sample design into account. Much of the planned analyses involves the estimation of hazard function models from weighted classical likelihood functions. Although this paper does not provide for the calculation of covariances for these types of models, the discussion of the consistency and normality of the maximum of the weighted likelihood function appears to provide justification for the use of case weights in the hazard function analysis. (Also see Hoem, 1985.)

Finally, linear and logistic regression analyses that are planned from the Survey of Income and Program Participation (SIPP) may find the calculations presented in this paper of some use (Nelson et al., 1985). Although it is not clear under what circumstances one would use SIPP case weights in such analyses (see the discussion of case weights, below), there may be significant dependence among observations within families or households.

The next two sections of this paper provide some background for the methodology presented and a brief discussion of the use of case weights in regression analyses. Then follow two sections that provide the formulas for the calculation of covariance matrices and the corresponding SAS computer programs. These are followed by a brief discussion of hypothesis testing in this analytic context. Finally, there is a conclusion section that outlines the work remaining to be done to complete the development of this methodology.

BACKGROUND

The estimation of models in the social and economic sciences usually proceeds under the assumption that the estimated model parameters are derived from observations that are self weighting (i.e., if the observations constitute a probability sample from some finite or hypothetically infinite population, then each observation had the same probability of selection) and stochastically independent. The data consist of n pairs of random variables $\{(y_k,x_k), k=1,...,n\}$ where y_k is an outcome variable that depends on x_k, a vector of regressors. (The x_k may not be random variables but fixed in the design and without measurement error.) The model specifies the relationship between outcome and regressors in some functional form up to a vector of unknown constants, β. The joint probability of the observations is taken to be the product of the probabilities of the individual observations: each observation is assumed to be drawn independently from some hypothetical distribution that depends on the unknown parameters, and each observation has the same probability of selection.

There are many types of data sets used in the social sciences, obtained from observational studies rather than designed experiments, where one or both of these sampling assumptions are violated. Household survey data are a prominent example where neither assumption may hold. Often the samples are not self-weighting, resulting in the use of case weights to compute unbiased finite population estimates. Also dependence among observations sometimes exists. If observations come from a sample of households and a number of interviews conducted in each household (comprising n observations overall), then observations of household members may be correlated because of shared characteristics that are not made explicit in the model specification. Interviewer effects may also result in associations among responses that would not be present otherwise.

Retrospective studies provide a second type of data collection design where unequal sample selection probabilities often arise. Most observational data collection plans are prospective, working forward from explanatory variables to responses. Prospective plans include those that use stratified samples where the stratification factors are limited to all or a subset of the explanatory variables (or the stratification variables happen to be of no consequence in the model). Most household survey samples with complex sample designs lead to prospective analyses even when the data collected are historical. Retrospective studies are different because they start with outcomes of the process under study and then work backwards to identify the precedent causes.

Retrospective sampling arises quite naturally in nonexperimental epidemiological studies (Farewell (1979), Prentice and Pyke (1979)). As an example, samples are chosen of disease and disease-free cases; measurements are taken on the possible causal factors; and comparisons are made on the differences in distributions of the causal factors between the groups. In these types of studies it is not unusual for the selection probabilities to be different for the outcome groups (albeit observations within group are taken independently): taking all or most of the group with the characteristic of interest (often a relatively small part of the population) and a small sample of the group without the characteristic. When samples are stratified and sampled disproportionately on the outcome (dependent) variable, model-based analyses must often take the complex sample design into account (Manski and Lerman (1977), Manski and McFadden (1981)). This is particularly true when the population has been stratified on both dependent and independent variables and sampled disproportionately to stratum size. When this is done, associations between dependent and independent variables that do not take the sample design into account are artifacts of the sample design itself.

One approach to the estimation of regression models when the sample design is complex is to incorporate the variation in selection probabilities through the case weights in the estimation of the model parameters. Manski and McFadden (1981) provide a number of ways of doing this for analyses with categorical dependent variables. One of their approaches is incorporated into the methodology discussed below.

A third type of social science data collection design–one that often results in lack of independence between self-weighted observations–is the pooled cross-section and time-series design. Pooled cross-section and time-series data consist of repeated observations at a number of points in time for each of a sample of individuals. The dependence among observations is often introduced into the model estimation as autocorrelation over time among disturbance terms (Tuma and Hannan, (1984), Dielman, (1983)). The correlations are thought to arise from omitted factors that are constant over time. When the outcome of interest is categorical–labor force participation, marital status, government program participation–the dependence among observations for a sample unit is often attributable to the arbitrary measurement periods (for example, calendar years) for a process that evolves continuously in time. When the primary focus of the analysis is in the relationship of outcome and regressor variables, the time dependence among the repeated measurements is not of interest and usually is not made explicit in the model specification. Liang and Zeger(1986) provide a number of ways of computing covariance matrices for estimated model parameters when observations are not independent. Bye and Riley(1989) applied one of their results to linear and logistic regression analysis. This approach will be combined with that of Manski and McFadden to provide covariance computations for case weighted analyses with dependent observations.

The methodology often used for the estimation of unknown model parameters for self-weighting samples with independent observations is the method of maximum likelihood. Maximum likelihood estimators (MLEs) have good statistical properties in large samples. For a large class of analysis problems, MLEs are consistent (that is, as the sample size becomes large, the estimate converges to the true value in some appropriate sense), asymptotically normally distributed, and have minimum variance over a wide class of estimators. When properly specified models are estimated in prospective studies with independent observations and where the stratification variables have been included in the set of regressors (or the outcomes are independent of the stratification variables given the other regressors), the classical maximum likelihood estimator is not influenced by the complexity of the sample design. Even the clustering often found in the sampling mechanism of large-scale national surveys can be ignored as long as there is no reason to suspect that the dependence arising in the sampling mechanism affects the stochastic aspects of the measurement process of the dependent variable.

On the other hand, Manski and Lerman (1977) have shown that the MLE is not consistent in most cases where disproportionate sampling on the outcome variable is used. One alternative estimator that they propose (also see Manski and McFadden(1981)), the Weighted Maximum Likelihood Estimator (WMLE), is obtained by introducing the case weights into the analysis as exponents on the individual terms of the classical likelihood function. They show that the WMLE is consistent and asymptotically normal; and they provide a consistent estimator for its covariance matrix.

When observations are not independent, Liang and Zeger (1986) have described sufficient conditions under which the classical MLE, assuming independence of observations, continues to provide consistent estimates that are asymptotically normally distributed. However, the covariance matrix of the estimated parameters is no longer consistently estimated by the inverse of the matrix of second order partial derivatives of the likelihood function. They provide an alternative estimator for the covariance matrix.

The purpose of this Working Paper is to describe an approach that combines the results in these two areas for certain sampling plans that exhibit either or both disproportionate sampling and lack of independence. The result is a WMLE with the covariance matrix of the estimated parameters adjusted for dependence among observations. The estimated covariance matrix can be obtained by closed-form calculations, and therefore, may be preferred over other variance estimation approaches that involve resampling, such as jackknife and balanced repeated replication (Wolter, 1985).

THE CASE WEIGHT DEBATE

The discussion presented here is based largely on presentations in Manski and McFadden (1981), DuMouchel and Duncan (1983), and Kott (1991).

Linear Regression Models

In the case of ordinary linear models estimated from (nonexperimental) data obtained from complex samples, there has been some disagreement between analysts interested primarily in the particular population sampled and those taking a more classical position to model estimation as to whether case weights, derived from the reciprocals of the probabilities of selection of the sample cases, should be used in estimating β coefficients. Actually two sets of coefficients are distinguished in these discussions. β is used to represent the coefficients of the classical regression model, $y=x\beta+\varepsilon$, where ε is a normally distributed disturbance with mean 0 (i.e. the model is correctly specified) and constant variance σ^2 (i.e. the error terms are homoskedastic). The maximum likelihood estimate of β is given by

$$\hat{\beta} = (X^TX)^{-1}X^TY. \qquad (1)$$

But there is another quantity β_w that is of interest to finite population samplers. It is referred to as the "census" coefficient and consists of the coefficients that would have been obtained if the regression had been estimated from the entire finite population from which the complex sample was drawn. One interpretation of β_w is that $x\beta_w$ provides the best linear predictor of y in the sense of minimizing mean squared error of prediction in the finite population sampled. β_w is estimated by,

$$\hat{\beta}_w = (X^T W^{-1} X)^{-1} X^T W^{-1} Y. \qquad (2)$$

W is a diagonal matrix with the case weights on the diagonal. If the error variances are heteroskedastic and the case weights are proportional to the variances, then (2) is the classical Generalized Least Squares (GLS) estimate of β.

When the model is properly specified and the sampling is not endogenous, there seems to be general agreement that (1) is preferred to (2). When there is an omitted predictor variable, then (2) may be preferable. If the omitted variable is uncorrelated with the other regressors in the population, but a correlation has been induced by the sample design, then (2) will provide an unbiased estimate of β, but (1) will not (Kott,1991). If there is an omitted variable and prediction is all that is desired, then (2) may be preferable to (1) (DuMouchel and Duncan, 1983). If the sample has been stratified on the dependent variable, y, then (2) may not only be preferred, but required. This last case is highly unusual in connection with "continuous" as opposed to categorical dependent variables.

Analysts who are searching for structure or interested in testing certain social science theories are, presumably, interested in β rather than β_w. A misspecified model can cause problems of interpretation. The bias in one or more estimated coefficients due to an omitted independent variable is a function of the correlations of the omitted variable with y and other relevant independent variables. Arguments concerning the impact of the unobserved variables on the estimated coefficients may or may not be altered by the sample design, again suggesting that (2) might be preferred in some cases.

Consider for example Census-type designs in which the stratification is essentially only geographic. If the correlational structures are independent of geography, then there would appear to be no advantage of (2) over (1). This would correspond in DuMouchel and Duncan (1983) to the case where the omitted variable is not related to the case weights. If omitted variables are related to the weights, DuMouchel and Duncan (1983) suggest that a comparison of weighted and unweighted regression results might help identify important omitted variables. It is not clear whether social scientists would want to approach model specification in this way.

In using (2), however, one should keep in mind that the variance of the weighted estimator is most likely larger than the unweighted estimator. If biases are small, (1) may provide estimates with smaller mean squared errors than (2).

Logistic Regression Models

A definitive exposition on the use of case weights in models with categorical outcomes appears to have been made by Manski and McFadden. If the sample design is exogenous--that is, not a function of the dependent variable--then design parameters factor from the likelihood function and do not influence MLEs; and therefore, case weights can be ignored. If there is stratification on the dependent variable only and sampling is disproportionate, then the sample design should generally be taken into account in obtaining unbiased estimates of model coefficients. Logistic regression may be one exception to this rule in that only the constant term is biased under stratification solely on the dependent variable.

Logistic regression parameters can be estimated without case weights (and the constant can be adjusted manually, if desired.)

If the population is stratified on both dependent and independent variables and sampling is disproportionate, then, case weights must be used in all analyses including logistic regression. These sample designs have induced artificial associations among dependent and independent variables.

The literature on the influence of the sample design in logistic regression analysis assumes that the model specification is correct. No discussion about the use of case weights when sampling is exogenous has been found. The absence of debate may be due to the fact that, because the MLEs for these analyses are nonlinear, none of the omitted variables arguments found in connection with linear models have analogs even in unweighted analyses. The omission of a variable that is associated with the outcome variable can bias the other coefficients even if the omitted variable is uncorrelated with the included variables. (See Bye and Dykacz (1987) for examples and the literature on unobserved mixture models (Folman and Lambert(1989)).

CORRECTED COVARIANCE MATRICES

The purpose of this section is to present the derivation of formulas that can be used to compute estimates of covariances from weighted data with dependence among the observations under the assumption that the estimator of the coefficients is consistent. The approach taken follows closely that of Bye and Riley (1989) after noting that the weighted density for linear and logistic regression models can be brought into exponential form; and therefore, the derivation of the equations for the covariance matrix can follow the approach used in that article. The development begins with a brief review of estimation by the method of maximum likelihood. Next, a class of exponential density functions is defined for which Liang and Zeger's results apply. Their results are then stated and applied to the weighted likelihood function for linear and logistic regression models. (Note that the proof of consistency and normality of the weighted estimator with dependence requires references, not yet obtained, to a strong law of large numbers and a central limit theorem for dependent observations.) Along the way, comparisons are made with equations for weighted data in Manski and Lerman (1977), Manski and McFadden (1981), and DuMouchel and Duncan (1983).

Maximum Likelihood Estimation

Assume that there is a sample of n independent observations (y_k,x_k) where y_k is the observed value of the dependent variable and x_k is a row vector of regressors of length p for the kth sample case. Let $f(y_k,x_k,\beta)$ be the density of (y_k,x_k) which is assumed to be known up to a column vector of unknown coefficients, β, also of length p. The joint density function, $L((y_1,x_1),\ldots,(y_n,x_n),\beta)$, of the sample of n observations, when regarded a function of β for a fixed sample, is referred to as the likelihood function of the sample. When the observations of the individual sample members are independent, the likelihood function is the product of the contributions of each sample member. That is,

$$L(\beta) = L((y_1,x_1),\ldots,(y_n,x_n),\beta) = \prod_{k=1}^{n} f(y_k,x_k,\beta) \qquad (3)$$

The value of β that maximizes (3), β^*, is found by first taking the natural logarithm of the likelihood function,

$$\ln L(\beta) = \sum_{k=1}^{n} \ln f(y_k,x_k,\beta) \qquad (4)$$

and then finding the β^* that satisfies the likelihood equations,

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{k=1}^{n} \left(\frac{1}{f(y_k,x_k,\beta)} \frac{\partial f(y_k,x_k,\beta)}{\partial \beta_j} \right) = 0, \; j = 1,\ldots,p \qquad (5)$$

The theory of the likelihood function asserts that under suitable regularity conditions for $f(y_k,x_k,\beta)$, the maximum, β^*, of ln L(β) is consistent, asymptotically normal, and has minimum variance over a wide class of estimators. The covariance matrix is given by the inverse of the Fisher information matrix,

$$V_{\beta^*} = \left[-\, nE \left(\frac{\partial^2 \ln f(y_k, x_k, \beta)}{\partial\beta_j \, \partial\beta_{j'}} \right) \right]^{-1} \qquad (6)$$

where j and j' index the elements of β. The matrix in the brackets is consistently estimated by the inverse of the negative of the matrix of second-order partial derivatives ln L(β) evaluated at β* (Serfling, 1980, p146). That is,

$$\hat{V}_{\beta^*} = \left[-\sum_{k=1}^{n} \frac{\partial^2 \ln f(y_k, x_k, \beta^*)}{\partial\beta \, \partial\beta'} \right]^{-1} \qquad (7)$$

The result in equation (6) is derived from a first order approximation of the likelihood equations,

$$0 = \left(\frac{\partial \ln L(\beta)}{\partial\beta} \right)_{\beta_n^*} \cong \left(\frac{\partial \ln L(\beta)}{\partial\beta} \right)_{\beta} + \left(\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'} \right)_{\beta} (\beta_n^* - \beta) \qquad (8)$$

Then

$$\sqrt{n}(\beta_n^{*}-\beta) \cong \left[-\frac{1}{n}\left(\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'}\right)_{\beta}\right]^{-1}\left[\frac{1}{\sqrt{n}}\left(\frac{\partial \ln L(\beta)}{\partial\beta}\right)_{\beta}\right] \tag{9}$$

and as n becomes large (and since β_n^* is a consistent estimate of β), the right hand side of (9) is asymptotically normal with mean 0 and covariance matrix

$$\Lambda^{-1}C\Lambda^{-1}, \ \textit{where } \Lambda = \left[E\left(\frac{\partial^2 f(y_k,x_k,\beta)}{\partial\beta\partial\beta'}\right)_{\beta}\right] \tag{10}$$

$$C = \left[E\left(\frac{\partial \ln f(y_k,x_k,\beta)}{\partial\beta}\right)_{\beta}\left(\frac{\partial \ln f(y_k,x_k,\beta)}{\partial\beta'}\right)_{\beta}^{T}\right] \tag{11}$$

and

C is the covariance matrix of the first order partial derivatives of $f(y_k,x_k,\beta)$ evaluated at β.

When $L(\beta)$ is in fact the likelihood function, then $\Lambda = -C$, whence equation (6). (Huber, 1967)

To look ahead, when the function maximized is not the likelihood function because the observations are not independent or case weights have been introduced or both, then (10) does not reduce to (6); and (10) is the general form of the covariance matrices of the regression coefficients.

Exponential Densities

If the density of y, $f(y_k, x_k, \beta)$, has a certain exponential form, Liang and Zeger (1986) show that the classical MLE is consistent and asymptotically normal even when observations are not independent. However, the covariance of the estimated parameters is no longer given by (6). As will be seen below, the advantage of restricting the densities to exponential form is that the C matrix in (10) is an explicit function of the covariances of the individual observations and, thus, provides a way of entering specific information about those dependencies into the calculation of the covariance matrix of the coefficients.

In order to present Liang and Zeger's results and then incorporate case weights into them, the notation used thus far must be expanded so that it permits the description of clusters of (possibly) dependent observations.

Let $Y_k=(y_{k1},...,y_{knk})$ **[note: nk should be n_k]** be an n_k x 1 row vector of outcome values and $X_k = (x_{k1},...,x_{knk})^T$ be the n_k x p matrix of values of the independent variables for the kth sample case. (The components of x_{ki} make up the ith row of X_k.) Let K be the number of clusters, and $n_1+...+n_K = n$ be the total sample size. This notation could describe any collection of groups or clusters of individual observations in which the first subscript denotes the cluster and the second subscript the members of the cluster. For example, k could denote household, and n_k could be the number of household members.

The marginal density function, $f(y_{kt,}x_{kt},\beta)$ has exponential form if it can be written as,

$$f(y_{kt},x_{kt},\beta) = \exp\left[[y_{kt}\theta_{kt} - a(\theta_{kt}) + b(y_{kt})]\phi\right] \qquad (12)$$

In this expression, $\theta_{kt} = h(\eta_{kt})$, where $\eta = x_{kt}\beta$; that is, θ_{kt} is a scaler function of the scaler variable $x_{kt}\beta$ and does not involve the values of the dependent variables y_{kt}. The function $a(\theta_{kt})$ is a scaler function of θ_{kt}; $b(y_{kt})$ is a scaler function of the observed values of the dependent variables that does not involve the unknown parameters β, and ϕ is a constant.

When the density takes the form in (12), the likelihood equations (5) can be written in matrix form as follows:

$$\phi\left(\frac{\partial \ln L(\beta)}{\partial\beta_1},\ldots,\frac{\partial \ln L(\beta)}{\partial\beta_p}\right)^T = \phi\sum_{k=1}^{K} X_k^T\Delta_k S_k = 0 \qquad (13)$$

where $\Delta_k = \mathrm{diag}\{h'(x_{kt}\beta)\}$ of order n_k x n_k and $S_k = ([y_{k1}\text{-}a'(\theta_{k1})],\ldots,[y_{knk}\text{-}a'(\theta_{knk})])^T$.

The covariance matrix (6) is given by:

$$V_{\beta^*} = \left[- n\phi \; E\left(a^{//}(\theta_k) \frac{\partial\theta_k}{\partial\beta_j} \; \frac{\partial\theta_k}{\partial\beta_{j/}} \right) \right]^{-1} \qquad \textbf{(14)}$$

which can be consistently estimated by

$$\hat{V}_{\beta^*} = \left(\phi \sum_{k=1}^{K} X_k^T \Delta_k A_k \Delta_k X_k \right)^{-1} \qquad \textbf{(15)}$$

where A_k = diag $\{a''(\theta_{kt})\}$ of order n_k x n_k;

When observations are not independent, (14) is not the correct covariance matrix. Liang and Zeger (1986, p.15) assert that the correct asymptotic covariance matrix of β is given by,

$$Z_{\beta^*} \left(\sum_{k=1}^{K} X_k^T \Delta_k cov(Y_k) \Delta_k X_k \right) Z_{\beta^*} \quad where Z_{\beta^*} = \phi V_{\beta^*} \qquad \textbf{(16)}$$

The correct covariance matrix can be consistently estimated by,

$$\hat{Z}_{\beta^*} \left(\sum_{k=1}^{K} X_k^T \hat{\Delta}_k \hat{S}_k \hat{S}_k^T \hat{\Delta}_k X_k \right) \hat{Z}_{\beta^*} \qquad (17)$$

substituting β^* for β where appropriate. Note that the estimation of ϕ is not necessary for (16) which is implicitly a function of ϕ through the matrix $S_k S_k^T$, the estimate of $\mathrm{cov}(Y_k)$.

Weighted Likelihood -- Consistency and Normality

If there is a case weight w_{kt} associated with each observation, the weighted likelihood function is defined to be,

$$WL(\beta) = \prod_{k=1}^{K} \prod_{t=1}^{n_k} f(y_{kt}, x_{kt}, \beta)^{w_{kt}} \qquad (18)$$

and the logarithm of the likelihood function is,

$$\ln WL(\beta) = \sum_{k=1}^{K} \sum_{t=1}^{n_k} w_{kt} \ln f(y_{kt}, x_{kt}, \beta) \qquad (19)$$

A proof of the consistency and normality of the maximum of (19) can be obtained by generalizing the argument present in Manski and Lerman (ManLer). (Manski and McFadden use the same argument but provide less details.) For this working paper, the consistency and normality arguments will be outlined, rather than formally developed with all of ManLer's machinery. In order to follow the outline, the reader must be fairly familiar with ManLer's setup.

The extension of the argument procedes as follows. First, the general definition of the stratified sampling process is extended to the sample space from which the observations (y,x) have been drawn. When y is not categorical, it is assumed that that dimension of the sample space is divided into a finite number of mutually exclusive parts (as one does with "continuous" x stratification variables). The case weights are (almost surely) the reciprocals of the selection probabilities under the stratified sampling process (ManLer, pp. 1980, 1982).

Now ManLer's consistency theorem (Theorem 1 (ManLer, p. 1983)) can be stated for any density f(y,x,β) (with f(y,x,β) replacing P(i,z,θ) in ManLer's notation, β and θ taken to be the true values of the parameters). The theorem identifies a set of sufficient conditions under which the maximum of the weighted likelihood function converges strongly to the true value. The conditions are quite general and would be true for many analyses associated with the family of exponential densities which will be the focus of the remainder of this paper. The conditions are: (1) the sample space from which the (y,x) were drawn is compact, (2) the space from which β is drawn is compact, (3) f(y,x,β) is continuous in all of its arguments, (4) selection probabilities for all strata in the sampling plan are non-zero, and (5) $E(w_{kt}\ln f(y_{kt},x_{kt},\beta))$ possess a unique maximum over the space that contains β, where the expectation is taken with respect to the stratified sampling process.

The consistency proof proceeds as follows.

(i) By the strong law of large numbers, the arithmetic average of the weighted densities from (19) converges to $E(w_{kt}\ln f(y_{kt},x_{kt},\beta))$.

(ii) The expectation in (i) involves the stratified selection probabilities and f evaluated at the true β. The selection probabilities (almost surely) eliminate the weights, algebraically (since they are reciprocals). Thus, the expectation in (i) is (almost surely) equal to $E(\ln f(y_{kt},x_{kt},\beta))$ where the expectation is taken with respect to f evaluated at the true β.

(iii) $E(\ln f(y_{kt},x_{kt}.\beta))$ attains its maximum at the true value of β (ManLer Lemma 1, p. 1982) Thus the weighted likelihood converges (almost surely) to a quantity which attains its maximum at the true value of β.

(iv) The maxima of the weighted likelihood functions are themselves measurable functions (ManLer Lemma 2, p. 1982 first part), and the weighted likelihood converges to its expected value uniformly (ManLer Lemma 3, p. 1983). Thus, the maximum of the weighted likelihoods converges to the maximum of the expected value, that is the true value of β (ManLer Lemma 2, p. 1982 second part).

This completes the consistency proof. Note that the argument holds for any density meeting the stated conditions with independent draws from a stratified sample design. In (ii) the complexity of the stratified sample design is canceled by the use of the weights, as one might expect intuitively.

The proof can be extended to sample designs where observations are not independent by invoking an appropriate strong law of large numbers in (i) above (but an appropriate reference has not yet been identified).

As suggested earlier, there may not be many analyses for which such an estimator ought to be used other than analyses that involve discrete dependent variables with endogenous sampling. (An exception might be the estimation of survival models from endogenous samples--Hoem, 1985). However this argument does establish the consistency of the weighted estimator for exogenous samples (for those analysts who continue to use such estimators) although classical theory suggests that the weighted estimator is not the best estimator, asymptotically.

Given the consistency of the estimator, the normality of the estimator follows from the central limit theorem or an appropriate version of such a theorem for dependent observations (reference not yet identified). Computation of consistent estimates of the covariance matrix of the maximum of the weighted likelihood is the subject of the remainder of this paper.

Weighted Likelihood -- Covariance Estimation

If the density is exponential, the likelihood equations (13) become,

$$\sum_{k=1}^{K} X_k^T W_k \Delta_k S_k = 0 \qquad (20)$$

and, assuming that the solutions to (20) provide a consistent estimator, the asymptotic covariance matrix of the estimated parameters is given by a matrix of the form,

$$D^{-1} E D^{-1} \qquad (21)$$

where,

$$D = \left(\sum_{k=1}^{K} X_k^T W_k \Delta_k A_k \Delta_k X_k \right) \qquad (22)$$

and

$$E = \left(\sum_{k=1}^{K} X_k^T W_k \Delta_k cov(Y_k) \Delta_k W_k X_k \right) \qquad \textbf{(23)}$$

where W_k = diag{w_{kt}} is an n_k x n_k diagonal matrix for the kth cluster with the n_k weights on the diagonal. The elements of the remaining matrices are defined as above.

Note that weighting the exponential likelihood preserves the property that the covariance matrix is an explicit function of cov(Y_k), thus permitting computations for analyses that involve both case weights and dependence.

Note that if $W_k = I_k$ for all k, then (22) and (23) are the same as (15) and (16); but if the weights are not equal to 1 uniformly and there are no covariances, i.e. cov(Y_k) is a diagonal matrix for all k, the matrix, $D^{-1}ED$ does not reduce to a simpler form. In particular, this matrix does not reduce to $(\phi D)^{-1}$ which would be the correct covariance matrix if the weights indicated repeated independent observations for the sample cases and not reciprocals of selection probabilities.

As above, the matrices in (22) and (23) can be consistently estimated by replacing β with β^* where appropriate and estimating cov(Y_k) by $S_kS_k^T$, as in (17).

Logistic and Linear Regression Analysis

The elements needed to produce estimates of the matrices in (22) and (23), involve only y_{kt}, w_{kt}, x_{kt}, $h'(x_{kt}\beta)$, $a'(\theta_{kt})$, and $a''(\theta_{kt})$ where $\theta_{kt} = h(x_{kt}\beta)$. Thus, if a specific likelihood, $f(y_{kt}, x_{kt}, \beta)$, can be written in the form of (12) and the functions a and h specified, it is not difficult to compute the matrices in (22) and (23).

Logistic Regression

The formulation of the estimator for the logistic regression model is developed as a special case of the general binary response model. For the binary response model, y_{kt} takes on the value 1 is the event of interest occurs and 0 otherwise. The likelihood of an individual observation has the form,

$$f(y_{kt}, x_{kt}, \beta) = P_{kt}^{y_{kt}}(1 - P_{kt})^{(1 - y_{kt})} \qquad (24)$$

where P_{kt} is $\text{prob}(y_{kt} = 1 \mid x_{kt}, \beta)$.

The right side of (24) can be reformatted in the form of (12) as,

$$f(y_{kt}, x_{kt}, \beta) = \exp[y_{kt} \ln(P_{kt}/(1 - P_{kt}) - \ln(1 - P_{kt})] \qquad (25)$$

then set

One is now generally free to choose the functional form for P_{kt}; however, when the linear logistic form is selected, the analysis simplifies considerably because, in this formulation, θ_{kt} is the logit of y_{kt},. That is,

$$\theta_{kt} = \ln(P_{kt}/(1-P_{kt})), \quad a(\theta_{kt}) = -\ln(1-P_{kt}), \quad b(y_{kt}) = 0, \quad \phi = 1. \qquad (26)$$

$$\theta_{kt} = \ln(P_{kt}(1-P_{kt})) = x_{kt}\beta. \qquad (27)$$

and

$$\begin{gathered} h(x_{kt},\beta) = x_{kt}\beta, \quad h^{\prime}(x_{kt}\beta) = 1, \quad \Delta_k = I_k \\ a(\theta_{kt}) = -\ln(1-P_{kt}) = -\ln(1/(1-\exp(x_{kt}\beta)), \quad a^{\prime}(x_{kt}\beta) = P_{kt} \\ a^{\prime\prime}(\theta_{kt}) = P_{kt}(1-P_{kt}), \quad \phi=1 \end{gathered} \qquad (28)$$

Thus, an estimator for the covariance matrix of β is given by,

$$\left[\sum_{k=1}^{K} X_k^T W_k A_k X_k\right]^{-1} \left[\sum_{k=1}^{K} X_k^T W_k S_k S_k^T W_k X_k\right] \left[\sum_{k=1}^{K} X_k^T W_k A_k X_k\right]^{-1} \qquad (29)$$

with β^* substituted for β where appropriate.

Note that if the observations are in fact independent, (29) corresponds to the formulas given by Manski and Lerman (1977;eq.(8),p. 1984: with their P(i,z,θ) taken to be binary response density (24) and P_{kt} with the logistic form as in (27)). (There appears to be an error in Manski and McFadden (1981,eq. (1.20), p. 18) where the weights, Q(i)/H(i), are not properly represented in their formulas.) Also note that if $W_k = I_k$ for all k and the observations are independent, then (29) reduces to,

$$\left[\sum_{k=1}^{K} X_k^T A_k X_k\right]^{-1} \qquad (30)$$

which is the standard covariance matrix for the logistic regression coefficients (Maddala, 1983).

Linear Regression

The standard model for linear regression is given by,

$$y_{kt} = x_{kt}\beta + \epsilon_{kt}, \quad where\ \epsilon_{kt} \sim N(0,\sigma^2) \qquad (31)$$

The density of the observation (y_{kt}, x_{kt}, β) is normal with

$$f(y_{kt}, x_{kt}, \beta) = (1/\sqrt{2\pi}\sigma)\ \exp[-(1/2)((y_{kt} - x_{kt}\beta)^2/\sigma^2] \qquad (32)$$

Equation (32) can be rewritten as,

$$f(y_{kt}, x_{kt}, \beta) = \exp\Big[[y_{kt}(x_{kt}\beta) - (1/2)(x_{kt}\beta)^2 - (1/2)y_{kt}^2 + \sigma^2 \ln\sqrt{2\pi}\sigma](1/\sigma^2)\Big] \quad (33)$$

Equation (33) can be interpreted in the form of (12) by setting,

$$\begin{aligned} \theta_{kt} &= h(x_{kt}\beta) = x_{kt}\beta, \quad h^{l}(x_{kt}\beta) = 1, \Delta_k = I_k \\ a(\theta_{kt}) &= (1/2)(x_{kt}\beta)^2, \quad a^{l}(\theta_{kt}) = x_{kt}\beta, \\ a^{ll}(\theta_{kt}) &= 1, \quad b(y_{kt}) = -(1/2)(y_{kt}^2 + \sigma^2 \ln(\sqrt{2\pi}\sigma), \quad \phi = 1/\sigma^2 \end{aligned} \qquad (34)$$

Thus an estimator for the covariance matrix of β is given by,

$$\left[\sum_{k=1}^{K} X_k^T W_k X_k\right]^{-1} \left[\sum_{k=1}^{K} X_k^T W_k S_k S_k^T W_k X_k\right] \left[\sum_{k=1}^{K} X_k^T W_k X_k\right]^{-1} \qquad (35)$$

where β^* replaces β as appropriate.

If the observations are independent, then $E(S_k S_k^T) = I_k\sigma^2$ for all k, and (35) becomes,

Note that (36) is the same as that given in DuMouchel and Duncan (1983,eq. 4.2, p. 538).

$$\left[\sum_{k=1}^{K} X_k^T W_k X_k\right]^{-1} \left[\sum_{k=1}^{K} X_k^T W_k^2 X_k\right] \left[\sum_{k=1}^{K} X_k^T W_k X_k\right]^{-1} \sigma^2 \qquad (36)$$

Thus under the assumption that the ε are independent and homoskedastic, (35) and (36), although not algebraically equivalent, have the same expected value. If in fact the ε is heteroskedastic but the variances have not been incorporated into the weights, then (35) would appear to be a more appropriate estimator since the heteroskedasticity is picked up in the estimate of $S_k S_k^T$. If in (36) $W_k = I_k$ for all k, then the covariance matrix is the familiar,

$$\left[\sum_{k=1}^{K} X_k^T X_k\right]^{-1} \sigma^2 \qquad (37)$$

SAS COMPUTER PROGRAMS

This section presents the SAS (Statistical Analysis System) computer programs that can be used to compute the covariance matrices for the logistic regression coefficients (29) and the linear regression coefficients (35). In each case two programs are provided: (1) a program that computes covariances when observations are not independent, with or without case weights, and (2) a program that computes covariances for weighted independent observations.

These programs are preceded by SAS code that would be required to implement the programs. There are typically three steps before the covariance programs can be run: (1) run the regression and save the predicted values, (2) prepare a data set that identifies the sets of observations that are possibly dependent, and (3) prepare a data set containing dependent and independent variables, predicted values of the dependent variables, and weights specifically for input to the covariance programs.

In all examples below, the assumption is that regressions are being run for adult respondents in the Survey of Income and Program Participation (SIPP). There is a SAS data set, SIPP.ONE, that contains the relevant variables as described below.

The dependent variable for the logistic regression analysis is "receipt of Title II benefits"--**yoasdi**=1 if benefits were received, **yoasdi**=2 otherwise (note that ordinarily the dependent variable would be set to 0 if the event did not occur, but SAS proc logistic requires a 2 because it develops a model to predict the smallest value of the dependent variable that it encounters). The dependent variable for the linear regression is the amount of Title II benefits, **yamt**. The weight variable is **fnlwgt**.

In both analyses, the independent variables are age and gender (**age, sex**). The assumption is that there may be dependence among observations in the same family (**suseqnum** is the same). The data set is assumed to be sorted by the variable(s) that determine possible dependence (in this example **suseqnum** only).

The data set inputted into the covariance program must have its variables organized in the following order: dependent variable, predicted dependent variable, weight, independent variables.

Covariances for Logistic Regression Coefficients

1. This code runs the logistic regression.

```
proc logistic data=sipp.one;      * input data set is sipp.one;
model yoasdi=age sex/ covb;   * covb is optional;
weight fnlwgt;                          * specifies the weight variable.
output out=work.data p=phat;  * creates an output data set work.data with input
run;                                          variables and predicted yoasdi in phat;
```

2. This code creates a data set, work.count1, in which the variable _freq_ contains the counts of the number of cases in each dependent group in the order that they appear in the data set sipp.one.

```
proc summary data=sipp.one;by suseqnum;output out=work.count1;run;
```

3. This code creates a data set, work.data2, with the variables in the order required by the covariance program.

```
data work.data2;
set work.data (keep=yoasdi phat fnlwgt age sex);
yoasdi1=yoasdi;phat1=phat;fnlwgt1=fnlwgt;age1=age;sex1=sex;
if yoasdi1=2 then yoasdi1=0;  * the covariance program requires that the values of
                                    the dependent variable be 1 and 0;
drop yoasdi phat fnlwgt age sex;
run;
```

4. This program computes the covariance matrix when observations may be dependent. If the data are unweighted then the weight variable in work.data2 should be set to 1. The covariance matrix is printed and outputted to a SAS data set, work.cov.

(AIS.P1171.$4848.SASPROG(FINLOGCL))

```
proc iml;
use work.count1;
read all var {_freq_} into cnt;
N=nrow(cnt);
ncount=0;
use work.data2;
do i=1 to n;
nn=cnt[i];
test=j(1,nn,0);
```

```
do j=1 to nn;
test[j]=ncount+j;
end;
read  point test  into z;
np=ncol(z);
nvar1=np-2;
if i=1 then do; vb=j(nvar1,nvar1,0);vtt=j(nvar1,nvar1,0);end;
ncount=ncount+nn;
y=z[,1];
ph=z[,2];
w=diag(z[,3]);
nvar=np-3;
npz=j(3,nvar,0);
inp=i(nvar);
txx=z*(npz//inp);
x=j(nn,1,1)||txx;
ones=j(nn,1,1);
q=ones-ph;
d=diag(ph#q);
t=x`*d*w*x;
vb=vb+t;
s=y-ph;
```

```
tt=x`*w*s*s`*w*x;
vtt=vtt+tt;
end;
vbi=inv(vb);
cov=vbi*vtt*vbi;
create work.cov from cov;
append from cov;
print  cov;
quit;
```

5. This program operates in the same manner as that in 4. above, except that it assumes that all observations are independent. Note that in this case the program at step 2. above is not required.

(AIS.P1171.$4848.SASPROG(FINLOGWT))

```
proc iml;
use work.data2;
i=0;nn=1;
do data;
i=i+1;
read next into z;
```

```
np=ncol(z);

nvar1=np-2;

if i=1 then do; vb=j(nvar1,nvar1,0);vtt=j(nvar1,nvar1,0);end;

y=z[,1];

ph=z[,2];

w=diag(z[,3]);

nvar=np-3;

npz=j(3,nvar,0);

inp=i(nvar);

txx=z*(npz//inp);

x=j(nn,1,1)||txx;

ones=j(nn,1,1);

q=ones-ph;

d=diag(ph#q);

t=x`*d*w*x;

vb=vb+t;

s=y-ph;

tt=x`*w*s*s`*w*x;

vtt=vtt+tt;

end;

vbi=inv(vb);

cov=vbi*vtt*vbi;
```

```
create work.cov from cov;
append from cov;
print  cov;
quit;
```

Covariances for Linear Regression Coefficients

1. This code subsets sipp.one, keeping only those cases for which the OASDI amounts exist, and runs the linear regression program.

```
data sipp.two;
if yamt>0;
run;

proc reg data=sipp.two;
model yamt=age sex/ covb;weight fnlwgt;
output out=work.data p=yhat;
run;
```

2. This code creates a data set, work.count1, in which the variable _freq_ contains the counts of the number of cases in each dependent group in the order that they appear in the data set sipp.two.

```
proc summary data=sipp.two;by suseqnum;output out=work.count1;run;
```

3. This code creates a dataset, work.data2, with the variables in the order required by the covariance program.

```
data work.data2;
set work.data (keep=yamt yhat fnlwgt age sex);
yamt1=yamt;yhat1=yhat;fnlwgt1=fnlwgt;age1=age;sex1=sex;
drop yamt yhat fnlwgt age sex;
run;
```

4. This program computes the covariance matrix when observations may be dependent. If the data are unweighted then the weight variable in work.data2 should be set to 1. The covariance matrix is printed and outputted to a SAS data set, work.cov.

(AIS.P1171.$4848.SASPROG(FINREGCL))

```
proc iml;
use work.count1;
read all var {_freq_} into cnt;
N=nrow(cnt);
```

```
ncount=0;

use work.data2;

do i=1 to n;

nn=cnt[i];

test=j(1,nn,0);

do j=1 to nn;

test[j]=ncount+j;

end;

read  point test into z;

np=ncol(z);

nvar1=np-2;

if i=1 then do; vb=j(nvar1,nvar1,0);vtt=j(nvar1,nvar1,0);end;

ncount=ncount+nn;

y=z[,1];

yh=z[,2];

w=diag(z[,3]);

nvar=np-3;

npz=j(3,nvar,0);

inp=i(nvar);

txx=z*(npz//inp);

x=j(nn,1,1)||txx;

ones=j(nn,1,1);
```

```
t=x`*w*x;
vb=vb+t;
s=y-yh;
tt=x`*w*s*s`*w*x;
vtt=vtt+tt;
end;
vbi=inv(vb);
cov=vbi*vtt*vbi;
create work.cov from cov;
append from cov;
print  cov;
quit;
```

5. This program operates in the same manner as that in 4. above, except that it assumes that all observations are independent. Note that in this case the program at step 2. above is not required.

(AIS.P1171.$4848.SASPROG(FINREGWT))

```
proc iml;
use work.data2;
i=0;nn=1;
```

```
do data;

i=i+1;

read  next  into z;

np=ncol(z);

nvar1=np-2;

if i=1 then do; vb=j(nvar1,nvar1,0);vtt=j(nvar1,nvar1,0);end;

y=z[,1];

yh=z[,2];

w=diag(z[,3]);

 nvar=np-3;

npz=j(3,nvar,0);

inp=i(nvar);

txx=z*(npz//inp);

x=j(nn,1,1)||txx;

ones=j(nn,1,1);

t=x`*w*x;

vb=vb+t;

s=y-yh;

tt=x`*w*s*s`*w*x;

vtt=vtt+tt;

end;

vbi=inv(vb);
```

```
cov=vbi*vtt*vbi;

create work.cov from cov;

append from cov;

print  cov;

quit;
```

Covariance Computations in PROC LOGISTIC AND PROC REG

Neither PROC LOGISTIC nor PROC REG provide for corrections to the covariance matrix when observations are not independent. Both procedures permit weighted data and maximize the weighted likelihood function when case weights are made available to them. The covariance matrix given by PROC LOGISTIC is,

$$\left[\sum_{k=1}^{K} X_k^T W_k A_k X_k\right]^{-1} \qquad (38)$$

with β^* substituted for β where appropriate. This formulation assumes that the case weights represent the presence of repeated independent observation in the sample, which is usually not the case. Thus the covariance estimates provided by PROC LOGISTIC for weighted data are usually not correct, equation (29) being the correct estimate. (The estimates provided tend to substantially understate the covariances.)

At this point it is not clear what computations are being done by PROC REG. The manual says that the weighted computations in PROC REG are the same as those in PROC GLIM; but the PROC GLIM writeup does not show the formulas. (Some preliminary tests indicate that the covariances provided have the same order of magnitude as those obtained from (35) but are not the same. The best guess is that they are providing estimates based on (36)).

HYPOTHESIS TESTING

The standard errors obtained from taking the square roots of the diagonal entries of the covariance matrix for a set of estimated regression coefficients can be used to test simple hypotheses about the coefficients, such as $\beta_j = 0$, for some j.

In analyses that do not involve case weights or dependent observations, complex hypotheses--such as, several coefficients being zero or being equal to other coefficients-- are usually tested using the Likelihood-Ratio Chi Square test. These tests can be performed using the SAS regression PROCs by estimating models with and without the constraints and computing the likelihood- ratio test statistic by taking twice the difference of the value of the log-likelihood functions (evaluated at β^*) provided by the programs. Under the null hypothesis, this test statistic will have a chi square distribution with the number of degrees of freedom equal to the difference in the number of parameters estimated between the two models.

When case weights and/or dependent observations are part of the analysis, it appears that likelihood-ratio chi square tests are not appropriate because the distribution of the likelihood ratio statistic under the null hypothesis is not known. (In particular, this calls into question the appropriateness of stepwise variable selection procedures offered in PROC LOGISTIC and PROC REG.) Instead, the use of Wald statistics as described by Grizzle **et al.** (1969) is recommended. This methodology involves the computation of chi square-type test statistics from the estimated coefficients, their covariance matrix and a contrast matrix that reflects the complex hypothesis being tested. This approach requires only the asymptotic normality of the estimated coefficients and a consistent estimate of the covariance matrix. These conditions have been assumed for the methodology described here.

A computer program, GENCAT, was developed by Landis **et al.** (1976) to implement this methodology. Fortunately, the main features of the GENCAT program have been incorporated into SAS PROC CATMOD. (See SAS/STAT User's Guide: Volume 1, p423 (among others).) Thus, if the array of coefficients is saved from the regression PROC, and the covariance matrix saved from the programs provided above, these data can be inputted to PROC CATMOD together with the other information needed to formulate the hypothesis; and PROC CATMOD will produce the necessary test statistics.

CONCLUSION

This paper has presented an approach to the calculation of estimated variances for linear and logistic regression coefficients in those analyses for which the sample design requires the use of case weights, the observations may not be independent, or both. Consistency and normality proofs have been outlined.

There are concerns as to the efficiency of the weighted likelihood estimators and estimators that do not take dependence into account. The major part of Liang and Zeger's work is the presentation of alternative approaches to the classical MLE that provide for increased efficiency, based on additional assumptions concerning the correlational structure. Concerning weighted estimators, Manski and McFadden present a number of alternative estimators for stratified sampling designs. Little is known about the relative efficiency of these estimators.

Even if the coefficient estimates are consistent and not terribly inefficient, nothing is known about the small sample biases in the estimates or the estimated covariances. (This seems to be the case with most classical MLEs.) For self-weighting samples of independent observations, the calculations presented in this paper are not algebraically equivalent to those that are ordinarily used. If the observations are, in fact, independent and self weighting, the classical formulas for calculation covariance matrices are presumed to be better since they incorporate these assumptions directly.

One alternative to the computational approach presented here is to compute the covariance matrices by some sort of resampling approach, such as half-sample replication. Sets of orthogonal half samples have been developed for the SIPP and the NBF. The advantage of using a half-sample estimator for the covariance of the regression coefficients is that special sophisticated computer programs are not required.

One disadvantage of using a half-sample approach is that the half samples must be large enough so that regression coefficients can be computed in each half-sample. This may constrain model specification. A second disadvantage is that for nonlinear estimators, the estimated covariance matrix may not be consistent even if the estimated coefficients are. On the other hand, for the sample sizes usually encountered, covariance estimates based on resampling methods may have smaller mean square errors than those provided by the calculations suggested here. Simulation studies need to be developed that can provide some information about the effectiveness of alternative approaches to covariance estimation.

REFERENCES

Bye, B. and J. Dykacz (1987) "The effects of misspecification in logistic regression models," **American Statistical Association, 1987, Proceedings of the Social Statistics Section**:98-102.

Bye, B. and G. Riley (1989) "Model estimation when observations are not independent: application of Liang and Zeger's methodology to linear and logistic regression analysis,"

Sociological Methods and Research, 17(4):353-375.

Dielman, T. (1983) "Pooled cross-section and time series data: a survey of current statistical methodology," **The American Statistician** 37(2):111-122.

DuMouchel, W. and G. Duncan (1983) "Using sample survey weights in multiple regression analysis of stratified samples," **Journal of the American Statistical Association** 78(383):535-543.

Farewell, J. (1979) "Some results on the estimation of logistic regression models based on retrospective data." **Biometrica** 66, 1:27-32.

Follman, D. and D. Lambert (1989) "Generalized logistic regression by nonparametric mixing," **Journal of the American Statistical Association** 84:295-300.

Hoem, J. (1985) "Weighting, misclassification, and other issues in the analysis of survey samples of life histories," pp. 249-293 in Heckman and Singer (eds.) **Longitudinal Analysis of Labor Market Data**. New York: Campridge Univ. Press.

Kott, P. (1991) "A model-based look a linear regression with survey data," **The American Statistician** 45(2):107-112.

Liang, K. and S. Zeger (1986) "Longitudinal data analysis using generalized linear models," **Biometrika** 73(1):13-22.

Manski C. and S. Lerman (1977) "The estimation of choice probabilities from choice based samples," **Econometrics** 45:1977-1988.

Maddala, G. (1983) **Limited-Dependent and Qualitive Variables in Econometrics**, New York: Cambridge University Press.

Manski C. and D. McFadden (1981) "Alternate estimator and sample designs for discrete choice analysis," in C. E. Manski and D. McFadden (eds.) **Structural Analysis of Discrete Data with Econometric Applications**. Cambridge, MA: M.I.T.

Nelson, D., McMillen, D. and D. Kasprzyk (1985) "An overview of the Survey of Income and Program Participation," **SIPP Working Paper Series** 8401(update 1), Bureau of the Census, Department of Commerce

Serfling, R. (1980) **Approximation Theorems of Mathematical Statistics**. New York: John Wiley and Sons

Prentice, R. and R. Pyke (1979) "Logistic disease incidence models and case control studies." **Biometrics** 66, 3:403-411.

Smith C. (1989) "The Social Security Administration's Continuous Work History Sample," **Social Security Bulletin** 52(10):21-28

Tuma, N. and M. Hannan (1984) **Social Dynamics: Models and Methods**. Orlando, Fl: Academic Press

Wolter, K. (1985) **Introduction to Variance Estimation**. New York: Springer-Verlag