NONLINEAR REGRESSION ANALYSIS FOR COMPLEX SURVEY DATA

Hsien-Ming Hung, Wichita State University

1. Introduction

A great deal of attention has been given to performing regression analyses for survey data, especially in social, health, and agricultural studies. Often such data arise from a complex survey for the need to collect the data as efficiently as possible within cost constraints. Kish and Frankel (1974) performed empirical studies on the behavior of regression coefficients from a cluster sample. Fuller (1975) investigated large sample properties of sample regression coefficients under the assumption that a finite population is generated from an infinite superpopulation. Hidiroglou et al. (1980) designed a computer package SUPERCARP containing algorithms for regression analyses for complex surveys. Other related work can be found in, for instance, Fuller and Battese (1973), Holt, Smith, and Winter (1980), Scott and Holt (1982), and the references therein. In their studies, the relationship between the dependent variable and the independent variables is described by a linear model.

A variety of nonlinear models have been proposed to describe the relationship among variables in many areas. In clinical studies the probability that an individual develops coronary heart disease within a given time period was assumed by Walker and Duncan (1967) to be a logistic function of independent variables such as age, systolic blood pressure, and serum cholestrol. Ratkowsky (1983) indicated that numerous nonlinear models such as a Weibull-type model had been used to model sigmoidal growth curves widespread in biology, agriculture, engineering, and economics. In the context of standard nonlinear regression analysis, a correct regression model is often assumed to exist; that is, the mean of residuals in the model is zero. Statistical methods for estimating the model parameters have been developed under the assumption of independent observations. A good reference is Draper and Smith (1981). Asymptotic properties of least squares estimators have been discussed extensively in the literature such as Jennrich (1969), Fuller (1976), Gallant (1987), and Wu (1981).

In this study we shall consider the cases where a given function that is of known functional form but depends upon an unknown parameter θ may be employed to reveal the relationship among the variables in a stratified finite population. The finite population parameter corresponding to θ will be defined as a solution of an estimating function defined by a properly chosen loss function. The estimation procedure will be presented in Section 2. Use of this function in constructing a regression estimator of finite population mean is discussed in Section 4. Large sample properties of estimators will be investigated under the assumption that the stratified finite population is generated from infinite superpopulations. No regression model is assumed for the superpopulations. However, the asymptotic theory is applicable to the cases where some type of overall nonlinear regression model arises. Discussions of the assumptions and mathematical results will be given in Section 5.

2. Parameters of interest and estimation

Consider a finite population $\mathcal{U}$ which is partitioned into L strata and contains N clusters in total. The h-th stratum contains N_h clusters and the i-th cluster of stratum h contains M_{hi} elements (h=1, ..., L; i=1, ..., N_h). Associated with the j-th element in the i-th cluster of stratum h is the vector $(Y_{hij}, X'_{hij})'$ of a $(p+1)$-dimensional characteristic $(Y, X')'$, where Y is the dependent variable and X is a vector of p explanatory variables.

Let $f(x;\theta)$ be a real valued function which is defined on a subspace $\chi \times \Theta$ of a $(p+q)$-dimensional Euclidean space and has a known parametric form. Suppose that this function can be used as a reasonable approximation to describe the relation between Y and X and the parameter θ may be to understand the approximate dependencies among the variables. The adoption of this function may be based upon prior knowledge or existing scientific evidence which specifies the form that population data ought to follow. It might suggest that the finite population is generated from infinite superpopulations wherein an overall nonlinear regression model with the parameter θ is well defined. The finite population parameter corresponding to θ could be defined as a point in Θ that minimizes a suitably chosen loss function. That is, one pretends that observations could be made on all M $(= \sum_{h=1}^{L} \sum_{i=1}^{N_h} M_{hi})$ elements of the finite population. A natural loss function for θ is

$$Q_N(\theta) = N^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (Y_{hi} - f_{hi}(\theta))'(Y_{hi} - f_{hi}(\theta)),$$

where $Y_{hi} = (Y_{hi1}, \ldots, Y_{hiM_{hi}})'$ and $f_{hi}(\theta) = (f(X_{hi1};\theta), \ldots, f(X_{hiM_{hi}};\theta))'$. The finite population parameter corresponding to θ is the vector θ_N in Θ that satisfies $Q_N(\theta_N) = \inf_{\theta\in\Theta} Q_N(\theta)$. When $f(x;\theta)$ is linear in θ, the parameter θ_N is the finite population vector of regression coefficients as defined in Fuller (1975). If a nonlinear regression model exists in the superpopulation, θ_N becomes the ordinary least squares estimator of the model parameter θ, which is obtained by assuming that the entire finite population acts as a sample. At times the superpopulation covariance structure of residuals, $Y_{hi} - f_{hi}(\theta)$, might be partially known and estimable from samples; then, $Q_N(\theta)$ could be modified using weighted residuals.

In estimating θ_N, a sample of n clusters is selected from $\mathcal{U}$ through a specific design by selecting a two-stage cluster sample from each stratum. The sampling is carried out independently in different strata. For stratum h, let n_h be the total number of clusters drawn and let m_{hi} be the total number of elements drawn from the i-th sampled cluster. For ease of presentation, the sample data is assumed to be $\{(Y_{hij}, X'_{hij})' : h = 1, \ldots, L; i = 1, \ldots, n_h; j = 1, \ldots, m_{hi}\}$, where $n = \sum_{h=1}^{L} n_h$. An estimator of θ_N can be constructed by finding θ_n in Θ to minimize the sample-based loss function

$$Q_n(\theta) = n^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} (Y_{his} - f_{his}(\theta))' V_{his}^{-1} (Y_{his} - f_{his}(\theta)),$$

where $Y_{his} = (Y_{hi1}, \ldots, Y_{him_{hi}})'$, $f_{his}(\theta) = (f(X_{hi1}; \theta), \ldots, f(X_{him_{hi}}; \theta))'$, and V_{his} are properly chosen weight matrices. For a general sampling design, the V_{his} are often to be diagonal matrices with elements proportional to selection probabilities π_{hij}, provided that $\pi_{hij} > 0$. If simple random nonreplacement sampling is carried out at each stage within each stratum, the inclusion probabilities become $\pi_{hij} = N_h^{-1} n_h M_{hi}^{-1} m_{hi}$. The existence of θ_n and θ_N can be established by following Jennrich (1969) under some mild conditions.

3. Large sample properties

The limiting behavior of θ_n will be developed for a single-stage stratified cluster sampling under a general framework analogous to that of Fuller (1984). It will be assumed that the finite population is generated from infinite superpopulations. The framework will not assume any nonlinear regression model like those appearing in classical regression analysis. Conditions similar to those of Fuller (1984) and of Krewski and Rao (1981) build the foundation of such a development.

Let $\{\mathcal{U}_r : r = 1, 2, \ldots\}$ be a sequence of finite populations, where $\mathcal{U}_r$ is partitioned into L_r strata, $L_r \geq L_{r-1}$. The h-th stratum of $\mathcal{U}_r$ contains N_{rh} clusters and the i-th cluster of stratum h contains M_{rhi} elements. The characteristic vectors for the r-th finite population are $(Y_{rhij}, X'_{rhij})'$, $h = 1, \ldots, L_r$, $i = 1, \ldots, N_{rh}$, $j = 1, \ldots, M_{rhi}$, and $N_r = \sum_{h=1}^{L_r} N_{rh}$ is the total number of clusters in $\mathcal{U}_r$. Let θ_r be the value in Θ that minimizes

$$Q_{N_r}(\theta) = N_r^{-1} \sum_{h=1}^{L_r} \sum_{i=1}^{N_{rh}} (Y_{rhi} - f_{rhi}(\theta))'(Y_{rhi} - f_{rhi}(\theta)),$$

where $(Y'_{rhi}, f'_{rhi}(\theta))$ are defined similarly as $(Y'_{hi}, f'_{hi}(\theta))$ in the previous section. To estimate θ_r, a simple random nonreplacement sample of clusters is selected from each stratum and the sampling is carried out independently in different strata. Let sample data be $\{(Y_{rhij}, X'_{rhij})' : h = 1, \ldots, L_r; i = 1, \ldots, n_{rh}; j = 1, \ldots, M_{rhi}\}$, where $n_{rh} \geq 2$, $n_{rh} > n_{r-1,h}$, and $n = \sum_{h=1}^{L_r} n_{rh}$. An estimator of θ_r is the value $\hat{\theta}_r$ in Θ that minimizes

$$Q_{n_r}(\theta) = \sum_{h=1}^{L_r} \sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1} (Y_{rhi} - f_{rhi}(\theta))'(Y_{rhi} - f_{rhi}(\theta)),$$

where $W_{rh} = N_r^{-1} N_{rh}$.

3.1. *Consistency*

The finite population in the h-th stratum of $\mathcal{U}_r$ is assumed to be a random sample of size $N_{rh} > N_{r-1,h}$ clusters selected from an infinite population ξ_{rh}. Strong consistency and weak consistency of $\hat{\theta}_r$ will be established under the following assumptions:

(i) Θ is a compact subspace of a q-dimensional Euclidean space.

(ii) χ is a p-dimensional Euclidean space. The function $f(x; \theta)$ is continuous on $\chi \times \Theta$ and has continuous partial derivatives of order through two with respect to $\theta \in \Theta$. Let $F(x; \theta)$ denote the $q \times 1$ vector of the first partial derivatives; let $L(x; \theta)$ denote the $(1 + q + q^2) \times 1$ vector of partial derivatives of order from zero through two. Let $R(x)$ be a $(1 + q + q^2) \times 1$ vector of dominating functions defined on χ such that for any $(x, \theta) \in \chi \times \Theta$ and for any k, $| L_k(x; \theta) | \leq R_k(x)$, where L_k and R_k are the k-th components of $L(x; \theta)$ and $R(x)$, respectively.

(iii) In ξ_{rh}, the cluster totals $\sum_{j=1}^{M_{rhi}} (1, |Y_{rhij}|, R_1(X_{rhij}), \ldots, R_{(1+q+q^2)}(X_{rhij}))$ have $2+2\delta$ $(\delta > 0)$ moments which are bounded uniformly in (r, h).

(iv) As $r \to \infty$, $N_r^{-1} \sum_{h=1}^{L_r} N_{rh} \mu_{rh}(\theta)$ converges to a positive function $Q(\theta)$ uniformly for θ in Θ, where $Q(\theta)$ has a unique minimum at an interior point θ_0, $\mu_{rh}(\theta) = E_{rh}\{(Y_{rhi} - f_{rhi}(\theta))'(Y_{rhi} - f_{rhi}(\theta))\}$, and E_{rh} denotes the expectation with respect to ξ_{rh}.

(v) $\sup_{1 \leq h \leq L_r} W_{rh} w_{rh}^{-1} = O(1)$, as $r \to \infty$, where $w_{rh} = n_r^{-1} n_{rh}$.

Some remarks on these conditions are as follows. In assumption (ii), a possible candidate for $R(x)$ is $\sup_{\theta \in \Theta} L(x; \theta)$ because Θ is compact. Under assumptions (ii) and (iii), the operations of differentiation and taking expectations are allowed to commute. With assumption (iv), the two assumptions also provide a basis for establishing strong or weak consistency of $\hat{\theta}_r$. Assumption (iv) is made to regulate the residual mean square errors in all the stratum superpopulations. Finally, assumption (v) provides a way of selecting sample clusters within each stratum relative to the total number of clusters in the stratum finite population. It implies that as $r \to \infty$, the total number of sampled clusters increases to infinity and $\sup_{1 \leq h \leq L_r} n_r W_{rh}^2 n_{rh}^{-2}$ converges to zero. It is possible to replace assumption (v) by C2 and C3 of Krewski and Rao (1981).

Theorem 1. Let the sequence of finite populations and samples be as stated. Under assumptions (i)-(v) with $\delta > 0$ in assumption (iii),

$$\operatorname{plim}_{r \to \infty} \theta_r = \theta_0,$$
$$\operatorname{plim}_{r \to \infty} \hat{\theta}_r = \theta_0,$$
$$\operatorname{plim}_{r \to \infty} (\hat{\theta}_r - \theta_r) = 0.$$

Proof: The sample-based loss function $Q_{n_r}(\theta)$ can be written as

$$Q_{n_r}(\theta) = n_r^{-1} \sum_{h=1}^{L_r} \sum_{i=1}^{n_{rh}} Z_{rhi},$$

where $Z_{rhi} = n_r W_{rh} n_{rh}^{-1}(Y_{rhi} - f_{rhi}(\theta))'(Y_{rhi} - f_{rhi}(\theta))$. By assumptions (ii) and (iii), for any $\theta \in \Theta$,

$$n_r^{-1} \sum_{h=1}^{L_r} \sum_{i=1}^{n_{rh}} E_{rh}(| Z_{rhi} |^{1+\delta}) \leq K \sup_{1 \leq h \leq L_r} (W_{rh} w_{rh}^{-1})^{1+\delta},$$

for some $K > 0$. By Lemma 3.2 of Krewski and Rao (1981), given $\varepsilon > 0$, there exists a $R_\varepsilon > 0$ such that if $r \geq R_\varepsilon$,

$$P(| Q_{n_r}(\theta) - N_r^{-1} \sum_{h=1}^{L_r} N_{rh} \mu_{rh}(\theta) | > \varepsilon/2) < \varepsilon/2,$$

for any $\theta \in \Theta$. It follows from assumption (iv) that $Q_{n_r}(\theta)$ converges in probability to $Q(\theta)$ uniformly for $\theta \in \Theta$.

Let $\{\hat{\theta}_{rt}\}_{t=1}^{\infty}$ be a subsequence of $\{\hat{\theta}_r\}_{r=1}^{\infty}$ corresponding to a realization of the vectors $\{(Y_{rhij}, X'_{rhij})$: $j = 1, \ldots, M_{rhi}$, $i = 1, \ldots, n_{rh}$, $h = 1, \ldots, L_r$, $r = 1, 2, \ldots\}$. Since Θ is compact there exists a subsequence $\{\hat{\theta}_{rts}\}_{s=1}^{\infty}$ and a limit point $\dot{\theta}$ such that $\lim_{s \to \infty} \hat{\theta}_{rts} = \dot{\theta}$.

Since $Q_{n_{rt}}(\theta)$ converges in probability to $Q(\theta)$ uniformly for $\theta \in \Theta$, $\{Q_{n_{rts}}(\theta)\}_{s=1}^{\infty}$ has a subsequence $\{Q_{n_{rtsl}}(\theta)\}_{l=1}^{\infty}$ which converges almost surely to $Q(\theta)$ uniformly for $\theta \in \Theta$. Except for the realizations belonging to a set A with $P(A) = 0$,

$$\begin{aligned} 0 \leq Q(\dot{\theta}) - Q(\theta_0) &= \lim_{l \to \infty} [Q_{n_{rtsl}}(\hat{\theta}_{rtsl}) - Q(\theta_0)] \\ &\leq \lim_{l \to \infty} [Q_{n_{rtsl}}(\theta_0) - Q(\theta_0)] = 0, \end{aligned}$$

and hence $\dot{\theta} = \theta_0$. Therefore, $\hat{\theta}_{rts}$ converges almost surely to θ_0 and $\hat{\theta}_r$ converges in probability to θ_0. The remaining results follow accordingly. ▮

Theorem 2. Let the sequence of finite populations and samples be as described. Under assumptions (i) - (v) with $\delta > 1$ in assumption (iii),

$$\begin{aligned} &P(\lim_{r \to \infty} \theta_r = \theta_0) = 1, \\ &P(\lim_{r \to \infty} \hat{\theta}_r = \theta_0) = 1, \\ &P(\lim_{r \to \infty} (\hat{\theta}_r - \theta_r) = 0) = 1. \end{aligned}$$

Proof: Let $\mu_r(\theta) = N_r^{-1} \sum_{h=1}^{L_r} N_{rh} \mu_{rh}(\theta)$ for $\theta \in \Theta$. By Lemma 3.2 of Krewski and Rao (1981), assumptions (ii)-(iv) imply that given $\varepsilon > 0$,

$$P(| Q_{n_r}(\theta) - \mu_r(\theta) | > \varepsilon) = O(n_r^{-(1+\delta)/2}),$$

uniformly for $\theta \in \Theta$. For any r, $n_r \geq r$ and hence given $\varepsilon > 0$,

$$\sum_{r=1}^{\infty} P(| Q_{n_r}(\theta) - \mu_r(\theta) | > \varepsilon) < \infty,$$

for any $\theta \in \Theta$, which implies, by Borel-Cantelli Lemma, $Q_{n_r}(\theta) - \mu_r(\theta)$ converges almost surely to zero uniformly for $\theta \in \Theta$. The results can be shown in a similar manner as we prove Theorem 1. ▮

3.2. *Asymptotic normality*

Two additional regularity conditions given in the next theorem are needed to establish the asymptotic normality of $n_r^{1/2}(\hat{\theta}_r - \theta_r)$. Assumption (vi) is necessary for some components of the asymptotic variance of $n_r^{1/2}(\hat{\theta}_r - \theta_r)$ to be well defined. It also ensures that a consistent estimator of the asymptotic variance of $\hat{\theta}_r$ may be obtained. For each r, let θ_{r0} be a point in Θ that minimizes $N_r^{-1} \sum_h N_{rh} \mu_{rh}(\theta)$. The second part of assumption (vii) is like the condition C3 of Francisco and Fuller (1986), which ensures that

$$n_r^{1/2} \sum_{h=1}^{L_r} \sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1} F'_{rhi}(\theta_{r0})(Y_{rhi} - f_{rhi}(\theta_{r0}))$$

has a covariance matrix with a determinant bounded away from zero. Let Σ_{rh} be the covariance matrix of $F'_{rhi}(\theta_{r0})(Y_{rhi} - f_{rhi}(\theta_{r0}))$. The second part of assumption (vii) may be replaced by the somewhat stronger condition $\lim_{r \to \infty} n_r \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \Sigma_{rh} = \Lambda$ (positive definite).

Theorem 3. Let the assumptions of Theorem 2 hold. Also, assume:

(vi) As $r \to \infty$, $N_r^{-1} \sum_{h=1}^{L_r} N_{rh} (\partial^2 / \partial\theta\partial\theta') \mu_{rh}(\theta)$ converges to $A(\theta)$ uniformly for $\theta \in \Theta$. The limiting matrix $A(\theta)$ is nonsingular at $\theta = \theta_0$.

(vii) For each r, $0 \leq f_{rh} < U_f < 1$, for some $U_f > 0$, where $f_{rh} = N_{rh}^{-1} n_{rh}$, and there exist L_Σ and U_Σ such that

$$0 < L_\Sigma <| n_r \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \Sigma_{rh} |< U_\Sigma.$$

Then, as $r \to \infty$,

$$\begin{aligned} \{\hat{V}(\hat{\theta}_r - \theta_{r0})\}^{-1/2}(\hat{\theta}_r - \theta_{r0}) &\xrightarrow{\mathcal{L}} N(0, I), \\ \{\hat{V}(\hat{\theta}_r - \theta_r)\}^{-1/2}(\hat{\theta}_r - \theta_r) &\xrightarrow{\mathcal{L}} N(0, I), \end{aligned}$$

where

$$\hat{V}(\hat{\theta}_r - \theta_{r0}) = \hat{A}_r^{-1} \{\sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \hat{\Sigma}_{rh}\} \hat{A}_r^{-1},$$

$$\hat{V}(\hat{\theta}_r - \theta_r) = \hat{A}_r^{-1} \{\sum_{h=1}^{L_r} W_{rh}^2 (1 - f_{rh}) n_{rh}^{-1} \hat{\Sigma}_{rh}\} \hat{A}_r^{-1},$$

$$\begin{aligned} \hat{\Sigma}_{rh} &= (n_r - 1)(n_r - q)^{-1}(n_{rh} - 1)^{-1} \\ &\quad \times \sum_{i=1}^{n_{rh}} (d_{rhi} - \bar{d}_{rh.})(d_{rhi} - \bar{d}_{rh.})', \end{aligned}$$

$$\bar{d}_{rh.} = n_{rh}^{-1} \sum_{i=1}^{n_{rh}} d_{rhi},$$

$$d_{rhi} = \sum_{j=1}^{M_{rhi}} F(X_{rhij}; \hat{\theta}_r)(Y_{rhij} - f(X_{rhij}; \hat{\theta}_r))$$

$$\begin{aligned}\hat{A}_r = \sum_{h=1}^{L_r}\sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1} \sum_{j=1}^{M_{rhi}} \{&F(X_{rhij}; \hat{\theta}_r)F'(X_{rhij}; \hat{\theta}_r) \\ &- G(X_{rhij}; \hat{\theta}_r)(Y_{rhij} - f(X_{rhij}; \hat{\theta}_r))\},\end{aligned}$$

$$G(x;\theta) = (\frac{\partial}{\partial\theta'})F(x;\theta).$$

Proof: A Taylor series expansion of $(\partial/\partial\theta)Q_{n_r}(\hat{\theta}_r)$ with respect to θ_{r0} leads to

$$\begin{aligned}&\sum_{h=1}^{L_r}\sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1} F'_{rhi}(\hat{\theta}_r)(Y_{rhi} - f_{rhi}(\hat{\theta}_r)) \\ &= \sum_{h=1}^{L_r}\sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1} F'_{rhi}(\theta_{r0})(Y_{rhi} - f_{rhi}(\theta_{r0})) \\ &\quad + \Big(\sum_{h=1}^{L_r}\sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1}\{\sum_{j=1}^{M_{rhi}} G(X_{rhij}; \bar{\theta}_r) \\ &\quad \times (Y_{rhij} - f(X_{rhij}; \bar{\theta}_r)) \\ &\quad - F'_{rhi}(\bar{\theta}_r)F_{rhi}(\bar{\theta}_r)\}\Big)(\hat{\theta}_r - \theta_{r0}),\end{aligned} \tag{3.1}$$

where $\bar{\theta}_r$ is on the line segment joining $\hat{\theta}_r$ and θ_{r0}. Now $\hat{\theta}_r$ converges almost surely to θ_0 and it can be easily shown that $\lim_{r\to\infty}\theta_{r0} = \theta_0$. Thus, in the second term of (3.1),

$$\begin{aligned}&\sum_{h=1}^{L_r}\sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1}\{F'_{rhi}(\bar{\theta}_r)F_{rhi}(\bar{\theta}_r) \\ &\quad - \sum_{j=1}^{M_{rhi}} G(X_{rhij}; \bar{\theta}_r)(Y_{rhij} - f(X_{rhij}; \bar{\theta}_r))\} \\ &= A(\theta_0) + o_p(1),\end{aligned}$$

by assumptions (ii) and (vi). Therefore, the results follow by showing the asymptotic normality of

$$n_r^{1/2}\sum_{h=1}^{L_r}\sum_{i=1}^{n_{rh}} W_{rh} n_{rh}^{-1} F'_{rhi}(\theta_{r0})(Y_{rhi} - f_{rhi}(\theta_{r0})). \quad \blacksquare$$

It is worth noting that the asymptotic normality presented in the theorem is proved under the assumption of δ being greater than one. This assumption ensures that $\hat{\theta}_r$ is strongly consistent for θ_0. It can be shown that if $0 < \delta \leq 1$, the asymptotic normality is still obtainable provided that $F'_{rhi}(\theta_{r0})(Y_{rhi} - f_{rhi}(\theta_{r0}))$ possess uniformly bounded $2 + \eta$ moments ($\eta > 0$).

4. Estimation of finite population mean

Proper use of $f(x,\theta)$ may largely improve the estimation of finite population mean of Y per cluster, provided that the population data on X are available. For practical applications, the reader is referred to Hung and Fuller (1987) and the references therein, where the discussions rest upon the use of estimated $f(X,\theta)$ as an auxiliary variable in regression estimation of finite population mean of Y per cluster. A motivation behind such utilities is that a good choice of $f(x,\theta)$ may well approximate the conditional mean of Y given X which is often unknown, and hence can take best advantage of the linear correlation between Y and X.

To avoid unnecessary notational duplication, let the sequence of finite populations and samples be as described in the previous section. Let the finite population mean of Y be given by

$$\bar{Y}_r = \sum_{h=1}^{L_r} W_{rh} N_{rh}^{-1} \sum_{i=1}^{N_{rh}} Y_{rhi+},$$

where Y_{rhi+} is the cluster sum of Y_{rhij}. The combined regression estimator of $\bar{Y}_r$ made using $f(X,\hat{\theta}_r)$ is

$$\hat{\bar{y}}_{rlc} = \bar{y}_r + \hat{b}_{rc}(\hat{\bar{Z}}_r - \hat{\bar{z}}_r),$$

where for each (r,h,i,j), $\hat{Z}_{rhij} = f(X_{rhij},\hat{\theta}_r)$, $\hat{Z}_{rhi+}$ is the cluster sum of $\hat{Z}_{rhij}$, $(\bar{y}_{rh}, \hat{\bar{z}}_{rh})$ is the sample mean of $(Y_{rhi+}, \hat{Z}_{rhi+})$ per cluster,

$$(\bar{y}_r,\ \hat{\bar{z}}_r) = \sum_{h=1}^{L_r} W_{rh}(\bar{y}_{rh},\ \hat{\bar{z}}_{rh}),$$

$$\hat{\bar{Z}}_r = \sum_{h=1}^{L_r} W_{rh} N_{rh}^{-1} \sum_{i=1}^{N_{rh}} \hat{Z}_{rhi+},$$

$$\begin{aligned}\hat{b}_{rc} = \{&\sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1}(1 - f_{rh})(n_{rh} - 1)^{-1} \\ &\times \sum_{i=1}^{n_{rh}}(\hat{Z}_{rhi+} - \hat{\bar{z}}_{rh})^2\}^{-1} \\ &\times \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1}(1 - f_{rh})(n_{rh} - 1)^{-1} \\ &\times \sum_{i=1}^{n_{rh}}(\hat{Z}_{rhi+} - \hat{\bar{z}}_{rh})Y_{rhi+}.\end{aligned}$$

The following theorem generalizes the asymptotic results of Hung (1985) and of Hung and Fuller (1987) to cover the cases of stratified cluster random sampling as in this study.

Theorem 4. Given the assumptions of Theorem 3,

$$\hat{\bar{y}}_{rlc} - \bar{Y}_r = \bar{y}_{rlc} - \bar{Y}_r + o_p(n_r^{-1/2}), \qquad \text{as } r \to \infty.$$

where $\bar{y}_{rlc}$ is the combined regression estimator of $\bar{Y}_r$ constructed using $f(X,\theta_{r0})$. Furthermore, if for all r,

$$n_r\sum_{h=1}^{L_r} W_{rh}^2(1-f_{rh})n_{rh}^{-1}E_{rh}\{Y_{rhi+}-E_{rh}(Y_{rhi+})\}^2(1-\rho_{rc}^2)$$

are uniformly bounded away from zero, where $Z_{rhi+} = \sum_{j=1}^{M_{rhi}} f(X_{rhij},\theta_{r0})$, $(\mu_{Yrh}, \mu_{Zrh}, \mu_{YZrh}) = E_{rh}(Y_{rhi+},$

$Z_{rhi+}, Y_{rhi+}Z_{rhi+})$,

$$\rho_{rc} = \Big(\sum_{h=1}^{L_r} W_{rh}^2(1 - f_{rh})n_{rh}^{-1}E_{rh}\{Z_{rhi+} - \mu_{Zrh})\}^2\Big)^{-1/2}$$
$$\times \Big(\sum_{h=1}^{L_r} W_{rh}^2(1 - f_{rh})n_{rh}^{-1}E_{rh}\{Y_{rhi+} - \mu_{Yrh})\}^2\Big)^{-1/2}$$
$$\times \sum_{h=1}^{L_r} W_{rh}^2(1 - f_{rh})n_{rh}^{-1}E_{rh}\{Y_{rhi+}Z_{rhi+} - \mu_{YZrh})\},$$

then,

$$\{\hat{V}(\hat{\bar{y}}_{rlc})\}^{-1/2}(\hat{\bar{y}}_{rlc} - \bar{Y}_r) \xrightarrow{\mathcal{L}} N(\ 0,\ I\ ),$$

where

$$\hat{V}(\hat{\bar{y}}_{rlc}) = \sum_{h=1}^{L_r} W_{rh}^2(1 - f_{rh})n_{rh}^{-1}(n_{rh} - q - 1)^{-1}$$
$$\times \sum_{i=1}^{n_{rh}}\{Y_{rhi+} - \bar{y}_{rh} - \hat{b}_{rc}(\hat{Z}_{rhi+} - \hat{\bar{z}}_{rhi+})\}^2.$$

Proof is similar to that of Theorem 3.1 of Hung (1985).

Note that ρ_{rc}^2 is always between 0 and 1. When the relation between Y and X is nearly linear and the ratio of the variance of Y_{rhi+} to that of Z_{rhi+} is constant for each stratum, ρ_{rc}^2 is close to one. Theorem 5 basically tells us that the estimation of θ in $f(X, \theta)$ does not inflat the asymptotic variance of the combined regression estimator when θ is estimated by the estimator $\hat{\theta}_r$. In fact, the result is still true when θ is estimated by any estimator $\ddot{\theta}_r$ satisfying $n_r^{1/2}(\ddot{\theta}_r - \theta) = O_p(1)$.

Another type of regression estimator is constructed by computing a separate regression estimator for each stratum mean and then taking the weighted sum of these stratum estimators. That is, the separate regression estimator of $\bar{Y}_r$ is given by

$$\hat{\bar{y}}_{rls} = \sum_{h=1}^{L_r} W_{rh}(\bar{y}_{rh} + \hat{b}_{rh}(\hat{\bar{Z}}_{rh} - \hat{\bar{z}}_{rh})),$$

where $\hat{\bar{Z}}_{rh}$ is the population average of $\hat{Z}_{rhi+}$ per cluster, and $\hat{b}_{rh}$ is the sample regression coefficient in the regression of Y_{rhi+} on $\hat{Z}_{rhi+}$ with intercept for the hth stratum. The difference between $\hat{\bar{y}}_{rls}$ and $\bar{Y}_r$ can be written as

$$\hat{\bar{y}}_{rls} - \bar{Y}_r = \sum_{h=1}^{L_r} W_{rh}(\bar{y}_{rh} + b_{rh}(\bar{Z}_{rh} - \bar{z}_{rh})) - \bar{Y}_r$$
$$+ \sum_{h=1}^{L_r} W_{rh}b_{rh}(\hat{\bar{Z}}_{rh} - \bar{Z}_{rh} - \hat{\bar{z}}_{rh} + \bar{z}_{rh})$$
$$+ \sum_{h=1}^{L_r} W_{rh}(\hat{b}_{rh} - b_{rh})(\hat{\bar{Z}}_{rh} - \bar{Z}_{rh} - \hat{\bar{z}}_{rh} + \bar{z}_{rh})$$
$$+ \sum_{h=1}^{L_r} W_{rh}(\hat{b}_{rh} - b_{rh})(\bar{Z}_{rh} - \bar{z}_{rh}), \qquad (4.1)$$

where b_{rh} is the sample regression coefficient in the regression of Y_{rhi+} on Z_{rhi+} with intercept for the hth stratum. The first term on the right-hand side of the equation is the difference between $\bar{Y}_r$ and the traditional separate regression estimator $\bar{y}_{rls}$ constructed using Z_{rhi+}. The stochastic orders of remaining terms do not appear evident even though the assumptions of Theorem 5 are made. However, these terms seem to remain of order in probability less than $n_r^{-1/2}$, provided that a large number of sample clusters are allowed to be taken within each strata. Motivation is given as follows. Let $T_{rhi+} = \sum_{j=1}^{M_{rhi}} F(X_{rhij}, \theta_{r0})$; let $\bar{T}_{rh}$ and $\bar{t}_{rh}$ be the averages of T_{rhi+} per cluster for the population and the sample, respectively. Then, under the previously given assumptions, it can be shown that

$$\sum_{h=1}^{L_r} W_{rh}b_{rh}(\hat{\bar{Z}}_{rh} - \bar{Z}_{rh} - \hat{\bar{z}}_{rh} + \bar{z}_{rh})$$
$$=\{\sum_{h=1}^{L_r} W_{rh}b_{rh}(\bar{T}_{rh} - \bar{t}_{rh})\}(\hat{\theta}_r - \theta_{r0}) + O_p(n_r^{-1}).$$

Suppose that each b_{rh} has second moments and let $\beta_{rh} = E_{rh}(b_{rh})$. Then as $r \to \infty$, $\sum_{h=1}^{L_r} W_{rh}\beta_{rh}(\bar{T}_{rh} - \bar{t}_{rh}) = o_p(1)$. Moreover, for any $j \in \{1, \ldots, q\}$,

$$E\{\sum_{h=1}^{L_r} W_{rh}(b_{rh} - \beta_{rh})(\bar{T}_{rh(j)} - \bar{t}_{rh(j)})\}$$
$$\leq \sum_{h=1}^{L_r} W_{rh}n_{rh}^{-1}(\nu_{rh}\sigma_{Trh(j)}^2)^{1/2}$$
$$\leq \sup_{1 \leq h \leq L_r} n_{rh}^{-1}N_r^{-1}\sum_{h=1}^{L_r} N_{rh}(\nu_{rh}\sigma_{Trh(j)}^2)^{1/2},$$

where $T_{rhi+(j)}$, $\bar{T}_{rh(j)}$, and $\bar{t}_{rh(j)}$ are the j-th elements of T_{rhi+}, $\bar{T}_{rh}$, and $\bar{t}_{rh}$, respectively,

$$\nu_{rh} = (E_{rh}\{(Z_{rhi+} - E_{rh}(Z_{rhi+}))^2\})^{-2}$$
$$\times E_{rh}\{(Z_{rhi+} - E_{rh}(Z_{rhi+}))^2(Y_{rhi+} - \beta_{rh}Z_{rhi+})^2\},$$
$$\sigma_{Trh(j)}^2 = E_{rh}\{(T_{rhi+(j)} - E_{rh}(T_{rhi+(j)}))^2\}.$$

Thus, if $\inf_{1 \leq h \leq L_r} n_{rh} \to \infty$ as $r \to \infty$, then

$$\sum_{h=1}^{L_r} W_{rh}b_{rh}(\hat{\bar{Z}}_{rh} - \bar{Z}_{rh} - \hat{\bar{z}}_{rh} + \bar{z}_{rh}) = o_p(n_r^{-1/2}),$$

and so are the third and the fourth terms on the right-hand side of (4.1). In other words, for the type of large-scale surveys in which all the strata are large and a large number of sample clusters are selected within each stratum, the estimation of θ will not change the asymptotic variance of the separate regression estimator.

5. Discussion

A basic assumption in our study is that a given function $f(x, \theta)$ involving the unknown parameter θ is available for approximating the interdependencies among the variables of interest in a finite population. The knowledge of the parametric form for $f(x, \theta)$ might arise from a plot of sample data. Perhaps more frequently, it comes from prior information leading to the postulate that the finite population is generated from an infinite superpopulation with $f(X, \theta)$ being the conditional mean of the dependent variable given the independent variables X. In both cases the definition of the finite population paramster θ_N given in Section 2 seems natural. An important point is that the function $f(x, \theta)$ must be determined prior to any attempt of defining θ_N. Therefore, for our results to be useful, the function $f(x, \theta)$ must be at least a reasonable approximation to the relationship between Y and X. In such instances, the estimated version of $f(X, \theta)$ can be used as an auxiliary variable in regression estimation of finite population mean of Y per cluster, provided that the population data on X are available. As noted in Section 4, there is no asymptotic cost due to estimating $f(X, \theta)$ in constructing a combined regression estimator for the finite population mean per cluster and in some cases it is also true for a separate regression estimator.

The proposed estimator for θ_N takes sample weights into account. In general survey designs may have complicate effects on the variance of a survey estimator. An example provided by Fuller (1984) demonstrates that in estimating regression coefficients, unit weights do not always produce smaller variances than sample weights. When the effects are ignored in the finite population inference, our estimation procedure can be modified by properly adjusting the weight matrices in the sample-based loss function. The large sample results presented in Section 3 are still valid under the aforementioned regularity conditions with slight modification and some additional conditions to regulate weight matrices; for example, some norms for the weight matrices are uniformly bounded away from zero.

The large sample results obtained in previous sections are also applicable to the surveys in which large numbers of strata with relatively few clusters are selected within each stratum but no strata are of disproportionate size. See the conditions of Krewski and Rao (1981). The framework for the development of asymptotic properties can be extended to cover the cases of multi-stage random subsampling within clusters by following Appendix A of Fuller (1975). Extension to various forms of unequal probability sampling is also possible. In such cases the weight matrices in the loss function $Q_n(\theta)$ involve only the inclusion probabilities. Therefore the condition similar to that the probabilities are uniformly bounded away from zero is required.

6. References

1. Draper, N. R. and Smith, H. (1981), Applied Regression Analysis. John Wiley, New York.
2. Francisco, C. A. and Fuller, W. A. (1986), Estimation of the distribution function with a complex survey, Presented at Joint Meetings, American Statistical Association, August 1986.
3. Fuller, W. A. (1975), Regression analysis for sample surveys, *Sankhya C*, **37**, 117-132.
4. Fuller, W. A. (1976), Introduction to Statistical Time Series. John Wiley, New York.
5. Fuller, W. A. (1984), Least squares and related analyses for complex survey designs, *Survey Methodology*, **10**, 97-125.
6. Fuller, W. A. and Battese, G. (1973), Transformations for estimation of linear models with nested error structure, *J. Amer. Statist. Assoc.*, **68**, 626-632.
7. Gallant, A. R. (1987), Nonlinear Statistical Models. John Wiley, New York.
8. Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), SUPERCARP (6th ed.), Statistical Laboratory, Iowa State University, Ames, Iowa.
9. 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.
10. Hung, H. M. (1985), Regression estimation with transformed auxiliary variates, *Statistics and Probability Letters*, **5**, 239-243.
11. Hung, H. M. and Fuller, W. A. (1987), Regression estimation of crop acreages with transformed auxiliary variables, *Journal of Business and Economic Statistics*. (To appear.)
12. Jennrich, R. I. (1969), Asymptotic properties of nonlinear least squares estimators, *Ann. Math. Stat.*, **40**, 633-643.
13. Kish, L. and Frankel, M. R. (1974), Inference from complex samples (with discussion), *J. R. Statist. Soc. B*, **36**, 1-37.
14. Krewski, D. and Rao, J. N. K. (1981), Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods, *Ann. Stat.*, **9**, 1010-1019.
15. Ratkowsky, D. A. (1983), Nonlinear Regression Modelling. Marcel Dekker, New York.
16. Scott, A. J. and Holt, D. (1982), The effect of two-stage sampling on ordinary least squares methods, *J. Amer. Statist. Assoc.*, **77**, 848-854.
17. Walker, S. H. and Duncan, D. B. (1967), Estimation of the probability of an event as function of several independent variables, *Biometrika*, **54**, 167-179.
18. Wu, C.-F. (1981), Asymptotic theory of nonlinear least squares estimation, *Ann. Stat.*, **9**, 501-513.