REGRESSION ANALYSIS WITH TRANSFORMED AUXILIARY VARIABLES

Hsien-Ming Hung, University of Minnesota

1. INTRODUCTION

In recent decades, regression analysis for survey data has received a great deal of attention. Two common problems often considered in the literature are those of estimating domain means and of deriving regression equations for survey data. In investigating limiting properties of the regression estimator of a finite population mean, Scott and Wu (1981) embedded the finite population in a sequence of fixed numbers possessing some well defined properties following Hajek (1960). Hidiroglou (1974) developed central limit theorem for regression coefficients estimated for a finite population. Fuller (1975) established large sample results for regression coefficients, assuming that the finite population is generated from an infinite superpopulation.

In some types of surveys, some functions of auxiliary information are often used as the best possible auxiliary variables to improve the estimation. For instance, in estimating the true proportion of individuals possessing an attribute of interest, one often utilizes the conditional probability of 'Y=1' given X, where 'Y=1' indicates that an individual has the attribute and X is an observable vector of auxiliary variables. A natural reason is that the expected value conditional on X of a zero-one variable Y is known as the probability of the event 'Y =1' conditional on X. In agricultural surveys, Sigman et al (1978) described the use of the U.S. Department of Agriculture pixel classifier as an auxiliary variable in regression estimation of crop acres. The pixel classifier is a set of classification functions classifying individual spectral readings for each picture element to a probable crop type. Hung and Fuller (1984) investigated the use of the estimated probability that a point with a satellite value X is from crop j in improving the crop acreage estimation.

This study investigates the use of functions of a vector X as auxiliary variables in the regression analysis for survey data. Functions of the vector X are estimated by estimating the unknown transformation parameters. As in Hung and Fuller (1984), the parameters of the transformation might be those (e.g., location or shape parameters) of the distribution of X for the categories indexed by Y. Hung (1985) studied the cases where transformation parameters are independent of Y and hence are estimated based only upon X. A regression estimator of the finite population mean of a survey variable Y and the estimation of regression equations are considered in Section 2 and 3, respectively. Large sample properties of estimators are developed under the framework of Fuller (1975). The effects of estimating the transformed variables on the estimators are investigated. The results are extended in Section 4 to cover a broad class of transformation functions. Discussions are provided in Section 5.

In subsequent sections, the finite population is taken to be a set of N individuals indexed by the subscript t. Associated with the t-th unit is a vector (Y_t, X_t'), where Y is a primary survey variate and X is a px1 vector of auxiliary information. The values of X for individual population units are known at the design stage. The qx1 vectors used as the auxiliary are denoted by $g(X_t; d_o)$, $t = 1,\dots, N$, where the rx1 parameter vector d_o is an interior point of Θ, a compact subset of r-dimensional Euclidean space R^r; $g: R^{p+r} \to R^q$. Let $Z_t = [1, g'(X_t; d_o)]'$ for all t. At the estimation stage, observations are available for a simple random nonreplacement sample of n units. For convenience, the sample units are assumed to be labeled from 1 to n. The finite population is assumed to be a random sample of an infinite superpopulation and $E\{\cdot\}$ is used to denote expectation with respect to the infinite population.

2. ESTIMATION OF FINITE POPULATION MEANS

Consider the regression estimation of the finite population mean $\bar{Y}_N$. The unknown vector d_o is estimated by an estimator d_n which is a function of (Y_t, X_t'), $t = 1,\dots, n$. The transformed auxiliary vectors $\hat{Z}_t$ are obtained by substituting d_n for d_o in Z_t for the whole finite population. The regression estimator of $\bar{Y}_N$ constructed by use of $g(X; d_n)$ as the auxiliary is

$$\hat{Y}_N = \bar{Y}_n + \hat{B}_n'(\hat{\bar{Z}}_N - \hat{\bar{Z}}_n), \qquad (2.1)$$

where $\bar{Y}_n$ is the sample mean of Y, and

$$\hat{\bar{Z}}_K = K^{-1} \sum_{t=1}^{K} \hat{Z}_t \qquad K = n, N,$$

$$\hat{A}_n = n^{-1} \sum_{t=1}^{n} \hat{Z}_t \hat{Z}_t' \quad ,$$

$$\hat{H}_n = n^{-1} \sum_{t=1}^{n} \hat{Z}_t Y_t \quad ,$$

$$\hat{B}_n = \hat{A}_n^{-1} \hat{H}_n \quad .$$

In deriving the large sample distribution of the estimator $\hat{Y}_N$, let $\{\zeta_n: n = 1,2,\dots\}$ be a sequence of finite populations, where ζ_n is a random sample of size N_n, $N_n > N_{n-1}$, selected from an infinite population. Let a simple random nonreplacement sample of size n be taken from the n-th finite population, n = 1,2,.... We assume:

(i) $d_n - d_o = n^{-1} \sum_{t=1}^{n} F(X_t, Y_t; d_o) + o_p(n^{-1/2})$, for some function F, where $E\{F(X, Y; d_o)\} = 0$.

(ii.a) $g(x; d)$ is continuous and has continuous partial derivatives of order up to three with respect to d on Θ. Let D_i, i=1,...,q, be the column random vector of partial derivatives of order up to three for the i-th element $g_i(x; d)$ of $g(x; d)$ with respect to d, evaluated at $(X; d_o)$. Let $D = (D_1', \ldots, D_q')'$.

(iii.a) Let the infinite population be such that the vector (Y, U', D', F') has finite fourth moments, where $U = g(X; d_o)$. All the covariance matrices in the multivariate population are positive definite.

The following theorem investigates the effect of estimating Z_t on the regression estimator of $\bar{Y}_N$. The proof is similar to the proof of Theorem 3.1 in Hung (1985). For simplicity, we drop the subscript n from N_n in the sequel.

Theorem 2.1. Let the assumptions given above hold. Then

$$\hat{Y}_N - \bar{Y}_N = \tilde{Y}_N - \bar{Y}_N + o_p(n^{-1/2}),$$

where

$$\bar{Z}_K = K^{-1} \sum_{t=1}^{K} Z_t \qquad K = n, N,$$

$$B_n = \left(\sum_{t=1}^{n} Z_t Z_t'\right)^{-1} \sum_{t=1}^{n} Z_t Y_t \qquad (2.2)$$

$$\tilde{Y}_N = \bar{Y}_n + B_n'(\bar{Z}_N - \bar{Z}_n).$$

Moreover, as n, $N \to \infty$, and $\lim_{n\to\infty}(nN^{-1}) = f$, where $0 \leq f < 1$,

$$n^{1/2}(\hat{Y}_N - \bar{Y}_N) \xrightarrow{\mathcal{L}} N[0, (1-f)V],$$

where

$$V = E\{[Y_t - B'Z_t]^2\} - [E\{Y_t - B'Z_t\}]^2$$

$$B = [E(Z_1 Z_1')]^{-1} E(Z_1 Y_1).$$

A consistent estimator of the variance of $\hat{Y}_N$ can be constructed by estimating V by

$$\hat{V} = L^{-1} \sum_{t=1}^{n} [Y_t - \bar{Y}_n - \hat{B}_n'(Z_t - \bar{Z}_n)]^2,$$

where $L = n - q - 1$.

When q = 1 and $g(x; d_o) > 0$, it can be easily shown that the results of Theorem 2.1 are also applied to the usual ratio estimator of $\bar{Y}_N$. In the prediction approach, one often considers the model

$$Y_t = B'Z_t + e_t \qquad t = 1, \ldots, N,$$

where Z_t and B are defined as above; e_t are random errors satisfying

$$E(e_t) = 0,$$

$$E(e_t e_s) = \sigma^2 v_t(d_o) \quad t = s, \qquad = 0 \quad t \neq s;$$

$\{e_1, e_2, \ldots, e_N\}$ and $\{X_1, X_2, \ldots, X_N\}$ are independent; $v_t(d)$, $t = 1, \ldots, N$, are known positive functions indexed by d. When d_o is known, the best model unbiased linear predictor of $\bar{Y}_N$, conditional on $Z_1, Z_2, \ldots, Z_N$, is given by

$$\bar{Y}_G = N^{-1} n \bar{Y}_n + (1 - N^{-1} n) \hat{B}_{GLS}' \bar{Z}_{N-n},$$

where

$$\bar{Z}_{N-n} = (N - n)^{-1} \sum_{t=n+1}^{N} Z_t,$$

$$\hat{B}_{GLS} = \left(\sum_{t=1}^{n} [v_t(d_o)]^{-1} Z_t Z_t'\right)^{-1} \times \sum_{t=1}^{n} [v_t(d_o)]^{-1} Z_t Y_t.$$

When d_o is unknown, it seems natural to consider the predictor $\hat{Y}_G$ by substituting d_n for d_o. Surprisingly, the results of Theorem 2.1 in general do not apply to the predictor $\hat{Y}_G$; that is, the effect of estimating d_o on the predictor is at least of order $1/\sqrt{n}$ under the above conditions. If $v_t(d_o)$ is an element of Z_t, one can prove that $\bar{Y}_G = \bar{Y}_n + \hat{B}_{GLS}'(\bar{Z}_N - \bar{Z}_n)$. Hence there is no limiting cost due to estimating Z_t; that is, $\hat{Y}_G$ and Y_G have the same limiting variance. If the original model does not contain $v_t(d_o)$ in Z_t, $\hat{Y}_G$ and Y_G will not have the same limiting variance except for a special class of transformation functions.

3. ESTIMATION OF REGRESSION COEFFICIENTS

Consider the estimation of the regression equation for a sample selected from the finite population. The finite population regression coefficient is defined to be

$$B_N = \left(\sum_{t=1}^{N} Z_t Z_t'\right)^{-1} \sum_{t=1}^{N} Z_t Y_t.$$

In case d_o is completely known, an effective estimator of B_N is the well-known least squares estimator B_n defined in (2.2). When auxiliary variables are estimated, one estimator worthy of consideration is $\hat{B}_n$ defined in (2.1). The following theorem demonstrates the asymptotic distribution of $\hat{B}_n$ and the effects of estimating Z_t on the estimator of B_N. The proof is similar to the proof of Theorem 4.1 in Hung (1985).

Theorem 3.1. Let the assumptions of Theorem 2.1 hold.

$$n^{1/2}(\hat{B}_n - B_N) = n^{1/2}(B_n - B_N) + O_p(1),$$

Moreover, as n, $N \to \infty$ and $\lim_{n\to\infty}(nN^{-1}) = f$, where $0 \leq f < 1$,

$$n^{1/2}(\hat{B}_n - B_N) \xrightarrow{\mathcal{L}} N(0, V_{zz}^{-1} V_o V_{zz}^{-1}),$$

where Q_{it} is a vector of the first partial derivatives of $g(X; d)$ with respect to the i-th component d_i of d, evaluated at (X_t, d_o), $t=1,\ldots, N;$ and $e_1 = Y_1 - Z_1'B$,

$$V_o = (1 - f)E[Z_1Z_1'e_1^2] - 2W_1 + W_2,$$

$$V_{zz} = E\{Z_1Z_1'\},$$

$$W_1 = \sum_{j=1}^{r} G_{13(j)}[V_{zQ_j}B + V_{zQ_j}'B - V_{Q_jY}]',$$

$$W_2 = \sum_{i=1}^{r}\sum_{j=1}^{r} G_{33(ij)}[V_{zQ_i}B + V_{zQ_i}'B - V_{Q_iY}][V_{zQ_j}B + V_{zQ_j}'B - V_{Q_jY}]',$$

$G_{13(j)}$ = the j-th column of $E\{Z_1F'(X_1, Y_1; d_o)e_1\}$,

$G_{33(ij)}$ = the ij-th element of $E\{F(X_1, Y_1; d_o)F'(X_1, Y_1; d_o)\}$,

$$V_{zQ_i} = E\{Z_1Q_{i1}'\},$$

$$V_{Q_iY} = E\{Q_{i1}Y_1\}, \quad i,j = 1,2,\ldots, r.$$

In regard to the estimation of the infinite population coefficient B, it follows from Theorem 3.1 that

$$n^{1/2}(\hat{B}_n - B) \xrightarrow{\mathcal{L}} N[0, V_{zz}V^*V_{zz}^{-1}],$$

where $V^* = V_o + fE\{ZZ'(Y - B'Z)^2\}$. The large sample variance of $\hat{B}_n$ can be consistently estimated by estimating the components V_{zz}, W_1, and W_2 in V_o.

It is worth noting that one can employ d_n and $\hat{B}_n$ as initial estimates to obtain, if exists, the Gauss-Newton iterative estimator of B, when additional information on the superpopulation leads to the model

$$Y_t = B'Z_t + e_t \quad t = 1,\ldots,N, \quad (3.1)$$

where e_t, $t = 1,\ldots, N$, are i.i.d. with mean zero and constant variance; e_t and X_t are independent. The asymptotic variance of the Gauss-Newton iterative estimator, B_{GN}, does not depend upon B. Hence, as the true value B falls in a certain range, the estimator $\hat{B}_n$ can be at least as good as the Gauss-Newton estimator under the model (3.1) because of the dependence of $var(\hat{B}_n)$ on B; that is, $var(B_{GN}) - var(\hat{B}_n)$ is positive semidefinite for some B.

When the null hypothesis H_o: B = 0 is tested in the model (3.1), the probability of Type I error for $\hat{B}_n$ will be nearly equal to that for B_n in large samples. However, in principal component regression models where d_o is a location or a scale parameter of the distribution of X and hence is estimated based only on X, the construction of sample principal components will possibly largely increase the variance of $\hat{B}_n$. These results are shown in Corollary 4.1 of Hung (1985).

4. IRREGULAR TRANSFORMATION FUNCTIONS

We extend the study to the case where the transformation function $g(x; d)$ is not necessarily continuous. The estimator $\hat{B}_n$ in (2.1) possesses an interesting characteristic in the sense that it involves substituting estimates for nuisance parameters. Sukhatme (1958) investigated the asymptotic normality of a U-statistic with estimated nuisance parameters. Randles (1982) extended the results of Sukhatme to a broad class of statistics with estimated parameters.

To cover a much broader class of transformation functions, we specify the following assumptions:

(ii.b) There is a constant $M_o > 0$ and a neighborhood of d_o, denote it by $M(d_o)$, such that if $d \in M(d_o)$ and a sphere $S(d; h)$ centered at d with radius h such that $S(d; h)$ is a subset of $M(d_o)$, then

$$E\{\sup_{e\in S(d;h)}[\max_{1\leq j\leq r}|g_j(X_1; e) - g_j(X_1; d)|]\}^2 \leq M_0h^2, \quad (A)$$

$$\lim_{h\to 0}E\{\sup_{e\in S(d;h)}[\max_{1\leq j\leq r}|g_j(X_1;e) - g(X_1;d)|]\}^4 = 0. \quad (B)$$

(iii.b) Let the infinite population be such that the vector (Y, U', F') has finite fourth moments, where $U = g(X; d_o)$. All the covariance matrices in the multivariate population are positive definite. For any j, $E[g(X; d)]$ has a finite supremum over Θ. The functions, $E[g(X; d)Y]$ and $E[g(X; d)g'(X; d)]$, have differentials at $d = d_o$.

Note that any continuous function $g(x; d)$ described in Section 2 apparently satisfies the above conditions. If Y and $g(X; d)$ are uniformly bounded, then Conditions (A) and (B) can be replaced by

$$E\{\sup_{e\in S(d;h)}[\max_{1\leq j\leq r}|g_j(X_1; e) - g_j(X_1; d)|]\} \leq M_oh.$$

Certain classification functions such as those constructed in Hung and Fuller (1984) satisfy these conditions.

Theorem 4.1. Let the sequence of samples and finite populations stated in Section 2 satisfy the conditions (i), (ii.b), and (iii.b). Then, the results of Theorem 3.1 follow, provided that in W_1 and W_2 of $var(\hat{B}_n)$ the term

$$V_{zQ_j}B + V_{zQ_j}'B - V_{Q_jY}$$

is replaced by the qx1 vector of the first derivatives of $E\{g(X; d)[Y - g'(X; d)B]\}$ with respect to d_j, evaluated at d_o, $j = 1,\ldots, r$.

Proof. Let $\eta > 0$. Condition (i) asserts that we can find a bounded sphere S centered at the origin such that

$$P(n^{1/2}(\hat{d}_n - d_o) \notin S) < \eta/2,$$

for every n. Let

$$Z_{ti}(s) = g_i(X_t;\ d_o + n^{-1/2}s),$$
$$T_{ij}(s) = E[Z_{1i}(s)Z_{1j}(s)],$$
$$Z_{ti} = Z_{ti}(0),$$
$$T_{ij} = T_{ij}(0),$$

$$R_{nij}(s) = n^{-1/2}\sum_{t=1}^{n}[Z_{ti}(s)Z_{tj}(s) - T_{ij}(s) - Z_{ti}Z_{tj} + T_{ij}].$$

for $i,j = 1,2,\ldots,r$; $t = 1,2,\ldots,N$; and $s \in R^r$. Let $h = \eta/\sqrt{16M_oM_2}$, where $M_2 = \max_{1\le j\le r}\sup_{d\in\Theta}[E|g_j(X_1;\ d)|^2]$.

Let $S_{a,h} = \{t:\ ah \le |t| \le (a+1)h\}$, $a=0,1,\ldots$

Let $Z^*_{ti,ah} = \sup_{s\in S_{a,h}}|Z_{ti}(s) - Z_{ti}(ah)|$, for any t, i. Ignoring the term of order n^{-1}, we obtain from condition (ii.b) that for any a,

$$E[\sup_{s\in S_{a,h}}|Z_{ti}(s)Z_{tj}(s) - Z_{ti}(ah)Z_{tj}(ah)|]$$
$$\le E[Z^*_{ti,ah}Z^*_{tj,ah}] + E[|Z_{ti}(ah)|Z^*_{tj,ah}] + E[|Z_{tj}(ah)|Z^*_{ti,ah}]$$
$$\le 2M_o^{1/2}M_2^{1/2}hn^{-1/2}.$$

Therefore, for any $s \in S_{a,h}$,

$$|R_{nij}(s) - R_{nij}(ah)|$$
$$\le n^{-1/2}\sum_{t=1}^{n}\{\sup_{s\in S_{a,h}}|Z_{ti}(s)Z_{tj}(s) - Z_{ti}(ah)Z_{tj}(ah)| - E[\sup_{s\in S_{a,h}}|Z_{ti}(s)Z_{tj}(s) - Z_{ti}(ah)Z_{tj}(ah)|]\} + \eta/2.$$

Under Conditions (ii.b) and (iii.b),

$$E[n^{-1/2}\sum_{t=1}^{n}\{\sup_{s\in S_{a,h}}|Z_{ti}(s)Z_{tj}(s) - Z_{ti}(ah)Z_{tj}(ah)| - E[\sup_{s\in S_{a,h}}|Z_{ti}(s)Z_{tj}(s) - Z_{ti}(ah)Z_{tj}(ah)|]\}]^2$$
$$\le E[Z^{*2}_{ti,ah}Z^{*2}_{tj,ah}] + E[Z^2_{ti}(ah)Z^{*2}_{tj,ah}] + 2E[|Z_{tj}(ah)|Z^{*2}_{ti,ah}Z^*_{tj,ah}] + 2E[|Z_{ti}(ah)||Z_{tj}(ah)|Z^*_{ti,ah}Z^*_{tj,ah}] + E[Z^2_{tj}(ah)Z^{*2}_{ti,ah}] + 2E[|Z_{ti}(ah)|Z^{*2}_{tj,ah}Z^*_{ti,ah}]$$
$$= o(1).$$

Thus, $\sup_{s\in S_{a,h}}|R_{nij}(s) - R_{nij}(ah)| = o_p(1)$ and $|R_{nij}(ah)| = o_p(1)$, for any a.

It follows that $\sup_{s\in S_{a,h}}|R_{nij}(s)| = o_p(1)$. Hence,

$$n^{-1}\sum_{t=1}^{n}\hat{Z}_t(Y_t - \hat{Z}_t'B)$$
$$= n^{-1}\sum_{t=1}^{n}Z_t(Y_t - Z_t'B) + n^{-1}\sum_{t=1}^{n}(\hat{Z}_t - Z_t)Y_t - n^{-1}\sum_{t=1}^{n}(\hat{Z}_t\hat{Z}_t' - Z_tZ_t')B$$
$$= n^{-1}\sum_{t=1}^{n}Z_t(Y_t - Z_t'B) - E[\hat{Z}_1(Y_1 - \hat{Z}_1'B)] + o_p(n^{-1/2}).$$

Following the same arguments as in the proof of Theorem 3.1, we can complete the proof. □

<u>Theorem 4.2</u>. Let the sequence of samples and finite populations stated in Section 2 satisfy the assupmtions (i), (ii.b), and (iii.b). Then the results of Theorem 2.1 follow.

<u>Proof</u>. Following the same arguments as in the proof of Theorem 4.1, we can show that

$$\hat{\bar{Z}}_N - \bar{Z}_N - \hat{\bar{Z}}_n + \bar{Z}_n = o_p(1),$$

and

$$\hat{B}_n - B = o_p(1).$$ □

5. SUMMARY AND DISCUSSIONS

We provided some discussions on the results presented in previous sections.

In estimating the population mean $\bar{Y}_N$, the estimation of the transformed variable Z will affect the regression estimator by a term of order in probability smaller than $1/\sqrt{n}$. Hence, the large sample variance will not be increased when the unknown parameters in the transformation are estimated with the error of order no larger than $1/\sqrt{n}$. However, under a general multiple regression model, the effect of estimating Z on the best model unbiased linear predictor of $\bar{Y}_N$ may be at least of order $1/\sqrt{n}$. If the model contains the variance of the random error as a part of independent variables, then the best model unbiased linear predictor will be affected only by a negligible term as indicated in Theorem 2.1.

In estimating the finite population regression coefficient B_N, the estimation of Z will affect the estimator by a term of order $1/\sqrt{n}$. The order of the effect is the same as that of the error of B_n. The first component in the asymptotic variance of $\hat{B}_n$ is the variance of B_n, the usual sample regression coefficient. The inflation component has the same order as the variance of B_n does, and hence in general the effects of estimating d_o can not be neglected. In cases where the finite population data can be characterized by the model (3.1), the Gauss-Newton estimator of B may be constructed with the large sample variance independent of the true value B. Noneless, $\hat{B}_n$ may be more efficient than the Gauss-Newton estimator, depending on the true value B. When prior knowledge about the range of B is available, it is advised to take $\hat{B}_n$ into account.

6. ACKNOWLEDGMENT

This work was supported in part by the 1984 Graduate School Faculty Summer Research Fellowship of the University of Minnesota.

7. REFERENCES

Fuller, W. A. (1975)."Regression Analysis for Sample Survey," Sankhya, Ser. C, 37, 117-132.

Hajek, J. (1960). "Limiting Distributions in Simple Random Sampling from a Finite Population," Periodica Mathmatica, 5, 361-374.

Hidiroglou, M. (1974). "Estimation of Regression Parameters for Finite Populations," Ph.D. thesis, Iowa State University, Ames, Iowa.

Hung, H. M. (1985)."Regression Estimation with Transformed Auxiliary Variates," Statistics & Probability Letters. To appear.

Hung, H. M., and Fuller, W. A. (1984). "Regression Estimation of Crop acreages with Transformed Landsat Data as Auxiliary Variables," Submitted to Journal of American Statistical Association.

Randles, H. R. (1982). "On the Asymptotic Normality of Statistics with Estimated Parameters," Annals of Statistics, 10, 462-474.

Scott, A. J., and Wu, C. F. (1981). "On the Asymptotic Distribution of Ratio and Regression Estimators," Journal of American Statistical Association, 76, 98-102.

Sigman, R. S., Hanuschak, G. A., Craig, M. E., Cook, P. W., and Cardenas, M. (1978). "The Use of Regression Estimation with LANDSAT and Probability Ground Sample Data," Survey Section Proceedings, 1978 American Statistical Association.

Sukhatme, B. V. (1958). "Testing the Hypothesis that Two Populations Differ Only in Location," Annals of Mathematical Statistics, 29, 60-78.