Survey Methodology, December 1986
Vol. 12, No. 2, pp. 121–138
Statistics Canada

Regression Analysis Using Survey Data with Endogenous Design

ARIE TEN CATE[1]

ABSTRACT

This paper discusses the influence of the sampling design on the estimation of a linear regression model. Particularly, sampling designs will be discussed which are dependent on the values of the endogenous variable in the population: endogenous (or "informative") designs. A consistent estimator of the regression coefficients is given. Its variance is the sum of a sampling design component and a disturbance term component. Also, model-free regression is briefly discussed. The model-free regression estimator is the same as the model estimator in the case of an endogenous design.

KEY WORDS: Regression; Survey sampling; Endogenous design.

1. INTRODUCTION

The heart of any statistical model is the assumption that the value of one or more variables is generated by drawing from some probability distribution; for example, a regression model with normally distributed disturbances. In this paper a finite set of elements which behave according to such a model will be considered. This set is called the population. Next, a sample is drawn from this population, without replacement. The subject of this paper is the influence of the sampling design on the estimation of the parameters of the model. This influence depends mainly on whether the design is exogenous or endogenous with respect to the model. In the case of an endogenous (or "informative") design, the sampling probabilities depend on the value of the endogenous ("dependent") variables. Then, the design should not be ignored in the estimation of the model parameters. The nature of the problem is indicated in Figure 1, where a stratified sampling design is shown. There are 3 strata, defined in the endogenous variable of a regression model. The middle stratum has a higher sampling fraction than the other two. The diagram shows that the slope of the regression line estimated using the sampled data points only, is biased downwards if one ignores the design. This bias does not vanish in large samples. This can be seen in an intuitive manner by imagining that every white and black dot in Figure 1 denotes a large number of identical data points. Even if this large number tends to infinity, the slope of the estimated regression line will be biased downwards, because the shape of the scatter will remain the same.

There is a rapidly growing body of literature on the application of regression techniques in finite population sampling. This literature deals with a variety of problems. One problem is, how to use regression techniques in order to estimate a finite population total. Another problem concerns the estimation of population parameters such as $\Sigma xy/\Sigma x^2$, where the summation runs over all elements of the finite population. Reviews of the literature about these problems are given by Nathan (1981) and Smith (1981). A third problem is the estimation of the parameters of a regression model, using a sample from a finite population. This problem can be solved relatively easily in the case of a exogenous design. See Porter (1973, Section 1.2), DuMouchel and Duncan (1983), and textbooks such as Cramer (1971, p. 143). Texts

[1] Arie ten Cate, Central Planning Bureau, 2585 JR 's-Gravenhage, Van Stolkweg 14, The Hague, The Netherlands.

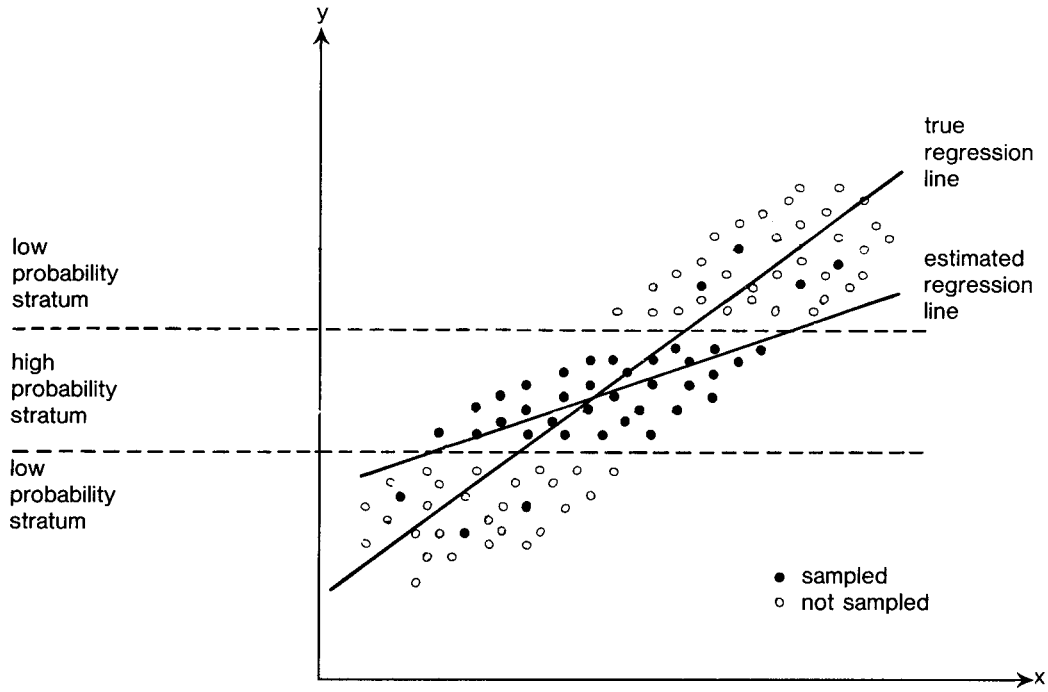

Figure 1. The Effect of Endogenous Stratification on the Estimated Regression Line

such as Kmenta (1971, Section 8.3) and Johnston (1972, Section 9.2) discuss the closely related topic of stochastic regressors. See also White (1980a) for non-linear regression. Our topic, regression analysis with endogenous design, is more complicated. Hausman and Wise (1981) discuss stratified endogenous designs in a very simple case: two strata and a regression model consisting of a constant term only. Jewell (1985) gives some iterative estimators for the case of endogenous stratification.

Regression analysis with endogenous design is related to the problem of endogenous non-response in regression analysis (see Heckman (1979)). However, we have a lesser problem here, since the probabilities involved in the sampling process are assumed to be known: they constitute the chosen design. On the other hand, as we shall see in Subsection 6.1, variance estimation with an endogenous design is in general rather difficult.

Regression analysis with endogenous design may be compared with logit analysis with endogenous design, also called logit analysis with choice based sampling or case-control sampling. See Manski and McFadden (1981, Chapters 1 and 2) and Breslow and Day (1980, Section 6.3).

The contents of the rest of the paper are as follows. In Sections 2 and 3 the main theorems are given. These theorems give a consistent estimator of the parameters of a linear regression model, using a sample with an endogenous design. Consistency is defined here in a similar way as in the discussion of the bias in the example above, though slightly more subtle: the x-values are replicated a large number of times and the y-values behave according to the regression model. In Sections 4 and 5 the variance of the estimator of the regression coefficients is studied. Section 6 discusses the estimation of this variance. Section 7 deals with model-free regression, Section 8 discusses the various motives for weighted regression and finally, Section 9 concludes the paper.

2. THE MODEL, THE SAMPLE AND A REGRESSION ESTIMATOR

In this section the asymptotic properties of an estimator of a regression model are studied within the framework of finite population sampling without replacement. Asymptotic theory for samples drawn without replacement from a finite population may seem a contradiction since such a sample must be bounded. This contradiction is solved by increasing both the population size and the sample size, without bound, at the same rate. The dependence between the inclusions of population elements in the sample constitutes another problem, especially in the case of complex sampling designs. Here we use an idea of Brewer (1979). In Brewer's system, limit theorems on sequences of independent variables can be used, while the results may still be applied to complex designs. Basically, this system consists of the replica idea already introduced informally above. This replica idea will be used extensively throughout the rest of this paper. For another approach, see Robinson (1982).

First, the structure of the population and the model are given. Consider a finite set of N_0 elements. Each element has r real-valued exogenous non-stochastic characteristics, together forming an $(N_0 \times r)$-matrix X_0. One of the fundamental assumptions of this paper is the following. The population consists of K replicas of this set of N_0 elements, having $N \equiv KN_0$ elements. Its matrix of exogenous variables is X, with

$$X = \iota_K \otimes X_0. \tag{1}$$

Here, ι_K is the K-vector with all elements equal to unity and $\otimes$ denotes the Kronecker matrix product. Aymptotic results will be derived by allowing K to tend to infinity.

The model assumptions describe the standard linear model. Each of the N elements of the population has a score on a stochastic, endogenous, variable. Together they form an N-vector y. It is assumed that

$$E_\xi(y) = X\beta \tag{2}$$

for some fixed, unknown r-vector β. E_ξ denotes the expectation over all $y \in R^N$. Next we define

$$\varepsilon = y - X\beta. \tag{3}$$

It is assumed that the N elements of ε are i.i.d. It follows from (2) that all elements of ε have expectation zero. Their variance is σ^2, that is,

$$E_\xi(\varepsilon\varepsilon') = \sigma^2 I. \tag{4}$$

Sampling is done without replacement here, as is common practice. The sample is described by a diagonal $(N \times N)$-matrix T, such that

$$t_{ii} = \begin{cases} 1 \text{ if population element } i \text{ is in the sample} \\ 0 \text{ otherwise} \end{cases}$$

for all $i = 1, \ldots, N$. Obviously, T is idempotent. The sample space S is the set of all such matrices T. This set is finite. The sampling design is some probability distribution over the elements of the sample space S. The sampling design is endogenous here, meaning that it depends on y. Hence, the sampling design itself is stochastic. (A design which does not depend on y is called exogenous, or uninformative.) Let T be partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. Let T_k be the k-th diagonal block, related to the k-th replica. Similarly, let y be partitioned in K N_0-vectors, such that $y' = (y_1', y_2', \ldots, y_k', \ldots, y_K')$. It is assumed that the sampling design depends on y in the following sense: the K pairs $(T_1, y_1), \ldots, (T_K, y_K)$ are i.i.d.

The expectation over all elements of S, conditional on y (or ε), plays an important role in this paper. It is denoted by E_p. Then we define

$$\Pi \equiv E_p(T). \tag{5}$$

It is assumed that Π is known. The diagonal elements of Π are called inclusion probabilities: the probabilities that the population elements are included in the sample. The matrix Π is partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. Let Π_k be the k-th diagonal block, related to the k-th replica. Note that each Π_k is stochastic because it depends on y_k. By the above assumption, the $\Pi_1, \ldots, \Pi_k$ are i.i.d. The dependence of the Π_k on y is denoted by a function F, such that

$$\Pi_k = F(y_k) \tag{6}$$

for all $k = 1, \ldots, K$. It is assumed that $F(y_k)$ is non-singular for every y_k. In other words, the inclusion probabilities are always positive.

This framework and Brewer's (1979) differ in somewhat. Brewer has no endogenous variables and therefore all his Π_k are nonstochastic and equal. One may also compare this approach with the idea of "constant in repeated samples" in the econometric literature; see e.g. Theil (1971, p. 364).

The stage is now set for the estimation of β. The stochastic properties of estimators will be considered over all pairs $(y, T) \in (R^N \times S)$. The corresponding expectation will be denoted by $E_\xi E_p$. We shall consider a generalized least square estimator of β, say $\hat{\beta}$, with weights equal to the square roots of the inclusion probabilities, as follows,

$$\begin{aligned} \hat{\beta} &\equiv [(\Pi^{-\frac{1}{2}}X)'T(\Pi^{-\frac{1}{2}}X)]^{-1}(\Pi^{-\frac{1}{2}}X)'T(\Pi^{-\frac{1}{2}}y) \\ &= (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}Ty. \end{aligned} \tag{7}$$

Recall that the matrix Π is known. Note that X and y relate to the population, but T effectuates summation over the sampled elements. As an alternative to considering $\hat{\beta}$ as a generalized least squares estimator, assume that all elements of Π^{-1} are integer numbers. Then, if each observation i in the sample is copied π_{ii}^{-1} times, $\hat{\beta}$ is the ordinary least squares estimator applied to this inflated sample. In this view, no square roots of the probabilities are involved. See also Hausman and Wise (1981, p. 373). The main theorem of this paper is:

Theorem 1. Under the assumptions made above ((1), (2) and the distribution of ε and T), the generalized least squares estimator $\hat{\beta}$, defined in equation (7) is consistent for $K \to \infty$.

The rest of the section is devoted to the proof of this theorem. The following lemma will be used in this proof and the proof of subsequent theorems.

Lemma 1. Consider an N-vector z, such that $z = \iota_K \otimes z_0$, where z_0 is some fixed N_0-vector. Consider also an N-vector η, partitioned such that $\eta' = (\eta_1', \eta_2', \ldots, \eta_K')$. Each η_k has N_0 elements. Assume that each η_k is a function of X_0, β and ε_k, all functions being the same. Then

$$\operatorname*{plim}_{K\to\infty} \left(\frac{1}{K} z' \Pi^{-1} T \eta\right) = z_0' E_\xi(\eta_0), \tag{8}$$

where $E_\xi(\eta_0)$ is the expectation of any η_k, being equal for all k.

Proof of lemma 1: Consider the expectation of $\Pi_k^{-1} T_k \eta_k$:

$$E_\xi E_p(\Pi_k^{-1} T_k \eta_k) = E_\xi[\Pi_k^{-1} E_p(T_k)\eta_k] = E_\xi(\eta_k), \tag{9}$$

for all k. Since the distribution of η_k is the same for each k, one may write

$$E_\xi E_p(\Pi_k^{-1} T_k \eta_k) = E_\xi(\eta_0) \tag{10}$$

for all k. Also, the K vectors $z_0' \Pi^{-1} T_k \eta_k$ are i.i.d. Thus, Khintchine's theorem applies as follows,

$$\operatorname*{plim}_{K\to\infty} \left(\frac{1}{K} z' \Pi^{-1} T \eta\right) = \operatorname*{plim}_{K\to\infty} \left(\frac{1}{K} \sum_k z_0' \Pi_k^{-1} T_k \eta_k\right) = E_\xi E_p(z_0' \Pi_1^{-1} T_1 \eta_1)$$

$$= z_0' E_\xi E_p(\Pi_1^{-1} T_1 \eta_1). \tag{11}$$

Substitution of (10) in (11) gives the lemma. The proof of theorem 1 is now straightforward.

Proof of theorem 1: The generalized least squares estimator of the theorem can be written as

$$\hat{\beta} = (X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T y = \beta + (X' \Pi^{-1} T X)^{-1} X' \Pi^{-1} T \varepsilon. \tag{12}$$

Thus,

$$\operatorname*{plim}_{K\to\infty} \hat{\beta} = \beta + \left[\operatorname*{plim}_{K\to\infty} \left(\frac{1}{K} X' \Pi^{-1} T X\right)\right]^{-1} \operatorname*{plim}_{K\to\infty} \left(\frac{1}{K} X' \Pi^{-1} T \varepsilon\right)$$

$$= \beta + (X_0' X_0)^{-1} X_0' 0 = \beta. \tag{13}$$

The expression $X_0' X_0$ is formed by repeated application of lemma 1, substituting the columns of X for both z and η. Notice that $E_\xi(X_0) = X_0$ since X_0 is a constant. The expression $X_0' 0$ is formed by repeated application of lemma 1, substituting the columns of X for z and ε for η.

3. THE ESTIMATION OF THE DISTURBANCE VARIANCE

The regression model described in Section 2 has two parameters: β and σ^2. Theorem 1 considered estimation of β; in this section the estimation of σ^2 will be considered. The result of this section is given in the following theorem.

Theorem 2. The disturbance variance σ^2 is estimated consistently by the weighted sample variance of the residuals of y if these weights are equal to the inverse of the square root of the inclusion probabilities.

Proof: The variance estimator of the theorem is

$$\hat{\sigma}^2 = (\iota_N' \Pi^{-1} T \iota_N)^{-1} \tilde{e}' \tilde{e} \tag{14}$$

with

$$\tilde{e} \equiv \Pi^{-\frac{1}{2}} T (y - X\hat{\beta}). \tag{15}$$

Let

$$\tilde{y} \equiv \Pi^{-\frac{1}{2}} T y, \tag{16}$$

$$\tilde{X} \equiv \Pi^{-\frac{1}{2}} T X, \tag{17}$$

and

$$\tilde{\varepsilon} \equiv \Pi^{-\frac{1}{2}} T \varepsilon. \tag{18}$$

Then

$$\tilde{e} = \tilde{y} - \tilde{X}\hat{\beta} = \tilde{y} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \tag{19}$$

and

$$\begin{aligned} \tilde{e}'\tilde{e} &= \tilde{y}' [I_N - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'] \tilde{y} = (\tilde{X}\beta + \tilde{\varepsilon})' [I_N - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'] (\tilde{X}\beta + \tilde{\varepsilon}) \\ &= \tilde{\varepsilon}'\tilde{\varepsilon} - \tilde{\varepsilon}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\varepsilon}. \end{aligned} \tag{20}$$

The first term in the right-hand side (RHS) of (20) converges in probability as follows

$$\begin{aligned} \operatorname*{plim}_{K\to\infty} \left(\frac{1}{K}\tilde{\varepsilon}'\tilde{\varepsilon}\right) &= \operatorname*{plim}_{K\to\infty} \left(\frac{1}{K}\varepsilon'\Pi^{-1}T\varepsilon\right) = \operatorname*{plim}_{K\to\infty} \left[\frac{1}{K}\iota_N'\Pi^{-1}T \operatorname{diag}(\varepsilon)\varepsilon\right] \\ &= \iota_{N_0}'(\sigma^2\iota_{N_0}) = N_0\sigma^2. \end{aligned} \tag{21}$$

Here, diag(ε) indicates the diagonal matrix with as the diagonal. Lemma 1 has been applied with ι_N substituted for z and diag(ε)ε for η, using model equation (4). Next, consider the second term in the RHS of (20).

$$\operatorname*{plim}_{K\to\infty}\left[\frac{1}{K}\tilde{\varepsilon}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\varepsilon}\right]$$

$$= \left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{X}'\tilde{\varepsilon}\right)\right]'\left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{X}'\tilde{X}\right)\right]^{-1}\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{X}'\tilde{\varepsilon}\right)$$

$$= \left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}T\varepsilon\right)\right]'\left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}TX\right)\right]^{-1}\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}T\varepsilon\right)$$

$$= 0'(X_0'X_0)^{-1}0 = 0. \tag{22}$$

In the derivation of (22), use has been made of lemma 1 in the same manner as in the derivation of (13). The combination of (20), (21) and (22) gives

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{e}'\tilde{e}\right) = N_0\sigma^2. \tag{23}$$

Finally, lemma 1 is applied to the first factor in (14), with ι_N substituted both for z and η. This gives

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\iota_N'\Pi^{-1}T\iota_N\right) = N_0. \tag{24}$$

With (23) and (24) we have

$$\operatorname*{plim}_{K\to\infty}(\hat{\sigma}^2) = \sigma^2, \tag{25}$$

which proves the theorem. Finally it may be useful to note, as a corollary of (23), that

$$\left(\frac{1}{N}\right)\tilde{e}'\tilde{e} \tag{26}$$

is also a consistent estimator of σ^2.

4. THE VARIANCE OF $\hat{\beta}$

In this section the asymptotic variance of the estimator $\hat{\beta}$ is given.

Theorem 3. The asymptotic variance of $\hat{\beta}$ is given by

$$\operatorname{Var}(\hat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}, \tag{27}$$

with

$$V \equiv E_\xi[\operatorname{diag}(\varepsilon)\Pi^{-1}P\Pi^{-1}\operatorname{diag}(\varepsilon)], \tag{28}$$

and

$$P \equiv E_p(T\iota\iota' T). \tag{29}$$

The elements of P are the so-called second order inclusion probabilities: the probability for any pair of elements of the population of being included in the sample. The diagonal of P is equal to the diagonal of Π. The rest of this section is devoted to a proof of this theorem.

Proof: Consider the asymptotic distribution for $K \to \infty$ of

$$\begin{aligned} K^{1/2}(\hat{\beta} - \beta) &= K^{1/2}[(X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}Ty - \beta] \\ &= K^{1/2}(X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}T\varepsilon. \end{aligned} \tag{30}$$

Since

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}TX\right) = X_0'X_0, \tag{31}$$

the asymptotic distribution of $K^{1/2}(\hat{\beta} - \beta)$ is equal to the asymptotic distribution of δ, with

$$\delta \equiv K^{-1/2}(X_0'X_0)^{-1}X'\Pi^{-1}T\varepsilon = K^{-1/2}(X_0'X_0)^{-1}\sum_k X_0'\Pi_k^{-1}T_k\varepsilon_k = K^{-1/2}\sum_k \delta_k, \tag{32}$$

and

$$\delta_k \equiv (X_0'X_0)^{-1}X_0'\Pi_k^{-1}T_k\varepsilon_k, \tag{33}$$

for all $k = 1, \ldots, K$. (See e.g. Rao (1973), p. 122). Since the vector δ_k $(k = 1, \ldots, K)$ are i.i.d. and also

$$\begin{aligned} E_\xi E_p(\delta_k) &= (X_0'X_0)^{-1}X_0'E_\xi E_p(\Pi_k^{-1}T_k\varepsilon_k) \\ &= (X_0'X_0)^{-1}X_0'E_\xi[\Pi_k^{-1}E_p(T_k)\varepsilon_k] \\ &= (X_0'X_0)^{-1}X_0'E_\xi(\varepsilon_k) = 0, \end{aligned} \tag{34}$$

the variance of δ, say $\mathrm{Var}(\delta)$, is equal for all K and also equal to the variance of the asymptotic distribution of δ for $K \to \infty$. This variance can be written as

$$\mathrm{Var}(\delta) = E_\xi E_p(\delta_k\delta_k') \tag{35}$$

for any $k \in \{1, \ldots, K\}$. Since the vectors δ_k are i.i.d. this may be rewritten as

$$\text{Var}\ (\delta) = \frac{1}{K}\sum_k E_\xi E_p(\delta_k\delta_k')$$

$$= \frac{1}{K}(X_0'X_0)^{-1}\left[E_\xi E_p\left(\sum_k X_0'\Pi_k^{-1}T_k\varepsilon_k\varepsilon_k'T_k\Pi_k^{-1}X_0\right)\right](X_0'X_0)^{-1}$$

$$= K(X'X)^{-1}[E_\xi E_p(X'\Pi^{-1}T\varepsilon\varepsilon'T\Pi^{-1}X)\,](X'X)^{-1}$$

$$= K(X'X)^{-1}X'\{E_\xi E_p[\text{diag}\,(\varepsilon)\Pi^{-1}T\iota\iota'T\Pi^{-1}\text{diag}\,(\varepsilon)\,]\}X(X'X)^{-1}$$

$$= K(X'X)^{-1}X'\{E_\xi[\text{diag}\,(\varepsilon)\Pi^{-1}E_p(T\iota\iota'T)\Pi^{-1}\text{diag}(\varepsilon)\,]\}X(X'X)^{-1}. \quad (36)$$

Division of $\text{Var}(\delta)$ by K gives $\text{Var}(\hat{\beta})$ and completes the proof.

5. A DECOMPOSITION OF VAR $(\hat{\beta})$

The variance formula (27) can be rewritten as

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1} + (X'X)^{-1}X'V^*X(X'X)^{-1} \quad (37)$$

with

$$V^* \equiv E_\xi[\text{diag}(\varepsilon)\,(\Pi^{-1}P\Pi^{-1} - \iota\iota')\,\text{diag}(\varepsilon)\,], \quad (38)$$

using (4). The first term in the RHS of (37) might reasonably be called the ξ-component of the variance of $\hat{\beta}$. This component would contain all the variance of $\hat{\beta}$ if the whole population was sampled. It is entirely due to the variation in the disturbance ε and it is the familiar expression for that case. The second term in the RHS of (37) might be called the p-component of the variance of $\hat{\beta}$. This component contains the matrices Π and P, which describe the sampling design. This component looks like the variance formula of the estimator of a total or average of a finite population. The theory of such estimators will be discussed briefly in the rest of this section, as an aid in the interpretation of the p-component of $\text{Var}(\hat{\beta})$.

Consider a finite population of N elements. (No replica structure is assumed here). Each element of this population has a score on some real non-stochastic variable, collected in an N-vector x. From this population a sample without replacement is taken. The sample is described by the diagonal matrix T, as before. Also as before,

$$\Pi \equiv E_p(T) \quad (39)$$

and

$$P \equiv E_p(T\iota\iota'T), \quad (40)$$

the first order and second order inclusion probabilities, respectively. There is no regression model here, so Π and P are fixed known matrices. Horvitz and Thompson (1952) suggested to estimate the population total $X'\iota$ by

$$\hat{X} = x'\Pi^{-1}T\iota \quad (41)$$

Obviously this is an unbiased estimator, in view of (39). The variance of $\hat{X}$ is

$$\begin{aligned} \mathrm{Var}(\hat{X}) &= E_p(\hat{X}^2) - [E_p(\hat{X})]^2 = E_p(x'\Pi^{-1}T\iota\iota'T\Pi^{-1}x) - x'\iota\iota'x \\ &= x'(\Pi^{-1}P\Pi^{-1} - \iota\iota')x. \end{aligned} \tag{42}$$

The last member of equation (42) is the variance formula of the Horvitz-Thompson estimator, which can be found in textbooks on sampling, such as Cochran (1977), though usually not in matrix format. The expression in parentheses in the last member of (42) is equal to the expression in parentheses in (38), the definition of V^*. The latter is contained in the formula of the p-component of $\mathrm{Var}(\hat{\beta})$. Thus, the diagonal elements of the p-component of the variance matrix $\mathrm{Var}(\hat{\beta})$ can be considered as the ξ-expectation of the p-variance of the Horvitz-Thompson estimator of the row totals of $(X'X)^{-1}X'$ diag (ε). These totals are the elements of the vector $(X'X)^{-1}X'\varepsilon$.

6. THE ESTIMATION OF VAR($\hat{\beta}$)

6.1 The General Case

In this section the estimation of the asymptotic variance $\mathrm{Var}(\hat{\beta})$ is considered. Consistent estimation of $\mathrm{Var}(\hat{\beta})$ is rather difficult, since this requires knowledge of the relationship F between y and the sampling design, as it appears in the matrix V. In practice, only the sampling design for the actual values of y will be known. In general, it is difficult to tell from this design only, what the design would be like if y took on different values. In a sense not only a regression model is involved, but also a model of the designer himself!

For the moment we assume that the function F is known, and therefore V is a known function of X and the parameters of the model. (See Subsection 6.2 for a special case). This is expressed as follows.

$$V = V(\beta, \sigma^2; X), \tag{43}$$

It is assumed that $V(\beta, \sigma^2; X)$ is a continuous function. For the sake of brevity, $\hat{V}$ is defined as

$$\hat{V} \equiv V(\hat{\beta}, \hat{\sigma}^2; X), \tag{44}$$

where and $\hat{\beta}$ and $\hat{\sigma}^2$ are consistent estimators of β and σ^2 respectively. The rest of this subsection gives a theorem on consistent variance estimation, and its proof. Consistent estimation of $\mathrm{Var}(\hat{\beta})$ by $\hat{\mathrm{var}}(\hat{\beta})$ is interpreted here as follows:

$$\operatorname*{plim}_{K\to\infty} K\hat{\mathrm{var}}(\hat{\beta}) = \lim_{K\to\infty} K\,\mathrm{Var}(\hat{\beta}). \tag{45}$$

Theorem 4. Under the assumptions made above, the asymptotic variance $\mathrm{Var}(\hat{\beta})$ is estimated consistently by

$$\hat{\mathrm{var}}(\hat{\beta}) = (X'\Pi^{-1}TX)^{-1}X'T\left(\frac{\hat{V}}{P}\right)TX(X'\Pi^{-1}TX)^{-1}, \tag{46}$$

where $(\hat{V}/P)$ denotes the matrix consisting of the elements of $\hat{V}$ divided by the corresponding elements of P.

Proof: First the structure of V will be considered. Let V be partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. The (k, r)-th off-diagonal block of V is equal to

$$\begin{aligned} & E_\xi \left[\operatorname{diag}(\varepsilon_k) \Pi_k^{-1} E_p(T_k \iota \iota' T_r) \Pi_r^{-1} \operatorname{diag}(\varepsilon_r) \right] \\ & = E_\xi \left[\operatorname{diag}(\varepsilon_k) \Pi_k^{-1} E_p(T_k) \iota \iota' E_p(T_r) \Pi_r^{-1} \operatorname{diag}(\varepsilon_r) \right] \\ & = E_\xi (\varepsilon_k \varepsilon_r') = 0, \end{aligned} \tag{47}$$

using the assumed replica structure of the population and the sampling design. The diagonal blocks of V are identical and depend on X_0. Thus, $V(\beta, \sigma^2; X)$ can be written as

$$V(\beta, \sigma^2; X) = I_K \otimes V_0(\beta, \sigma^2; X_0), \tag{48}$$

where $V_0(\beta, \sigma^2; X_0)$ is an $N_0 \times N_0$ matrix function. Together with (1), equation (48) can be used to rewrite $K\operatorname{Var}(\hat{\beta})$ as follows.

$$K\operatorname{Var}(\hat{\beta}) = (X_0' X_0)^{-1} X_0' V_0 X_0 (X_0' X_0)^{-1}, \tag{49}$$

where V_0 denotes $V_0(\beta, \sigma^2; X_0)$. The RHS of (49) is independent of K and therefore equal to its limit as K tends to infinity. Next, the LHS of (45) is considered.

$$K\hat{\operatorname{var}}(\hat{\beta}) = \left(\frac{1}{K} X' \Pi^{-1} T X \right)^{-1} \left[\frac{1}{K} X' T \left(\frac{\hat{V}}{P} \right) T X \right] \left(\frac{1}{K} X' \Pi^{-1} T X \right)^{-1}. \tag{50}$$

Earlier, in the derivation of (13) and (22), use has already been made of

$$\operatorname*{plim}_{K \to \infty} \left(\frac{1}{K} X' \Pi^{-1} T X \right) = X_0' X_0. \tag{51}$$

It follows from the assumption that $V(\beta, \sigma^2; X)$ is a continuous function, that

$$\operatorname*{plim}_{K \to \infty} \hat{V}_0 = V_0, \tag{52}$$

where $\hat{V}_0$ denotes $V_0(\hat{\beta}, \hat{\sigma}^2; X_0)$. Using (1), (48) and (52) gives

$$\begin{aligned} \operatorname*{plim}_{K \to \infty} \frac{1}{K} X' T \left(\frac{\hat{V}}{P} \right) T X &= \operatorname*{plim}_{K \to \infty} \frac{1}{K} \sum_k \left[X_0' T_k \left(\frac{\hat{V}_0}{P_0} \right) T_k X_0 \right] \\ &= \operatorname*{plim}_{K \to \infty} \frac{1}{K} \sum_k \left[X_0' T_k \left(\frac{V_0}{P_0} \right) T_k X_0 \right] = X_0' V_0 X_0. \end{aligned} \tag{53}$$

Here P_0 denotes $E_p\,(T_k\iota\iota'\,T_k)$, which is the same for all $k = 1, \ldots, K$. The last equality sign results from the application of Khintchine's theorem, since the terms in the second summation over k in (53) are i.d.d. with p-expectation equal to $X_0' V_0 X_0$. Finally, the combination of (50), (51) and (53) gives

$$\operatorname*{plim}_{K\to\infty} K\hat{\mathrm{var}}(\hat{\beta}) = (X_0'X_0)^{-1}X_0' V_0 X_0 (X_0'X_0)^{-1}, \tag{54}$$

which is the same expression as the RHS of (49).

6.2 Stratified Sampling

In this subsection the computation of the matrix $T(\hat{V}/P)T$ is given for a special case: (1) the disturbances are normally distributed, and (2) the sampling design is an endogenously stratified sampling design, such that the inclusion probability π_{ii} of element i of the population is a function f of only the i-th element of y, say $y_{(i)}$. Thus,

$$\pi_{ii} = f(y_{(i)}), \tag{55}$$

for $i = 1, \ldots, N$. As an example, consider the stratified sample which was shown in Figure 1. The design contains three strata there. The elements in the middle stratum have the highest inclusion probability. Figure 2 shows the corresponding function f.

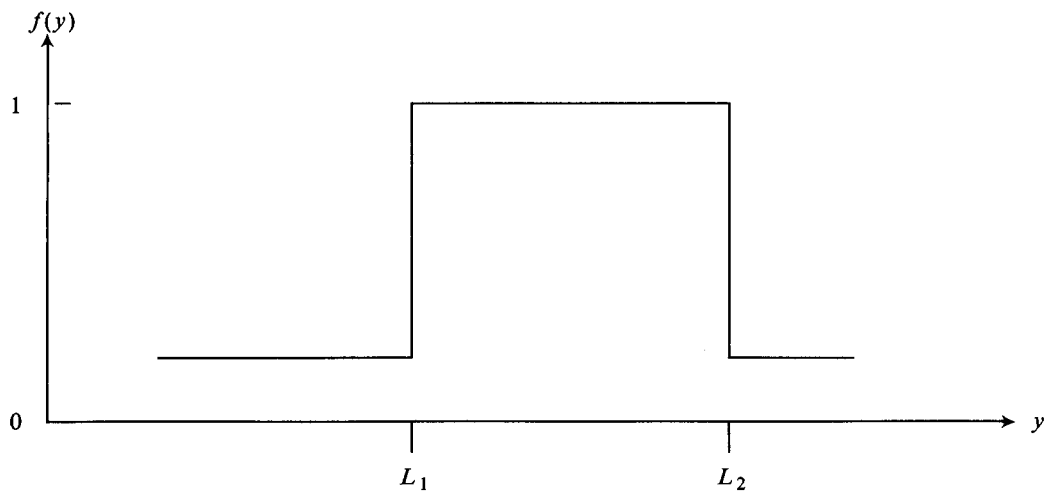

Figure 2. The Probability Function f Corresponding to Figure 1

In general, let there be H strata, indicated by $h = 1, \ldots, H$. Let the boundaries of these strata be $L_0, L_1, \ldots, L_H$. Typically, $L_0 = -\infty$ and $L_H = +\infty$. Let $\pi_{(h)}$ be the inclusion probability of the population elements in stratum h. More formally, the function $f(\cdot)$ is such that $f(y)$ equals $\pi_{(h)}$ if $L_{h-1} \leq y < L_h$. The values of $\pi_{(h)}$ and L_h are usually known in practice, since the actual sampling design depends on their values.

In stratified sampling, the second order inclusion probability of any two population elements not in the same stratum equals the product of their respective first order inclusion probabilities: their inclusions in the sample are independent. For any two population elements in the same stratum this holds approximately. Thus, approximately the off-diagonal elements of P are equal to the off-diagonal elements of $\Pi\iota\iota'\Pi$. The diagonal of P is equal to the diagonal of Π, as before. Thus, approximately,

$$P = \Pi\iota\iota'\Pi - \Pi^2 + \Pi. \tag{56}$$

Then

$$V = E_\xi[\operatorname{diag}(\varepsilon)(\iota\iota' - I + \Pi^{-1})\operatorname{diag}(\varepsilon)]$$

$$= E_\xi[\varepsilon\varepsilon' - \operatorname{diag}^2(\varepsilon) + \operatorname{diag}^2(\varepsilon)\Pi^{-1}] = E_\xi[\operatorname{diag}^2(\varepsilon)\Pi^{-1}], \tag{57}$$

in view of assumption (4). Thus V is a diagonal matrix here. Then

$$T\left(\frac{V}{P}\right)T = T\Pi^{-1}E_\xi[\operatorname{diag}^2(\varepsilon)\Pi^{-1}], \tag{58}$$

which is also a diagonal matrix. Now consider a population element i, which is included in the sample. Then, using (58) and assuming normally distributed disturbances,

$$\left[T\left(\frac{\hat{V}}{P}\right)T\right]_{ii} = \frac{1}{\pi_{ii}}\sum_{h=1}^{H}\frac{1}{\pi_{(h)}}\int_{L_{h-1}-x_i'\hat{\beta}}^{L_h-x_i'\hat{\beta}}\varphi(\varepsilon_i;\hat{\sigma}^2)\varepsilon_i^2 d\varepsilon_i$$

$$= \frac{\hat{\sigma}^2}{\pi_{ii}}\left\{\frac{1}{\pi_{(H)}} + \sum_{h=1}^{H-1}\left(\frac{1}{\pi_{(h)}} - \frac{1}{\pi_{(h+1)}}\right)\Psi[(L_h - x_i'\hat{\beta})/\hat{\sigma}]\right\}. \tag{59}$$

Here, $\phi(\cdot;\hat{\sigma}^2)$ indicates the normal density with mean zero and variance $\hat{\sigma}^2$. The function $\Psi(\cdot)$ is defined as

$$\Psi(x) \equiv \int_{-\infty}^{x}\varphi(\varepsilon;1)\varepsilon^2 d\varepsilon = \Phi(x) - x\varphi(x;1), \tag{60}$$

where $\Phi(\cdot)$ denotes the cumulative density function for the standard normal distribution. In the derivation of (59), use has been made of $\Psi(L_0) = 0$ and $\Psi(L_H) = 1$.

7. MODEL-FREE REGRESSION

7.1 Consistent Estimation

As a digression from the main theme of this paper, model-free regression will be considered in this section. Firstly, model-free regression can be usefully applied in the case of doubt about the validity of a linear model. See Fuller (1975), who studies model-free regression for some specific designs. Van Praag (1981, 1982) studies model-free regression in the

case of repeated sampling from some probability distribution. See also DuMouchel and Duncan (1983). White (1980b, Section 3) studies related problems. Secondly, the so-called regression estimator of a population total uses model-free regression. See textbooks such as Cochran (1977), the review papers mentioned above by Nathan (1981) and Smith (1981) and Bethlehem and Keller (1983).

The purpose of model-free regression is the estimation of the population parameter vector

$$b \equiv (X'X)^{-1}X'y, \tag{61}$$

without assumptions about the probability distribution of y. In fact, both X and y are considered non-stochastic. Further, the same replica structure as in Section 2 is used, as follows.

$$X = \iota_K \otimes X_0, \tag{62}$$

and

$$y = \iota_K \otimes y_0, \tag{63}$$

where y_0 is some fixed N_0-vector. As before, the K diagonal matrices T_k $(k = 1, \ldots, K)$ are i.i.d. These matrices describe the sample as in Section 2. Together the matrices T_k form the matrix T. No additional assumptions are made concerning the distribution of T.

It is proved relatively easily, along the same lines as in Section 2, that the weighted estimator $\hat{\beta}$ defined before in (7), is a consistent estimator of b defined in (61). See also Jönrup and Rennermalm (1976), who indicates $\hat{\beta}$ as an "approximately unbiased" estimator of b, and Van Praag (1982, Section 4d), where "selectivity bias" with known inclusion probabilities is studied for the model-free case.

It follows in the same manner as in Section 4 that in the model-free case the asymptotic variance of $\hat{\beta}$, say $\mathrm{Var}_{\mathrm{MF}}(\hat{\beta})$, equals

$$\mathrm{Var}_{\mathrm{MF}}(\hat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}, \tag{64}$$

with

$$e \equiv y - Xb, \tag{65}$$

$$V = \mathrm{diag}(e)\Pi^{-1}P\Pi^{-1}\mathrm{diag}(e), \tag{66}$$

and with P defined as before in (29). Notice that V in (66) differs from V in (28) in the omission of the ξ-expectation and the substitution of e for ε.

It is interesting to rewrite $\mathrm{Var}_{\mathrm{MF}}(\hat{\beta})$ in the same way as $\mathrm{Var}(\hat{\beta})$ was rewritten in Section 5. In doing so, use will be made of

$$X'e = 0, \tag{67}$$

which follows directly from (61) and (65). The $\mathrm{Var}_{\mathrm{MF}}(\hat{\beta})$ can be rewritten as

$$\begin{aligned}\mathrm{Var}_{\mathrm{MF}}(\hat{\beta}) &= (X'X)^{-1}X'\mathrm{diag}(e)\,(\Pi^{-1}P\Pi^{-1} - \iota\iota')\,\mathrm{diag}(e)X(X'X)^{-1} \\ &+ (X'X)^{-1}X'ee'X(X'X)^{-1} \\ &= (X'X)^{-1}X'\mathrm{diag}(e)\,(\Pi^{-1}P\Pi^{-1} - \iota\iota')\,\mathrm{diag}(e)X(X'X)^{-1}.\end{aligned} \tag{68}$$

The last member of (68) corresponds with the p-component of the decomposition of $\text{Var}(\hat{\beta})$ in (37). It may be concluded from (68) that in model-free regression the variance of the estimator of the regression coefficients consists of the p-component, while the ξ-component vanishes.

Notice finally that, using the discussion at the end of Section 5, the last member of (68) can be written as

$$(X'X)^{-1}\Sigma(X'X)^{-1}, \tag{69}$$

where the matrix Σ is the p-variance-covariance of the row totals of $X'\text{diag}(e)$. A similar result was reached by Binder (1983, Section 4), though along different lines.

8. DISCUSSION

In this section some practical considerations are given concerning the use of weights in regression analysis. Several motives for the use of weights are discussed shortly, related to the preceding technical sections of this paper.

First of all, it must be noted that the difference between weighted and unweighted regressions may be of some significance. An important example is the case where business firms are the unit of study - either farms, industrial enterprises of any other kind of business firms varying considerably in the number of employees. At the Netherlands Central Bureau of Statistics, for instance, the classification by number of employees is a standard stratification variable in sampling designs of business firms, giving a considerable range of inclusion probabilities - the large units chosen with relatively large probabilities. In studies with employment as the endogenous variable, such a sampling design is endogenous, which calls for weighted regression; the large units receiving small weights.

Secondly, in the case of units varying widely in size, a major problem with regression analysis is the heteroscedasticity of the error term. This calls for weighted regression, of the same sort as the weighting due to an endogenous design discussed in Section 2: large units receiving small weights.

Finally, there is a third motive for the weighting of sampled data: the notion of a model free regression, as discussed in Section 7 above. Again, the weights here are of the same sort as the weights in Section 2.

Summing up, there seems to be no reason not to incorporate the sampling design in regression analysis.

9. CONCLUSIONS

In this paper the estimation of a regression model with survey sample data has been studied. In particular, samples drawn with an endogenous design have been studied; for example, a sample stratified on the endogenous variable. It has been shown that for such a sample the weighting of the observations with the inverse of the square root of the sampling fractions gives a consistent estimator. The concept of consistency used here is a modification of Brewer (1979). The asymptotic variance of the estimator has been given, as well as a consistent estimator of this variance. The variance is the sum of a sampling component and a model component.

Also, model-free regression has been considered. Model-free regression requires the same weighting as endogenous stratification. The variance of the estimator of the model-free regression coefficients contains only the sampling component, and not the model component.

Finally, some practical considerations relative to the weighting of the data have been given.

ACKNOWLEDGEMENT

The author thanks Abby Israëls and Albert Verbeek and several anonymous referees for their comments on previous versions of this paper.

REFERENCES

BETHLEHEM, J.G., and KELLER, W.J. (1983). Weighting sample survey data using linear models. Internal Report, Department for Statistical Methods, Netherlands Central Bureau of Statistics, Voorburg.

BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review,* 51, 279-292.

BRESLOW, N., and DAY, N.E. (1980). *Statistical Methods in Cancer Research, Volume 1: the Analysis of Case-Control Studies.* International Agency for Research on Cancer, Lyon.

BREWER, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association,* 74, 911-915.

COCHRAN, W.G. (1977). *Sampling Techniques.* New York: Wiley.

CRAMER, J.S. (1971). *Empirical Econometrics.* Amsterdam: North-Holland.

DuMOUCHEL, W.H., and DUNCAN, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association,* 78, 535-543.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā,* C37, 117-132.

HAUSMAN, J.A., and WISE, D.A. (1981). Stratification on endogenous variables and estimation: the Gary income maintenance experiment. In *Structural Analysis of Discrete Data with Econometric Applications,* (Eds., C.F. Manski and D. McFadden), Cambridge: MIT Press.

HECKMAN, J.J. (1979). Sample selęction bias as a specification error. *Econometrica,* 47, 153-161.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47, 663-685.

JEWELL, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika,* 72, 11-21.

JOHNSTON, J. (1972). *Econometric Methods.* Tokyo: McGraw-Hill Kogakusha.

JONRUP, H., and RENNERMALM, B. (1976). Regression analysis in samples from finite populations. *Scandinavian Journal of Statistics,* 33-36.

KMENTA, J. (1978). *Elements of Econometrics.* New York: McMillan.

MANSKI, C.F., and McFADDEN, D. (eds.) (1981). *Structural Analysis of Discrete Data with Econometric Applications.* Cambridge: MIT Press.

NATHAN, G. (1981). Notes on inference based on data from complex sample designs. *Survey Methodology,* 7, 110-129.

PORTER, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement,* 2, 141-158.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications.* New York: Wiley.

ROBINSON, P.M. (1982). On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics,* 24, 234-238.

SMITH, T.M.F. (1981). Regression analysis for complex surveys. In *Current Topics in Survey Sampling,* (Eds. D. Krewski, R. Platek, and J.N.K. Rao), New York: Academic Press, 267-292.

THEIL, H. (1971). *Principles of Econometrics.* New York: Wiley.

VAN PRAAG, B.M.S. (1981). Model-free regression. *Economics Letters,* 7, 139-144.

VAN PRAAG, B.M.S. (1982). The population-sample decomposition with an application to minimum distance estimators. Report 8218, Center for Research in Public Economics, Leyden University.

WHITE, H.,(1980a). Nonlinear regression on cross section data. *Econometrica,* 48, 721-746.

WHITE, H., (1980b). Using least squares to approximate unknown regression functions. *International Economic Review,* 12, 149-170.