SCB STATISTISKA CENTRALBYRÅN

1983-06-20

PROMEMORIOR FRÅN P/STM

NR 9

REGRESSION ANALYSIS AND RATIO ANALYSIS FOR DOMAINS, A RANDOMIZATION THEORY APPROACH

AV EVA ELVERS, CARL ERIK SÄRNDAL,
JAN H WRETMAN OCH GÖRAN ÖRNBERG

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1983:9. Regression analysis and ratio analysis for domains : a randomization theory approach / Eva Elvers m.fl.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1983-9

SCB STATISTISKA CENTRALBYRÅN 1983-06-20

PROMEMORIOR FRÅN P/STM

NR 9

REGRESSION ANALYSIS AND RATIO ANALYSIS FOR DOMAINS, A RANDOMIZATION THEORY APPROACH

AV EVA ELVERS, CARL ERIK SÄRNDAL,
JAN H WRETMAN OCH GÖRAN ÖRNBERG

MARCH, 1983

REGRESSION ANALYSIS AND RATIO ANALYSIS FOR DOMAINS, A RANDOMIZATION THEORY APPROACH

by

Eva ELVERS*, Carl Erik SÄRNDAL**, Jan H. WRETMAN* and Göran ÖRNBERG***

*Statistics Sweden, Stockholm

**Université de Montréal

***Department of Mathematical Statistics, University of Stockholm

ABSTRACT

In most surveys, inference for domains poses a difficult problem because of data shortage. This paper presents a design-inference (or randomization theory) approach to some common types of statistical analysis for domains of a surveyed population. Simple and multiple regression analysis, and analysis of ratios are considered. Two new methods are constructed and explored which, with the aid of auxiliary information, can improve substantially over the ordinary method based on straight π-inverse (product) sums. The theoretical conclusions are supported by empirical results from Monte Carlo experiments.

Key words: Domains, survey sampling, design inference, regression analysis, analysis of ratios.

- This work was supported in part by the Natural Sciences and Engineering Research Council of Canada

1. THE RESEARCH QUESTION

Inference for domains is required in most surveys, and extensive research efforts are currently directed to this problem area. The question that initiated our work was: given survey data from a population divided into many domains, how do we make inference, domain by domain and in the standard randomization theory fashion, about measures of relationship between a criterion variable y and explanatory variables $x_1,\ldots,x_r$, such as simple ratios or (multiple) regression slopes? The possible shortage of observations in any given domain poses a difficulty which is overcome, in our methods below, by exploiting auxiliary information.

Let $U = \{1,\ldots,k,\ldots,N\}$ denote a finite population of labelled units and let U be divided into nonoverlapping domains $U_{d\cdot}$ of sizes $N_{d\cdot}$, $d = 1,\ldots,D$; $N = \Sigma_{d=1}^{D} N_{d\cdot}$. If U is a country, and the units households, the $U_{d\cdot}$ may represent a possibly large number of geographical subdivisions of U. A sample survey is carried out on U according to a perhaps complex survey design. A random, often small number of observations will fall in a given domain $U_{d\cdot}$. With only one x-variable, we have in mind the estimation of parameters measuring the rate of change in y given x, such as, for $d = 1,\ldots,D$,

$$R_d = \Sigma_{U_{d\cdot}} y_k / \Sigma_{U_{d\cdot}} x_k \tag{1.1}$$

$$B_d = \Sigma_{U_{d\cdot}} (x_k - \bar{x}_{U_{d\cdot}})(y_k - \bar{y}_{U_{d\cdot}}) / \Sigma_{U_{d\cdot}} (x_k - \bar{x}_{U_{d\cdot}})^2 \tag{1.2}$$

(Σ_A denotes sum over k in the set A). More generally we seek to estimate multiple regression coefficients for the d:th domain. Frequently in surveys one wishes not only to estimate such rate-of-change measures, but also to test for their significant differences between domains. Only the estimation part is dealt with below, but the paper contains the basis for futher work on the hypothesis testing problem.

We work under the randomization theory principle of adjustment for varying inclusion probabilities by "π-inverse weighting" of units. In this spirit, Kish and Frankel (1974) studied the estimation of (multiple) regression coefficients for the entire population. Fuller (1975), Shah, Holt and Folsom (1977) contributed further to the theory. They did not examine regression analysis for domains, and the use auxiliary information was not discussed. Important is that these authors see the regression slopes as descriptive finite population parameters, not as superpopulation model parameters. Here we share that view; extensions to discriminant analysis and logistic regression are reported in Binder (1982).

Whether the finite population parameter or the superpopulation parameter perspective should be adopted depends on the situation. The latter view is held in the interesting model-based regression analysis of Nathan and Holt (1980), Holt, Smith and Winter (1980), Smith (1981). In Section 8 we discuss their approach, which permits auxiliary information to be incorporated, but again does not consider the domain estimation problem.

2. STATEMENT OF PROBLEM AND GENERAL PROCEDURE

Associated with unit $k(k = 1,\ldots,N)$ is the vector (y_k, x_k, δ_k), where $x_k = (x_{k1},\ldots,x_{ki},\ldots,x_{kr})'$, and domain membership is indicated by the D-vector δ_k with typical element $\delta_{dk} = 1$ if $k \in U_d$. and $\delta_{dk} = 0$ otherwise. In regression with an intercept, $x_{k1} = 1$ for $k = 1,\ldots,N$. Prior to sampling, (y_k, x_k) and often δ_k, too, are unknown. However in two methods discussed below (F and P), knowledge about domain membership is assumed to permit improvement of the basic C-method.

The problem is to estimate, for $d = 1,\ldots,D$, the regression coefficient vector of y on x defined for the d:th domain as

$$B_d = (B_{d1},\ldots,B_{dr})' = \Lambda_{dxx}^{-1}\,\Lambda_{dxy} \qquad (2.1)$$

where $\Lambda_{dxx} = \Sigma_1^N \delta_{dk} w_k x_k x_k'$ and Λ_{dxy} is analogous, that is, the r-vector $x_k y_k$ replaces the $r \times r$ matrix $x_k x_k'$ in Λ_{dxx}. The known constants w_k, if not all equal to one, are to permit weighted regression.

Our interest in B_d is in line with the concern often present in surveys to estimate descriptive quantities for the finite population, rather than parameters in models. One notes that B_d is the weighted least squares estimator of β that would arise in the hypotehtical "census fit" of the superpopulation model $y_k = x_k'\beta + \xi_k$ to all $N_{d.}$ points of $U_{d.}$, if the ξ_k are independant with mean 0 and variances w_k^{-1}. Simple examples of (2.1) are (1.1), arising when $r = 1$ and $x_{1k} = x_k = w_k^{-1}$, and (1.2), arising when $r = 2$, $x_k = (1,x_k)'$, $w_k = 1$. Their estimation is dealt with in Sections 5 and 6.

A probability sample s of fixed size n is drawn from U by a sampling design $p(s)$ with strictly positive inclusion probabilities $\pi_k = P(k\epsilon s)$, $\pi_{k\ell} = P(k,\ell\epsilon s)$. The part of s that happens to fall within $U_{d.}$ is denoted $s_{d.}$, of random size $n_{d.}$, where $n = \Sigma_d\, n_{d.}$. The estimators will be built on the data (y_k,x_k) for $k \epsilon s$. For estimation of B_d, we examine three methods, called C, F and P, each containing a Step 1 for constructing the estimator $\hat{B}_d$, and a Step 2 for constructing an estimated variance-covariance matrix, $\hat{V}_p(\hat{B}_d)$, of $\hat{B}_d$'s theoretical variance-covariance matrix, $V_p(\hat{B}_d)$.

STEP 1. First estimate Λ_{dxx} and Λ_{dxy} by, respectively, $\hat{\Lambda}_{dxx}$ and $\hat{\Lambda}_{dxy}$, which then define the B_d-estimator as $\hat{B}_d = \hat{\Lambda}_{dxx}^{-1}\hat{\Lambda}_{dxy}$.

STEP 2. Calculate an estimated (design-based) variance-covariance matrix of $\hat{B}_d$ by

$$\hat{V}_p(\hat{B}_d) = \hat{\Lambda}_{dxx}^{-1} YG_d \hat{\Lambda}_{dxx}^{-1} \tag{2.2}$$

where the ij:th element $(i,j = 1,\ldots,r)$ of the $r \times r$ matrix YG_d is the Yates-Grundy type quantity

$$\sum_{k<\ell \atop \epsilon s} \sum \Delta_{k\ell}(\pi_k^{-1} u_{dki} - \pi_\ell^{-1} u_{d\ell i})(\pi_k^{-1} u_{dkj} - \pi_\ell^{-1} u_{d\ell j})$$

where u_{dki} is to be defined and $\Delta_{k\ell} = (\pi_k \pi_\ell - \pi_{k\ell})/\pi_{k\ell}$.

The three methods propose different $\hat{\Lambda}_{dxx}$, $\hat{\Lambda}_{dxy}$ in Step 1 and will give rise to different u_{dki} in Step 2. Computationally, Methods F and P also involve a preliminary Step 0. A brief statement of the three computational procedures is given in Section 3; discussion is saved for later sections. For interval estimation of B_{di} with $100(1-\alpha)\%$ confidence,

$$\hat{B}_{di} \pm Z_{1-\alpha/2}\{\hat{V}(\hat{B}_{di})\}^{1/2} , \tag{2.3}$$

$Z_{1-\alpha/2}$ being the normal score, is recommended until more accurate methods have been explored. If n is very large and the domain not too small, (2.3) gives roughly the right coverage rate for the procedures stated in Section 3. However, our Monte Carlo studies show that the normal score if often too modest to reach the nominal $100(1-\alpha)\%$ coverage of the true slope B_{di} in repeated samples drawn by the fixed design. For a given total sample size n , the achieved coverage rate deteriorates with smallness of domain; further work may improve the confidence interval procedure.

3. SKELETON OUTLINE OF PROCEDURE, METHODS C , F AND P

The C method (C for Common) uses straight π-inverse weighting in estimating each (product) sum. When applied to the full population, the C-method

is found in Fuller (1975), Shah, Holt and Folsom (1977). The extension to domain estimation is a minor modification. The two steps require no auxiliary information:

Step 1. Calculate the estimator of B_d as $\hat{B}_{Cd} = \hat{\Lambda}^{-1}_{Cdxx}\hat{\Lambda}_{Cdxy}$ with $\hat{\Lambda}_{Cdxx} = A_{dxx}$, $\hat{\Lambda}_{Cdxy} = A_{dxy}$ where, by definition,

$$A_{dxx} = \Sigma_s \delta_{dk} w_k x_k x_k'/\pi_k \; ; \; A_{dxy} = \Sigma_s \delta_{dk} w_k x_k y_k/\pi_k \qquad (3.1)$$

Step 2. Calculate the estimated variance-covariance matrix of $\hat{B}_{Cd}$ by (2.2) with $u_{dki} = u_{Cdki}$, where, by definition, $u_{Cdki} = \delta_{dk} w_k x_{ki} e_{Cdk}$ with $e_{Cdk} = y_k - x_k'\hat{B}_{Cd}$.

The remaining two methods seek to improve the B_d-estimator by incorporating other information. Let M-vector $z_k = (z_{k1},\ldots,z_{kM})'$ be known for $k = 1,\ldots,N$. Knowledge of domain membership of each unit k is also assumed. Here we explore two improved methods, both of which use the principle of generalized regression estimation by means of the known z_k-vectors, Cassel, Särndal and Wretman (1976, 1977); Särndal (1980):

In the F method (F for First order variable), each of $x_1,\ldots,x_r$ and y are explained, in the preliminary Step 0, by a regression fit on z . The steps are:

Step 0. Calculate predictions of x_k and y_k as respectively, $\tilde{x}_k' = z_k'\tilde{L}_x$ and $\tilde{y}_k = z_k'\tilde{L}_y$ $(k = 1,\ldots,N)$, where the $M \times r$ matrix $\tilde{L}_x$ is

$$\tilde{L}_x = (\Sigma_s a_k z_k z_k'/\pi_k)^{-1}\Sigma_s a_k z_k x_k'/\pi_k$$

and the M-vector $\tilde{L}_y$ is analogous, the scalar y_k replacing the r-vector x_k' in $\tilde{L}_x$. The known constants a_k are to permit differential weighting, if

desired. If the first x-variable is a constant one indicating an intercept in y's regression on x, then define $\tilde{x}_{k1} = x_{k1} = 1$ for all k.

STEP 1. Calculate the estimator of B_d as $\hat{B}_{Fd} = \hat{\Lambda}^{-1}_{Fdxx}\hat{\Lambda}_{Fdxy}$ with $\hat{\Lambda}_{Fdxx} = A_{dxx} + \Lambda_{d\tilde{x}\tilde{x}} - A_{d\tilde{x}\tilde{x}}$; $\hat{\Lambda}_{Fdxy} = A_{dxy} + \Lambda_{d\tilde{x}\tilde{y}} - A_{d\tilde{x}\tilde{y}}$. Here A_{dxx} and A_{dxy} are given by (3.1),

$$\Lambda_{d\tilde{x}\tilde{x}} = \Sigma_1^N \delta_{dk} w_k \tilde{x}_k \tilde{x}_k' \; ; \; A_{d\tilde{x}\tilde{x}} = \Sigma_s \delta_{dk} w_k \tilde{x}_k \tilde{x}_k'/\pi_k$$

while $\Lambda_{d\tilde{x}\tilde{y}}$ and $A_{d\tilde{x}\tilde{y}}$ are analogous, $\tilde{x}_k\tilde{y}_k$ replacing $\tilde{x}_k\tilde{x}_k'$ in both $\Lambda_{d\tilde{x}\tilde{x}}$ and $A_{d\tilde{x}\tilde{x}}$. Domain membership, δ_{dk}, must be known for $k = 1,\ldots,N$ in order to calculate $\Lambda_{d\tilde{x}\tilde{x}}$.

STEP 2. Calculate the estimated variance-covariance matrix of $\hat{B}_{Fq}$ by (2.2), using $\hat{\Lambda}_{dxx} = \hat{\Lambda}_{Fdxx}$ and $u_{dki} = u_{Fdki}$ defined by

$$u_{Fdki} = \delta_{dk} w_k (x_{ki} e_{Fdk} - \tilde{x}_{ki} \tilde{e}_{Fdk}) \tag{3.2}$$

with $e_{Fdk} = y_k - x_k'\hat{B}_{Fd}$; $\tilde{e}_{Fdk} = \tilde{y}_k - \tilde{x}_k'\hat{B}_{Fd}$.

The P method (P for Product variable) considers the $r(r+1)/2 + r$ product variables $t_{ij} = x_i x_j$, $t_{io} = x_i y$ $(i \le j = 1,\ldots,r)$, each of which is explained in Step 0 by a regression fit on z. (If the regression of y on x has an intercept, then $t_{11} \equiv 1$, $t_{1j} = x_j$ $(j = 2,\ldots,r)$, and $t_{1o} = y$.) The steps are:

STEP 0. Let $t_{kij} = x_{ki}x_{kj}$, $t_{kio} = x_{ki}y_k$, and calculate, for $i = 1,\ldots,r$, predictions of $t_{ki} = (t_{ki1},\ldots,t_{kir})'$ and t_{kio} by, respectively, $\tilde{t}_{ki}' = z_k'\tilde{J}_{ixx}$ and $\tilde{t}_{kio} = z_k'\tilde{J}_{ixy}$ where $\tilde{J}_{ixx}$ $(M \times r)$ is given by

$$\tilde{J}_{ixx} = (\Sigma_s\ a_k z_k z_k'/\pi_k)^{-1}\ \Sigma_s\ a_k z_k t_{ki}'/\pi_k$$

and $\tilde{J}_{ixy}$ (M×1) is analogous, the scalar t_{kio} replacing the r-vector t_{ki}' in $\tilde{J}_{ixx}$. For the intercept case, define $\tilde{t}_{kll} = t_{kll} = 1$ for $k = 1,\ldots,N$.

STEP 1. Calculate the estimator of B_d as $\hat{B}_{Pd} = \hat{\Lambda}_{Pdxx}^{-1}\hat{\Lambda}_{Pdxy}$ with

$$\hat{\Lambda}_{Pdxx} = A_{dxx} + \Lambda_{d(\tilde{x}\tilde{x})} - A_{d(\tilde{x}\tilde{x})}\ ;\ \hat{\Lambda}_{Pdxy} = A_{dxy} + \Lambda_{d(\tilde{x}\tilde{y})} - A_{d(\tilde{x}\tilde{y})}$$

Here, A_{dxx} and A_{dxy} are given by (3.1); $\Lambda_{d(\tilde{x}\tilde{x})}$ and $A_{d(\tilde{x}\tilde{y})}$ are $r \times r$ matrices whose i:th rows are given, respectively, by $\Sigma_1^N\ \delta_{dk} w_k \tilde{t}_{ki}'$ and $\Sigma_s\ \delta_{dk} w_k \tilde{t}_{ki}'/\pi_k$ while the $r \times 1$ columns $\Lambda_{d(\tilde{x}\tilde{y})}$ and $A_{d(\tilde{x}\tilde{y})}$ are analogous, their i:th element having the scalar $\tilde{t}_{kio}$ in place of the r-vector $\tilde{t}_{ki}'$. As in Method F, domain membership, δ_{dk}, must be known for $k = 1,\ldots,N$.

STEP 2. Calculate the estimated variance-covariance matrix of $\hat{B}_{Pd}$ by (2.2) using $\hat{\Lambda}_{dxx} = \hat{\Lambda}_{Pdxx}$ and $u_{dki} = u_{Pdki}$ defined by

$$u_{Pdki} = \delta_{dk} w_k (x_{ki} e_{Pdk} - (\tilde{x}\tilde{e})_{Pdki})$$

with

$$x_{ki} e_{Pdk} = x_{ki}(y_k - x_k'\hat{B}_{Pd}) = t_{kio} - t_{ki}'\hat{B}_{Pd}\ ;\ (\tilde{x}\tilde{e})_{Pdki} = \tilde{t}_{kio} - \tilde{t}_{ki}'\hat{B}_{Pd}.$$

Our Monte Carlo experiments so far (see Section 8) have not shown any great differences in efficiency between Methods F and P, both of which can however improve greatly over Method C.

4. DERIVING VARIANCE-COVARIANCE ESTIMATES

The variance-covariance estimates in Step 2 of the C- , F- and P-methods rest on approximations. For the C-method, reference to Fuller (1975) suffices. As for the other two methods, we choose to illustrate how $\hat{V}_p(\hat{B}_{Fd})$ is obtained.

In the F-method, $\tilde{L}_x$ estimates its population analogue

$$L_x^o = (\Sigma_1^N a_k z_k z_k')^{-1} \Sigma_1^N a_k z_k x_k' .$$

The j:th column of L_x^o is made up of the weighted least squares regression coefficients that would arise in the "census fit" of x_j on z using the N data points of the whole population. Let the resulting r-vector of fitted values for unit k be $x_k^{o'} = z_k' L_x^o$. Similarly, the k:th fitted y-value would be $y_k^o = z_k' L_y^o$, where L_y^o is analogous to L_x^o , with y_k in place of x_k' . Set

$$\Lambda_{dx^o x^o} = \Sigma_1^N \delta_{dk} w_k x_k^o x_k^{o'} \; ; \; A_{dx^o x^o} = \Sigma_s \delta_{dk} w_k x_k^o x_k^{o'} / \pi_k$$

We express the error of the F-estimator as $\hat{B}_{Fd} - B_d = \hat{\Lambda}_{Fdxx}^{-1}(F_1+F_2)$ where the r-vectors F_1 and F_2 are

$$F_1 = (A_{dxy} - \Lambda_{dxy}) - (A_{dxx} - \Lambda_{dxx})B_d$$

$$- \{(A_{dx^o y^o} - \Lambda_{dx^o y^o}) - (A_{dx^o x^o} - \Lambda_{dx^o x^o})B_d\} ,$$

$$F_2 = (A_{dx^o y^o} - \Lambda_{dx^o y^o}) - (A_{dx^o x^o} - \Lambda_{dx^o x^o})B_d$$

$$- \{(A_{d\tilde{x}\tilde{y}} - \Lambda_{d\tilde{x}\tilde{y}}) - (A_{d\tilde{x}\tilde{x}} - \Lambda_{d\tilde{x}\tilde{x}})B_d\} .$$

We arrive at $\hat{V}_p(\hat{B}_{Fd})$ by way of an approximation of the theoretical matrix, $V_p(\hat{B}_{Fd})$. In large samples, F_2 will be near zero in probability (as determined by the sampling design) for the twofold reason: (a) each A-matrix is close to its Λ-counterpart, and (b) $\tilde{L}_x$ and L_x^o are close, as are $\tilde{L}_y$ and L_y^o, so that $A_{dx^oy^o}$ and $A_{d\tilde{x}\tilde{y}}$ are close, as are $\Lambda_{dx^oy^o}$ and $\Lambda_{d\tilde{x}\tilde{y}}$, etc. The design variance contributed by F_2 is expected to be small relative to that coming from F_1, which is near zero in probability for only the first of the two reasons. Further, in a first approximation to variance, $\hat{\Lambda}_{Fdxx}^{-1}$ may be replaced by its target, Λ_{dxx}^{-1}. Thus, we take $\hat{B}_{Fd} - B_d \doteq \Lambda_{dxx}^{-1}F_1$ as the basis for an approximate variance-covariance calculation. Alternatively,

$$F_1 = \Sigma_s(\mu_{Fdk}/\pi_k) - \Sigma_1^N \mu_{Fdk}$$

where $\mu_{Fdk} = (\mu_{Fdk1},\ldots,\mu_{Fdkr})'$ is determined by

$$\mu_{Fdki} = \delta_{dk}w_k(x_{ki}\varepsilon_{Fdk} - x_{ki}^o\varepsilon_{Fdk}^o) \; ; \; \varepsilon_{Fdk} = y_k - x_k'B_d \; ; \; \varepsilon_{Fdk}^o = y_k^o - x_k^{o\prime}B_d \; .$$

Now F_1 involves simple Horvitz-Thompson estimators, so directly

$$V_p(\hat{B}_{Fd}) = E_p(\hat{B}_{Fd} - B_d)^2 \doteq \Lambda_{dxx}^{-1}V_{Fd}^o\Lambda_{dxx}^{-1}$$

where V_{Fd}^o is $r \times r$ with ij-element

$$\sum_{k<\ell}^{N}\sum_{\ell}^{N} \pi_{k\ell}\Delta_{k\ell}(\pi_k^{-1}\mu_{Fdki} - \pi_\ell^{-1}\mu_{Fd\ell i})(\pi_k^{-1}\mu_{Fdkj} - \pi_\ell^{-1}\mu_{Fd\ell j}) \; .$$

This contains the unknowns Λ_{dxx}, B_d, L_x^o and L_y^o. For an estimated matrix, replace each of these by its sample couterpart, $\hat{\Lambda}_{Fdxx}$, $\hat{B}_{Fd}$, $\tilde{L}_x$ and $\tilde{L}_y$, which implies replacement of μ_{Fdki} by u_{Fdki} given by (3.2). Replacement of

$$\sum_{k<\ell}^{N}\sum^{N} \pi_{k\ell}\Delta_{k\ell}(\cdot)(\cdot) \quad \text{by} \quad \sum_{k<\ell}\sum_{\epsilon s} \Delta_{k\ell}(\cdot)(\cdot)$$

completes the procedure. We have obtained the variance-covariance estimate given in the F-method's Step 2.

5. DOMAINWISE ESTIMATION OF A RATIO OF AGGREGATES

A special case of the foregoing is to estimate, domain by domain, the ratio $R_{d\cdot}$, formula (1.1), of two domain sums. Important applications are of the type where one seeks to estimate "acreage under wheat (y) to total acreage (x)" for farms belonging to the d:th of D geographical subdivisions of a sampled larger region. Now $R_{d\cdot}$ is the ratio of two scalars, the linear sums $\Lambda_{dly} = \Sigma_1^N \delta_{dk}y_k$ and $\Lambda_{dlx} = \Sigma_1^N \delta_{dk}x_k$. Their estimation is simple in the C-method. The F-method as given in Section 3 also provides for estimation of linear sums, since the r-vector x_k may contain the constant one.

The C-method gives simply

$$\hat{R}_{Cd} = (\Sigma_{s_{d\cdot}} y_k/\pi_k)/(\Sigma_{s_{d\cdot}} x_k/\pi_k)$$

with estimated design variance

$$\hat{V}_p(\hat{R}_{Cd}) = \sum_{k<\ell}\sum_{\epsilon s} \Delta_{k\ell}(\pi_k^{-1}u_{Cdk}-\pi_\ell^{-1}u_{Cd\ell})/(\Sigma_{s_{d\cdot}} x_k/\pi_k)^2$$

where $u_{Cdk} = \delta_{dk}(y_k-x_k\hat{R}_{Cd})$.

In the F-method's Step 0, we have

$$\tilde{x}_k = z_k'\tilde{L}_x = z_k'(\Sigma_s a_k z_k z_k'/\pi_k)^{-1}\Sigma_s a_k z_k x_k/\pi_k$$

and analogously for $\tilde{y}_k$, with y_k replacing x_k in $\tilde{x}_k$. In carrying out Step 1, look upon Λ_{dlx} as $\Sigma_1^N \delta_{dx} x_{ko} x_k$, where $x_{ko} = 1 = \tilde{x}_{ko}$ for all k, and analogously for Λ_{dly}. We get

$$\hat{R}_{Fd} = \hat{\Lambda}_{Fdly}/\hat{\Lambda}_{Fdlx}$$

where

$$\hat{\Lambda}_{Fdlx} = \Sigma_1^N \delta_{dk}\tilde{x}_k + \Sigma_s \delta_{dk}(x_k - \tilde{x}_k)/\pi_k$$

and analogously for $\hat{\Lambda}_{Fdly}$, with y_k and $\tilde{y}_k$ instead of x_k and $\tilde{x}_k$. The estimated variance, from Step 2, is

$$\hat{V}_p(\hat{R}_{Fd}) = \sum_{k<\ell \atop \epsilon s}\sum \Delta_{k\ell}(\pi_k^{-1}u_{Fdk} - \pi_\ell^{-1}u_{Fd\ell})^2/(\hat{\Lambda}_{Fdlx})^2$$

with

$$u_{Fdk} = \delta_{dk}(e_{Fdk} - \tilde{e}_{Fdk}) \; ; \; e_{Fdk} = y_k - \hat{R}_{Fd}x_k \; ; \; \tilde{e}_{Fdk} = \tilde{y}_k - \hat{R}_{Fd}\tilde{x}_k$$

The performance of the two methods has been tested in Monte Carlo experiments on which we comment in Section 8.

6. DOMAINWISE SIMPLE REGRESSION ANALYSIS

To enrich the brief statement in Section 3, let us discuss our three methods with another simple application in mind: We seek to estimate, for each of the D domains, the slope B_d, given by (1.2), of a simple regression with intercept of y on x. Frequently one needs to compare slopes in different domains, so the standard error calculations given below are important.

In practice one often exploits the homogeneity gained by an a priori known categorization of the units. Let $z_k = (z_{k1},\ldots,z_{km},\ldots,z_{kM})'$ required in Step 0 be a category-indicator, with $z_{km} = 1$ if $k \in U_{\cdot m}$ and $z_{km} = 0$

otherwise; $m = 1,\ldots,M$. The mutually exclusive categories $U_{\cdot m}$ cut across the domains $U_{d\cdot}$. Possibilities discussed in Section 7 are: (1) the categories are $M = G$ "postgroups" (called groups in the following) that do not participate in the sampling design but are exploited after sampling to reduce the estimator's variance; (2) the categories are the $M = H$ strata in a stratified random sampling design; (3) the categories are the $M = GH$ cells resulting from a crossclassification of G groups with H strata. In the third case there will thus be three dimensions involved: if units are households, the domains could be smaller administrative areas of a sampled country, the strata could be larger geographical areas, and the groups family types.

Domains crossed with the categories of z divides the population into DM cells U_{dm} of sizes N_{dm} . These latter are the auxiliary quantities that must be known, from census or other reliable sources, in order to make the F- and P-methods work. We have

$$N = \sum_{d}^{D} N_{d\cdot} = \sum_{m}^{M} N_{\cdot m} = \sum_{d}^{D}\sum_{m}^{M} N_{dm} \tag{6.1}$$

The respective parts of the sample s that happen to fall in $U_{\cdot m}$ and U_{dm} are denoted $s_{\cdot m}$, of size $n_{\cdot m}$, and s_{dm} , of size n_{dm} . Then (6.1) holds with small n's substituted for the capital N's . In the general formulas of Section 3, let $x_k = (1,x_k)'$, and assume equal weights: $w_k = a_k = 1$ for all k . Let I indicate the method used; $I = C$, F or P . The slope estimator is in all three cases of the form

$$\hat{B}_{Id} = \hat{\Sigma}_{Idxy}/\hat{\Sigma}_{Idxx} \tag{6.2}$$

where $\hat{\Sigma}_{Idxy}$ and $\hat{\Sigma}_{Idxx}$ are defined below, and the estimated variance is

$$\hat{V}_p(\hat{B}_{Id})_{\cdot} = \sum_{\epsilon s}\sum_{k<\ell} \Delta_{k\ell}(\pi_k^{-1}u_{Idk} - \pi_\ell^{-1}u_{Id\ell})^2/(\hat{\Sigma}_{Idxx})^2 \tag{6.3}$$

Details for each method are as follows: Set $\hat{N}_{dm} = \Sigma_{s_{dm}} 1/\pi_k$. Use the symbol $\sim$ to indicate "π-means", that is, π-inversely weighted averages such as

$$\tilde{x}_{s_{d\cdot}} = (\Sigma_{s_{d\cdot}} x_k/\pi_k)/(\Sigma_{s_{d\cdot}} 1/\pi_k) \quad \text{or} \quad \tilde{x}_{s_{\cdot m}} = (\Sigma_{s_{\cdot m}} x_k/\pi_k)/(\Sigma_{s_{\cdot m}} 1/\pi_k)$$

and analogously for several other π-means used below.

In the C-method, use $I = C$ in (6.2) and (6.3), where

$$\hat{\Sigma}_{Cdxx} = N^{-1}\Sigma_{s_{d\cdot}}(x_k-\tilde{x}_{s_{d\cdot}})^2/\pi_k \; ; \; \hat{\Sigma}_{Cdxy} = N^{-1}\Sigma_{s_{d\cdot}}(x_k-\tilde{x}_{s_{d\cdot}})(y_k-\tilde{y}_{s_{d\cdot}})/\pi_k$$

Letting $e_{Cdk} = y_k - \tilde{y}_{s_{d\cdot}} - \hat{B}_{Cd}(x_k-\tilde{x}_{s_{d\cdot}})$, we have

$$u_{Cdk} = \delta_{dk}(x_k-\tilde{x}_{s_{d\cdot}})e_{Cdk} \tag{6.4}$$

To simplify the F- and P-methods, create centered variables X_d , Y_d for the d:th domain as

$$X_{dk} = x_k - x^*_{s_{d\cdot}} \; ; \; Y_{dk} = y_k - y^*_{s_{d\cdot}} \; ; \; k \in U_{d\cdot}$$

where

$$x^*_{s_{d\cdot}} = N_{d\cdot}^{-1}\{\Sigma_{s_{d\cdot}} x_k/\pi_k + \sum_{m=1}^{M}(N_{dm}-\hat{N}_{dm})\tilde{x}_{s_{\cdot m}}\}$$

is the estimate of the domain mean $\bar{x}_{U_{d\cdot}} = \Sigma_{U_{d\cdot}} x_k/N_{d\cdot}$ produced by both the F- and P-methods, and $y^*_{s_{d\cdot}}$ is analogous.

In the F-method, use $I = F$ in (6.2) and (6.3), where

$$\hat{\Sigma}_{Fdxx} = \{\Sigma_{s_{d\cdot}} X_{dk}^2/\pi_k + \sum_{m=1}^{M}(N_{dm}-\hat{N}_{dm})(\tilde{X}_{ds_{\cdot m}})^2\}/N$$

and $\hat{\Sigma}_{Fdxy}$ is analogous, with $X_{dk}Y_{dk}$ and $\tilde{X}_{ds_{\cdot m}}\tilde{Y}_{ds_{\cdot m}}$ replacing X_{dk}^2 and $(\tilde{X}_{ds_{\cdot m}})^2$. Letting $e_{Fdk} = Y_{dk} - \hat{B}_{Fd}X_{dk}$, with π-mean $e_{Fds_{\cdot m}}$ in $s_{\cdot m}$, we have for k in $s_{\cdot m}$

$$u_{Fdk} = \delta_{dk}(X_{dk}e_{Fdk} - \tilde{X}_{ds_{\cdot m}}\tilde{e}_{Fds_{\cdot m}}) \tag{6.5}$$

In the P-method, use $I = P$ in (6.2) and (6.3), where

$$\hat{\Sigma}_{Pdxx} = \{\Sigma_{s_{d\cdot}} X_{dk}^2/\pi_k + \sum_{m=1}^{M}(N_{dm} - \hat{N}_{dm})(\widetilde{X^2})_{ds_{\cdot m}}\}/N$$

where $(\widetilde{X^2})_{ds_{\cdot m}}$ is the π-mean of X_{dk}^2 for k in $s_{\cdot m}$. Further, $\hat{\Sigma}_{Pdxy}$ is analogous, with $X_{dk}Y_{dk}$ and $(\widetilde{XY})_{ds_{\cdot m}}$ replacing X_{dk}^2 and $(\widetilde{X^2})_{ds_{\cdot m}}$. With $e_{Pdk} = Y_{dk} - \hat{B}_{Pd}X_{dk}$, we have for k in $s_{\cdot m}$

$$u_{Pdk} = \delta_{dk}\{X_{dk}e_{Pdk} - (\widetilde{Xe})_{Pds_{\cdot m}}\} \tag{6.6}$$

where $(\widetilde{Xe})_{Pds_{\cdot m}}$ is the π-mean of $X_{dk}e_{Pdk}$ over k in $s_{\cdot m}$.

The u-quantities (6.4), (6.5) and (6.6) explain heuristically why the F- and P-methods are often superior to the C-method, as discussed in the next section.

7. DISCUSSION OF SIMPLE REGRESSION ANALYSIS

In the C-method, u_{Cdk} is (apart from the δ-factor) structured as "centered x-variable times residual". By contrast, in the F- and P-methods u_{Fdk} and u_{Pdk} have (apart from δ) the form "centered x-variable times residual minus category mean adjustment", the latter being $\tilde{X}_{ds_{\cdot m}}\tilde{e}_{Fds_{\cdot m}}$ and $(\widetilde{Xe})_{Pds_{\cdot m}}$, respectively. This difference explains the variance reductions realizable by the F- and P-methods, as seen more clearly for some simple designs:

(i) Simple random sampling (srs) and G groups. The z-vector indicates membership in one of $M = G$ groups labelled $g = 1,\ldots,G$. The

estimated variance (6.3) becomes

$$\hat{V}_{srs}(\hat{B}_{Id}) = N^2(1-f)v_s(u_{Id})/n(\hat{\Sigma}_{Idxx})^2$$

where $f = n/N$ and

$$v_s(u_{Id}) = \Sigma_s(u_{Idk}-\bar{u}_{Ids})^2/(n-1)$$

is the sample variance of the u_{Idk}. The respective group mean adjustments contained in u_{Fdk} and u_{Pdk} are $\bar{X}_{ds.g}\bar{e}_{Fds.g}$ and $(\overline{Xe})_{Pds.g}$, the overbar indicating straight average over the group part, $s_{.g}$, of the whole sample, s. In methods F and P, the sample variance of u thus consists essentially of within group components. It will be substantially smaller than in the C-method, if the groups are efficient so that their mean adjustments differ markedly. With $G = 1$ group only, the F- and P-methods are identical, but the two are not identical to the C-method, as a first guess may have been. However, essentially no variance reduction will be realized by the $F = P$ method, since the one and only group adjustment applies to every unit k.

(ii) Stratified random sampling (strs) and G groups. The strata, labelled $h = 1,\ldots,H$, ordinarily cut across the G groups $(g = 1,\ldots,G)$ as well as the D domains $(d = 1,\ldots,D)$. Let $N_{..h}$, $n_{..h}$ and $f_h = n_{..h}/N_{..h}$ denote stratum population size, stratum sample size and stratum sampling fraction. Then (6.3) becomes

$$\hat{V}_{strs}(\hat{B}_{Id}) = \sum_{h=1}^{H} N_{..h}^2(1-f_h)v_{s_h}(u_{Id})/n_{..h}(\hat{\Sigma}_{Idxx})^2 \qquad (7.1)$$

where $v_{s_h}(u_{Id})$ is the variance of u_{Idk} over k in the sample s_h selected randomly from stratum h. Here too, the F- and P-methods will lead to substantial variance reductions over the C-method, if the grouping strongly supplements, rather than closely copies, the stratification. The argument is

as in case (i) above: u_{Fdk} and u_{Pdk} contain a group mean adjustment, which can strongly reduce the stratum variance of u if the grouping is efficient. There are at least two possibilities to code the z-vector necessary in Step 0 of the F- and P-methods: (a) simply let z be the G-vector that indicates group membership; (b) let z be the extended vector of dimension GH indicating one of the cells in the crossclassification of groups with strata. For each of the three methods separately, (a) and (b) will not cause much difference in efficiency. The possible gains from stratification are essentially discounted already in method (a), through the stratumwise buildup of (7.1).

(iii) strs without grouping. Let z_k indicate stratum membership. Clearly here one does not expect the F- and P-methods to be superior to the C-method. This can be seen more formally from (7.1), where $v_{s_h}(u_{Id})$ is the stratum variance of u_{Idk}, or, equivalently, of $u_{Idk} - \bar{u}_{Ids_h}$, which in all three methods is "centered x-variables times residual minus stratum mean adjustment". In the F- and P-methods, the stratum mean adjustment is already present in u itself and $\bar{u}_{Fds_h} = \bar{u}_{Pds_h} = 0$; in the C-method, the stratum adjustment is created through $\bar{u}_{Cds_h}$. For a given unit k, $u_{Idk} - \bar{u}_{Ids_h}$ is not exactly the same number in the C-, F- and P-methods, but the calculated values of $v_{s_h}(u_{Id})$ will be essentially equal in the three methods.

Our methods have the generality of permitting more complex designs, including cases of two or more stages. The variance reducing effects of Methods F and P will continue to be strong when the z-vector contains essential extraneous information.

8. MONTE CARLO EXPERIMENTS

Our Monte Carlo experiments were designed primarily to assess the

variance reductions realizable by the F- and P-methods over the simple C-method, and to see if the F- and P-methods differ by much in efficiency. It must be kept in mind that the results of any simulation depend entirely on the nature of the population chosen for study. A detailed account of our simulations will be given elsewhere. We simulated the estimation of the ratio R_d given by (1.1) through the methods of Section 5, and the estimation of the slope B_d given by (1.2) through the methods of Sections 6-7. Many conclusions are similar. Here we give only a brief summary of the results concerning B_d .

Three populations, called REAL , ART1 and ART2, were used. REAL consisted of real data on 1202 Swedish households divided into $D = 24$ domains by Sweden's major admistrative regions (län) and into $G = 5$ groups by size and age characteristics of a household; y and x were, respectively, disposable household income and taxable household income. The groups were a rather weak explanatory factor for x as well as for y , so a priori the structure of REAL does not strongly favour the F- and P-methods. The artificial populations ART1 and ART2 were therefore created to have a smaller within group variance, relative to the between group variance, in x as well as in y .

The ART1 population shared a number of features with REAL : the cell frequencies N_{dg} , and the x-means, y-means and x-to-y correlations in each group. Group by group, each (x,y)-point of REAL was replaced by a new randomly generated point with the objective to reduce the within group variance. ART2 was created to provoke a situation where extremely large efficiency gains are expected from the F- and P-methods. The cell frequencies N_{dg} were as in REAL , but the (x,y)-values were chosen so that the within group to between postgroup variance was very small, and in addition the regression of y on x was markedly non-linear.

The domains varied in size from 20% to about 1% of the total population. 1000 repeated simple random samples of size $n = 300$ (and $n = 600$ in a second round of experiments) were drawn, so the situation was that of case (i) in Section 7. For each domain and sample we calculated the B_d-estimate and the confidence interval (2.3), by each of methods C , F and P as given Section 3. Summary statistics for the 1000 repetitions were calculated, including each estimator's mean, variance, average estimated variance and coverage rate (= the percentage of the 1000 samples with a confidence interval covering the true slope B_d) .

The three methods C , F and P shared the following features: (1) The design bias of each estimator is very small; (2) The variance of the estimates agrees well with the average of the estimated variances in the larger domains, but the two differed markedly in some of the smaller domains; (3) The achieved coverage rates were close to (but always somewhat short of) the nominal 95% or 90% in the larger domains, but considerably less in the smallest domains, when $n = 300$. The increase in sample size to $n = 600$ improved these trailing coverage rates.

The following emerged in the comparison of the three methods: (4) The F- and P-methods performed very similarly for all three populations, in terms of variance as well as coverage rate. There was some indication that the P-method is prone to more erratic behavior for the smallest domains; (5) The variance reductions realized by the F- and P-methods over the C-method were modest for most domains (30%-0%) in the REAL population, strong in virtually all domains (60%-20%) in the ART1 population, and dramatically large in all domains (over 90%) in the ART2 population.

9. IMPLICATIONS FOR THE ESTIMATION OF COVARIANCE MATRICES

In this section we disregard the important issue of domain estimation and compare our approach with some earlier work. Consider the estimation of the finite population covariance matrix

$$\Sigma_{xx} = \Sigma_1^N (x_i - \bar{x})(x_i - \bar{x})'/N$$

where $\bar{x} = \Sigma_1^N x_k/N$. Let the earlier y-variable be included with other variables in the x-vector, which in this section is assumed not to contain the constant one indicating intercept. Indirectly, we dealt with domainwise covariance estimation in Section 6. The F- and P-methods of the earlier sections estimate product sums, rather than covariances directly. But

$$\Sigma_{xx} = N^{-1}\Lambda_{xx} - \bar{x}\bar{x}' \tag{9.1}$$

where $\Lambda_{xx} = \Sigma_1^N x_k x_k'$ and $\bar{x} = \Sigma_1^N x_i/N$. Therefore we can carry out the F- or P-method's Steps 0 and 1 for the intercept case to obtain estimators of Λ_{xx} and $\bar{x}$ which, substituted in (9.1), yield an estimator of Σ_{xx} .

For the F-method, this works as follows: let z_k and $\tilde{x}_k' = z_k'\Gamma_x$ be the auxiliary vector and the prediction constructed in Step 0. The F-estimator of Σ_{xx} is approximately design unbiased and given by

$$\hat{\Sigma}_{Fxx} = N^{-1}\hat{\Lambda}_{Fxx} - x^*x^{*\prime} \tag{9.2}$$

where $x^* = \{\Sigma_s(x_k/\pi_k) + \Sigma_1^N \tilde{x}_k - \Sigma_s(\tilde{x}_k/\pi_k)\}/N$ and

$$\hat{\Lambda}_{Fxx} = \Sigma_s(x_k x_k'/\pi_k) + \Sigma_1^N \tilde{x}_k\tilde{x}_k' - \Sigma_s(\tilde{x}_k\tilde{x}_k'/\pi_k) \ . \tag{9.3}$$

As for the P-method's estimator, $\hat{\Sigma}_{Pxx}$, x^* is the same, but $\hat{\Lambda}_{Fxx}$ is replaced by $\hat{\Lambda}_{Pxx}$ structured as (9.3) but with the obvious changes implied by the P-method's Step 0.

Nathan and Holt (1980), Holt, Smith and Winter (1980), Smith (1981), Skinner (1982) use a model-based approach to estimating a mean vector and a covariance matrix when a selection procedure may cause bias in the straightforward estimators, unless corrective steps are taken. They use a result, described by Birnbaum, Paulson and Andrews (1950) and going back to Lawley (1943) and to Karl Pearson, about random vectors x and q following some multidimensional (superpopulation) distribution: We have made N observations on q. For a selected subset of n, we have also observed x. The selection of the n from the N may depend on q. Under assumptions that (a) the regression of x given q is linear and (b) the conditional variances and covariances of x given q do not depend on q, one arrives at "selection-corrected" estimators of the superpopulation's mean vector of x and its covariance matrix of x. One can proceed to study estimators of superpopulation regression coefficients and their model-based confidence intervals, as done, in part by Monte Carlo techniques, in Nathan and Holt (1980), Holt, Smith and Winter (1980).

The method, translated into the survey sampling setting, leads to the following estimator of the superpopulation's covariance matrix of x, see Holt, Smith and Winter (1980), Smith (1981)

$$S_{xxs} + B'_{qx}(S_{qqU}-S_{qqs})B_{qx} \qquad (9.4)$$

Here, $S_{xxs} = \Sigma_s(x_k-\bar{x}_s)(x_k-\bar{x}_s)'/n$ is the simple uncorrected sample matrix, and the corrective second term is defined by

$$S_{qqU} = \Sigma_1^N(q_k-\bar{q}_U)(q_k-\bar{q}_U)'/N \; ; \; B_{qx} = S_{qqs}^{-1}S_{qxs}$$

$$S_{qqs} = \Sigma_s(q_k-\bar{q}_s)(q_k-\bar{q}_s)'/n \; ; \; S_{qxs} = \Sigma_s(q_k-\bar{q}_s)(x_k-\bar{x}_s)'/n$$

where $\bar{q}_s$, $\bar{x}_s$, $\bar{q}_U$ are vectors of straight sample or population means. Since

the reasoning is model-based, no inclusion probabilities appear in these formulas. Instead, the correction term is meant to remove the biasing effect of, say, a non-proportional statified selection through the inclusion in q of the "design variables". They consist, in the stratified case, of a vector indicating stratum membership, and q may contain other known variables.

For an example, let q be the GH-vector indicating membership in one of the cells, labelled gh, in the crossclassification of G groups with H strata sampled by varying selection fractions. This is the situation ii(b) of Section 7, except that domains no longer exist. The estimator (9.4) can after rearrangement be partitioned as a within-cell term plus a between-cell term,

$$\sum_g \sum_h (n_{gh}/n) S_{xxs_{gh}} + \sum_g \sum_h (N_{gh}/N)(\bar{x}_{s_{gh}} - \bar{x}_{st})(\bar{x}_{s_{gh}} - \bar{x}_{st})' \qquad (9.5)$$

where

$$S_{xxs_{gh}} = \Sigma_{s_{gh}} (x_k - \bar{x}_{s_{gh}})(x_k - \bar{x}_{s_{gh}})'/n_{gh} \; ; \; \bar{x}_{s_{gh}} = \Sigma_{s_{gh}} x_k/n_{gh} \; ; \; \bar{x}_{st} = \sum_g \sum_h (N_{gh}/N)\bar{x}_{s_{gh}}$$

To compare, let us find the F-method's estimator in the same situation. In (9.2), let $\pi_k = n_{.h}/N_{.h} = f_h$ for k in stratum h, and let the z-vector indicate membership in one of the GH cells. Defining $\hat{N}_{gh} = N_{.h}\, n_{gh}/n_{.h}$, we arrive at

$$\hat{\Sigma}_{Fxx} = \sum_g \sum_h (\hat{N}_{gh}/N) S_{xxs_{gh}} + \sum_g \sum_h (N_{gh}/N)(\bar{x}_{s_{gh}} - \bar{x}_{st})(\bar{x}_{s_{gh}} - \bar{x}_{st})' \qquad (9.6)$$

In the P-method, we obtain

$$\hat{\Sigma}_{Pxx} = \sum_g \sum_h (N_{gh}/N) S_{xxs_{gh}} + \sum_g \sum_h (N_{gh}/N)(\bar{x}_{s_{gh}} - \bar{x}_{st})(\bar{x}_{s_{gh}} - \bar{x}_{st})' \qquad (9.7)$$

To set the comparison straight, note that the design-based estimators (9.6) and (9.7) were conceived to estimate the finite population parameter (9.1), which has the partitioning

$$\Sigma_{xx} = \sum_g \sum_h (N_{gh}/N) S_{xxU_{gh}} + \sum_g \sum_h (N_{gh}/N)(\bar{x}_{U_{gh}} - \bar{x}_U)(\bar{x}_{U_{gh}} - \bar{x}_U)' \qquad (9.8)$$

where U_{gh} indicates mean vector or covariance matrix for the population cell U_{gh}. By contrast, the model-based (9.5) rather estimates the superpopulation's covariance matrix, so we do not have full comparability. All three estimators (9.5), (9.6) and (9.7) require that the N_{gh} be known. They differ only in the weights attached to the within-cell matrix $S_{xxs_{gh}}$; n_{gh}/n, $\hat{N}_{gh}/N$ and N_{gh}/N.

In large samples, the common between-cell component of (9.5), (9.6) and (9.7) converges in design probability to its finite population counterpart in (9.8). Moreover, $S_{xxs_{gh}}$ converges to $S_{ssU_{gh}}$ for each cell. A priori, $\bar{\Sigma}_{Pxx}$ seems the more natural estimator since it applies the known proportions N_{gh}/N as weights for the $S_{xxs_{gh}}$. However, the weights $\hat{N}_{gh}/N$ used by $\hat{\Sigma}_{Fxx}$ gives the same large sample performance, so that both (9.6) and (9.7) converge to (9.8). By contrast, the weight n_{gh}/n of $S_{xxs_{gh}}$ causes design bias in the model-based method (9.5); it does not conform to the design-based standards of this paper, when allocation to strata is non-proportional. However, the weight n_{gh}/n seems natural under the assumed model that the covariance structure of x given q does not depend on q.

REFERENCES

Binder, D. (1982). On the variance of asymptotically normal estimators from complex surveys. Report, Institutional and Agriculture Survey Methods Division, Statistics Canada.

Birnbaum, Z.W., Paulson, E. and Andrews, F.C. (1950). On the effect of selection performed on some coordinates of a multidimensional population. Psychometrika, 15, 191-204.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63, 615-620.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley.

Fuller, W.A. (1975). Regression analysis for sample survey. Sankhya C, 37, 117-132.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. J. Roy. Statist. Soc. A, 143, 474-487.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. J. Roy. Statist. Soc. B, 36, 1-37

Lawley, D.N. (1943). A note on Karl Pearson's selection formula. Proc. Roy. Soc. Edin., Sect A, 62, 28-30.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. J. Roy. Statist. Soc. B, 42, 377-386.

Särndal, C.E. (1980). On π-inverse weighting versus best linear unbiased weighted in probability sampling. Biometrika, 67, 639-650.

Shah, B.V., Holt, M.M. and Folsom, R.E. (1977). Inference about regression models from sample survey data. Bull. Internat. Statist. Inst., 47 (3), 43-57.

Skinner, C.J. (1982). Multivariate prediction from selected samples. Department of Social Statistics, University of Southampton.

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

G-90693-30861-001-K-H

CATARINA ELFFORS

A/BF-S

Tidigare nummer av Promemorior från P/STM:

Nr

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognoser av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

Kvarvarande exemplar av ovanstående nummer kan rekvireras från SCB, P/STM, Elseliv Lindfors, 115 81 Stockholm, eller per telefon 08 14 05 60 ankn 4178.

Garnisonstryckeriet ★ Stockholm ★ 1983