SOFTWARE FOR INFERENCE ON LINEAR MODELS FROM SURVEY DATA

B. V. Shah and Lisa Morrissey LaVange, Research Triangle Institute

1. INTRODUCTION

Survey statisticians concerned with inference in regression models are faced with the difficult task of selecting the most appropriate statistical model and parameters. A variety of models and interpretations have been suggested, e.g., Konijn (1962), Godambe and Thompson (1971), Royall (1971), Kish and Frankel (1974), Fuller (1974), and Folsom (1974). However, in this paper, the discussion will be limited to an approach proposed by Kish and Frankel (1974), Fuller (1974), and Folsom (1974). The approach involves estimating regression coefficients

$$b = (x'wx)^{-1}x'wy, \tag{1.1}$$

where x'wx and x'wy are weighted sums of squares and cross products, w is the diagonal matrix of weights, x is the matrix of independent variables, and y is the vector of values of the dependent variable. An application of Wald's theorem (1943) for testing hypotheses requires a consistent estimator of the variance-covariance matrix of the estimator b. Further, the text requires inverting the matrix $\hat{V}(b)$. If the two matrices x'wx and $\hat{V}(b)$ are nonsingular and "well-conditioned" (Stewart 1973), then there are no problems. However, if either of the matrices x'wx or $\hat{V}(b)$ is singular or "ill-conditioned," the usual weighted least squares approach involving matrix inversions is not possible. The problem arises in practice when the number of dependent variables is moderately large (10 to 50) and the number of PSU's (primary sampling units) or number of pseudo-replications is not large (<50).

The objective of this paper is to review some alternative approaches to deal with singularity (or near-singularity) of x'wx or $\hat{V}(b)$ and to provide some suggestions for improvements in software for regression models from sample survey data.

2. THE MODEL AND THE NOTATION

The most popular model for infinite populations is assumed to be

$$Y = XB + e \tag{2.1}$$

where e represents random errors, B is a vector of parameters of interest, Y and X consist of observable variates for each individual in a population. For finite populations, the parameter of interest B is defined as

$$B = (X'X)^{-1}X'Y, \tag{2.2}$$

where the Y vector and X matrix are population quantities. This study is restricted to inferences about linear functions of B. The parameter B is a function of the second-order moments of the finite population of interest. If the $X'X$ and $X'Y$ matrices are replaced by second-order moments of a multivariate normal distribution, then the vector B will be equal to the population coefficients for regression of Y on X.

For the purpose of this paper, a stratified, two-stage sample design is considered. For this class of design, the population is divided into h = 1, . . . ,H strata by various geographic, demographic, and socioeconomic characteristics. For stratum-h, n(h) primary sampling units are selected with unequal probabilities without replacement. The second-stage observations consist of row vectors X(hij) consisting of the independent variable values for the second-stage unit (SSU) j from PSU i and stratum h, plus the corresponding scalar responses Y(hij). If the second-stage units are selected with equal probabilities without replacement and P(hij) represents the overall (PSU $\times$ SSU) selection probability for unit hij, and w(hij) is the weight defined as the reciprocal of P(hij), then the unbiased Horvitz-Thompson estimators for $X'X$ and $X'Y$ are x'wx and x'wy, respectively, where

$$x'wx = \sum_{h=1}^{H} \sum_{i=1}^{n(h)} \sum_{j=1}^{n(hi)} x'(hij)x(hij)w(hij) \tag{2.3}$$

$$x'wy = \sum_{h=1}^{H} \sum_{i=1}^{n(h)} \sum_{j=1}^{n(hi)} x'(hij)y(hij)w(hij) \tag{2.4}$$

These estimates can be utilized to provide an estimate of B, namely

$$b = (x'wx)^{-1}(x'wy). \tag{2.5}$$

As shown by Shah (1981), for probability samples from finite populations if the order of sample selection is ignored, then the Horvitz-Thompson (1952) estimator is the maximum likelihood estimate of the population parameter. Since b is a function of the Horvitz-Thompson estimators x'wx and x'wy, b is a maximum likelihood estimator of B. The application of Wald's theorem to test hypotheses about b requires a consistent estimator for $\hat{V}(b)$. The two most commonly used approaches for computing $\hat{V}(b)$ are (a) balanced repeated replications (BRR), and (b) Taylor series linearization (TSL). The test statistic for the hypothesis HB = 0 is:

$$t = b'H'\{HV(b)H'\}^{-1}Hb/r(H), \tag{2.6}$$

where r(H) is the rank of the matrix H. The t is approximately distributed as Hotelling's T^2 statistic. The formulas for estimating V(b) are as follows:

$$Z(hij) = (x'wx)^{-1}[x'(hij)\{y(hij) - x(hij)b\}]w(hij)$$

$$Z(hi) = \sum_{j=1}^{n(hi)} Z(hij)$$

$$ZZ'(h) = \sum_{i=1}^{n(h)} Z(hi)Z'(hi) \tag{2.7}$$

$$Z(h) = \sum_{i=1}^{n(h)} Z(hi)$$

$$\hat{V}(b) = \sum_{h=1}^{H} \{n(h)ZZ'(h) - Z(h)Z'(h)\}/\{n(h) - 1\},$$

The available software works well in all cases in which x'wx and $\hat{V}(b)$ are nonsingular and well-conditioned. In a balanced repeated replication approach, the rank of $\hat{V}$ is less than the number of pseudo-replications. For the Taylor series linearization, if the mean square between PSU's within stratum is used to estimate variances, then the rank of $\hat{V}(b)$ is less than or equal to the number of PSU's minus the number of strata.

The primary cause of singularity in x'wx is the overspecification of the model, which produces some independent variables that are linear functions of the other independent variables. The near singularity or ill-conditioning occurs when the number of independent variables is many and the sample values do not exhibit sufficient spread over the range of the variables.

In the next section, we present the approach taken in the SURREGR program developed at the Research Triangle Institute. In the subsequent sections, we present some suggestions for improvement in the software. It is hoped that this paper will initiate some discussion of these suggestions leading to improved software in the future.

3. SURREGR APPROACH

The basic approach for testing the hypothesis HB = 0 in the SURREGR program is as follows:

(a) Find a matrix H_1 such that (1) the space spanned by the columns of H_1 is a subspace of the space spanned by the columns of H, and (2) H_1B is estimable.
(b) Find a matrix H_2 such that (1) H_2 is in the same space as spanned by the columns of H_1, and (2) $H_2V(b)H_2'$ is positive definite.
(c) Perform a test of the hypothesis $H_2B = 0$ that satisfies (a) and (b) and has the maximum possible rank.
(d) If no such H_2 exists, then report the hypothesis as "not testable."

The computational details are as follows.

In order to solve for the regression coefficients, b, it is necessary to compute the inverse of x'wx. In the procedure SURREGR, an algorithm based on the Cholesky decomposition (Stewart 1973) is used when x is of full column rank. For the singular case, a generalized inverse, A^-, is computed that satisfies the following conditions:

$$\begin{aligned} &\text{(i)} \quad A = A A^- A \\ &\text{(ii)} \quad A^- = A^- A A^-. \end{aligned} \tag{3.1}$$

The maximum relative difference between the matrices on either side of these equations is reported to the user. Thus, the user is warned of any numerical inaccuracies due to the ill-conditioning of x'wx.

To test the hypothesis HB = 0, it is first necessary to find a linear function H_1 of H such that H_1B is estimable. Using the appropriate inverse of x'wx defined above, a matrix M is found that satisfies

$$MH\{I - (x'wx)^{-1}(x'wx)\} = 0. \tag{3.2}$$

(If x'wx is nonsingular, then M = I and no subspace need be used.) If we let $H_1 = MH$, then clearly the columns of H_1 form a subspace of the column space of H, and H_1B is estimable. In addition, the columns of M are eigenvectors of $H\{I - (x'wx)^{-1}(x'wx)\}$. Therefore, the singular value decomposition can be applied to obtain M. If no such M can be found (i.e., there is no estimable subspace of HB), then the program reports that the hypothesis is "not testable."

Assuming that a solution M has been obtained, the hypothesis now being tested is

$$H_1B = MHB = 0. \tag{3.3}$$

Due to previous numerical inaccuracy or ill-conditioning of the problem, the variance-covariance matrix of the estimates may be nonpositive definite. In this case, it is necessary to find $H_2 = LH_1$ such that $Var(H_2b) = H_2Var(b)H_2'$ is positive definite. By applying the Cholesky decomposition to $H_1Var(b)H_1'$, H_2 can be obtained with maximum possible rank.

The Wald statistic defined in the previous section for testing $H_2B = 0$ is then computed. A test of the hypothesis follows from the well-known asymptotic properties of Wald statistics (Wald 1943). If there is no linear function, H_2 of H_1 such that $Var(H_2b)$ is positive definite, then the program reports that the hypothesis is "not testable."

4. ESTIMABILITY

The general approach in SURREGR produces some unpleasant side effects. For example, if one tries to test the hypothesis regarding the main effects of A in the presence of the interaction AB; then no H_2 is available and hence the main effect of A is not testable. An extensive discussion of what alternate hypotheses can be tested in such a case is given by Speed et al. (1978). There is no easy solution but some facilities for the user can be provided for specifying: (a) what linear functions are to be tested, or (b) what linear restrictions on the parameters are acceptable. Alternatively, the output data set may contain the quantities x'wx and $\hat{V}(b)$, so that the user may perform further analysis using other tools such as IMSL, PROC MATRIX, and GENCAT.

5. VARIANCE ESTIMATOR

In computing the estimated variance of a nonlinear statistic, the most common simplifying assumption is that the PSU's were selected with replacement within each stratum. This is true for the SURREGR procedure. This assumption is implicitly inherent in BRR techniques. This approach produces acceptable results when the rank of x is small; but results are unacceptable if the rank of x is large, especially when the rank of x is greater than the number of PSU's.

One alternative will be to use the formula appropriate to "without replacement" sampling. The general formula is complex. If the units within PSU's were selected with equal probability and the PSU's within strata were selected with unequal probabilities, then the appropriate equation (using TSL approach) for an approximate estimator of the variance covariance matrix of b is

$$\hat{V}(b) = \sum_{h=1}^{H} \sum_{i \neq j}^{n(h)} \pi_i \pi_j \pi_{ij}^{-1} - 1)\{(Z_{hi} - Z_{hj})(Z_{hi} - Z_{jh}) \quad (5.1)$$

$$- f_{2hi} W_{hi} - f_{2hj} W_{hj}\} + \Sigma\Sigma f_{2hi} W_{hi}$$

where

$$f_{2hi} = \{1 - m(hi)/M(hi)\} \quad (5.2)$$

$$w_{hi} = \{m_{hi} ZZ'(hi) - Z(hi)Z'(hi)\}/\{m(hi) - 1\} \quad (5.3)$$

$$ZZ'(hi) = \sum_{j=1}^{m(hi)} Z(hij)Z'(hij),$$

and π_i, π_{ij} are probabilities for selection of the PSU(i), for the joint selection of PSU's i and j. This formula reduces to a simple form if the number of PSU's per stratum is 2.

Another simple expression results if the PSU's are selected with equal probability; the corresponding formula is

$$\hat{V}(b) = \sum_{h=1}^{H} [\{1 - n(h)/N(h)\}B$$

$$+ n(h)/N(h) \sum_{i=1}^{n(h)} \{1 - m(hi)M(hi)\}W]$$

where

$$B = \sum_{h=1}^{H} \{n(h)ZZ'(h) - Z(h)Z'(h)\}/\{n(h) - 1\}.$$

$$W = \sum_{h=1}^{N} \sum_{i=1}^{n(h)} \{m(hi)ZZ'(hi) - Z(hi)Z'(hi)\}/\{m(hi) - 1\}.$$

This approach is not readily applicable to BRR. Some techniques for estimating within PSU variance components have been suggested by Bean and Schnack (1977), and Schindler and Kulpinski (1981); however, it is not clear how to combine these to form an estimator of the variance under the assumption of "without replacement" sampling.

6. COMMENTS ON ACCURACY

The singular value decomposition technique (SVDT) may produce unreliable results when the matrix is ill-conditioned. It is sufficient here to quote from LINPACK by Dongarra et al. (1979):

"Finally, suppression of singular values below the error level will stabilize least squares solutions only if the significant singular values are well above the error level. What to do with small but significant singular values is a difficult and unsolved problem."

It may be preferable to avoid SVDT whenever x'wx and V(b) happen to be of full rank and well-conditioned. There is a need to include better diagnostics in regression programs (see Belsey et al., 1980). If approaches could be developed to inform the user about the group of parameters that contain either ill-conditioning or singularities, as well as the rank of x, the user could then input "acceptable" restrictions on the parameters (see Gallant and Gerig, 1980) to resolve the singularities. The final computations could then be made through the Cholesky decomposition or a similar algorithm with a more stable performance compared to the SVDT.

In conclusion, until improved software becomes available, the user must exercise care in dealing with regression problems involving near-singularities.

REFERENCES

Bean, Judy A., and Schnack, George A. (1977), "An Application of Balanced Repeated Replication to the Estimation of Variance Components," *Proceedings of the Social Statistics Section of the American Statistical Association*, 938-942.

Belsey, D., Kuh, E., and Welsch, R. (1980), *Regression Diagnostics*, John Wiley, New York.

Ben-Israel, A., and Greville, T. (1975), *Generalized Inverses: Theory and Applications*, John Wiley, New York.

Berk, Kenneth (1977), "Tolerance and Condition in Regression Computations," *Journal of the American Statistical Association*, Vol. 72, No. 360.

Dongarra, J. J., Moler, C. B., Bunch, J. R., and Stewart, G. W. (1979), *LINPAK User's Guide*, SIAM, Philadelphia.

Dwivedi, T., Srivastava, V., and Hall R. (1980), "Finite Sample Properties of Ridge Estimators," *Technometrics*, 22, 205-212.

Folsom, R. E. (1974), *National Assessment Approach to Sampling Error Estimation, Sampling Error Monograph.* Prepared for National Assessment of Educational Progress, Denver.

Fuller, Wayne A. (1974), "Regression Analysis for Sample Surveys." A report prepared for the U. S. Bureau of the Census on work conducted under the Joint Statistical Agreement, Iowa State University, Ames, Iowa.

Gallant, A., and Gerig, T. (1980), "Computations for Constrained Linear Models," *Journal of Econometrics*, 12, 59-84.

Godambe, V. P., and Thompson, M. E. (1971), "Bayes, Fiducial, and Frequency Aspects of Statistical Inference in Regression Analysis in Survey Sampling." *Journal of the Royal Statistical Society*, B, 33, 361-390.

Hajek, J. (1960), "Limiting Distributions in Simple Random Sampling From a Finite Population," *Pub. Math. Inst., Hungarian Acad. Sci.*, 5, 361-374.

Hidiroglou, Michael A., Fuller, Wayne A., and Hickman, Roy D., (1976), *SUPERCARP*, Survey Section, Statistical Laboratory, Iowa State University.

Hoerl, A., and Kennard, R. (1970), "Ridge Regression Biased Estimation for Non-orthogonal Problems," *Technometrics*, 12, 55-67.

Holt, Mary M. (1977), "SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data." Research Triangle Institute, Research Triangle Park, N. C.

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.

Kapland, Bruce, Francis, I., and Sedransk, J. (1979), "Criteria for Comparing Programs for Computing Variances of Estimators From Complex Sample Surveys." *Proceedings of the 12th Annual Symposium on the Interface*, Waterloo, Ontario.

Kish, L., and Frankel, M. R. (1974), "Inference From Complex Samples," *Journal of Royal Statistical Society*, B, 36, 1-37.

Konijn, H. (1962), "Regression Analysis in Sample Surveys," *Journal of the American Statistical Association*, 57, 590-605.

Maurer, Kurt, Jones, G., and Bryant, E. (1978), "Relative Computational Efficiency of the Linearized and Balanced Repeated Replication Procedures for Computing Sampling Variances." Presented at the Survey Research Methods Section of the ASA, San Diego.

OSIRIS IV User's Manual (1981), Seventh Edition, Survey Research Center Computer Support Group, the Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.

Royal, R. M. (1971), "Linear Regression Models in Finite Population Sample Theory," in V. P. Godambe and D. A Sprott, eds., *Foundations of Statistical Inference*, Holt, Rinehart and Winston, Ontario.

Schindler, Eric, and Kulpinski, Stanley (1981), "Components of Variance by Replicated BRR," *American Statistical Association 1981 Proceedings of the Section on Survey Research Methods*, 200-203.

Shah, B. V. (1981), "A Logic of Inference in Sample Survey Practice," *ASA 1981 Proceedings of the Section on Survey Research Methods*, p. 638-643.

Shah, B. V. (1978), "Variance Estimates for Complex Statistics From Multi-stage Sample Surveys," in K. N. Namboodiri, ed., *Survey Sampling and Measurement*, Academic Press, New York, pp. 24-35.

Shah, B. V. (1977), "Inference About Regression Models From Sample Survey Data," *Bulletin of the International Statistical Institute*, 47(3), 43-57.

Speed, F. M.; Hocking, R. R.; and Hackney, O. P. (1978), "Methods of Analysis of Linear Models With Unbalanced Data," *Journal of the American Statistical Association*, Vol. 73, No. 361.

Stewart, G. W. (1973), *Introduction to Matrix Computations*, Academic Press, New York.

Sukhatme, P. B., and Sukhatme, B. (1970), *Sampling Theory of Survey With Applications*, Iowa State University Press, Ames, Iowa.

Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large," *Transactions of the American Mathematical Society*, 54-426.