APPLICATION OF LINEAR AND LOG-LINEAR MODELS TO DATA FROM COMPLEX SAMPLES

Robert E. Fay[1]

Most sample surveys conducted by organizations such as Statistics Canada or the U.S. Bureau of the Census employ complex designs. The design-based approach to statistical inference, typically the institutional standard of inference for simple population statistics such as means and totals, may be extended to parameters of analytic models as well. Most of this paper focuses on application of design-based inferences to such models, but rationales are offered for use of model-based alternatives in some instances, by way of explanation for the author's observation that both modes of inference are used in practice at his own institution.

Within the design-based approach to inference, the paper briefly describes experience with linear regression analysis. Recently, variance computations for a number of surveys of the Census Bureau have been implemented through "replicate weighting"; the principal application has been for variances of simple statistics, but this technique also facilitates variance computation for virtually any complex analytic model. Finally, approaches and experience with log-linear models are reported.

1. INTRODUCTION

Statistics Canada has played a significant role in many of the methodological developments in the application of analytic methods to sample survey data. The intent of this paper is to review and to share some of the experience acquired by the U.S. Bureau of the Census with these same questions.

The "design-based" (also sometimes called "classical") mode of inference predominates in the analysis and presentation of data by most governmental statistical agencies, such as Statistics Canada and the U.S. Bureau of the Census, as well as by most large private survey organizations. The basis of

[1] Robert E. Fay, Statistical Methods Division, U.S. Bureau of the Census, Washington, D.C.

statistical inference with this approach is the randomization employed to select the sample from the finite population. Construction of confidence intervals and tests of hypotheses are based on a large-sample theory tied to this randomization rather than to a specific model. Standard texts such as those by Cochran [4], Kish [17], and Hansen, Hurwitz, and Madow [14] present the elements of this theory. Hansen, Madow and Tepping [15] recently argued the advantages of this approach to the problem of inference from survey data over "model-based" methods; Särndal [25] and Cassel, Särndal, and Wretman [3], have discussed the choice between the model and design-based approaches from a somewhat different point of view. Most of the original development of the design-based theory of inference was specifically for population totals, proportions, means, and ratios, and much of the corresponding literature for the model-based theory similarly concentrates on such basic statistics.

Common analytic models, such as linear regression, log-linear models, and generalized linear models, on the other hand, were initially developed in the context of explicit stochastic models, for example, the normal or multinomial distributions. "Classical" inference here has generally come to refer to statistical inferences based upon such distributional assumptions (where "classical" may include "Bayesian" in this discussion). Developments in "robust" estimation avoid specific distributional requirements, but often maintain assumptions not typically encountered in survey sampling, for example, that the error terms of the model are independent and selected from a symmetric population.

Many researchers familiar with one or more of these analytic models have applied them directly to sample survey data without recognition of the possible consequences of the sample design on the validity of inferences based on the usual distributional assumptions. The subject of this conference, of course, essentially concerns "design-based" alternatives that do reflect the effect of the design. Although all other sections of this paper will address "design-based" methods, the next section considers some of the theoretical and practical issues in choosing between these two approaches, and how these considerations appear manifested in practice at the Census Bureau.

The third section briefly describes some of our experience at the Census Bureau with design-based methods for linear regression. The fourth section discusses an approach taken in the computer implementation of replication

methods, using "replicate weights". Although principally intended for the computation of variance for the usual survey characteristics, this technique also facilitates computation of standard errors for complex models. This general approach may be particularly useful for less standard models, i.e., models other than the linear, log-linear, and other generalized linear models. Finally, some developments with respect to log-linear models are discussed, including specific computer software.

2. CHOOSING BETWEEN DESIGN-BASED AND MODEL-BASED INFERENCE FOR ANALYTIC MODELS

The choice between design-based and model-based inference may involve several factors, including effects of stratification, and existence or extent of dependence between sampled values ("clustering"). Many of the essential issues related to this general choice are enumerated by DuMouchel and Duncan [6] in their discussion of whether to incorporate survey weights in linear regression.

If $\underset{\sim}{Y}$ represents a column vector of observations Y_i, and $\underset{\sim}{X} = \{X_{ij}\}$, $j = 1, \ldots, p$ represents predictors for $\underset{\sim}{Y}$, the model

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \tag{2.1}$$

with $\underset{\sim}{\varepsilon} = \{\varepsilon_i\}$ composed of independent, identically distributed error terms $\varepsilon_i \sim N(0, \sigma^2)$, has as its maximum-likelihood estimate for $\underset{\sim}{\beta}$

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{Y}. \tag{2.2}$$

Typical survey estimation associates a weight W_i with each survey case i, based on the inverse of the probability of selection, often adjusted by factors for nonresponse and ratio estimation. If $\underset{\sim}{W}$ represents a diagonal matrix of W_i, then

$$\hat{\underset{\sim}{\beta}}_W = (\underset{\sim}{X}^T\underset{\sim}{W}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{W}\underset{\sim}{Y} \tag{2.3}$$

gives a design-consistent alternative incorporating the weights. Under the original stochastic model justifying the choice of (2.2), or, more generally, if the ε_i's are uncorrelated with zero expectations and equal variances, (2.3) has a larger sampling variance than (2.2). On the other hand, if these specific assumptions fail (particularly concerning the expectations of the ε_i's), (2.3) remains a design-consistent estimate of the census parameter, $\underset{\sim}{\beta}^*$, defined as the application of (2.2) to the values in the complete finite population, whereas computation of (2.2) for unweighted sample cases cannot guarantee consistent estimation of $\underset{\sim}{\beta}^*$.

DuMouchel and Duncan further elaborate on the issue of choosing between the variance advantage of (2.2) under the simple model and the consistency of (2.3) under model failure. Their presentation includes a number of citations to earlier commentary by others on both sides of this controversy, and can be recommended for its balanced perspective. Additionally, they propose a test, which can be performed with typical computer packages for linear regression, of whether the weighted and unweighted regressions are significantly different. If the test rejects the hypothesis that (2.2) and (2.3) are consistent estimates of the same set of coefficients, then the argument for consistency with the census value, $\underset{\sim}{\beta}^*$, favors (2.3). If the test does not reject, the authors prefer (2.2) with its (generally) lower variance.

If a researcher rejects (2.2) on the basis of the test proposed by DuMouchel and Duncan, and computes (2.3) instead, the implications of this choice are relatively clear: that (2.3) is selected over (2.2) for its consistency under failure of the model. If the test "accepts" the hypothesis, and (2.2) is used with its associated standard errors derived under the model, caution is nonetheless required in uncritically interpreting (2.2) and associated confidence intervals as statements about the census parameter $\underset{\sim}{\beta}^*$. In many applications, choice of (2.3) and its associated reliability could be defended as the only "safe" interpretation of the data as an estimate of $\underset{\sim}{\beta}^*$ when model failure is suspected, in spite of possible acceptance by the test of a hypothesis of no significant difference between the weighted and unweighted analyses.

The paper of DuMouchel and Duncan clearly illustrates the most essential consideration in choosing between model-based and design-based inference, namely, efficiency under a correctly specified model versus consistency under

failure of the assumptions of the model. Two footnotes may be added. Although ignoring survey weights is inconsistent under any design-based approach and can only be justified under model-based approaches, not all model-based inference requires ignoring the information represented in the weights.

Rubin [24] gave a concise explanation of this last point in his discussion of the paper of Hansen, Madow, and Tepping [15]. Referring to the more extensive work of Rosenbaum and Rubin [22], Rubin pointed out that a complete Bayesian interpretation of the observed data reflects not only consideration of the functional and distributional relationships in the total population (such as models like (2.1) for the complete population) but also the process by which the sample observations become observed. (In a randomized design, "propensity" to be included in the sample may be equated to probability of selection and the "propensity score" in Rosenbaum and Rubin [22].) On the basis of this consideration, Rubin [23] presented an interesting justification, from a Bayesian perspective, of the use of randomization in sample selection, a procedure that has been staunchly defended by proponents of design-based inference but treated with some disdain by many proponents of model-based inference. Consequently, Rubin advocates model-based inference tempered by careful analysis of the effects of selection or propensity to be included in the sample; these principles in some circumstances could lead to either (2.2) or (2.3), or perhaps alternatives to both.

As a second footnote, DuMouchel and Duncan explicitly restricted their attention to the issue of weighting for stratified simple random sampling. An equally important issue in many applications is the effect on inferences of clustering, that is, dependencies among sampled units due to their joint inclusion in the sample by design, such as persons in sampled households or persons in neighboring households jointly selected into sample. In self-weighting samples (where all sample cases have equal weight), design-based and model-based analyses may often produce the same estimates of the parameters of an analytic model but substantially different assessments of their reliability, unless the dependencies from clustering are explicitly incorporated into the model-based inference. Unlike the issue of the use of weights in stratified simple random samples, where a model-based approach may be defended if the error terms conform to the original full specification of the model, a known dependence among the observations due to clustering (to any serious

degree) inherently conflicts with any assumption of independence of errors that might be required by an overly simplified model. Hence, models that do not reflect known effects of clustering automatically fail to model the data properly.

Design-based inference is the institutional standard at the U.S. Bureau of the Census; yet, practice incorporates both modes of inference with respect to models. Researchers are most likely to adhere strictly to a design-based standard for inferences to national relationships based upon complex samples. When survey weights vary by only a modest degree or not at all, and the effects of clustering may be presumed small, model-based inferences for analytic models appear to enjoy acceptance. The attraction of model-based inference in these cases, no doubt, reflects less a philosophic choice than a practical one: model-based methods are more accessible and familiar than the design-based counterparts. (The author has encountered applications meeting such conditions on variation on the weights and effects of clustering where design-based methods simply duplicate model-based conclusions, thus justifying the substitution of model-based methods under similar favorable circumstances. When the weights do appreciably vary, or characteristics are subject to considerable clustering, however, examples are easily found where the two modes of inference substantially disagree, and where the model-based inference is highly questionnable.)

Specific areas of application at the Census Bureau appear almost exclusively model-based. Methods for imputation of missing data, in particular, some of which derive from explicit parametric models, characteristically avoid any consideration of design-based weights. Another specific field of study, estimation for small areas or domains, often reflects a mixed strategy of design- and model-based inference. Thus, practice at the Census Bureau appears to parallel the choice outlined by DuMouchel and Duncan: efficiency (and simplicity) under the assumed model versus consistency under model failure. Strict inference to national relationships are most likely to elicit design-based methods, while less formal analyses or analyses in which the model is hoped correct (missing data) often favor a model-based approach.

3. DESIGN-BASED INFERENCE FOR LINEAR REGRESSION AT THE U.S. CENSUS BUREAU

In general statistical practice, linear regression is probably the single most popular analytic technique. Most data collected by the Census Bureau, particularly for the "demographic areas" involving characteristics of persons or housing, are categorical: linear regression, in any form, is used relatively seldom at the Census Bureau by comparison.

Fuller [13] developed basic results in design-based inference for linear regression, using methods based upon Taylor-series expansions (linearization). These results are incorporated in the computer program SUPER CARP [16], whose development was partially supported by the U.S. Bureau of the Census. We can report successful use of the program ourselves, although it has been applied to only a few problems thus far. The report by Moore [26] is probably the most accessible illustration of the use of SUPER CARP at our institution.

The next section discusses the implementation of replication methods through replicate weights, and we have given preliminary thought, but not yet attempted to implement, alternative computer software specifically designed for this approach. No substantial philosophic difference with SUPER CARP is implied by these considerations, although replication methods tend to give slightly larger and thus more conservative standard errors than linearization. The intent in developing this software would be to take advantage of replication methods developed for some of our surveys, which can be made to reflect the effects of complex estimators more completely than programs implementing linearization.

4. COMPUTING DESIGN-BASED VARIANCES THROUGH REPLICATE WEIGHTS

Replication methods, such as jackknife, half-sample, and bootstrap techniques, represent the principal general alternative to linearization for design-based variance estimation for nonlinear statistics. Kish and Frankel [18] presented an early discussion of the use of replication for such purposes and much research has been conducted since.

The popularity of replication for variance estimation has gone through

cycles. Linearization is a powerful technique, of course, and relationships presented by Binder [1] facilitate its implementation for a wide class of analytic models. Census Bureau surveys tend to employ quite complex estimators, however, and fully representing the effect on the sampling variances of these estimators has frequently proven to consume large amounts of professional time, both by statisticians and, especially, experienced computer programmers. Recently, variance computations for a number of surveys have used replication methods achieved through a "replicate weighting" approach. The principal features of this method are to provide a unified approach to enable the computation of variances for a large number of survey characteristics and to simplify the estimation of variance for complex analytic statistics.

The replicate weighting approach is not a new discovery: some of its earlier history is reported in [5], which also describes experience acquired by the U.S. Bureau of Labor Statistics, Bureau of the Census, and Westat, Inc. The algorithm may be said to represent the variance from a (possibly complex) design and a (possibly complex) survey estimator in the form of data to be associated with the survey data file rather than as a set of (possibly complex) variance formulas requiring computer programming. Familiar replication methods, such as balanced half-samples and the jackknife, may be represented through replicate weights, but the algorithm also facilitates the implementation of a much wider class of resampling plans, as in [7]. In [10], it is shown that there exists a resampling plan (actually an infinite number of resampling plans) corresponding to essentially any familiar variance estimator for estimates of population totals, such as variance expressions for multi-stage designs, Yates-Grundy estimators, etc. By representing complex variance relationships as data, variance computation becomes accessible to a larger group of data users.

Estimation in many surveys assigns weights W_{i0} to each case i, so that for any characteristic X_i, estimates of total are given by the weighted sum of the characteristic times the survey weight

$$\hat{X}_0 = \sum_i W_{i0} X_i . \tag{4.1}$$

The product of the replicate weighting approach is a set of additional weights W_{ir}, r = 1, ..., R, for each survey case i, from which alternative estimates of total

$$\hat{X}_r = \sum_i W_{ir} X_i \tag{4.2}$$

may be computed. The estimate of variance is given by

$$\hat{Var}(\hat{X}_0) = \sum_{r=1}^{R} d_r (\hat{X}_r - \hat{X}_0)^2 \tag{4.3}$$

for predetermined d_r independent of the choice of survey characteristic X. (As an example, a simplified balanced half-sample estimate of variance, ignoring the effect of any complex survey estimation reflected in the weights W_{i0}, would be given by assigning weights W_{ir} equal either to $2W_{i0}$ or to 0 according to whether case i was included in half-sample r, and setting $d_r = 1/R$ for each r.) More generally, for a smooth function S that are functions of weighted population estimates of total $\hat{X}_0^{(1)}$, ..., $\hat{X}_0^{(k)}$, each of the form (4.1),

$$\hat{Var}\{S(\hat{X}_0^{(1)}, \ldots, \hat{X}_0^{(r)})\} = \sum_{r=1}^{R} d_r \{S(\hat{X}_r^{(1)}, \ldots, \hat{X}_r^{(k)}) - S(\hat{X}_0^{(1)}, \ldots, \hat{X}_0^{(k)})\}^2 \tag{4.4}$$

The estimator S in (4.4) may stand for the sometimes extremely complex estimators often used in survey estimation, incorporating noninterview adjustments and ratio or iterative ratio estimation. Furthermore, these forms of complex survey estimation, if incorporated in the weights W_i, may be included in the derivation of W_{ir} as well. Thus, variance computation with this approach falls naturally into three distinct steps or phases:

1. Generate replicate basic weights W_{ir}* for the simple unbiased (Horwitz-Thompson) weighting of the data given by the basic weights W_{i0}*.
2. Compute replicate (final) weights, W_{ir}, by applying the same noninter-

view and ratio estimators to the replicate basic weights, W_{ir}*, as the original estimation procedures used to compute W_{i0} from the W_{i0}*.

3. Apply (4.4) to the estimation of variance of simple or complex statistics.

The modularity of the preceding three phases is a key feature of this technique: general programs may be used to perform phases 1 and 2, or custom programs may be written to cover unusual circumstances as required. For a single survey, phases 1 and 2 need be performed only once. Programs for phase 3 need take no specific note of the design or estimator and can be run as needed by any user with access to the replicate weights W_{ir} produced in the second phase.

Although most applications of this method at the Census Bureau have been to estimate variances for basic survey characteristics such as means, totals, or proportions, (4.4) lends itself well to analytic purposes as well. This approach fully represents the effects of complex designs and estimators, whereas in practice implementation of linearization often is restricted to the more common and simple situations. Furthermore, although specific computer software may be developed to implement linearization for common analytic methods, such as linear regression, log-linear models, generalized linear models, etc., formula (4.4) enables researchers to compute variances for more specialized analytic models for which no linearization methods have been programmed, since (4.4) only requires that the researcher apply complete data algorithms to the alternative estimates produced by the replicate weights.

5. DESIGN-BASED INFERENCE FOR LOG-LINEAR MODELS

Log-linear models, which express the logarithm of the expected frequencies for categorical responses as a linear function of unknown parameters, encompass both factorial models for cross-classified categorical data, and logistic models for one or more dependent categorical variables as a function of any combination of categorical and continuous predictors. Bishop, Fienberg, and Holland [2] provided one of the earliest books in this rapidly expanding field.

Many log-linear models, particularly those for fully cross-classified categorical data, involve a large number of parameters. The three most typical problems of inference are:

1. To compute standard errors and confidence intervals for the individual estimated parameters,
2. To test the significance of the contribution of specific sets of parameters to the fit of a model,
3. To test the overall goodness-of-fit of the model.

In the context of simple random samples, standard results in maximum likelihood theory provides an answer to these questions, although the Pearson chi-square test rightfully enjoys greater popularity than the likelihood-ratio chi-square test as a solution to the third problem.

Koch, Freeman, and Freeman [19] extended the Weighted Least Squares (WLS) method to complex samples, thereby providing solutions to each of the three principal inferential problems. While this method has proven of substantial general use, it is limited in some applications by the necessity to produce highly precise estimates of the design-based covariance of the sample estimates before the asymptotic theory approximates the actual performance of the WLS procedures. (Further comments on the limitations of WLS are given in [8] and [11].)

Fellegi [12] made an early contribution to the development of alternative tests to WLS for specific situations. More recently, Rao and Scott [20], [21] have formulated and extended a set of related methods to cover the problem of testing for a general class of models including log-linear models. Development of these methods has been closely associated with Statistics Canada.

A less well-known "jackknife chi-square test" [11] gives an alternative approach to the general problem of design-based tests of hypotheses. This test is based upon replication, using (4.4) and a similar expression related to the approximation of the first-order bias (as in the usual jackknife) to draw approximate inferences about the null hypothesis distribution of the usual chi-square tests applied directly to the weighted survey estimates. The method shares much in common with those developed by Rao and Scott. Although a full comparison of the relative merits the jackknifed test and the tests

proposed by Rao and Scott has not been conducted, the preliminary suggestion is that both work well and neither entirely dominates the other. (Further comments are given in [11].)

The jackknifed tests do appear somewhat easier to implement, however, especially to tables involving a large number of cells. A FORTRAN computer program, CPLX (described in [8] and documented by [9]), implementing the jackknifed tests for factorial log-linear models for cross-classified data is now in the public domain. The program also computes replication-based standard errors for parameters of log-linear models, thus also addressing the first of the three problems of inference listed earlier. Although CPLX fits well into an environment in which other survey variances are also estimated through replication approaches, such as the replication weighting techniques described in the previous section, these circumstances are by no means necessary to use the program, and a number of researchers within and outside the Census Bureau have applied the program in a variety of settings.

In time, the author hopes to be able to incorporate the methodology of Rao and Scott into a program like CPLX in order to make both methods available. For the short term, however, the current version of CPLX should be of help to researchers seeking design-based inferences from survey data.

REFERENCES

[1] Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Samples. International Statistical Review 51: pp. 279-292.

[2] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis. Cambridge, MA: MIT Press.

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

[4] Cochran, W.G. (1977). Sampling Techniques (3rd ed.). New York: John Wiley.

[5] Dippo, C.S., Fay, R.E., and Morganstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. Prepared for presentation at the annual meetings of the American Statistical Association, Section on Survey Research Methods.

[6] DuMouchel, W.H., and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association 78: pp. 535-543.

[7] Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

[8] Fay, R.E. (1982). Contingency Tables for Complex Designs: CPLX. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 44-53.

[9] Fay, R.E. (1983). CPLX - Contingency Tables Analysis for Complex Sample Designs, Program Documentation. Unpublished report, Washington, D.C.: U.S. Bureau of the Census.

[10] Fay, R.E. (1984). Some Properties of Estimates of Variance based on Replication Methods. Prepared for presentation at the annual meetings of the American Statistical Association, Section on Survey Research Methods.

[11] Fay, R.E. (1984). A Jackknifed Chi-square Test for Complex Samples. To appear in the Journal of the American Statistical Association.

[12] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness-of-Fit Based on Stratified Multistage Samples. Journal of the American Statistical Association 75: pp. 261-268.

[13] Fuller, W.A. (1975). Regression Analysis for Sample Survey. Sankhyā C 37: pp. 117-132.

[14] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vols. I and II. New York: John Wiley.

[15] Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association 78: pp. 776-793.

[16] Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1978). Super Carp (3rd edition). Ames, IO: Statistical Laboratory, Iowa State University.

[17] Kish, L. (1965). Survey Sampling. New York: John Wiley.

[18] Kish, L., and Frankel, M.R. (1974). Inference from Complex Samples. Journal of the Royal Statistical Society, Ser. B 36: pp. 1-37.

[19] Koch. G.G., Freeman, D.H., and Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Samples. International Statistical Review 43: pp. 59-78.

[20] Rao, J.N.K.. and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness-of-Fit and Independence in Two-Way Tables. Journal of the American Statistical Association 76: pp. 221-230.

[21] Rao, J.N.K., and Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Annals of Statistics 12: pp. 46-60.

[22] Rosenbaum, P.R.R., and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies. Biometrika 70: pp. 41-55.

[23] Rubin, D.B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics 6: pp. 34-58.

[24] Rubin, D.B. (1983). Comment: Probabilities of Selection and Their Role for Bayesian Modeling in Sample Surveys. Journal of the American Statistical Association 78: pp. 803-805.

[25] Särndal, C.-E. (1978). Design-Based and Model-Based Inference in Survey Sampling. Scandinavian Journal of Statistics 5: pp. 27-52.

[26] U.S. Bureau of the Census (1982). Preliminary Evaluation Results Memorandum No. 31: Evaluating the Public Information Campaign for the 1980 Census - Results of the 1980 KAP Survey. Prepared by Jeffrey C. Moore, Washington, D.C.