Statistics Canada Statistique Canada

SURVEY METHODOLOGY

June 1984
Volume 10
Number 1

SPECIAL EDITION

Analysis of Survey Data – Issues and Methods

A Journal produced by
Statistics Canada

Canada

INTERNATIONAL SYMPOSIUM ON SMALL AREA STATISTICS

OTTAWA, CANADA MAY, 22-24, 1985

In recent years the demand for reliable information on small areas has greatly increased in many countries. Some important work, both theoretical and practical has been carried out by researchers at Universities and National Statistical Bureaus. The International Symposium will provide a forum where views, ideas and results of such work could be discussed and exchanged.

The symposium is jointly sponsored by Statistics Canada and the Laboratory for Research in Statistics and Probability of Carleton University and the Department of Mathematics and Statistics of the University of Montreal. The Symposium is expected to attract over 200 participants from different countries and different organizations (Universities, government and business). There will be 20 invited speakers who will present papers on the following topics:

Synthetic Estimation
Other modeling approaches to estimation
Demographic Methods
Application in Social and Economic Areas
Policy Aspects
Organizational Experiences

In addition, contributed sessions are also planned. Those who intend to present contributed papers should forward the abstracts by January 31st 1985. It is intended to publish the proceedings soon after the Symposium, therefore full paper will be needed by April 30, 1985. The organizing committee consists of:

R. Platek - Statistics Canada
J.N.K. Rao - Carleton University
C.E. Sarndal - University of Montreal
M.P. Singh - Statistics Canada.

Please forward abstracts to, R. Platek, Statistics Canada, 4-C7, Jean Talon Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6 (Canada).

SURVEY METHODOLOGY

June 1984 Vol. 10 No. 1

Special Edition

A Journal produced by Statistics Canada

CONTENTS

8-3200-501
Reference No.
~~Z — 079~~
#12-001
ISSN: 0714-0045

SURVEY METHODOLOGY

June 1984 Vol. 10 No. 1

Special Edition

A Journal produced by Statistics Canada

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints. data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed; however. the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.

PREFACE

This issue is devoted to presenting papers given at a symposium entitled **Analysis of Survey Data - Issues and Methods,** held at Statistics Canada on Thursday May 3, 1984.

The symposium was jointly sponsored by Methodology Research Committee at Statistics Canada and the Laboratory for Research in Probability and Statistics at Carleton and Ottawa Universities. The aim of this symposium was to demonstrate how recent developments in the area of analysis of data from complex surveys could be applied to analytic studies in Statistics Canada.

The symposium opened with remarks from the Chief Statistician, Martin B. Wilk, who emphasized the importance that Statistics Canada places in enhancing its research and development capacity and in the joint endeavours by the practitioners and academics on such issues. The symposium consisted of two sessions: - A morning session, chaired by Leslie Kish of the Institute for Social Research at the University of Michigan, which included contributions from Statistics Canada presented by D. Binder, P. Cholette, L. Heslop and S. Kumar, in addition to the presentation of an overview of the analysis issues by the Chairman.

The afternoon session chaired by the Deputy Chief Statistician, Ivan P. Fellegi started with brief remarks from the chair and included papers from R. Fay, U.S. Bureau of the Census and W. Fuller, Iowa State University. The session concluded with general discussion of the developments on the data analysis issues led by J.N.K. Rao, Carleton University. Well over 200 participants from various Universities and Federal and Provincial Government Departments attended the symposium.

A selected bibliography on the topic compiled by the Project Team on the Analysis of Data from Complex Surveys is also given at the end.

ON ANALYTICAL STATISTICS FROM COMPLEX SAMPLES[1]

Leslie Kish[2]

I want to plead the case that an important and urgent task facing mathematical statistics consists of providing useful expressions for analytical statistics for complex sample designs. I should like to describe these problems to mathematical statisticians who should find them interesting because they meet the criteria of all good problems: they are important, unsolved and solvable.

The most important and difficult problems of survey sampling still await adequate mathematical treatment: the textbooks are aimed almost entirely at producing good estimates of aggregates, means and ratio means. One may also deal with the differences of two of these, but there is only fleeting and occasional reference to this problem. However, with that we come to the end of the statistical tools available for complex samples.

As sampling theory developed, probability sampling has been capturing the field of respectable sampling practice with sample designs, which are often simultaneously economical and complex. One result has been an increasing volume of sample survey data which is of high quality and which researchers wish to put to more involved analytical use. But the mathematical statistics for doing this validly are lacking. The available analytical statistics assume independence among the selected elements: but this independence is lacking in complex sample designs. Thus the researcher may be forced to forego the analysis which he considers desirable and valuable. But if he is too impatient or too ignorant for that act of self-denial, he may go ahead and use the <u>srs</u> formulas he finds in books on statistics, which often result in very serious errors.

I hope that mathematical statisticians will be impressed with the importance of the unsolved problems of analytical statistics for data arising from complex sample designs. The lack of these is a more frequent source of gross mistakes than any other kind of departure from the usual assumptions.

---

[1] Overview talk for the symposium.

[2] Leslie Kish, Institute for Social Research, The Univeristy of Michigan.

These problems are important, unsolved and interesting. You may ask: are they solvable now? Supporting my affirmative answer are three sources of justification. First, we observe the great recent advances in statistical theory. Secondly, the rapid increases in the quantity and quality of electronic computing machines make the time ripe for the solution of some of these problems. There is new interest in a general method which holds promise of rapid advance toward useful approximations. At the Survey Research Center we are now introducing this method for computing estimates of variances for regression coefficients and other statistics for which formulas are not now available.

It seems to me that this procedure resembles that of Alexander when he "solved" the Gordian knot. From a theoretical viewpoint I don't know whether it constitutes a solution of the problem or its avoidance. But insofar as it promises to give good approximations for much needed variances the practicing statistician will welcome its development with enthusiasm and interest. In this way one may obtain estimates of the confidence intervals of some analytical statistics for which specific formulas are not now available.

All of the above is verbatim from my talk to a joint session of the American Statistical Association and the Institute of Mathematical Statistics in 1957. Since then our situation has changed but little. Our 1957 hopes for that cut of the Gordian knot is now much used as BRR or Balanced Repeated Replications (Kish and Frankel 1970, 1974). But my moving plea for distribution theory for doubly complex analytical statistics did not move the mathematical statisticians. I know now why not, since I am sadder and wiser now. First, statisticians like other scientists work not on what solutions are needed but on those that seem feasible at the time. (Like nuclear bombs, for example.) Second, distribution theory for complicated statistics for complex samples seems too difficult to solve. Third, the solutions would have too many parameters to be useful. Thus my views in (Kish 1978) and today are more sober: "New computational methods can give us approximate variances that appear satisfactory for practical purposes. However, it would be more satisfying to have mathematical distribution theory for analytical statistics (e.g. regression coefficients) without the assumptions of independence, but with complex correlations between sample observations. We may hope for some progress, but not for generally useful results, because of mathematical complexities, and even more because the numbers of needed parameters will prove too great for practical utility."

Here follow seven important points about complex samples put boldly. They are not all widely known or believed, but I ask you to know, believe, use and teach them, as I do.

1. The effects of complex designs must be considered separately for point estimates and for probability statements, like confidence intervals or tests of hypotheses. For point estimates we have for all sample designs consistent approaches to parameters from similar probability-weighted (H-T) estimators. But the probability statements like confidence intervals are highly subject to design effects, especially in cluster sampling.

a) "Statistics (means, regression coefficient, etc.) approach their population values as the sample size increases.

b) The approach is generally slowed by design effects.

c) The design effects differ for different statistics, for different variables and different sample designs." (Kish and Frankel, 1974).

That paper also presents the most convincing evidence for these points: and evidence is widespread; e.g. (Verma et al 1980). Nevertheless two famous statisticians completely misstated our position in discussions of our paper: "Here the authors make the important observation that the confidence interval statements for the unknown parameter are numerically not much affected by the lack of independence of observations introduced by complex survey techniques such as stratified cluster sampling." Alas, that mistake gets quoted by other theoreticians who fail to read our answer of survey samplers: "They misunderstand completely our principal and repeated message: that confidence interval statements are numerically greatly affected by the lack of independence of observations introduced by complex survey techniques such as stratified cluster sampling." (Kish and Frankel, 1974).

This misunderstanding shared by naive non-statisticians with sampling theorists causes troubles for us survey samplers: hence we are working on a clearer statement.

2. Do we need sampling errors for analytical statistics for data from complex surveys? Or have a few of us been devoted to a negligible even trivial problem? I feel like a St. Sebastian, the target practice for the

slings and arrows of diverse outrageous heathen. (Mixed metaphors are better than fixed or random.) First come the market researchers and pollsters who ignore us, though some have learned to put a $\sqrt{}$ between a 2 and (pq/n). Second, some demographers write that with their large samples and larger measurement errors they have no time for sampling errors. Third come the mathematical psychologists, econometricians and biometricians who take their linear models straight from mathematical statistics, and that hurts. Fourth, even more hurtful are the mathematical statisticians themselves, who either forget that their n's do not justify their means, or they invoke IID, or they use some Bayesian exorcism against the spirits of the sample design. Fifth and worst are sampling theorists who display theorems to prove that, with completely specified models of arbitrary superpopulations, we need not worry about whence or how our elements were selected, nor weight them for unequal selection probabilities. They even convince a few survey samplers that they can dwell on some Olympus with their models and not come down to earth where the population lives.

From these necessarily brief remarks you notice that I am an extremist for several reasons: a) Design effects for analytical statistics provide common evidence for imperfectly specified models for the best stratified samples: b) We frequently find the effects of selection weights on samples: c) Relations between predictor and predictand variables exist in actual individuals, and they in real populations, and these interact with sample designs. (I am developing these points in a book on Statistical Design for Social Research for Wiley, 1985.)

My philosophy is consistent, but in practice I am less dogmatic. I recognize that in practice: a) it is never possible to cover completely our target populations, hence we must always resort to models for inference; b) probability sampling is too costly and not feasible for most experiments; c) despite lack of randomization either in selection or in treatments, we often blunder our way to reliable results with care, replication, design, additivity and a little bit of luck.

3. Analytical statistics begin with subclasses and with their comparisons. In the last three decades much useful material has been published about variances and design effects for subclasses. There are masses of empirical results and several useful guiding rules based on them (Kish 1980, Kish et

al. 1976, Verma et al. 1980), also some recent theory (Rust 1984, Chapter 6).

a) Distinquish between proper domains and the more common crossclasses, on which we focus here.

b) Selection probabilities are preserved for crossclasses but sample sizes become highly variable.

c) Estimates of totals and means from complex samples are retained in ratio and conditional forms.

d) Design effects for crossclasses tend to approach to almost 1 proportionately as the subclass sizes per primary cluster approach 1. This approximate model needs care and qualification but it is preferable to all venerable alternatives about design effects: that it is simply 1, or some other constant, or the same as for the entire sample. The pooled model may be often better than separate and highly variable computations.

4. Comparisons of paired means tend to have design effects greater than 1 but considerably less than the sum of the two variances. These reductions due to positive covariances (hence to a kind of additivity) have been found widely and regularly for comparisons both of crossclasses and of periodic surveys (Kish 1965, 14.1, also the above).

5. For complex analytical statistics several methods exploit the potentialities of electronic computing: Taylor linearized (delta) methods, including machine differentiation, Balanced Repeated Replications and Jackknife Repeated Replications, all have been shown to yield useful estimates of variance and design effects for complex samples (Kish and Frankel 1970 and 1974; Woodruff and Causey 1978), Bootstrapping may also be added in the future (Rao 1984).

Analytical statistics consistently show design effects greater than 1, significantly greater in every sense, but also lower than design effects for means. The relations of design effects between diverse coefficients and comparisons with those for means show some regularities.

For useful guidance we need not only more empirical work but also more results from sampling theory and model building. I am disappointed frankly that since our early work we have not seen more publications in theory and models that would be directly useful for guiding inference for actual data. The empirical bases of design effects are necessary, but to satisfy our intellectual needs for understanding we need more theory and better models.

Furthermore, even our practical needs remain unsatisfied with merely empirical design effects, because they are functions jointly of the variables, of the type of estimates, of the sample design used and of the population basis for the data. That four-dimensional source of variation is too complex and we need theory to construct models for greater simplicity.

6. Categorical data analysis is an important area, rapidly developing, and several contributions have been made to apply these methods to complex survey data (Fay 1982; Landis et al. 1982: Koch et al. 1975). These also have implications for analysis of variance where some of the earliest models were started, but not followed (Kempthorne and Wilk, 1955; Tukey and Cornfield).

7. As for the future I am hopeful about contributions from theory to applications but for two exceptions. First, mathematical statistics has not and will not give us complete distribution theories that will be useful directly, because there are too many parameters in the double complexity of analytical statistics from complex surveys. Second, model builders cannot make those complexities vanish. They will however guide us toward better and more comprehensive inference. Also toward better utilization and presentation of analytical statistics from complex surveys.

REFERENCES

[1] Fay, R. (1982). Contingency table analysis for complex sample designs: CPLX. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 44-53.

[2] Kish, L. (1957). Some unsolved problems of complex samples. Paper for Joint Meeting of the American Statistical Association and Institute for Mathematical Statistics.

[3] Kish, L., and Frankel, M.R. (1970). Balanced repeated replications for standard errors. JASA, 65, pp. 1071-94.

[4] Kish, L., and Frankel, M.R. (1974). Inference from complex samples. JRSS (B), 36, pp. 1-74.

[5] Kish, L. (1980). Design and estimation for domains. The Statistician (London), 29, pp. 209-22.

[6] Kish, L., Groves R.M., and Krotki (1976). Sampling errors for fertility surveys. Occasional paper 17, London: World Fertility Surveys, 61 pages.

[7] Koch, G., Freeman, D., and Freeman, J. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 43, pp. 53-59.

[8] Landis, J.R., Lepkowski, J., Eklund, S., and Stehouwer, S. (1982). A statistical methodology for analyzing data from a complex sample survey. Vital and Health Statistics, Series 2 - No. 92. DHHS Publ. No. 82-1366. Public Health Service, Washington, U.S. Government Printiq Office.

[9] Rao, J.N.K. (1984). Bootstrap inference with stratified samples. (Submitted for Publication).

[10] Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys. The University of Michigan, Ph.D. dissertation.

[11] Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. JRSS (a) 143, pp. 431-73.

AN INTRODUCTION TO LINEAR MODELS AND GENERALIZED LINEAR MODELS: CONCEPTS AND METHODS

David A. Binder[1]

Univariate statistical models, linear regression models and generalized linear models are briefly reviewed. Examples of a two-way analysis of variance, a three-way analysis of variance and logistic regression for a three way layout are given.

1. INTRODUCTION

The purpose of this presentation is to give a bird's-eye view of some of the concepts used in statistical applications for modelling data

The use of data sampled from a population to estimate means and proportions is now a common practice. In Section 2 we briefly review this concept and describe the interval estimates obtained from constructing confidence intervals.

Linear regression and analysis of variance models are often used to reduce multi-dimensional data to a model consisting of a few parameters. This tool is a valuable device for the analyst looking for a deeper understanding of a complex data set. These methods are reviewed in Section 3.

The concepts of linear regression methods can be extended to a much wider class of models through the generalized linear models described by Nelder and Wedderburn (1972). This is particularly useful when the dependent variable is categorical as opposed to continuous. In Section 4 we review the structure of these models.

Brief mention of appropriate diagnostics to guard against model failure and to detect multicollinearities is given in Section 5.

[1] David A. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada.

2. UNIVARIATE MODELS

2.1 Binomial Models

Suppose we have a large population from which we will select a sample and we take an observation from each selected unit. If the sample size is n, we denote the observations by $Y_1, Y_2, \ldots, Y_n$. The purpose of collecting this data is that we would like to make some inferences about the population based on this sample. For example, our population could be residents of Canada and our data are defined as

$$Y_j = \begin{cases} 1 \text{ if the person was born in Canada} \\ 0 \text{ if the person was born outside of Canada,} \end{cases}$$

for the j-th individual selected. Based on this sample we would like to make some inferences on the proportion of people in the population who were born in Canada.

If a simple random sample of $n = 5000$ residents is selected and the actual proportion of persons born in Canada is $p = 0.85$, then the number of persons in our sample who are born in Canada will be a random variable with a binomial distribution given by

$$f(y) = \binom{5000}{y}(.85)^y(.15)^{n - y}; \; y = 0, 1, \ldots, 5000.$$

In this case, since we know $p = .85$, we can completely describe the properties of $Y = \Sigma Y_j$, the total in our sample who are born in Canada. For most statistical applications, though, we do not know all the characteristics of the population and we use our sample to make inferences about this population. For example, suppose we do not know the value of p in the previous example. Then we can say that the number of persons in our sample who were born in Canada will be a binomial random variable having a distribution given by

$$f(y) = \binom{5000}{y}p^y(1 - p)^{5000 - y}; \; y = 0, 1, \ldots, 5000.$$

Now, the usual estimator for p, based on this data is $\hat{p} = \bar{Y} = \Sigma Y_j/5000$. We let $s(\hat{p}) = \{\hat{p}(1-\hat{p})/(5000)\}^{\frac{1}{2}}$. This is our estimate of the standard error of $\hat{p}$. Now, it turns out that $\hat{p} \pm 1.96\, s(\hat{p})$ is a random interval which has a 95% chance of including the true unknown value of p. This interval is called a 95% confidence interval. By changing the value of 1.96 we would either shorten or lengthen the confidence interval, thus changing the coefficient from 95% to some other value. These coefficients can be obtained from probabilities associated with the standard normal distribution.

We have described the binomial model via a simple random sample from a large population. Thus, all our inferences pertain to that population. However, in many contexts we would like our inferences to relate to other populations which we believe have been generated under similar conditions. For example, the number of deaths in Canada from a particular age-sex group in a given year may be thought of as a single realization from a binomial model, where each individual has the same probability of dying and the individual deaths are essentially independent. If this probability of dying is constant over a number of years then the number of deaths in one year can be used to make inferences for other years, even though the populations are different. (Life insurance companies and their actuaries rely on these types of assumptions in their calculations.) Providing that individual deaths are independent, assumptions about constancy of the probability of death are testable using these binomial models.

It should be pointed out that by using some generalized linear models to be described in Section 4, it may be possible to improve on the assumption of constant probabilities for all individuals, by allowing the probabilities to depend on other factors such as age, sex, health status, smoking habits, weight, etc.

2.2 Normal Models

An important distribution used in modeling data is the normal distribution given by

$$f(y) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}; \; -\infty < y < \infty.$$

The population mean is μ and is usually the parameter of interest. The population variance is σ^2.

If we observe data Y_1, Y_2, ..., Y_n from this population, our usual estimator for μ is $\hat{\mu} = \bar{Y} = \Sigma Y_j/n$. Our estimator for the standard error of $\hat{\mu}$ is given by $s(\hat{\mu}) = s/n^{\frac{1}{2}}$, where

$$s^2 = \Sigma (Y_j - \hat{\mu})^2/(n - 1).$$

As in the case of the binomial model, for large samples the 95% confidence interval is given by $\hat{\mu} \pm 1.96s(\hat{\mu})$. This is a random interval which has a 95% chance of including the true value of μ. For small samples (e.g. $n < 60$), the value 1.96 may be replaced by the appropriate value from the t distribution for more accurate intervals. Other confidence coefficients may also be obtained by changing the value 1.96 to the appropriate percentile from the standard normal or t distribution.

In some applications, the assumption of constant variance is unrealistic, particularly in the linear models to be discussed in Section 3. A simple extension of this model is to assume that the variance of X_i is given by σ_i^2 where $\sigma_i^2 = \sigma^2/w_i$. Here we assume that w_1, w_2, ..., w_n are known weights. In this case $\hat{\mu} = \Sigma w_j Y_j/\Sigma w_j$, a weighted average of the data. Also $s(\hat{\mu}) = s/(\Sigma w_j)^{\frac{1}{2}}$, where

$$s^2 = \Sigma w_j(Y_j - \hat{\mu})^2/(n - 1).$$

Confidence intervals for μ are obtained analogously. It should be pointed out here that the weights, w_1, ..., w_n are based on the normal model specification and are usually unrelated to sampling weights which are derived from complex survey designs from finite populations. When fitting models to finite populations based on data from a complex survey design, the analyst <u>may</u> wish to incorporate both the model weights as well as the sampling weights in the estimation.

2.3 Exponential Family Models

The binomial and normal models just described can be viewed as special cases of a much wider class of models known as the exponential family. The general form which we will use for this model is given by:

$$f(y_j) = \exp[\kappa_j \{y_j\theta - b(\theta)\} + c(y_j, \kappa_j)],$$

where y_j takes values which do not depend on θ.

We assume $\kappa_j = \kappa w_j$ where $w_1, \ldots, w_n$ are known. In many cases κ will also be known.

Example 1 (Binomial Proportion)

We let $\bar{y}_j = y_j/n_j$ be the sample proportion from a binomial model based on n_j observations. Therefore we have:

$$f(\bar{y}_j) = \binom{n_j}{n_j\bar{y}_j} p^{n_j\bar{y}_j}(1-p)^{n_j(1-\bar{y}_j)}; \ \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \ldots, 1,$$

$$E(\bar{y}_j) = p, \ \mathrm{Var}(\bar{y}_j) = p(1-p)/n_j,$$

$$\theta = \log[p/(1-p)].$$

$$\kappa_j = n_j,$$

$$b(\theta) = \log(1 + e^{\theta}).$$

Example 2 (Normal)

Suppose y_j is normally distributed with mean μ and variance σ_j^2. We have:

$$f(y_j) = (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\frac{y_j - \mu}{\sigma_j})^2\}: \ -\infty < y_j < \infty$$

$$E(y_j) = \mu, \qquad \mathrm{Var}(y_j) = \sigma_j^2,$$

$$\theta = \mu,$$

$$\kappa_j = 1/\sigma_j^2,$$

$$b(\theta) = \mu^2/2.$$

Example 3 (Poisson Mean)

Suppose y_j is Poisson with mean $n_j\lambda$. Letting $\bar{y}_j = y_j/n_j$, we have:

$$f(\bar{y}_j) = e^{-n_j\lambda} (n_j\lambda)^{n_j\bar{y}_j}/(n_j\bar{y}_j)!; \qquad \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \ldots,$$

$$E(\bar{y}_j) = \lambda, \qquad \text{Var}(\bar{y}_j) = \lambda/n_j,$$

$$\theta = \log \lambda,$$

$$\kappa_j = n_j,$$

$$b(\theta) = e^{\theta}.$$

Example 4 (χ^2)

Suppose y_j has a $\sigma^2\chi^2_{\nu_j}/\nu_j$ distribution. This is common for analysis of variance and variance components models, where y_j is the mean-square. Then, we have:

$$f(y_j) = y_j^{(\nu_j - 2)/2} \left(\frac{\nu_j}{2\sigma^2}\right)^{\nu_j/2} \exp\{- y_j\nu_j/(2\sigma^2)\}/\Gamma(\nu_j/2); \quad y_j \geq 0,$$

$$E(y_j) = \sigma^2, \qquad \text{Var}(y_j) = 2\sigma^4/\nu_j,$$

$$\theta = - 1/\sigma^2,$$

$$\kappa_j = \nu_j/2,$$

$$b(\theta) = - \log(- \theta).$$

As we can see from these examples, the exponential family includes a wide variety of common distributions. In general, we have

$$E(y_j) = b'(\theta) = \mu, \qquad Var(y_j) = b''(\theta)/\kappa_j = V_j$$

where $b'(\cdot)$ and $b''(\cdot)$ denote the first and second derivatives of $b(\cdot)$.

If $y_1, \ldots, y_n$ are independent, then the maximum likelihood estimate of θ is given by the solution to:

$$\hat{\mu} = \Sigma\ \kappa_j y_j/\Sigma\ \kappa_j = \Sigma\ w_j y_j/\Sigma\ w_j$$

where $\hat{\mu} = b'(\hat{\theta})$. This implies that there is a large family of models where a weighted sample mean provides an efficient estimator of the population mean. The estimated variance of $\hat{\mu}$ is given by

$$\hat{V}(\hat{\mu}) = (\Sigma\ \kappa_j^2 \hat{V}_j)/(\Sigma\ \kappa_j)^2$$

$$= b''(\hat{\theta})/(\Sigma\ \kappa_j).$$

For large samples, the 95% confidence interval for μ is given by $\mu \pm 1.96 \times \{\hat{V}(\hat{\mu})\}^{\frac{1}{2}}$, providing the model is true.

In cases where $\kappa_j = \kappa w_j$ is known only up to the constant of proportionality κ, (e.g. normal model), it will be necessary to estimate the value of κ. The maximum likelihood estimate is given by the solution to:

$$\Sigma\ w_j [y_j \theta - b(\theta) + \frac{\partial c(y_j,\ \kappa_j)}{\partial \kappa_j}] = 0.$$

Alternatively, an unbiased estimator for $\hat{V}(\hat{\mu})$ which is less model-dependent is given by

$$\hat{V}_1(\hat{\mu}) = \frac{\Sigma\ w_j(y_j - \hat{\mu})^2}{(n - 1)(\Sigma\ w_j)} .$$

This may be used instead to create the confidence intervals for $\hat{\mu}$. The main assumption required for the validity of this approach is that $Var(y_j) \propto 1/w_j$.

3. LINEAR MODELS

3.1 One Way Analysis of Variance

A simple extension of the univariate normal models, described in Section 2.2, is the one-way analysis of variance (ANOVA) model. Here, in addition to observing one characteristic from each individual sampled, we also have a sub-population identifier. Some such identifiers could be age-sex groups, industry/occupation groups, etc. Here the model could be written as

$$y_{ij} = \mu_i + \varepsilon_{ij};\ i = 1, \ldots, I;\ j = 1, \ldots, n_i,$$

where the μ's are population means, which differ among subpopulations and the ε's are assumed to be independent normal with variances $\sigma^2_{ij} = \sigma^2/w_{ij}$, where the w_{ij}'s are known weights. In most applications the weights are constant.

The usual estimator for μ_i in this model is

$$\hat{\mu}_i = \sum_j w_{ij}y_{ij}/\sum_j w_{ij}.$$

Under the model assumptions, the estimated means are independent normal with $E(\hat{\mu}_i) = \mu_i$ and $Var(\hat{\mu}_i) = \sigma^2/\sum_j w_{ij}$. From this, confidence intervals for the individual means may be derived.

An alternative but equivalent description of this model is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where $\Sigma\Sigma\, w_{ij}\alpha_i = 0$. Here we have

$$\mu = \Sigma\Sigma\, w_{ij}\mu_i/\Sigma\Sigma\, w_{ij}$$

$$\alpha_i = \mu_i - \mu.$$

An extension of this representation is particularly useful for two-way and higher order analysis of variance models, to be discussed in Sections 3.2 and 3.3. One of the main questions of interest for these models is whether all the means are equal. This is equivalent to $\mu_1 = \mu_2 = \ldots = \mu_I$ or $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$. Standard ANOVA statistical packages (e.g. SAS, SPSS, etc.) are available to test these hypotheses. A related problem is: Which subpopulation means are equal, given that we have concluded already that not all means are equal? When we have no further structure (such as in a two-way ANOVA), this is known as the multiple comparison problems. Special treatments for this problem are available in many statistical packages.

3.2 Two-Way Analysis of Variance

The data of Table 1 has been taken from the 1975 Sri Lanka Fertility Survey (see Little, 1982). The cell means describe the average number of children ever born cross-classified by Marital Duration and Level of Education.

The row and column means seem to indicate that the average number of children increases with longer marriage durations and decreases with more schooling. Now, the two-way analysis of variance model may be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where the ε's are assumed to be independent normal with variances $\sigma^2_{ijk} = \sigma^2/w_{ijk}$. The w's are known weights. In most applications the weights are constant. In order to estimate the parameters of this model, it is necessary to impose constraints on these parameters, otherwise they are not unique. The usual side conditions are:

$$\sum_i \sum_j \sum_k w_{ijk}\alpha_i = 0,$$

$$\sum_i \sum_j \sum_k w_{ijk}\beta_j = 0,$$

$$\sum_i \sum_k w_{ijk}\gamma_{ij} = 0,$$

$$\sum_j \sum_k w_{ijk}\gamma_{ij} = 0.$$

The estimators are defined by the equations:

$$\sum_i \sum_j \sum_k w_{ijk}(y_{ijk} - \hat{\mu}_{ij}) \frac{\partial \hat{\mu}_{ij}}{\partial \hat{\theta}_\ell} = 0$$

where $\hat{\theta}_1$, $\hat{\theta}_2$, ... correspondent to the parameter estimates $\hat{\mu}$, $\hat{\alpha}_i$, etc. The α's are β's are referred to as main effects and the γ's are the two-way interactions. This results in the following estimators:

$$\hat{\mu} = \bar{y}_{...},$$

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} - \sum_j \sum_k w_{ijk}\hat{\beta}_j / \sum_j \sum_k w_{ijk},$$

$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...} - \sum_i \sum_k w_{ijk}\hat{\alpha}_i / \sum_i \sum_k w_{ijk},$$

$$\gamma_{ij} = \bar{y}_{ij.} - \bar{y}_{...} - \hat{\alpha}_i - \hat{\beta}_j,$$

where $\bar{y}_{ij.}$, $\bar{y}_{i..}$, etc. are the appropriate weighted averages.

Now, the additive model specifies that $\mu_{ij} = \mu + \alpha_i + \beta_j$. We have plotted the cell means from Table 1 in Figure 1. The additive model would specify that all the lines are parallel. If the data of Table 1 are fitted to the additive model, we obtain the adjusted mean values in Table 2. These are plotted in Figure 2. As we can see, the effect of the level of education has been dramatically reduced after fitting this model. This is because the more educated women were not married for as long, so that the years since first marriage proves to be the important factor. However, as the analysis of variance in Table 3 shows, all the main effects and the interactions are significant. Hence the additive model is rejected. However, only 0.4% of the total variation is explained by the Education-Marital Durations interactions, whereas 49.7% of the variation is explained by the additive model. We may surmise from this that the additive model has led to a better understanding of the data and that the Education effect is not as dramatic as it first seemed.

3.3 Regression Formulation

The above analysis of variance models can be considered as special cases of the multiple linear regression model, given by

$$y_j = \beta_0 X_{0j} + \beta_1 X_{1j} + \dots + \beta_r X_{rj} + \varepsilon_j,$$

where X_{0j}, X_{1j}, ..., X_{rj} are known constants and β_0, β_1, ..., β_r are unknown coefficients. We assume that the ε's are independent normal with variances $\sigma_j^2 = \sigma^2/w_j$, where the w_j's are known weights. For example, in the one way analysis of variance, we could let

$$X_{0j} = 1 \text{ for all } j$$

$$\begin{aligned} X_{ij} &= 1 \text{ if the } j\text{-th individual is in the } i\text{-th sub-population} \\ &= -\, a_i/a_I \text{ if the } j\text{-th individual is in the } I\text{-th sub-population} \\ &= 0 \text{ otherwise,} \end{aligned}$$

for $i = 1, \dots, I - 1$, where a_i is the sum of the weights for individuals in the i-th sub-population. In this case we have

$$\mu_i = \beta_0 + \beta_i \qquad \text{for } i = 1, \dots, I - 1,$$

$$\mu_I = \beta_0 - (a_1\beta_1 + \dots + a_{I-1}\beta_{I-1})/a_I.$$

Therefore $\mu = \beta_0$ and $\alpha_i = \beta_i$ for $i = 1, \dots, I - 1$.

A similar regression formulation is possible for two-way and higher order layouts as well.

Now, for the general regression model, the estimator for β_0, ..., β_r is given by $\hat{\beta}_0$, ..., $\hat{\beta}_r$, the solution to

$$\Sigma\, w_j(y_j - \hat{y}_j)X_{ij}, \qquad i = 0, 1, \dots, r$$

where $\hat{y}_j = \hat{\beta}_0 X_{0j} + \hat{\beta}_1 X_{ij} + \dots + \hat{\beta}_r X_{rj}$.

In order to test hypotheses, perform model-building and develop confidence intervals for the β's, we need the covariance matrix of the $\hat{\beta}$'s. This is given by

$$\mathrm{Var}(\hat{\underset{\sim}{\beta}}) = \sigma^2 A^{-1}$$

where A is the matrix with (k,ℓ)-th entry being $\sum_j w_j X_{kj} X_{\ell j}$. To estimate σ^2, we use $\hat{\sigma}^2 = \sum_j w_j (y_j - \hat{y}_j)^2/(n - r - 1)$.

Many statistical packages routinely perform various hypothesis tests on $\hat{\beta}$ using the estimated covariance matrix $\hat{\sigma}^2 A^{-1}$ and the critical values from the appropriate F-distribution (e.g. PROC REG, PROC ANOVA and PROC GLM in SAS).

For example, Koch, Gillings and Stokes (1980) give the data in Table 4 for the number of physician visits per person per year in 1973 in the U.S. cross-classified by size of city (SMSA = Standard Metropolitan Statistical Area vs. Non-SMSA), Income (3 groups) and Education (3 groups). This data is based on the 1973 Health Interview Survey, a survey using a complex probability sample. The data are illustrated in Figure 3.

By using a regression model and performing a number of statistical tests, the following reduced model was obtained:

$$E(Y_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j},$$

where X_{1j} = 1 if the j-th person is in an SMSA

= 0 otherwise,

X_{2j} = 1 if the j-th person has less than $5000 family income or more than 12 years education for the family head

= 0 otherwise.

The estimated parameters were $\hat{\beta}_0 = 4.18$ (standard error of 0.11), $\hat{\beta}_1 = 0.65$ (standard error of 0.11) and $\hat{\beta}_2 = 1.12$ (standard error of 0.09). The standard errors derived here were not those described above since the authors used the 18×18 estimated covariance matrix from the survey to obtain the standard errors. This approach removes the assumption of independent error terms in the model-fitting and is a common approach for analysing data from complex surveys.

In Table 5 we summarize the results. These are illustrated in Figure 4. We see that the model fit is quite good. We have reduced the data from 18 values to 3 summary statistics and also have smaller standard errors (hence higher precision) of the estimated values.

4. GENERALIZED LINEAR MODELS

4.1 Regression with a Dichotomous Dependent Variable

One of the difficulties often encountered with the linear models discussed in Section 3 is that the error terms were assumed to be normally distributed. It is true that analyses similar to those in Section 3 may be performed with non-normal errors, providing the variances of the errors still satisfy $\sigma_j^2 = \sigma^2/w_j$ and the errors are uncorrelated. In this case the estimators we have described yield the minimum variance linear unbiased estimates of the model parameters, however better estimators (i.e. non-linear estimators) may be available. These considerations have led to generalized linear models (see Nelder and Wedderburn, 1972) and robust estimators (see Huber, 1973). We concentrate here on the generalized linear models.

For example, suppose the dependent variable, y_j, can take on only two values, 0 or 1. We now want to model $p_j = Pr(Y_j = 1)$ as a function of the linear expression $X_{0j}\beta_0 + X_{1j}\beta_1 + \ldots + X_{rj}\beta_r$. There are three popular approaches for this problem. One is to let $\hat{\beta}_0, \ldots, \hat{\beta}_r$ be the usual estimate from a stardard regression model. This is analogous to discriminant analysis where the variables $X_{0j}, \ldots, X_{rj}$ are not considered fixed known constants, but are themselves random variables (multivariate normal with constant covariance matrix) whose mean depends on the value of Y_j. The problem with this approach is that $\hat{Y}_j = X_{0j}\hat{\beta}_0 + \ldots + X_{rj}\hat{\beta}_r$ cannot be used directly to predict the value of p_j. Also, in many applications the X_{ij}'s are categorical, (e.g. province, occupation, etc.), thus violating the assumption of multivariate normality.

Two other popular approaches are known as probit analysis and logistic regression. In probit analysis it is assumed that $p_j = \Phi(\sum_i X_{ij}\beta_i)$, where Φ is the cumulative distribution function of a standard normal random variable. In logistic regression, it is assumed that

$$\theta_j = \log[p_j/(1 - p_j)] = \sum_i X_{ij}\beta_i.$$

Both these approaches are valuable analytic tools, and are available in many statistical packages (e.g. SAS, BMDP). The two approaches may be viewed together by letting

$$\eta_j = g(p_j) = \sum_i X_{ij}\beta_i.$$

For probit analysis we have $\eta_j = \Phi^{-1}(p_j)$, whereas for logistic regression we have $\eta_j = \log\,[p_j/(1 - p_j)]$. The maximum likelihood estimate for $\beta_0, \ldots, \beta_r$ is the solution to

$$\sum_j \frac{(y_j - \hat{p}_j)X_{ij}}{\hat{p}_j(1 - \hat{p}_j)g'(\hat{p}_j)} = 0, \qquad \text{for } i = 0, \ldots, r,$$

where $g(\hat{p}_j) = \sum_i X_{ij}\hat{\beta}_i$. These equations often must be solved iteratively. For the probit analysis we have

$$g'(p_j) = \frac{1}{\phi[\Phi^{-1}(p_j)]}$$

where $\phi(\cdot)$ is the standard normal density function. For the logistic regression,

$$g'(p_j) = [p_j(1 - p_j)]^{-1}$$

so that the parameter estimate is given by the solution to

$$\sum_j (y_i - \hat{p}_j)X_{ij} = 0, \qquad \text{for } i = 0, \ldots, r.$$

The covariance matrix of $\hat{\beta}_0, \ldots, \hat{\beta}_r$ is A^{-1} where A is a matrix with (k, ℓ)-th entry given by

$$A_{k\ell} = \sum_j \frac{X_{kj}X_{\ell j}}{p_j(1 - p_j)\{g'(p_j)\}^2}$$

This can be used to construct confidence intervals and perform hypothesis tests and model-building.

For logistic regression, the covariance simplifies to

$$A_{k\ell} = \sum_j p_j(1 - p_j)X_{kj}X_{\ell j}.$$

As an example of the utility of these models, we consider an unpublished analysis performed by Dolson and Morin on the Canadian Health and Disability Survey. The dependent variable was whether or not a person would be screened in as potentially disabled using the Screening Test 2 of the January 1983 Labour Force Supplement on Disability. For details, see Dolson and Morin (1983). Analysis was restricted to males aged 15-64. Of the 13,897 respondents, 14.4% (unweighted) were screened in. The screened-in rates are cross-classified by age-groupings, labour force participation and a proxy/non-proxy variable (with 3 levels: non-proxy, proxy by male or proxy by female) in Table 6. (The fitted values from the model to be discussed below are also shown.) The data are illustrated in Figure 5.

The fitted model reduced the number of parameters from 30 to 11. The final model was given by

$$\log[p_{ijk}/(1 - p_{ijk})] = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij},$$

where $\Sigma\,\alpha_i = \Sigma\,\beta_j = \Sigma\,\gamma_k = 0$, $\sum_j \delta_{ij} = 0$, $\sum_i \delta_{ij} = 0$, for the i-th age group, j-th labour force status and k-th proxy status (2 levels: non-proxy vs. proxy). The following were the estimated parameters.

Parameter	Subscript	Estimate
μ		-1.43
α	Age 15-24	-1.12
	Age 25-34	-0.571
	Age 35-44	0.0143
	Age 45-54	0.629
	Age 55-64	1.05
β	In Labour Force	-0.576
	Not in Labour Force	0.576
γ	Non-proxy	0.0859
	Proxy	-0.0859
δ	Age 15-24, in L.F.	0.385
	Age 25-34, in L.F.	0.0938
	Age 35-44, in L.F.	-0.175
	Age 45-54, in L.F.	-0.243
	Age 55-64, in L.F.	-0.0612
	Age 15-24, not in L.F.	-0.385
	Age 25-34, not in L.F.	-0.0938
	Age 35-44, not in L.F.	0.175
	Age 45-54, not in L.F.	0.243
	Age 55-64, not in L.F.	0.0612

The fitted values are illustrated in Figure 6.

We see that even after adjusting for age and labour force status, there is a proxy effect on the screening rates. This proxy effect does not seem to depend on the sex of the proxy respondent. Also, there is no interaction between the proxy and the age/labour force status variables. This model does not necessarily imply a proxy bias, but it indicates that a proxy bias may potentially be present. Without a special study such as a re-interview program for the proxy respondent, it is impossible to definitively conclude the existence of a proxy bias.

4.2 Generalized Linear Models

In the previous section we discussed a large class of linear models related to the binomial model, of which probit analysis and logistic regression were special cases. We now extend these to the exponential family as proposed by Nelder and Wedderburn (1972).

As in Section 2.3, we assume y_j has probability function given by

$$f(y_j) = \exp[\kappa_j \{y_j\theta_j - b(\theta_j)\} + c(y_j, \kappa_j)],$$

where $\mu_j = E[Y_j] = b'(\theta_j)$ and $V_j = \mathrm{Var}[Y_j] = b''(\theta_j)/\kappa_j$.

We let $\eta_j = g(\mu_j) = \sum_i X_{ij}\beta_i$ be the linear component of the model, where $g(\cdot)$ is a known function.

Now the maximum likelihood estimates of $\underset{\sim}{\beta}$ are given by the solution to

$$\sum_j \frac{(y_j - \hat{\mu}_j)X_{ij}}{\hat{V}_j[g'(\hat{\mu}_j)]} = 0.$$

Nelder and Wedderburn (1972) have shown that a reasonable method for estimating $\underset{\sim}{\beta}$ is given by performing a number of weighted least-squares regressions, updating the weights and the dependent variables on successive iterations. This is called iteratively re-weighted least squares. In particular, the weights for the t-th iteration are given by

$$\hat{w}_j^{(t)} = \frac{1}{\hat{V}_j^{(t)}[g'(\hat{\mu}_j^{(t)})]^2}$$

and the dependent variables on the t-th iteration are given by

$$\hat{Z}_j^{(t)} = g(\hat{\mu}_j^{(t)}) + g'(\hat{\mu}_j^{(t)})(y_j - \hat{\mu}_j^{(t)}).$$

The (t + 1)-th iteration of $\hat{\underset{\sim}{\beta}}$ is then the solution to

$$\sum_j \hat{w}_j^{(t)}[\hat{Z}_j^{(t)} - \sum_\ell X_{\ell j}\hat{\beta}_\ell^{(t+1)}]\, X_{kj} = 0.$$

The estimated covariance matrix of $\hat{\underset{\sim}{\beta}}$ is given by A^{-1} where the (k, ℓ)-th entry for A is

$$A_{k\ell} = \sum_j \hat{w}_j X_{kj} X_{\ell j}.$$

This implies that many standard weighted least-squares packages could be invoked to perform analysis of these generalized linear models.

For example, a common analysis of contingency tables, called log-linear models assumes a basic Poisson model with $\log \mu_j = \sum_i X_{ij}\beta_i$. Here we have

$$V_j = \mu_j,$$

$$g(\mu_j) = \log \mu_j,$$

so that the iteratively reweighted solution is given by assigning

$$\hat{w}_j^{(t)} = \hat{\mu}_j^{(t)},$$

$$\hat{z}_j^{(t)} = \log \hat{\mu}_j^{(t)} + \frac{y_j - \hat{\mu}_j^{(t)}}{\hat{\mu}_j^{(t)}}.$$

Hence, models similar to those described in Section 3 can be analyzed analogously using the generalized linear model formulation.

5. DIAGNOSTICS

Linear regression methods have been known now for over a century; see Hocking (1983) for a review of developments over the last 25 years. In more recent years attention has been focused on difficulties encountered when there is multicollinearity in the variables (leading to large variances of the parameter estimates) and when the models may fail. Some of these diagnostics are now available in SAS and SPSS-X.

The methods discussed in this paper extend linear regression to a much wider class of problems. Newer diagnostic techniques for models of this sort are discussed in Landwehr, Pregibon and Shoemaker (1984).

In many statistical applications, the proposed model is only used as an approximation to reality. Therefore, the user of these models should employ these diagnostic tools in the course of the analysis.

REFERENCES

[1] Dolson, D. and Morin, J.-P. (1983). Disability data development project: Analysis of screening questionnaires. Technical Report, Health Division, Statistics Canada.

[2] Hocking, R.R. (1983). Developments in linear regression methodology: 1959-1982 (with discussion). Technometrics, 25, pp. 219-249.

[3] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist., 1, pp. 799-821.

[4] Koch, G.G., Gillings, D.B., and Stokes, M.E. (1980). Biostatistical implications of design, sampling and measurement to health science data analysis. Ann. Rev. Public Health, 1, pp. 163-225.

[5] Landwehr, J.M., Pregibon, D., and Shoemaker, A. (1984). Graphical methods for assessing logistic regression models (with discussion). J. Amer. Statist. Assoc., 79, pp. 61-83.

[6] Little, R.J.A. (1982). Direct standardization: A tool for teaching linear models for unbalanced data. Amer. Statist., 36, pp. 38-43.

[7] Nelder, J.A., and Wedderburn, R.W.M., (1972). Generalized linear models. J. Roy. Statist. Soc., Ser. A, 135, pp. 370-384.

Table 1: Mean Number of Children Ever Born, by Marital Duration and Education Level. Sri Lanka 1975 (from Little, 1982)

Years since First Marriage		Level of Education				
		No School	1 - 5 Years	6 - 9 Years	10+ Years	Row
0 - 4	Mean	0.96	0.88	0.95	0.92	0.92
	Count	112	376	442	351	1281
5 - 9	Mean	2.54	2.46	2.39	2.39	2.44
	Count	172	442	362	255	1231
10 - 14	Mean	3.87	3.91	3.73	3.14	3.76
	Count	197	482	293	145	1117
15 - 19	Mean	5.13	4.97	4.61	4.13	4.84
	Count	239	461	262	95	1057
20 - 24	Mean	6.22	5.87	5.22	4.47	5.79
	Count	292	377	184	40	893
25+	Mean	6.92	6.55	6.23	5.97	6.65
	Count	501	548	161	22	1232
Column	Mean	5.17	4.24	3.26	2.30	3.94
	Count	1513	2686	1704	908	6811

Table 2: Interactions for Mean Number of Children from Table 1

Years Since First Marriage		Level of Education				
		No School	1 - 5 Years	6 - 9 Years	10+ Years	Row
0 - 4	Raw Mean	0.96	0.88	0.95	0.92	0.92
	Adjusted Mean	1.31	1.07	0.86	0.71	1.02
	Interaction	-0.35	-0.19	0.09	0.21	
5 - 9	Raw Mean	2.54	2.46	2.39	2.39	2.44
	Adjusted Mean	2.78	2.54	2.33	2.18	2.49
	Interaction	-0.24	-0.08	0.06	0.21	
10 - 14	Raw Mean	3.87	3.91	3.73	3.14	3.76
	Adjusted Mean	4.06	3.82	3.61	3.46	3.77
	Interaction	-0.19	0.09	0.12	-0.32	
15 - 19	Raw Mean	5.13	4.97	4.61	4.13	4.84
	Adjusted Mean	5.11	4.87	4.66	4.51	4.82
	Interaction	0.02	0.10	-0.05	-0.38	
20 - 24	Raw Mean	6.22	5.87	5.22	4.47	5.79
	Adjusted Mean	6.01	5.77	5.56	5.41	5.72
	Interaction	0.21	0.10	-0.34	-0.94	
25+	Raw Mean	6.92	6.55	6.23	5.97	6.65
	Adjusted Mean	6.82	6.58	6.37	6.22	6.53
	Interaction	0.10	-0.03	-0.14	-0.25	
Column		5.17	4.24	3.26	2.30	3.94
		4.23	3.99	3.78	3.63	3.94

Table 3: Analysis of Variance of Data from Table 1

Source	Sum of Squares	Proportion of Total SS	DF	Mean Square	F	Signif. of F
Main Effects						
Marital Duration	27402.684	0.493	5	5480.537	1340.990	.000
Education/Duration	225.535	0.004	3	75.178	18.395	.000
Interactions						
Duration×Education	206.965	0.004	15	13.798	3.376	.000
Residual	27729.848	0.499	6787	4.986		
Total	55565.031		6810			

Table 4: Physician Visits per Person per Year by Residence Size, Family Income and Education of Family Head, U.S. 1973

Education in Years	Family Income		
	0 - 4999	5000 - 14999	15000 or more
SMSA			
Less than 12	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
12	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
More than 12	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
Non-SMSA			
Less than 12	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
12	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
More than 12	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)

Note: Bracketed figures indicate standard errors of estimate.

Table 5· Estimated Physician Visits from Table 4, Original and Fitted Values

Education (in Years)		Family Income 0 - 4999	5000 - 14999	15000 or more
SMA				
Less than 12	Original	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
	Fitted	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Difference	0.20	-0.10	-0.01
12	Original	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
	Fitted	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Difference	0.22	0.15	-0.13
More than 12	Original	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
	Fitted	5.95 (0.07)	5.95 (0.07)	5.95 (0.07)
	Difference	0.36	0.13	-0.29
Non-SMSA				
Less than 12	Original	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
	Fitted	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Difference	-0.22	-0.04	0.24
12	Original	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
	Fitted	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Difference	0.06	0.14	0.31
More than 12	Original	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)
	Fitted	5.30 (0.11)	5.30 (0.11)	5.30 (0.11)
	Difference	-0.72	-0.24	-0.82

Table 6· Unadjusted and Fitted Screened-in Rates from Test 2. Canadian Health and Disability Survey, Males Aged 15-64, by Labour Force Participation and Proxy Status, Canada January 1983 (Unweighted)

Age		Non-Proxy	Male Proxy	Female Proxy
In Labour Force				
15 - 24	Unadjusted	.065(.0067)	.055(.0143)	.056(.0069)
	Fitted	.065(.0051)	.056(.0044)	.056(.0044)
	Difference	.000	-.001	.000
25 - 34	Unadjusted	.085(.0058)	.058(.0252)	.069(.0069)
	Fitted	.085(.0048)	.071(.0046)	.071(.0046)
	Difference	.000	-.013	-.002
35 - 44	Unadjusted	.113(.0079)	.029(.0290)	.094(.0086)
	Fitted	.111(.0064)	.093(.0059)	.093(.0059)
	Difference	.002	-.064	.001
45 - 54	Unadjusted	.180(.0109)	.082(.0351)	.154(.0120)
	Fitted	.177(.0088)	.153(.0083)	.153(.0083)
	Difference	.003	-.071	.001
55 - 64	Unadjusted	.284(.0150)	.207(.0752)	.250(.0183)
	Fitted	.283(.0124)	.249(.0124)	.249(.0124)
	Difference	.001	-.042	.001
Not in Labour Force				
15 - 24	Unadjusted	.104(.0127)	.071(.0190)	.074(.0084)
	Fitted	.104(.0078)	.079(.0065)	.079(.0065)
	Difference	.000	-.008	-.005
25 - 34	Unadjusted	.146(.0239)	.367(.1450)	.227(.0365)
	Fitted	.192(.0213)	.167(.0194)	.167(.0194)
	Difference	-.046	.200	.060
35 - 44	Unadjusted	.348(.0372)	.455(.1501)	.324(.0544)
	Fitted	.359(.0309)	.320(.0299)	.320(.0299)
	Difference	-.011	.135	.004
45 - 54	Unadjusted	.534(.0361)	.625(.1712)	.454(.0505)
	Fitted	.525(.0293)	.483(.0301)	.483(.0301)
	Difference	.009	.142	-.029
55 - 64	Unadjusted	.571(.0220)	.563(.1240)	.591(.0420)
	Fitted	.585(.0194)	.543(.0217)	.543(.0217)
	Difference	-.014	.020	.048

NOTE: Bracketed figures are Standard Errors

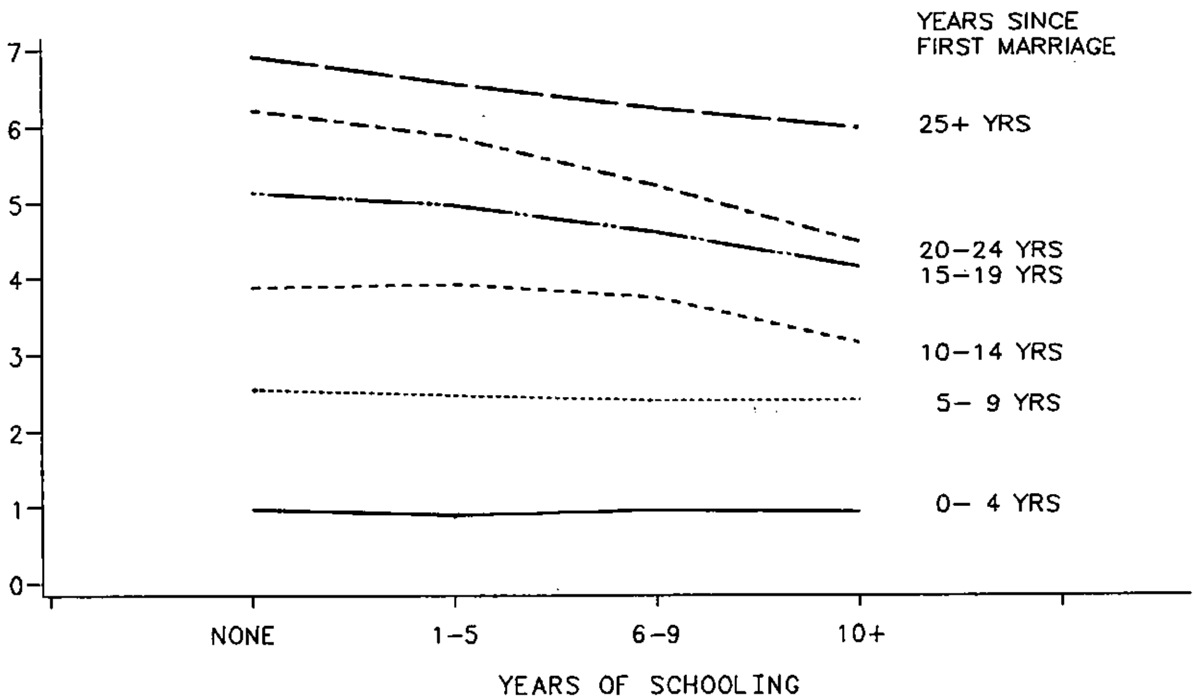

Figure 1: Observed Means from Sri Lanka Fertility Survey. 1975. Data source: Little (1982).

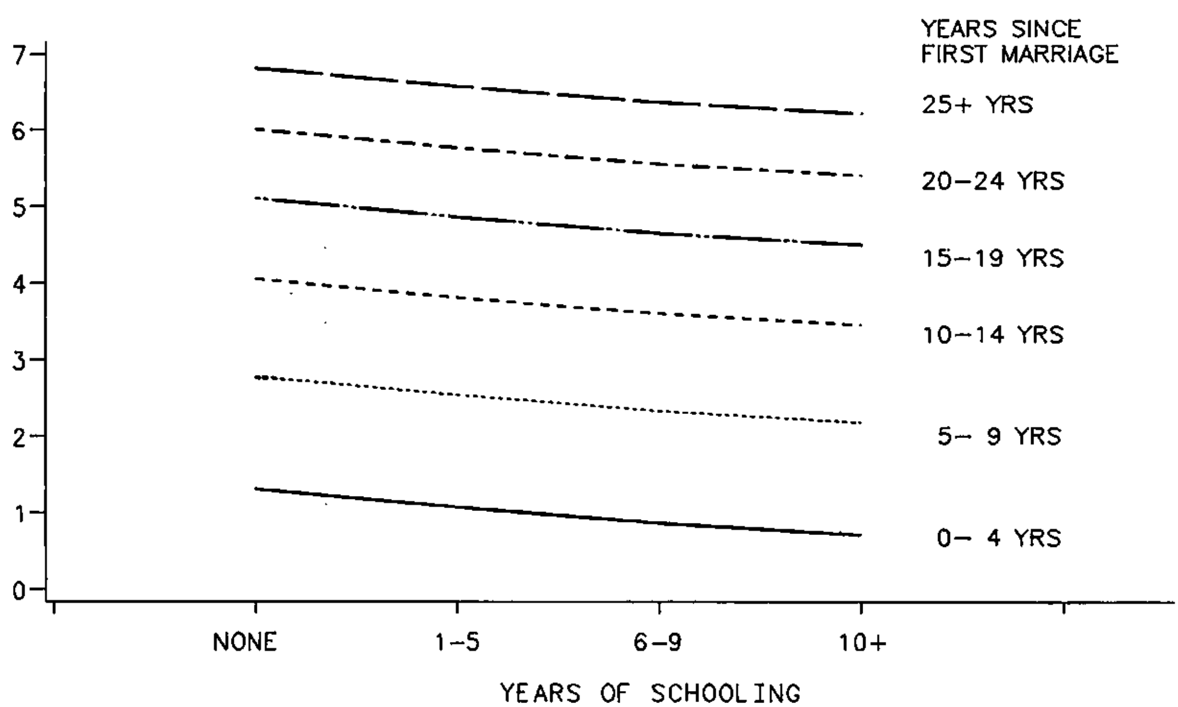

Figure 2: Adjusted Means from Sri Lanka Fertility Survey. 1975. Data source: Little (1982)

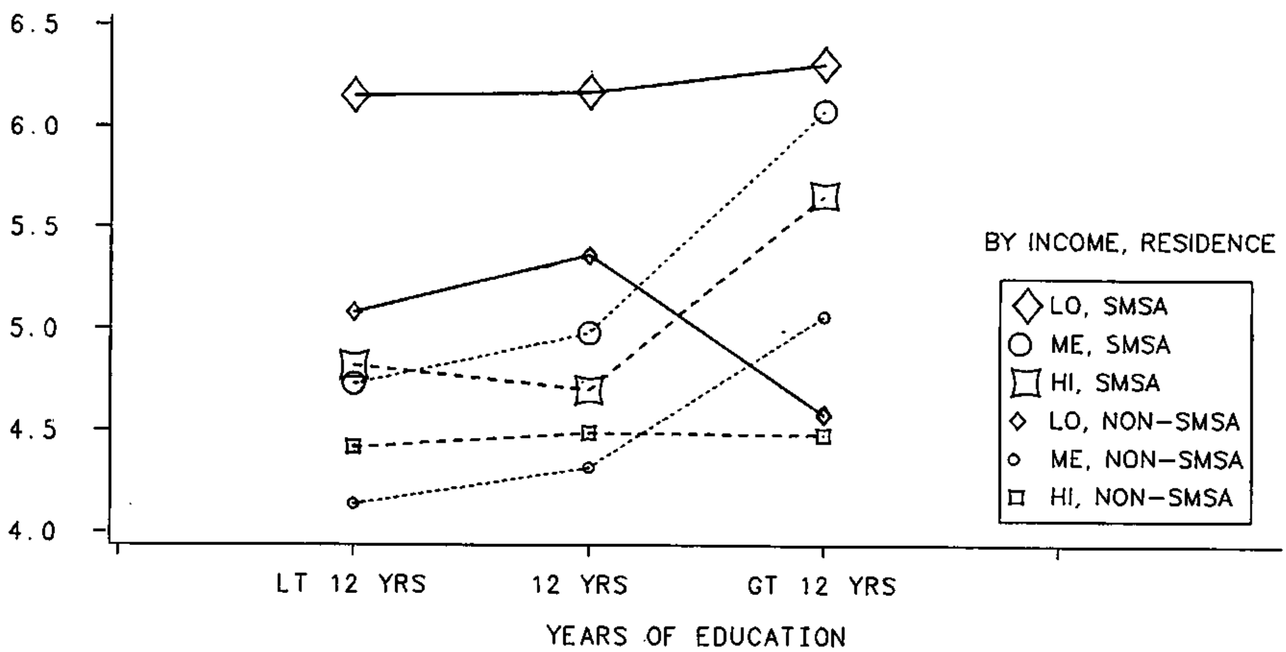

Figure 3: Observed Mean Number of Physician Visits per Person per Year. U.S.A., 1973.

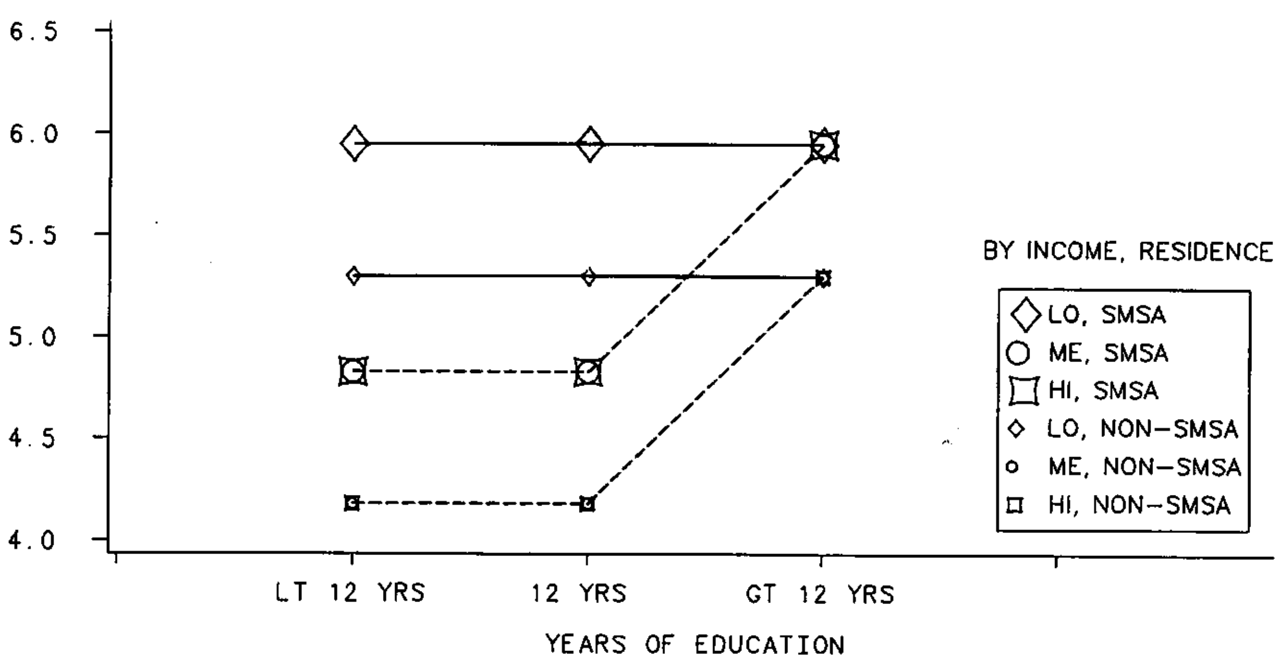

Figure 4: Model Predicted Mean Number of Physician Visits per Person per Year. U.S.A., 1973.

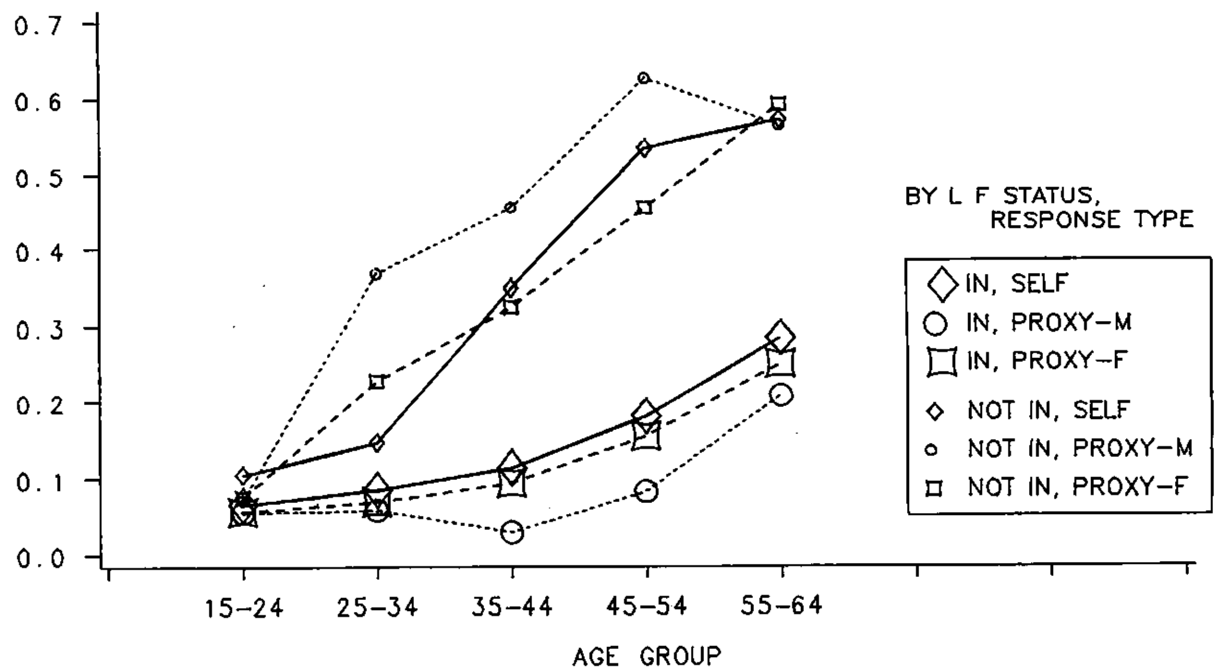

Figure 5: Observed Screening Rates, Disability Survey, January 1983, Males 15-64.

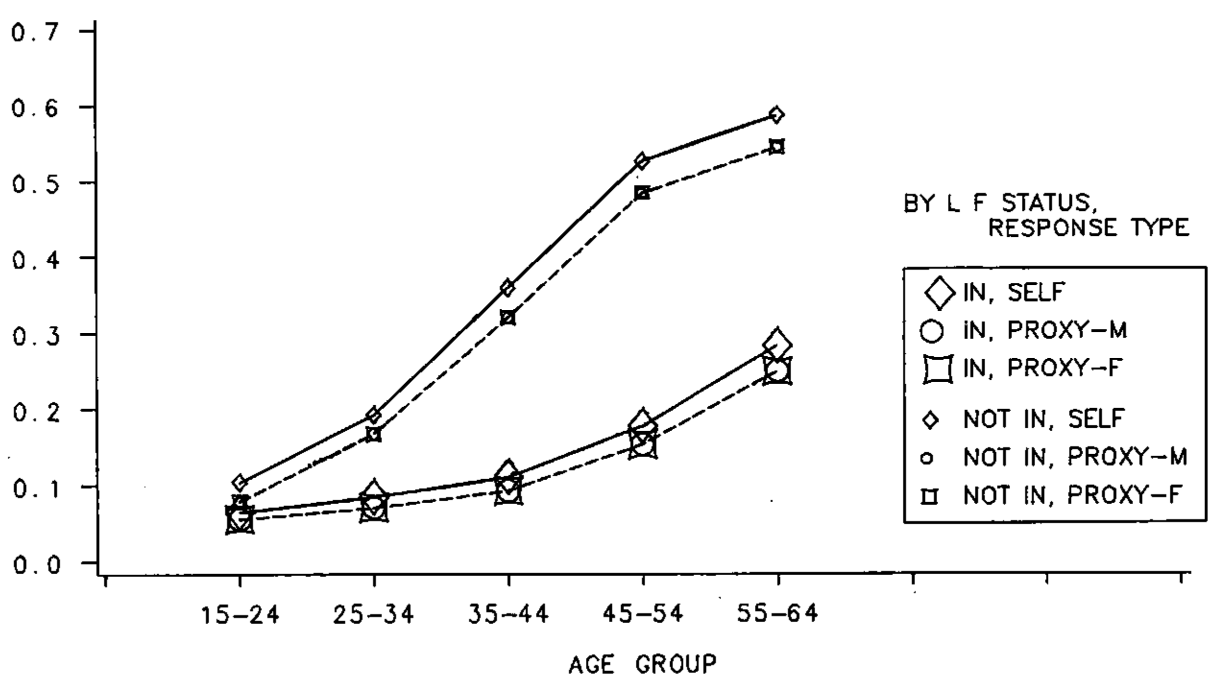

Figure 6: Predicted Screening Rates, Disability Survey. January 1983. Males 15-64.

ADJUSTING SUB-ANNUAL SERIES TO YEARLY BENCHMARKS

Pierre A. Cholette[1]

This paper proposes a modification to the method of Denton (1971) for adjusting sub-annual series to yearly totals. These totals originate from more reliable sources and constitute annual benchmarks. The benchmarked series derived according to the modified method is more parallel to the unbenchmarked series than this is the case with the original method. An additive and a proportional variant of the method are presented. These can easily be adapted for flow, stock and index series. Also presented are a few recommendations about the preliminary benchmarking of current data and the management of "historical" estimates of the series.

1. INTRODUCTION

In many cases, the statistician obtains sub-annual data of a series from one source of data (such as a sample survey); and, the corresponding annual benchmark values from another more reliable source of data (such as a census). The annual sums of the observed sub-annual values are generally not equal to the annual benchmark values. Such sub-annual series require adjustment to annual benchmarks, that is benchmarking.

The solution proposed by Denton (1971) (and generalized by Fernandez in 1981) consists of finding a sub-annual series which would display the movement of the available sub-annual series as much as possible and whose annual sums (or averages) would match the more reliable annual benchmarks. The level of the resulting series would then be given by the annual benchmarks, whereas its movement would be dictated by the original sub-annual series. In other words, the adjusted or benchmarked series should run as parallel as possible to the original, while still satisfying the annual benchmarks. This paper suggests a modification to Denton's specification which makes the original and the adjusted series even more parallel.

We follow the model of Ehrenberg (1982) for the presentation of scientific

[1] Pierre A. Cholette, Time Series Research and Analysis, Statistics Canada.

papers. The reader will be exposed to the illustrations and results first; and the methodological details, afterwards.

2. ILLUSTRATION OF THE RESULTS

Figure 1 shows the corrections ($x_t - z_t$) made to the original series z_t according to the additive solution (with first differences) of Denton and according to the corresponding solution proposed in this paper. Since the corrections are to be added to the original sub-annual series z_t, the adjusted series x_t will be completely parallel to the original series, if and only if the corrections are constant. In the figure, this happens only for the corrections derived under the method proposed in this paper.

Figure 1 presented a trivial and ideal case which allowed the solution of constant corrections: All the average annual discrepancies, the differences between the annual benchmarks and the annual totals of the original series (divided by the number of months per year), were constant. Figure 2 displays a more realistic case, where the five average annual discrepancies vary about 200. As in the first example, the corrections derived by the herein proposed method are much more constant, especially in the first year.

As explained below, Denton's method does not only minimizes the change in the corrections (to make them as constant as possible) but also the size of the first correction. This can be seen both in Figures 1 and 2, where the first corrections are close to zero. The alternative solution, on the other hand, only minimizes the change in the corrections. Graphically this consists of fitting a curve through the average annual discrepancies, which is as flat as possible and which spans the same annual surfaces as the average annual discrepancies.

3. KEEPING THE ORIGINAL AND THE BENCHMARKED SERIES PARALLEL

Resuming the additive first difference formulation of Denton as well as his notation, the desired series x_t minimizes the following objective function

$$p(x) = \sum_{t=1}^{n} (\Delta x_t - \Delta z_t)^2 = \sum_{t=1}^{n} (\Delta(x_t - z_t))^2, \; x_0 = z_0, \tag{1}$$

where z_t stands for the original sub-annual series at time t. This function is minimized subject to the equality constraints between the annual sums of the values obtained and the available benchmarks y_i:

$$\sum_{t=(i-1)k+1}^{ik} x_t = y_i, \; i = 1, 2, \ldots, m. \tag{2}$$

where k is the number of "months" per year.

Denton justifies hypothesis $x_0 = z_0$ claiming that it is legitimate to assume the equality of the last fitted and observed values prior to the estimation interval. Objective function (1) would then mean that the adjusted series x_t should have the same slope as the original series z_t; and therefore, that the slope of the differences between the two series should be minimized (subject to the constraints). However, after substituting $x_0 = z_0$, objective function (1) can be rewritten as:

$$p(x) = (x_1 - z_1)^2 + \sum_{t=2}^{n} (\Delta(x_t - z_t))^2. \tag{3}$$

This transformation emphasizes that the assumption $x_0 = z_0$ implies minimizing the size of the first correction. As illustrated in Figures 1 and 2, minimizing the first correction pulls the correction curve towards zero at the start of the series. This produces a wave in the first year which is transmitted to the other years. This wave in the corrections prevents, by definition, the maximum parallelism between the observed and adjusted series.

The specification proposed here simply refrains from postulating $x_0 = z_0$ and yields the following objective function

$$p(x) = \sum_{t=2}^{n} (\Delta(x_t - z_t))^2, \tag{4}$$

subject to the same constraints of equation (2).

In linear algebra, the constrained objective function is written

$$\underline{u}(\underline{x}, \underline{q}) = (\underline{x} - \underline{z})'\underline{A}\,(\underline{x} - \underline{z}) - 2\,\underline{q}'(\underline{y} - \underline{B}'\underline{x}), \tag{5}$$

where the vectors and matrices involved are:

$$\underset{n\times 1}{\underline{x}} = \begin{vmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{vmatrix}, \quad \underset{n\times 1}{\underline{z}} = \begin{vmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{vmatrix}, \quad \underset{m\times 1}{\underline{y}} = \begin{vmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{vmatrix}, \quad \underset{m\times 1}{\underline{q}} = \begin{vmatrix} q_1 \\ q_2 \\ \cdot \\ \cdot \\ \cdot \\ q_m \end{vmatrix}, \tag{6}$$

$$\underset{n\times n}{\underline{A}} = \underline{D}'\underline{D}, \quad \underset{(n-1)\times n}{\underline{D}} = \begin{vmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{vmatrix}, \tag{7}$$

$$\underset{n\times m}{\underline{B}} = \begin{vmatrix} \underline{j} & \underline{0} & \cdots \\ \underline{0} & \underline{j} & \cdots \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{vmatrix}, \quad \underset{k\times 1}{\underline{j}} = \begin{vmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \end{vmatrix}, \ (n = km). \tag{8}$$

Vector $\underline{q}$ contains the Lagrangian multipliers. Variables n (= mk), m and k respectively stand for the number of observations and of years in the series and the number of months per year.

The normal equations associated with objective function (5) are

$$\begin{aligned} d\underline{u}/d\underline{x} &= (\underline{A} + \underline{A}')(\underline{x} - \underline{z}) + 2\,\underline{B}\,\underline{q} = \underline{0} \\ d\underline{u}/d\underline{q} &= 2(\underline{B}'\underline{x} - \underline{y}) = \underline{0} \end{aligned} \tag{9}$$

and yield solution

$$\begin{vmatrix} \underline{x} \\ \underline{q} \end{vmatrix} = \begin{vmatrix} \underline{A} & \underline{B} \\ \underline{B}' & \underline{0} \end{vmatrix}^{-1} \begin{vmatrix} \underline{A} & \underline{0} \\ \underline{0} & \underline{I} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{y} \end{vmatrix} = \underset{(n+m)\times(n+m)}{\underline{W}} \begin{vmatrix} \underline{z} \\ \underline{y} \end{vmatrix}. \tag{10}$$

Substituting identity y = B'z + r, where r contains the m annual discrepancies, gives

$$\begin{vmatrix} \underline{x} \\ \\ \underline{q} \end{vmatrix} = \begin{vmatrix} \underline{A} & \underline{B} \\ \\ \underline{B}' & \underline{0} \end{vmatrix}^{-1} \begin{vmatrix} \underline{A} & \underline{0} \\ \\ \underline{B}' & \underline{I} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \\ \underline{r} \end{vmatrix} = \begin{vmatrix} \underset{n\times n}{\underline{I}} & \underset{n\times m}{\underline{W}_x} \\ \underset{m\times n}{\underline{0}} & \underset{m\times m}{\underline{W}_1} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \\ \underline{r} \end{vmatrix} \rightarrow \underline{x} = \underline{z} + \underline{W_x}\, \underline{r}. \qquad (11)$$

This reformulation of the solution reduces computing time in the application of the calculated weights compared to formulation (10). Also note that once the weights W_x are obtained, they can be used for any number of series having the same number of observations. Furthermore, we recommend (Cholette, 1978, section 6; 1979, 4.3) to compute W_x for a 5-year interval and to use it in a moving average manner (moving one year at the time) for series of 5 years and more. Apart from saving on calculations, this procedure generates only two revisions in the estimates (ceteris paribus) when new years of observations are added to the series.

Denton solves the inversion in equation (10) by parts. This is impossible here since matrix A is singular. The overall matrix however is not singular and can be inverted.

In fact, the method developped herein uses the solution proposed by Boot, Feibes and Lisman (1967) to interpolate between annual data in the absence of sub-annual information. Solution (11) exactly consists in interpolating between the annual discrepancies with the method of these authors and in adding the resulting estimates (the corrections) to the original sub-annual series.

4. PROPORTIONAL VARIANT

The proportional method now presented in this section is also a variant of Denton's proportional method, from which $x_0 = z_0$ was removed. As in Section 2, the objective function still minimizes the sum of the squared differences between the slopes of the original and desired sub-annual series (z_t and x_t). Each term in the sum is weighted however by the value of the corresponding sub-annual observation:

$$p(x) = \sum_{t=2}^{n} (\Delta(x_t - z_t)/z_t)^2 = \sum_{t=2}^{n} (\Delta(x_t/z_t))^2. \qquad (12)$$

This variant is suitable for series with strong seasonality, when it is thought that seasonal trough months cannot account for the annual discrepancy as much as seasonal peak months: The size of the corrections are proportional to the level of each observation, as illustrated in Figure 3. The low observations get smaller corrections than the seasonally higher observations, although the minimized proportional corrections x_t/z_t are as flat as permitted by the annual discrepancies. Note that with the proportional variant all observations must be positive and that all the adjusted values will also be positive.

It can also be shown (Cholette, 1978, Section 3; 1979, 3) that the proportional variant is a linear approximation of the strongly non-linear growth rate preservation method (Smith, 1977: Helfand et al., 1978), which would have the following objective function:

$$p(x) = \sum_{t=2}^{n} (x_t/x_{t-1} - z_t/z_{t-1})^2. \qquad (13)$$

The approximation is exact in situations of constant annual proportional discrepancies on the estimation interval.

In linear algebra, the constrained objective function associated to the proportional method is

$$\underline{u}(\underline{x}, \underline{g}) = (\underline{x} - \underline{z})' \, \underline{Z}^{-1}\underline{A} \, \underline{Z}^{-1}(\underline{x} - \underline{z}) - 2 \, \underline{g}'(\underline{y} - \underline{B}'\underline{x}), \qquad (14)$$

where Z^{-1} is a diagonal matrix with elements $1/z_1$, $1/z_2$, ... The solution has the same structure as the additive variant ($Z^{-1}A \, Z^{-1}$ replacing A in (11)) and writes:

$$\begin{vmatrix} \underline{x} \\ \underline{g} \end{vmatrix} = \begin{vmatrix} \underline{Z}^{-1}\underline{A} \, \underline{Z}^{-1} & \underline{B} \\ \underline{B}' & \underline{0} \end{vmatrix}^{-1} \begin{vmatrix} \underline{Z}^{-1}\underline{A} \, \underline{Z}^{-1} & \underline{0} \\ \underline{B}' & \underline{I} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{r} \end{vmatrix} = \begin{vmatrix} \underline{I} & \underline{W_x} \\ \underline{0} & \underline{W_1} \end{vmatrix} \begin{vmatrix} \underline{z} \\ \underline{r} \end{vmatrix}. \qquad (15)$$

Unlike the weights in the additive variant however, weights W_x of the proportional solution must be computed for each series and even for each application interval of a given series.

5. STOCK AND INDEX SERIES

The additive and proportional variants of the method presented above are designed for flow series, whose annual values correspond to the sum of the sub-annual values. The solutions can very easily be adapted for stock series, whose annual values are associated to only one sub-annual value (usually that of the last month); and for index series, whose annual values correspond to the average of the sub-annual values. For a quarterly stock series, for instance, one merely has to redefine the component vector $\underline{j}$ of matrix $\underline{B}$ as

$$\underset{1\times 4}{\underline{j}'} = [0\ 0\ 0\ 1];$$

and, for monthly index series as

$$\underset{1\times 12}{\underline{j}'} = [1/12\quad 1/12\quad \ldots\quad 1/12].$$

6. DISCUSSION

6.1 Historical Data

There is a lot of confusion regarding the interpretation of assumption $x_0 = z_0$ of Denton. In that respect, the author writes: "It is assumed that no adjustments are to be made to the original series for years outside the range from year 1 to m, inclusive." (p. 100, above equation (3.2)).

If these years are left untouched because they never had any benchmarks, the solution proposed by Denton is defendable: No corrections result for years -1 and 0; and small and gradually introduced corrections, at the start of year 1. (Remember that $x_0 = z_0$ implies minimizing the first correction.) The resulting adjusted series is then continuous as illustrated in Figure 4 by curve ADEB.

However, if the first years are left untouched because they were already benchmarked and are now considered "historical", we do not agree with assumption $x_0 = z_0$. Indeed, this assumption will generally produce a discontinuity between years 0 and 1, as shown in Figure 4 by curve A'CDEB. Years -1 and 0 have already received corrections of magnitude around CD, whereas the start of year 1 receives corrections which are as small as possible.

In order to "freeze" the historical data after a certain number of years, two solutions are possible. First, one can explicitely specify the freezing constraint in the objective function which becomes

$$p(x) = ((x_1 - z_1) - (x_0 - z_0))^2 + \sum_{t=2}^{n} (\Delta (x_t - z_t))^2, \tag{16}$$

where $(x_0 - z_0)$ is known and equal to the last correction used for historical year 0. This correction is generally not equal to zero (Cholette, 1979b, 1983). This specification amounts to determining the starting point of the correction curve.

Second, a less specific but equally effective solution consists of applying the methodology already proposed in this paper (additive or proportional versions) as a moving average, which moves one year at the time. With a 5-year estimation interval, for instance, the estimates automatically become final after two years of revision; and, after one year, in the case of a 3-year interval (Cholette, 1978, section 6 a; 1979, 4.3). The resulting benchmarked series is continuous, as illustrated in Figure 4 by curve A'CB.

6.2 Implementation

The practioners of benchmarking have a tendency to feed to the benchmarking programme the already benchmarked years of data followed by one year of unbenchmarked data (all accompanied by their benchmarks). For methodologists, it is obvious that one must always submit the unbenchmarked data (with the yearly benchmarks). Feeding benchmarked data will generally induce an artificial seasonal movement in the resulting benchmarked series (Cholette, 1978, Section 6b).

6.3 Preliminary Benchmarking of Current Data

A final comment is in order. During a current (uncompleted) year, one cannot calculate growth rates, for instance, between the benchmarked segment of the series (AB) and the unbenchmarked segment (CD). Doing so usually produces a discontinuity BC between the two segments AB and CD as illustrated in Figure 5 by curve ABCD.

Two solutions are then possible. One, the inter-temporal comparisons are based only on the unbenchmarked data. Two, the current data are preliminarily benchmarked by repeating the last available correction BC for the current year. (Note that including the incomplete current year in the objective function (4) (or 12) would yield identical preliminarily benchmarked values.) One can then compare the benchmarked segment AB with the preliminarily benchmarked segment BE as illustrated in Figure 5 by curve ABE. We favour this second alternative.

6.4 Relation with Other Methods

The Denton (1971) benchmarking method, the modified Denton method (presented in this paper), the methods of Glejser (1966), of Boot, Feibes and Lisman (1967), of Lisman and Sandee (1964), and of Bassie (1939) could be reffered to as univariate methods. No series other than that considered and its annual benchmarks enter the benchmarking process. On the contrary, the methods by Friedman (1962), by Chow and Lin (1971), by Somermeyer, Jansen and Louter (1976) and by Wilcox (1983) are multivariate. Auxiliary series are used in the computation of the desired series.

For instance, Chow and Lin (1971) proposed a method to obtain the desired sub-annual series from yearly totals and from related series. The movement of the resulting series is as much as possible similar to the movements of the related series (and the series obtained satisfies the annual constraints). Fernandez (1981) observes that the Chow and Lin method can produce movement discontinuities between the years. He then proposes a synthesis of the Chow-Lin and of the Denton methods. The combined method elimates the inter-annual discontinuities, but still relies on the hypothesis $x_0 = z_0$. As illustrated above, this hypothesis often introduces spurious fluctuations in the calculated series. We would think that it should be possible to refrain from the hypothesis in the case of Fernandez as in the case of Denton.

7. SUMMARY AND CONCLUSIONS

Denton (1971) intended to keep the original and benchmarked series as parallel as made possible by the annual discrepancies. This paper suggested a modification to the benchmarking method which makes the original and benchmarked series more parallel than is the case with the original method. This improvement holds both for the additive and the proportional variants of the method. We suspect that the generalized multivariate method by Fernandez could be improved in the same direction.

The method proposed can very easily be adapted for flow, stock as well as index series.

Before making intertemporal comparisons between the benchmarked and current data, it is essential to preliminarily benchmark the current data (in the manner proposed).

The suggested 5-year moving average implementation of the method will automatically "freeze" the past estimates after two years of revision.

REFERENCES

[1] Baldwin, A. (1978), "New Benchmarking Algorithms using Quadratic Minimization." National Product Division, Statistics Canada, Research Paper.

[2] Bassie, B.L. (1939), "Interpolation Formulae for the Adjustment of Index Numbers," Proceedings of the Annual Meetings of the American Statistical Association

[3] Boot, J.C.G., Feibes, W., Lisman, J.H.C. (1967), "Further Methods of Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 16, No. 1, pp.65-75.

[4] Cholette, P.A. (1978), "A Comparison and Assessment Various Adjustment Methods of Sub-Annual Series to Yearly Benchmarks," Time Series Research and Analysis, Statistics Canada, Research Paper 78-03-001B.

[5] Cholette, P.A. (1979a), "Adjustment Methods of Sub-Annual Series to Yearly Benchmarks," Proceedings of the Computer Science and Statistics, 12th Annual Symposium on the Interface, J.F. Gentleman Ed., University of Waterloo, pp. 358-36.

[6] Cholette, P.A. (1979b), "A Note on 'Freezing' Past Estimates when Benchmarking," Time Series Research and Analysis, Statistics Canada, Research Paper 79-06-002E.

[7] Cholette, P.A. (1982), "Minimum Quadratic Adjustment Program (MQAP-I) of Series to Annual Totals - Users Manual," Time Series Rsearch and Analysis, Statistics Canada, 82-11-003B.

[8] Cholette, P.A. (1983), "Benchmarking Series with Bi-Annual Benchmarks when Knowing the Ending Point," Time Series Research and Analysis, Statistics Canada, Research Paper 83-05-002B.

[9] Chow, G.C., Lin, An-loh (1971), "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series," Review of Economics and Statistics, Vol. 53, No. 4, pp. 372-375.

[10] Dagum, E.B. (1977), "Comparison of Various Interpolation Procedures for Benchmarking Economic Time Series," Time Series Research and Analysis, Statistics Canada, Research Paper 77-05-006E.

[11] Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization," J.A.S.A., Vol. 66, No. 333, pp. 99-102.

[12] Fernandez, R.B. (1981), "A Methodological Note on the Estimation of Time Series," Review of Economic and Statistics, Vol. 63, pp. 471-476.

[13] Friedman, M. (1962), "The Interpolation of Time Series by Related Series," J.A.S.A., Vol. 57, No. 300, pp. 729-757.

[14] Glejser, H. (1966), "Une méthode d'évaluation de données mensuelles à partir d'indices trimestriels ou annuels," Cahiers Economiques de Bruxelles, No. 19, 1er trimestre, pp. 45-64.

[15] Helfand, S.D. Monsour, N.J, Trager, M.L. (1978), "Historical Revision of Current Business Survey Estimates," U.S. Bureau of the Census, (Research Paper).

[16] Huot, G. (1975), "Quadratic Minimization of Montlhly Estimates to Annual Totals," Time Series Research and Analysis, Statistics Canada, Research Paper 75-11 010E.

[17] Lisman. J.H.C., Sandee, J. (1964), "Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 13, No. 2, pp. 87-90.

[18] Smith, P. (1977), "Alternative Method for Step Adjustment," Current Economic Analysis Division, Statistics Canada, Research Paper.

[19] Somermeyer, W.H, Jansen, R., Lauter, A.S. (1976), "Estimating Quarterly Values from Annually Known Variables in Quarterly Relationships," J.A.S.A. Vol. 71, No 355, pp. 588-595.

[20] Wilcox, J.A. (1983), "Disaggregating Data Using Related Series," Journal of Business and Economic Statistics, Vol. 1, No 3, pp. 187-191.

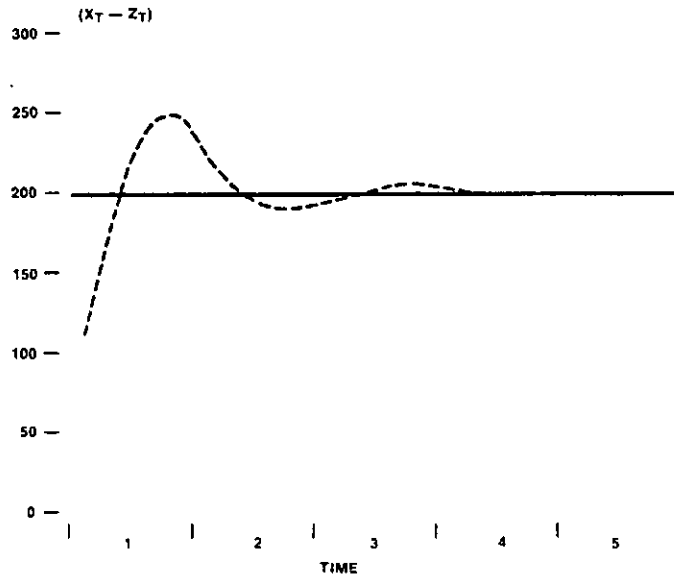

Figure 1: Corrections $(x_t - z_t)$ made to the unbenchmarked series according to Denton's method (dashed line) and according to the method proposed in this paper (solid) in an ideal situation of constant annual discrepancies.

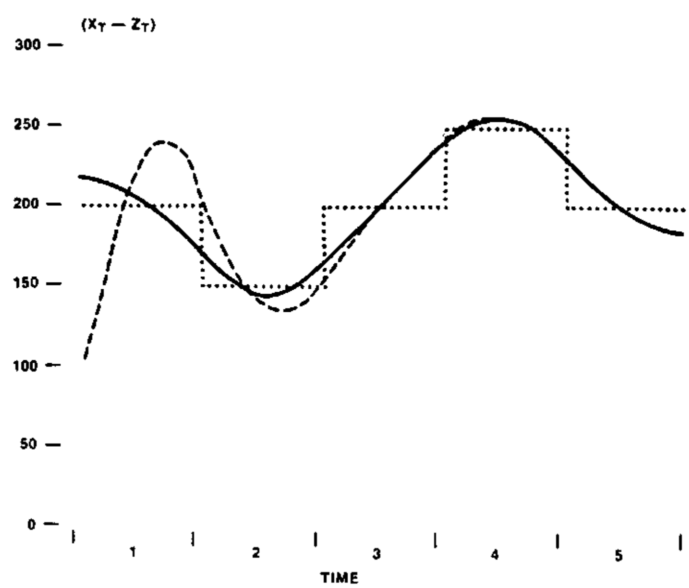

Figure 2: Corrections $(x_t - z_t)$ made to the unbenchmarked series according to Denton's method (dashed line) and according to the benchmarking method proposed in this paper (solid) in a situation of variable average annual discrepancies (dotted).

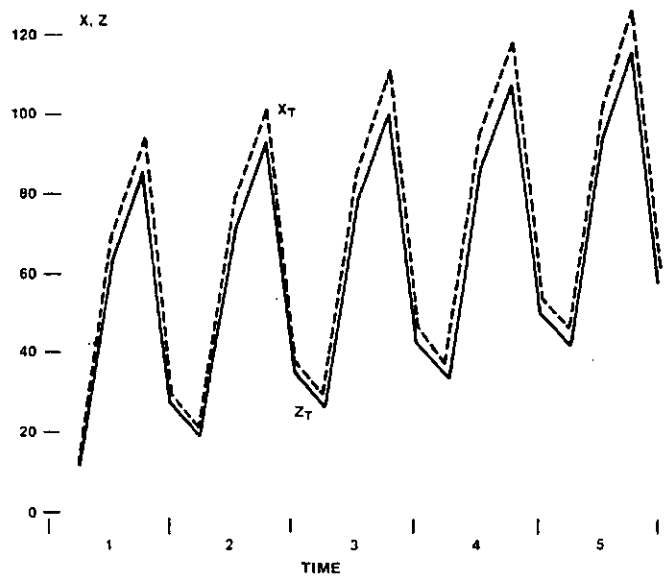

Figure 3: Original series (solid curve) and benchmarked series (dashed) according to the proportional variant of the benchmarking method proposed in this paper (in a situation of constant annual proportional discrepancies).

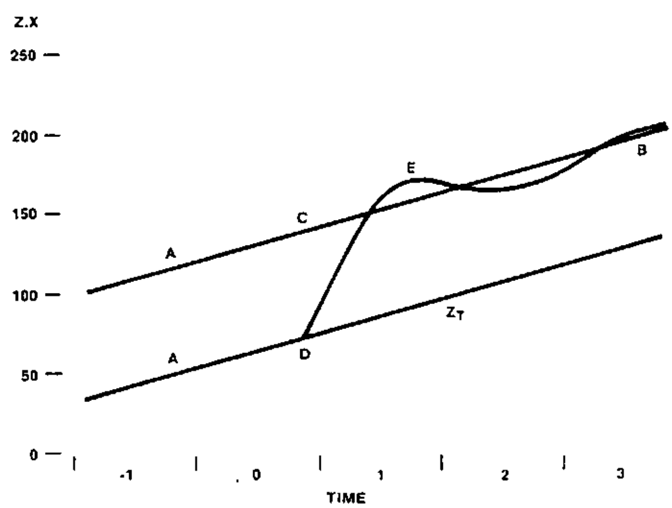

Figure 4: Benchmarked series according to Denton's method, when there are no benchmarks for year -1 and 0 (curve ADEB) and when there are benchmarks and year -1 and 0 were already benchmarked (A'CDEB): and according to the method proposed in this paper, applied in a moving average manner, when there are benchmarks for years -1 and 0 (A'B).

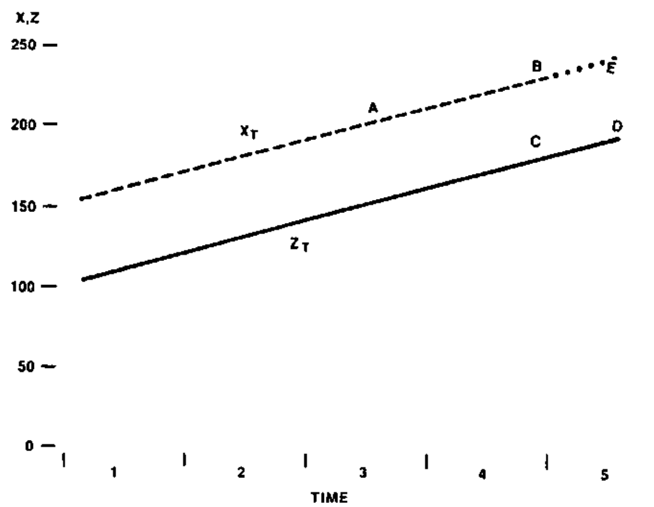

Figure 5: Continuity between the benchmarked series (dashed curve) and the preliminarily benchmarked series (dotted) and discontinuity BC between the benchmarked (dashed) and the unbenchmarked (solid) series.

EXAMINING EXPENDITURES ON ENERGY

Louise A. Heslop[1]

Using data from the Family Expenditures Surveys over time, consumer expenditures on in-home and transportation energy from 1969 to 1982 are being studied. This article briefly summarizes some of the procedures being used to explore the data, summarize it and develop insights into shifts in consumption for policy implications purposes. With such a complex data set and such a complex, multi-faceted subject for analysis some effort must be made to reduce information flows and at the same time increase the information content of each factor of both input and output in the analyses.

1. THE ENERGY ISSUE

To some, energy conservation may be a dead issue. There is no shortage of energy (maybe never was): prices for energy have stabilized.

Energy matters dominated the 1970's having major impacts on the world economic order and creating international strife. Domestically they impacted drastically on federal - provincial relations and business - government relations and on family budgets: caused the restructuring of the manufacturing base, the auto industry, etc. Despite its reported demise as an important issue, energy consumption and prices remain as high priority concerns of consumers, businesses and governments. Energy conservation has lost its sparkle but not its real value.

The research I will be reporting on briefly has been developed in consultation with policy makers in Consumer and Corporate Affairs Canada and Energy, Mines and Resources Canada which continue to run active research programmes on consumer energy use and conservation. The project structure has taken their interests, orientations and limitations into consideration.

Also, within the last five years an international group of social scientists has begun a series of research and information exchanges on consumer behaviour and energy use. As a member of that group I have been keenly aware

[1] Louise A. Heslop, Research and Analysis Division, Statistics Canada.

of the problems and prospects and the current state of knowledge and research techniques of that group.

2. PROBLEMS IN ENERGY RESEARCH

Perhaps the major problem in studies of consumer energy use has been to obtain reasonably reliable measures of use from sufficiently large and representative samples. Getting such data over a period of time, especially a time period spanning the infamous 1973 oil embargo period, would send a researcher into Nirvana. The Family Expenditure data collected by the Consumer Income and Expenditure Division of Statistics Canada come close enough to these requirements to at least set one's heart fluttering. It is a series of retrospective recall studies conducted for the years 1969, 1972, 1974, 1976, 1978 and 1982. So it covers the time period of interest for a large sample and the sampling technique used ensures that the design is representative of Canada for those areas studied, usually urban centres. Additionally it contains a great many other variables of interest in any study of energy use, e.g., home ownership, some house characteristics, vehicle and appliance ownership, family characteristics and expenditures on other categories of consumer goods and services, etc.

Most studies which attempt a measure of consumer expenditures rely on recall or file checking by respondents. There are obvious problems with the accuracy of such data on an individual basis. The problems are less restrictive with very large samples. For most independent studies, the costs of such large samples are prohibitive. However, FAMEX sample sizes are very large.

Only one major study in Canada has used independent record checking, obtaining records from suppliers by household with the permission of the householder, but through this technique was able to obtain electricity use records on less than half of its sample. Natural gas and oil records were obtained on only about one-third of the sample. This procedure of record checking is highly accurate, removes the problems associated with recall, especially over long periods of time, and of reporting bias of respondents. However, practically it is impossible to use for large samples across the country.

Although the FAMEX Study uses recall procedures, the information on energy expenditures are not likely to be as biased as in a study specifically designed to record energy behaviours since respondents are not sensitized to the subject of the study. Also the data from pre-energy crisis periods was collected in the same way as that since the crisis, again reducing the likelihood of response bias. So the FAMEX data set offers a unique opportunity to examine a very large set of samples during a very important period of time.

The data set is not without its problems, some because of the sampling procedure and some because of the inherent complexity of any study of energy use. Changes in expenditure categories and their contents, especially those other than energy, have required that we manipulate the data considerably to create consistency across years. It is not possible to track in-home energy expenditures for those families who do not pay for energy directly, i.e., apartment dwellers with central metering and roomers. Some researchers have imputed values to these households based on their rents but we chose not to, and instead have chosen to restrict our study to those households who have the ability to monitor and affect their own energy use. These households are the consumer groups who will be the focus of any government programmes to alter consumer comsumption.

There are several factors which make the study and the altering of energy consumption of households difficult:

- Capital commitments restrict the ability of the household to respond in the short-term and increase the cost of response - e.g., house size, number and type of appliances, size and number of vehicles. Some studies have noted that home characteristics alone may account for 24% of in-home energy consumption. Family size may be considered as a capital commitment as well.
- Flow feasibilities - There are restrictions in the ability to change the amount and types of fuels used depending on the technology and fuels available under different circumstances and for varying amounts of money, e.g., natural gas heating is not available to rural residents: instantaneous changes can not be made in the type of home heating fuel used.
- Exogenous factors affect the amount of energy needed for similar

performance in different situations, e.q., weather, distances between points in cities, etc.

3. SUMMARIZING INFORMATION INPUTS AND MAXIMIZING INFORMATION OUTPUTS

With such a complex data set and such a complex, multi-faceted subject for analysis some effort must be taken to reduce information flows and at the same time increase the information content of each factor of both input and output. There are several ways of doing this, some of which we will be usinq, they include:

a) Constructing Complex Input Variables - to reduce the number of factors being studied to the most salient ones.

i) Discontinuous complex input variables were created by combininq in-home and transportation energy consumption but not as continuous variables. Rather groupings were created to develop a set of typologies whose characteristics can then be examined for differences. In this case the qroupinqs were developed by creating expenditure quartiles for each energy category, collapsinq the two middle categories, and then combining the two resulting three cells into a nine cell matrix of interrelated categories (see Table 1, source: McDougall, Ritchie and Claxton). In particular, the corner cells are of interest in contrast to each other and to the middle cell. This typology was developed in an earlier study for Consumer and Corporate Affairs Canada. So comparing the output from the FAMEX data to the data set used in the CCA study will be of particular interest. Comparing the characteristics of these groups over time will also be of interest. For example, do the Churchmice continue to be impoverished Canadians (involuntary simplicity) or is there any indication that there is some voluntary embracinq of low energy, lifestyles? In Table 2 the characteristics of three cells of the typology from two different years are compared - the Churchmice, the Roadrunners and the Hippos. Looking first at the Churchmice, information on a selection of possible analysis variables is shown across two different years, 1974 and 1978. To simplify for this presentation only the rankings of the cell within the typology set of cells is given. Characteristically those consuming the least amount of energy have had the least resources in general, i.e., the lowest incomes, the lowest levels of education, the oldest. These characteristics are evident for the Churchmice in 1974, they also have the lowest levels of consumption for all the expenditure categories shown. Although they are the oldest group they do not have the lowest number of very young children. Probably this group consists of a mix of senior citizens and single parent households (probably headed by women) with young children. Note that this group also has the lowest number of full-time earners (F-T earners). In 1978 the general picture is still the same except that this group is no longer the oldest. In fact the oldest group is in the adjacent cell to the right in the typology (not shown here). It would seem that in 1978 the very old are consuming a relatively larger amount of in-home energy. Perhaps this group is financially better off in 1978 than in 1974 or perhaps they have been unable to hold the line on energy expenditures as prices have risen.

In 1974 the Hippos also fit expectations. They seem to be middle-aged with large numbers of children 5-16 years of age. The "full nest" family, they spend the largest amount on most expenditure categories. They are also the most highly educated. In 1978 this is no longer true as the education ranking of this cell has dropped. Also this group no longer has the highest shelter expenditure. Some suggestions for these observations may be that those with the largest homes and the highest education have begun to modify their homes to reduce energy expenditures.

The Roadrunners have changed also. In 1974 they were the youngest group with very small families. In 1978 they appear to be characterized as young families with young children. One of the most dramatic changes for this group has been that their alcoholic beverages and tobacco expenditures have dropped dramatically.

The significance of these changes can be determined with appropriate statistical tests. The purpose of this discussion was to introduce the idea of searching for meaningful typologies within the data. Pictures of the life-styles of the groups emerge which can be very useful in furthering conservation programmes directed at each group.

Further analysis may look not at level of expenditures but at percent of expenditures. Such an analysis will reveal the characteristics of those who are most heavily burdened with energy bills.

ii) Continuous complex input variables can be constructed to eliminate the effects of variables known to have very large effects, but ones which are difficult or impossible for consumers to manage.

In-home energy expenditures can be examined for factors related to them, but since one of the main determinants of in-home energy expenditures is house size, this size factor can be absorbed into the input variable to allow for examination of other more relevant (from a policy perspective) factors. So instead of in-home energy expenditures, in-home expenditures/room are examined. Taking this one step further, climate and weather variances from year to year may be controlled for by looking at expenditures/room/degree day. This last factor is added to the data set by city by year. Degree day data for each year for each city were obtained from Environment Canada. Table 3 indicates how the figures change as the factor studied becomes more complex again across two of the years of data. A comparison of the two years and differences in the measures of change between years suggests the importance of refining the measure to improve understanding of the process.

b) Constructing summary output variables to examine the structure of the data - Example of regression coefficients.

In Tables 4-6 some regression outputs are presented. Three models are examined. In each succeeding model the dependent variable becomes more complex. In so doing the factors known to impact significantly on energy consumption can be controlled for and the effects of the remaining variables examined more constructively for any significant explanatory power.

In these analyses no attempt has been made to deal with the problem of the complex sampling design. A future analysis will do so using the Taylor linearization procedure and results will be compared. However, the results from both a weighted and an unweighted sample are shown for 1974. As can be seen the values of the coefficients change very little and their significance or lack thereof does not change. Because of the restrictions indicated and also the fact that the very large sample sizes are used here produce significant results under conditions of very slight differences, it is advised that great care be taken in viewing these preliminary results for purposes of this discussion. I will only note the variables significant at the .01 level and beyond and then only their sign.

In the independent variable list dummy variables are used in the first and second models for city and in all three models for type of dwelling type. The unspecified condition is Ottawa for city and single detached house for dwelling type.

In 1974 house size, some city variables, total expenditures, age of head and family size and some house types are significant. Large families with high total expenditures living in single detached homes in St. John's consume the most. Western cities consume less than the east, and all other housing types consume less than detached houses, although duplexes not significantly so when number of rooms is controlled for. The unweighted results are similar to the weighted.

When the dependent variable is changed to \$/room and number of rooms is removed from the list of independent variables the general pattern remains. However, family size is no longer significant (probably closely tied to dwelling size only), and education of family head becomes significant with a negative sign. Those with less education consumed more, all other things being equal. Finally duplexes become significant with a positive sign, so when number of rooms is controlled for, duplexes use more energy than detached houses.

In model 3 climatic conditions are taken into account by controlling on degree days in the dependent variable and the list of cities is dropped from the independent variable set.

It should be noted that the value of the coefficients drops so dramatically because there are between 4000 and 7000 degree days in these cities. So the small value of the coefficients does not mean they are unimportant. Total expenditures remains significant as does education of the family head and the rowhouse effect. An important thing to note is the drop in the value of the adjusted R-squared. In fact the independent variables remaining in the equation do not do very much to help in explaining variance in the dependent variable. Other more useful variables should be sought.

When we compare just the unweighted 1974 and 1978 results, in model 1 some change in the Vancouver parameter can be noted and in the importance of semi-detached and duplex housing over detached houses.

In model 2 again the major change is in dwelling type effects. Finally in model 3 only the rowhouse variable shows any difference from the detached:

education of the head is again important, but in 1978 age of head is significant with a positive coefficient. Some improvement is seen in the R-squared for 1978, but it is still very low.

This cross-year comparison from a policy perspective suggests perhaps that improvements have been made in the quality of the detached housing stock in Canada. From a methodological perspective it indicates the importance of choosing the dependent variable with care.

As was earlier noted, much additional analysis and re-analysis will be done using the regression procedures available to refine these results and take the sampling design into account.

As I noted earlier the FAMEX data sets have their limitations but they also contain a wealth of important information which should be fruitfully explored.

REFERENCE

[1] McDougall, Gordon H.G., Ritchie, J.R. Brent, and Claxton, John D. (1979). "Energy Conservation and Conservation Patterns in Canadian Households: Overview." Behavioral Energy Research Group, 203-2053 Main Hall, University of British Columbia.

Table 1: Energy Consumption Taxonomy - Labels

		Level of In-Home Energy Consumption			
		Low 127 Mil. kJ	Medium 127-222 Mil. kJ	High 222 Mil. kJ	Total
Level of Automobile Gasoline Consumption	Low 1136 litre	CHURCH MOUSE 4.5% of sample	9.8% of sample	BEAR 2.5% of sample	16.8
	Medium 1136-4545 litre	14.5% of sample	BEAVER 33.7% of sample	12.3% of sample	60.5
	High 4546 litre	ROADRUNNER 4.0% of sample	12.6% of sample	HIPPO 6.1% of sample	22.7
	Total	23.0	56.1	20.9	100.0

Source: See reference list.

Table 2: Rank among Typology Cells

	Churchmice		Hippos		Roadrunners	
	1974	1978	1974	1978	1974	1978
Education of Head (low-hi)	1	1	9	7	7	8
Age (old - yng)	1	2	6	6	9	9
F-T Earners (low-hi)	1	1	8.5	9	7	6.5
Family Size (low-hi)	1	1	9	9	4	4
Child Less than 5 (low-hi)	3	1	4	2	1.5	7
Child 5-15 (low-hi)	1	2.5	7	6.5	5	2.5
Food at Stores (low-hi)	1	1	9	9	4	4
Food at Eating Places (low-hi)	1	1	9	9	6	6
Shelter (low-hi)	1	1	9	7	4	3
Clothing (low-hi)	1	1	9	9	6	5
Personal Care (low-hi)	1	1	9	9	5	4
Medical (low-hi)	1	1	8	8	4	4
Tobacco & Alcohol (low-hi)	1	1	9	9	7	4
Reading, Recreation, Education (low-hi)	1	1	9	8	8	9

Table 3: Average In-Home Energy Expenditures, 1974-78

	1974	1978	% Change
Average $ in-home energy expenditure	451	764	+69
Average $/room in-home energy expenditure	73	121	+66
Average $/room/dd in-home energy expenditure	.019	.029	+53

Table 4: Regression Analysis Results - Model 1 - $In-Home Energy

	1974 Unweighted	1974 Weighted	1978 Unweighted
Intercept	197.3 A	225.4 A	298.0 A
No. of Rooms	13.9 A	12.0 A	4.2 C
City - St. John's	193.9 A	204.9 A	341.1 A
Halifax	75.5 A	73.9 B	162.0 A
Montreal	12.2	22.7	-16.6
Toronto	-10.2	-3.0	50.5
Winnipeg	-127.1 A	-125.4 A	-72.2 C
Edmonton	-244.9 A	-243.2 A	-195.8 A
Vancouver	-22.9	-17.5	-71.9 C
Total Expenditures	.006 A	.006 A	.01 A
Age of Head	1.2 A	0.8 B	3.6 A
Family Size	13.2 A	12.1	21.6 B
Education of Head	0.7	0.6	-3.6
House Type - Semi Det.	-50.9 B	-49.0 A	-23.8
Rowhouse	-81.2 A	-88.9 A	-119.7 B
Duplex	-12.3	-13.7	-84.6 C
Adjusted R^2	0.43	0.34	0.38
F value (prob.)	118.5(.0001)	79.7(.0001)	74.6(.0001)

Note: A = prob. less than .0001, B = prob. less than .001, C = prob. less than .01

Table 5: Regression Analysis Results - Model 2 - $/Room

	1974 Unweighted	1974 Weighted	1978 Unweighted
Intercept	76.2 A	77.3 A	99.8 A
City - St. John's	30.4 A	32.0 A	74.8 A
Halifax	16.8 A	16.3 B	31.6 A
Montreal	4.5	6.7	6.5
Toronto	-3.5	-1.7	10.1
Winnipeg	-17.6 A	-16.3 A	-0.9
Edmonton	-37.9 A	-36.8 A	-26.4 A
Vancouver	0.3	0.8	-6.7
Total Expenditures	2.2×10^{-4} B	2.5×10^{-4} B	6.9×10^{-4} A
Age of Head	0.015	-0.03	0.33 B
Family Size	0.6	0.04	-0.63
Education of Head	-1.9 A	-1.4 B	-4.0 A
House Type - Semi Det.	-6.5 C	-7.1 B	3.1
Rowhouse	-11.5 A	-11.8 A	-11.0
Duplex	6.1 C	6.6 C	3.24
Adjusted R^2	.31	.19	.24
F value (Prob.)	73.85(.0001)	38.9(.0001)	41.4(.0001)

Note: A = prob. less than .0001, B = prob. less than .001, C = prob. less than .01.

Table 6: Regression Analysis Results - Model 3 - $/Room/DD

	1974 Unweighted	1974 Weighted	1978 Unweighted
Intercept	.017 A	.019 A	.02 A
Total Expenditures	8.01×10^{-8} B	9.4×10^{-8} A	1.4×10^{-7} A
Age of Head	1.8×10^{-5}	-7.0×10^{-6}	9.9×10^{-5} A
Family Size	-1.4×10^{-5}	-18.4×10^{-5}	27.0×10^{-5}
Education of Head	-5.3×10^{-4} A	-4.7×10^{-4} A	-7.8×10^{-4} B
House Type - Semi Det.	3.4×10^{-4}	-7.5×10^{-4}	24.8×10^{-4}
Rowhouse	-23×10^{-4} C	-35.9×10^{-4} A	-38.8×10^{-4} B
Duplex	16.9×10^{-4}	6.3×10^{-4}	11.6×10^{-4}
Adjusted R^2	.01	.02	.03
F value (Prob.)	5.6(.0001)	6.6(.0001)	9.5(.0001)

Note: A = prob. less than .0001, B = prob. less than .001, C = prob. less than .01

LOGISTIC REGRESSION ANALYSIS OF LABOUR FORCE SURVEY DATA

S. Kumar and J.N.K. Rao[1]

Standard chisquared (X^2) or likelihood ratio (G^2) tests for logistic regression analysis, involving a binary response variable, are adjusted to take account of the survey design. The adjustments are based on certain generalized design effects. The adjusted statistics are utilized to analyse some data from the October 1980 Canadian Labour Force Survey (LFS). The Wald statistic , which also takes the survey design into account, is also examined for goodness-of-fit of the model and for testing hypotheses on the parameters of the assumed model. Logistic regression diagnostics to detect any outlying cell proportions in the table and influential points in the factor space are applied to the LFS data, after making necessary adjustments to account for the survey design.

1. INTRODUCTION

Logistic regression models have been extensively used by researchers in social, behavioural and health sciences to analyse the variation in binomial proportions (see, for example, the books by Cox (1970) and McCullagh and Nelder (1983)). Due to clustering and stratification used in the survey design the statistical methods for binomial proportions, however, are often inappropriate for analysing sample survey data. For instance, the standard chisquared (X^2) or the likelihood ratio (G^2) tests greatly inflate the type I error rate (significance level). Hence, some adjustments to the classical methods that take account of the survey design are necessary in order to make valid inferences from survey data. In this article, we have utilized two simple adjustments to X^2 or G^2, based on certain generalized design effects (deffs) to analyse some data from the October 1980 Canadian Labour Force Survey (LFS) (Section 3). The Wald statistic, which also takes the survey design into account, is also examined.

[1] S. Kumar, Census and Household Survey Methods Division, Statistics Canada, and J.N.K. Rao, Department of Mathematics and Statistics, Carleton University.

In addition to formal statistical tests, it is essential to develop diagnostic procedures to detect any outlying cell proportions and influential points in the factor space. Regression diagnostics for the standard linear model have been extensively investigated in the literature (see the recent book by Cook and Weisberg (1982)). Pregibon (1981) recently developed similar methods for the logistic regression with binomial proportions. In Section 4 some of these methods have been applied to the October 1980 LFS data, after making necessary adjustments to account for the survey design.

2. THEORETICAL RESULTS

Suppose that the population of interest is partitioned into I cells (domains) according to the levels of one or more factors, and $\hat{N}_i$ denotes the survey estimate of the i-th domain size, N_i ($i = 1, 2, \ldots, I$; $\Sigma N_i = N$). The corresponding estimate of the i-th domain total, N_{i1}, of a binary (0, 1) response variable is denoted by $\hat{N}_{i1}$. The ratio estimate, $\hat{p}_i = \hat{N}_{i1}/\hat{N}_i$, is used to estimate the population proportion $\pi_i = N_{i1}/N_i$.

A logit model on the proportions π_i is given by $\pi_i = f_i(\underset{\sim}{\beta})$, where

$$\ln\{f_i/(1 - f_i)\} = \text{logit } f_i = \underset{\sim}{x}_i'\underset{\sim}{\beta}\,, \; i = 1, \ldots, I. \tag{1}$$

In (1), $\underset{\sim}{x}_i$ is an s-vector of known constants derived from the factor levels and $\underset{\sim}{\beta}$ is the s-vector of unknown parameters. Under independent binomial sampling in each domain, the maximum likelihood estimates (m.l.e.) are obtained from the following likelihood equations:

$$X'D(\underset{\sim}{n}/n)\hat{\underset{\sim}{f}} = X'D(\underset{\sim}{n}/n)\hat{\underset{\sim}{q}}, \tag{2}$$

where $X' = (\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_I)$, $D(\underset{\sim}{n}/n) = \text{diag}(n_1/n, \ldots, n_I/n)$, $\hat{\underset{\sim}{f}} = \underset{\sim}{f}(\hat{\underset{\sim}{\beta}}) = (\hat{f}_1, \ldots, \hat{f}_I)'$, and $\hat{\underset{\sim}{q}}$ is the vector of sample proportion $q_i = n_{i1}/n_i$, where n_i is the sample size from i-th domain ($\Sigma n_i = n$). For general sample designs, we do not have m.l.e. due to difficulties in obtaining appropriate likelihood functions. Hence, it is a common practice to use a "pseudo m.l.e." of $\underset{\sim}{\beta}$ or $\underset{\sim}{f}$

obtained from (2) by replacing n_i/n by the estimated domain relative size, $w_i = \hat{N}_i/\hat{N}$, and q_i by the survey estimate $\hat{p}_i$:

$$X'D(\underset{\sim}{w})\hat{\underset{\sim}{f}} = X'D(\underset{\sim}{w})\hat{\underset{\sim}{p}}. \tag{3}$$

The resulting estimates, $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{f}} = \underset{\sim}{f}(\hat{\underset{\sim}{\beta}})$, are asymptotically (i.e., in large samples) consistent. The equations (3) may also be written as

$$X'\hat{\underset{\sim}{N}}_1(m) = X'\hat{\underset{\sim}{N}}_1, \tag{4}$$

where $\hat{\underset{\sim}{N}}_1$ is the vector of estimated counts $\hat{N}_{i1}$, and $\hat{\underset{\sim}{N}}_1(m)$ is the vector of pseudo m.l.e., $\hat{N}_{i1}(m) = \hat{N}_i\hat{f}_i$, of the totals N_{i1}. The estimates $\hat{\underset{\sim}{\beta}}$, and hence $\hat{\underset{\sim}{f}}$ and $\hat{\underset{\sim}{N}}_1(m)$, are obtained from (3) or (4) by iterative calculations.

2.1 Estimated Variances and Covariances

Let $\hat{V}$ denote the estimated covariance matrix of $\hat{\underset{\sim}{p}}$, then the estimated covariance matrix of $\hat{\underset{\sim}{\beta}}$ is given by

$$\hat{D}(\hat{\underset{\sim}{\beta}}) = (X'\hat{\Delta}X)^{-1}(X'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})X)(X'\hat{\Delta}X)^{-1} \tag{5}$$

in large samples, where $\hat{\Delta} = \mathrm{diag}(w_1\hat{f}_1(1 - \hat{f}_1), \ldots, w_I\hat{f}_I(1 - \hat{f}_I))$. The diagonal elements of (5) provide the estimated variances of the estimates $\hat{\beta}_i$. Similarly, the estimated covariance matrix of the residual vector $\underset{\sim}{r} = \hat{\underset{\sim}{p}} - \hat{\underset{\sim}{f}}$ is given by

$$\hat{D}(\underset{\sim}{r}) = A\hat{V}A', \tag{6}$$

where

$$A = I - D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})X(X'\hat{\Delta}X)^{-1}X'D(\underset{\sim}{w}). \tag{7}$$

The diagonal elements $\hat{V}_{ii}(r)$ of (6) lead to standardized residuals $r_i/\mathrm{s.e.}(r_i)$ which are useful in detecting outlying cell proportions.

2.2 Goodness-of-Fit Tests

The standard chi-squared test of goodness-of-fit of the model (1) is given by

$$X^2 = n \sum_{i=1}^{I} \frac{(\hat{p}_i - \hat{f}_i)^2 w_i}{\hat{f}_i(1 - \hat{f}_i)} = \sum_{i=1}^{I} X_i^2. \tag{8}$$

The likelihood ratio test statistic is given by

$$G^2 = 2n \sum_{i=1}^{I} w_i \left\{ \hat{p}_i \ln \frac{\hat{p}_i}{\hat{f}_i} + (1 - \hat{p}_i) \ln \frac{(1 - \hat{p}_i)}{(1 - \hat{f}_i)} \right\} = \sum_{i=1}^{I} G_i^2 \tag{9}$$

Note that G_i^2 is also defined at $\hat{p}_i = 0$ and 1 as given by $-2nw_i \ln(1 - \hat{f}_i)$ and $-2nw_i \ln \hat{f}_i$ respectively. Under independent binomial sampling, it is well known that both X^2 and G^2 are asymptotically distributed as a χ^2 variable with $I - s$ degrees of freedom, but for general designs this result is no longer valid. In fact, X^2 (or G^2) is asymptotically distributed as a weighted sum $\Sigma\, \delta_i Z_i$, of independent χ^2 variables, Z_i, each with 1 d.f. where the weights δ_i $(i = 1, \ldots, I - s)$ are the eigenvalues of a "generalized design effects" matrix given by $\Sigma_0^{-1} \Sigma_\phi$, where

$$\Sigma_\phi = G'D(\underset{\sim}{\hat{f}})^{-1}D(\underset{\sim}{1} - \underset{\sim}{\hat{f}})^{-1}\hat{V}D(\underset{\sim}{\hat{f}})^{-1}D(\underset{\sim}{1} - \underset{\sim}{\hat{f}})^{-1}G, \tag{10}$$

$$\Sigma_0 = \frac{1}{n} G'\hat{\Delta}^{-1}G \tag{11}$$

and G is any $I \times (I - s)$ matrix of rank $I - s$ such that $G'X = 0$, i.e., G is orthogonal to X. Under binomial sampling, $\Sigma_0^{-1} \Sigma_\phi$ reduces to I, the identity matrix

A simple adjustment to X^2 (or G^2) is obtained (Roberts, 1984) by treating $X_c^2 = X^2/\delta$. or $G_c^2 = G^2/\delta$. as χ^2 with $I - s$ degrees of freedom (d.f.) under the hypothesis that the model is true, where

$$(I - s)\delta_{.} = n \sum_{i=1}^{I} \hat{V}_{ii}(r) w_i / [\hat{f}_i(1 - \hat{f}_i)]. \tag{12}$$

The adjusted statistic X_c^2 (or G_c^2) should be satisfactory excepting in those cases with a large coefficient of variation (C.V.) of the δ_i's. A better adjustment, based on the Satterthwaite approximation, treats $X_S^2 = X_c^2/(1 + a^2)$ or $G_S^2 = G_c^2/(1 + a^2)$ as χ^2 with $(I - s)/(1 + a^2)$ d.f., where

$$a^2 = \Sigma\,(\delta_i - \delta_{.})^2/[(I - s)\delta_{.}^2] \tag{13}$$

is the $(C.V.)^2$ of the δ_i's and

$$\Sigma\,\delta_i^2 = \sum_{i=1}^{I} \sum_{j=1}^{I} \hat{V}_{ij}^2(r)(nw_i)(nw_j) \;/\; [\hat{f}_i \hat{f}_j (1 - \hat{f}_i)(1 - \hat{f}_j)] \;, \tag{14}$$

where $\hat{V}_{ij}(r)$ is the (i, j)-th element of $\hat{D}(r)$. The statistics X_S^2 and G_S^2 take account of the variation in δ_i's.

A Wald statistic for goodness-of fit of the model (1) is given by

$$X_W^2 = \hat{\underset{\sim}{\nu}}' G \Sigma_{\hat{\phi}}^{-1} G' \hat{\underset{\sim}{\nu}}, \tag{15}$$

where $\hat{\underset{\sim}{\nu}}$ is the vector of logits $\hat{\nu}_i = \text{logit}\ \hat{p}_i$. The statistic X_W^2 is distributed as χ^2 with $I - s$ d.f., in large samples. The statistic X_W^2 is not defined if $\hat{p}_i = 0$ or 1 for some i. Moreover, it becomes unstable when any $\hat{p}_i$ is close to 1 (see Section 3), or when the degrees of freedom for $\hat{V}$ is not large compared to I - s (Fay, 1983).

2.3 Nested Hypothesis

Suppose the matrix X is partitioned as (X_1, X_2) where X_1 is $I \times r$ and X_2 is $I \times u$ $(r + u = s)$, then the model (1) may be written as

$$\underset{\sim}{\nu} = X\underset{\sim}{\beta} = X_1\underset{\sim}{\beta}_1 + X_2\underset{\sim}{\beta}_2, \tag{16}$$

where $\underset{\sim}{\beta}_1$ is r×1 and $\underset{\sim}{\beta}_2$ is u×1. We are often interested in testing the null hypothesis H: $\underset{\sim}{\beta}_2 = 0$ given the model (16). The "pseudo m.l.e." under H can be obtained from the equations

$$X_1'D(w)\hat{\hat{\underset{\sim}{f}}} = X_1'D(w)\hat{\underset{\sim}{p}} \tag{17}$$

again by iterative calculations, where $\hat{\hat{\underset{\sim}{f}}} = f(\hat{\hat{\underset{\sim}{\beta}}})$. The standard chisquared and likelihood ratio tests of H: $\underset{\sim}{\beta}_2 = 0$ are given by

$$X^2(2|1) = n\sum_{i=1}^{I}\frac{w_i(\hat{f}_i - \hat{\hat{f}}_i)^2}{\hat{\hat{f}}_i(1 - \hat{\hat{f}}_i)} \tag{18}$$

and

$$G^2(2|1) = 2n\sum_{i=1}^{I} w_i\left\{\hat{f}_i\ln\frac{\hat{f}_i}{\hat{\hat{f}}_i} + (1 - \hat{f}_i)\ln\frac{(1 - \hat{f}_i)}{(1 - \hat{\hat{f}}_i)}\right\} \tag{19}$$

respectively. Under binomial sampling, both $X^2(2|1)$ and $G^2(2|1)$ are asymptotically distributed as χ^2 with u d.f. when H is true, but for general designs this result is no longer valid. In fact $X^2(2|1)$ or $G^2(2|1)$ is asymptotically distributed as a weighted sum, $\Sigma\delta_i(H)Z_i$, of independent χ_1^2 variables Z_i, where the weights $\delta_i(H)$ (i =1, ..., u) are the eigenvalues of the design effects matrix.

$$(\widetilde{X}_2'\Delta\widetilde{X}_2)^{-1}(\widetilde{X}_2'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})\widetilde{X}_2), \tag{20}$$

where

$$\widetilde{X}_2 = [I - X_1(X_1'\Delta X_1)^{-1}X_1'\Delta]X_2, \tag{21}$$

(Roberts, 1984). In the binomial case, the design effects matrix (20) reduces to I, as in the previous case of goodness-of-fit.

A simple adjustment to $X^2(2|1)$ or $G^2(2|1)$ is obtained by treating $X_c^2(2|1) = X^2(2|1)/\delta_.(H)$ or $G_c^2(2|1)/\delta_.(H)$ as χ^2 with u d.f. under H, where

$$u\,\delta_{\cdot}(H) = n \sum_{i=1}^{I} \widetilde{V}_{ii}(r) w_i / \hat{\hat{f}}_i (1 - \hat{\hat{f}}_i) \tag{22}$$

and $\widetilde{V}_{ii}(r)$ is the i-th diagonal element of the covariance matrix of residuals, $r_i(H) = \hat{f}_i - \hat{\hat{f}}_i$, given by

$$\widetilde{V}(r) = D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})\widetilde{X}_2 A \widetilde{X}_2' D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}}) \tag{23}$$

where

$$A = (\widetilde{X}_2'\Delta\widetilde{X}_2)^{-1}[\widetilde{X}_2' D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})\widetilde{X}_2](\widetilde{X}_2'\Delta\widetilde{X}_2)^{-1} \tag{24}$$

The standardized residuals $(\hat{f}_i - \hat{\hat{f}}_i)/[\widetilde{V}_{ii}(r)]^{\frac{1}{2}}$ can also be computed. As in the case of goodness-of-fit. improved approximation. based on Satterthwaite's method, can also be obtained.

A Wald statistic of H: $\underset{\sim}{\beta}_2 = 0$ is given by

$$X_W^2(2|1) = \hat{\beta}_2'[\hat{D}(\hat{\underset{\sim}{\beta}}_2)]^{-1}\hat{\underset{\sim}{\beta}}_2. \tag{25}$$

where $\hat{D}(\hat{\underset{\sim}{\beta}}_2)$ is the principal submatrix in (5) corresponding to $\hat{\underset{\sim}{\beta}}_2$. Under H. $X^2(2|1)$ is asymptotically distributed as χ^2 with u d.f. In particular if β_2 is a scalar, we can treat $\hat{\beta}_2/\text{s.e.}(\hat{\beta}_2)$ as N(0.1)-variate under the hypothesis H: $\beta_2 = 0$ or $\hat{\beta}_2^2/\text{var}(\hat{\beta}_2)$ as χ^2 with 1 d.f.

2.4 Diagnostics

It is desirable to make a critical assessment of the logit fit by identifying any outlying cell proportions and influential points in the factor space. For this purpose. the vector of residuals and a projection matrix in the factor space provide useful tools. However. unlike in the case of the standard linear model. the residuals can be defined on different scales. The natural choice that takes account of the survey design is the vector of standardized residuals $e_i = r_i/[\hat{V}_{ii}(r)]^{\frac{1}{2}}$ given in section 2.1. Since the e_i's are approximately N(0, 1) under the model (1), the expected numbers of residuals e_i exceeding 1.96, 2.33 and 2.58 in magnitude are 0.05I, 0.02I and 0.01I respectively, where I is the number of residuals (cells). These expected numbers provide a rough guide to identify any outlying cells. Ignoring the design and hence using standardized residuals under binomial sampling could lead to misleading conclusions.

The standardized residuals e_i, however, become unreliable for those cells with $\hat{p}_i = 1$ or close to 1. Following Pregibon (1981), we suggest the use of components of X_c^2 or G_c^2, viz., $\tilde{X}_i = X_i/\delta_\cdot^{\frac{1}{2}}$ or $\tilde{G}_i = G_i/\delta_\cdot^{\frac{1}{2}}$, i = 1, ..., I, for residual analysis in order to circumvent this difficulty. In either case, large individual components should roughly indicate cells poorly accounted for by the model. Index plots (i.e., plots of $\tilde{X}_i$ vs i and $\tilde{G}_i$ vs i) are useful for displaying these components. Normal probabilities plot of $\tilde{X}_i$ or $\tilde{G}_i$ (i.e., the ordered values plotted against standard normal quantiles) is also useful to detect deviations from the model (i.e., deviations from a straight-line configuration).

Pregibon (1981) suggested the use of diagonal elements, m_{ii}, of the projection matrix

$$M = I - \hat{V}_b^{\frac{1}{2}}X(X'\hat{V}_bX)^{-1}X'\hat{V}_b^{\frac{1}{2}}$$

$$= I - H \text{ (say)} \tag{26}$$

to detect influential points, where $\hat{V}_b$ is the estimated covariance matrix under binomial sampling, viz., diag$[\hat{p}_1(1 - \hat{p}_1)/(nw_1), \ldots, \hat{p}_I(1 - \hat{p}_I)/(nw_I)]$ in the context of survey data. The matrix M arises naturally in solving likelihood equations (4) by iteratively reweighted least squares, and small values of m_{ii} call attention to extreme points in the factor space. Again, an index plot (m_{ii} vs i) would provide a useful display. It may be noted that the design effect does not come into picture with m_{ii} since we are using "pseudo m.l.e." based on binomial sampling. Another useful plot which effectively summarizes the information in the index plots $\tilde{X}_i$ vs i and m_{ii} vs i is given by the scatter plot of $\tilde{X}_i^2/X_c^2 = X_i^2/X^2$ vs h_{ii}, where h_{ii} is the i-th diagonal element of H given by (26) (see Pregibon, 1981).

The diagnostic measures e_i, $\tilde{X}_i$ or $\tilde{G}_i$ and m_{ii} are useful for detecting extreme points, but not for assessing their impact on various aspects of the fit including parameter estimates, $\hat{\underset{\sim}{\beta}}$, fitted values, $\hat{\underset{\sim}{f}}$, and goodness-of-fit measures $X^2/\delta_.$ or $G^2/\delta_.$ or others. Following Pregibon (1981) we suggest three measures which quantify the effect of extreme cells (points) on the fit.

(1) <u>Coefficient sensitivity</u>: Let $\hat{\beta}_j(-\ell)$ denote the pseudo m.l.e. of β_j obtained after deleting the ℓ-th cell data. Then the quantity $\Delta_j(\ell) = [\hat{\beta}_j - \hat{\beta}_j(-\ell)]/\text{s.e.}(\hat{\beta}_j)$ provides a measure of the j-th coefficient sensitivity to ℓ-th point. The index plots $\Delta_j(\ell)$ vs ℓ for each j provide useful displays but the task of looking at the index plots could become unmanageable if the number of coefficients in the model is large.

(2) <u>Sensitivity of fitted values</u>: Significant changes in coefficient estimates when ℓ-th point (cell) deleted does not necessarily imply that the fitted values $\hat{\underset{\sim}{f}}$ also vary significantly from $\hat{\underset{\sim}{f}}(-\ell)$, the vector of fitted values obtained after deleting the ℓ-th cell, i.e., $\|\hat{\underset{\sim}{f}} - \hat{\underset{\sim}{f}}(-\ell)\|$ could be small. We therefore use $[G^2 - \tilde{G}^2(-\ell)]/\delta_.$ or $[X^2 - \tilde{X}^2(-\ell)]/\delta_.$ to assess the impact of the ℓ-th point on the fitted values, where $\tilde{G}^2(-\ell)$ and $\tilde{X}^2(-\ell)$ are given by (9) and (8) respectively when $\hat{f}_i = f_i(\hat{\underset{\sim}{\beta}})$ is replaced by $\hat{f}_i(-\ell) = f_i(\beta(-\ell))$.

(3) <u>Goodness-of-fit</u>: A measure of goodness-of-fit sensitivity is given by $[G^2 - G^2(-\ell)]/\delta_.$ or $[X^2 - X^2(-\ell)]\delta_.$, where $G^2(-\ell)$ and $X^2(-\ell)$ are the likelihood ratio and chisquared statistics obtained after deleting the ℓ-th cell. (Note that $G^2(-\ell) \neq \tilde{G}^2(-\ell)$).

3. APPLICATION TO LFS

We have applied the previous methods to some data from the October 1980 Canadian Labour Force Survey (LFS). The sample consisted of males aged 15-64 who were in the labour force and not full-time students. We have chosen two factors, age and education, to explain the variation in unemployment rates via logit models. Age-group levels were formed by dividing the interval [15, 64] into ten groups with the j-th age group being the interval [10 + 5j, 14 + 5j], j = 1, 2, ..., 10, and then using the mid-point of each interval, A_j, as the value of the age for all persons in that age group. Similarly, the levels of education. E_k. were formed by assigning to each person a value based on the median years of schooling resulting in the following six levels = 7, 10, 12, 13, 14 and 16. Thus the age by education cross-classification provided a two-way table of I = 60 cell proportions, π_{jk}.

The LFS design employed stratified multi-stage cluster sampling with two stages in the self-representing (SR) urban areas and three or four stages in non-self-representing (NSR) areas in each province. The survey estimates, $\hat{p}_{jk}$, were adjusted for post-stratification, using the projected census age-sex distribution at the provincial level. The estimated covariance matrix $\hat{V}$ of the estimates $\hat{p}_{jk}$ is based on more than 450 first-stage units (psu's) so that the degrees of freedom for $\hat{V}$ are large compared to I = 60.

3.1 Formal Tests of Hypotheses.

Scatter plot of the logits $\hat{\nu}_{jk}$ vs age levels A_j at each education level E_k indicated that $\hat{\nu}_{jk}$ for given k generally increases with age to a maximum and then decreases (i.e., the graph is convex and upward to a maximum). Hence, the following model might be suitable to explain the variation in π_{jk}'s:

$$\nu_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k + \beta_4 E_k^2,$$

$$j = 1, \ldots, 10; \; k = 1, \ldots, 6. \tag{27}$$

Some previous work in sociological literature also supports such a model (Bloch and Smith, 1977). Applying the results of Section 2 we obtained the following values for goodness-of-fit statistics

$$X^2 = 98.9 \qquad G^2 = 101.2$$
$$X^2/\delta_. = 52.5 \qquad G^2/\delta_. = 53.7, \quad \delta_. = 1.88.$$

Since X^2 or G^2 is larger than $\chi^2_{0.05}(55) = 73.3$, the upper 5% point of χ^2 with I - s = 55 d.f., we would reject the model if the survey design is ignored. On the other hand, the value of $X^2/\delta_.$ or $G^2/\delta_.$ indicate that the model is adequate, the significance level (or P-value) being approximately equal to 0.52. The value of X_S^2 when adjusted to refer to $\chi^2_{0.05}(55)$ is equal to 47.7 which is also not significant. Moreover, in the present context with s(= 5) relatively small compared to I(= 60), the simple correction $\bar{d}$, the average cell deff, (see Fellegi, 1980), is very close to $\bar{\delta}$: $\bar{d}$ = 1.905 compared to $\bar{\delta}$ = 1.88: see Rao and Scott (1984) for a theoretical explanation.

The Wald statistic X_W^2 is not defined here since two of the cells have $\hat{p}_{jk}$ = 1, but we made minor perturbations to the estimated counts to ensure that $\hat{p}_{jk} < 1$ for all cells and then computed X_W^2. The resulting values of X_W^2 are all large compared to $X^2/\delta_.$ (at least 30 times larger than $X^2/\delta_.$) and vary considerably (1715 to 3061). Hence, the Wald statistic is very unstable for goodness-of-fit test in the present context. If the two cells having $\hat{p}_{jk} = 1$ are deleted, then $X_W^2 = 68.4 < \chi^2_{0.05}(53) = 71.0$, indicating that the model (27) is adequate. However, it is not a good practice to delete cells just to accomodate a chosen test statistic. The other problem with X_W^2, noted by Fay (1983), does not arise here since d.f. for $\hat{V}$ is large compared to the number of cells in the table.

The pseudo m.l.e., their s.e. and the corresponding s.e. under binomial sampling, all obtained under the model (27), are given in Table 1 along with Wald statistic $X_W^2(2|1)$ and G^2 statistic $G^2(2/1)/\delta_.(H)$ for the hypotheses H_i: $\beta_i = 0$, i=1, 2, 3, 4 given the model (27). As expected, the true s.e.'s are larger than the corresponding binomial s.e.'s. The hypothesis H_4: $\beta_4 = 0$ (i.e., coefficient of E_i^2 is zero) is not rejected at the 5% level either by the Wald statistic or G^2 statistic. On the other hand, the coefficient, β_2, of A_i^2 is highly significant. In testing the significance of individual coefficients we compare the values of $X_W^2(2|1)$ or $G^2(2|1)/\delta_.(H)$ to $\chi^2_{0.05}(1)$ = 3.84, the upper 5% point of χ^2 - variate with 1 d.f.

We have also tested the following nested hypotheses given model (27): H_{34}: $\beta_3 = \beta_4 = 0$ (i.e., no education effect); H_{24}: $\beta_2 = \beta_4 = 0$ (i.e., no quadratic effects). Both H_{34} and H_{24} are highly significant:

$$G^2(2|1)/\delta_.(H_{34}) = 282.2/1.64 = 172.1,\ X_W^2(2|1) = 165.6 \text{ for } H_{34}:$$

$$G^2(2|1)/\delta_.(H_{24}) = 242.2/2.28 = 106.3,\ X_W^2(2|1) = 162.1 \text{ for } H_{24} \text{ compared to}$$
$\chi^2_{0.05}(2) = 5.99$.

Table 1: Pseudo m.l.e. $\hat{\beta}_i$, s.e. $(\hat{\beta}_i)$, $X_W^2(2|1) = \hat{\beta}_i^2/\text{var}(\hat{\beta}_i)$ and $G^2(2|1)/\delta_.(H_i)$ Values for the LFS Data under Model (27).

	$\hat{\beta}_i$	s.e.$(\hat{\beta}_i)$ True	s.e.$(\hat{\beta}_i)$ Binomial	$X_W^2(2\|1)$	$G^2(2\|1)/\delta_.(H_i)$
0	-2.76	0.557		24.6	
1	0.209	0.0132	0.012	250.6	168.4
2	-0.00217	0.000173	0.000136	157.3	102.1
3	0.0913	0.0891	0.068	1.04	1.01
4	0.00276	0.00411	0.0030	0.45	0.46

Unlike in the case of goodness-of-fit, the Wald statistics is stable for testing nested hypotheses and leads to values close to the corresponding $G^2(2|1)/\delta_.(H)$ values.

By the above test of goodness-of-fit and tests of nested hypotheses we have arrived at the following simple model involving only four parameters:

$$\nu_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, \tag{28}$$

with $\hat{\beta}_0 = -3.10$, $\hat{\beta}_1 = 0.211$, $\hat{\beta}_2 = -0.00218$ and $\hat{\beta}_3 = 0.1509$ and corresponding standard errors are 0.247, 0.0130, 0.000172, and 0.0115. We will use the model (28) in Section 3.2 to develop logistic regression diagnostics.

3.2 Diagnostics

We now illustrate the use of diagnostics developed in Section 2.4.

(i) Residual Analysis

The 60 cells in the two-way table were numbered lexicographically, and the standardized residuals e_i were computed under the model (28) arrived through formal testing of hypotheses. Among the sixty e_i, cells numbered 6 and 54 with $\hat{p}_{jk} = 1$ lead to very large e_i values: 166.6 and 6.2 respectively. Among the remaining e_i, the residuals numbers 7, 27 and 59 have values 3.84, 2.73 and 2.52 respectively, whereas the expected number of $|e_i|$ exceeding 2.33 under model (28) is roughly $0.02 \times 60 = 1.2$. Hence, there is some indication that cells 7 and 27 could correspond to outlying cell proportions.

The normal probability plot of $\tilde{G}_i$ is displayed in FIG. 1; the plot of $\tilde{X}_i$ is not given to save space since it is similar to the plot of $\tilde{G}_i$. Figure 1 indicates no strong deviations from a straight line configuration. The index plot of $\tilde{G}_i$, Figure 2, is consistent with Figure 1. Hence, there is no evidence of outlying cell proportions when the components $\tilde{G}_i$ of G_c^2 are used for residual analysis.

(ii) Detection of Influential Cells.

The index plot of m_{ii} is displayed in Figure 3 which clearly points to cells 1 and 6. Figure 4 displays the plot of $\tilde{X}_i^2/X_c^2 = X_i^2/X^2$ vs h_{ii}, where the line with slope -1 is given by $X_i^2/X^2 + h_{ii} = 3\text{ave}(h^*_{ii})$. Here $h^*_{ii} = h_{ii} + X_i^2/X^2$, and the values of h^*_{ii} near unity corresponds to cells which are outlying or influential or both (Pregibon, 1981) and appear above the line in Figure 3. It is clear that cells 1 and 6, and to a lesser extent cells 7 and 58, warrant further examination.

(iii) Coefficient Sensitivity.

The index plots for measuring coefficient sensitivity ($\Delta_j(\ell)$ vs ℓ) are displayed in Figures 5, 6, 7, and 8 for β_0, β_1, β_2 and β_3 respectively. It is clear from the plots that cells 2 and 3 cause instability in $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, whereas $\hat{\beta}_3$ is affected by cell 7.

(iv) Sensitivity of Fitted Values

Figure 9 displays the plot of $[G^2 - \tilde{G}^2(-\ell)]/\delta_. = c$ vs ℓ for assessing the impact of individual cells on fitted values. Significant peaks in this figure correspond to cells 2 and 3 and to a lesser extent to cell 7. Following Cook (1977) and Pregibon (1981), it may be noted that the comparison of c to the percentage point of $\chi^2(s)$ ($s = 4$ in model (28)) gives a rough guide as to which contour of the confidence region the pseudo m.l.e. is displaced due to deletion of the ℓ-th cell. The value $c = 2.1$ for cell 2 roughly corresponds to 78% contour of the confidence region.

(v) Goodness-of-fit Sensitivity

Figure 10 displays the plot of $[G^2 - G^2(-\ell)]/\delta_.$ vs ℓ: the plot of $[X^2 - X^2(-\ell)]/\delta_.$ is similar and hence not displayed but the former plot is preferred (Pregibon, 1981). Significant peaks in this figure corresponds to cells 2, 3, 7, 27, 39 and 54 (values $\geq$ 3), the most significant being cell 7 with the value 5.4. By deleting cell 7 and recomputing the adjusted statistic $G_c^2(-\ell) = G^2(-\ell)/\delta_.(-\ell)$ where $\delta_.(-\ell)$ is the corresponding value of $\delta_.$, we get a value of 48.43 with 55 d.f. compared to $G^2/\delta_. = 55.3$ with 56 d.f.

Our investigation on the whole indicated that cells 7, 2 and 3 are possible candidates for deletion, but we feel that their impact is not significant enough to warrant their deletion - one would like to explain the variation among all cell proportions unless certain cells contribute heavily to the disagreement between the data and the fitted model.

ACKNOWLEDGEMENT

We wish to thank M. Gratton of Statistics Canada for producing the graphs included in the paper.

REFERENCES

[1] Bloch, F.E., and Smith, S.P. (1977). Human capital and labour market employment. J. Human Resources, 12, pp. 550-559.

[2] Cook, R.D. (1977). Detection of influential observations in linear regression. J. American Statistical Association, 72, pp. 169-174.

[3] Cook, R.D., and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, London.

[4] Cox, D.R. (1970). Analysis of Binary Data. Chapman and Hall, London.

[5] Fay. R.E. (1983). Replication approaches to the log-linear analysis of data from complex samples. Unpublished manuscript (courtesy of the author).

[6] Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. J. American Statistical Association, 75, pp. 261-268.

[7] McCullagh, P., and Nelder, J.A. (1983). Generalized Linear Models. Chapman and Hall, London.

[8] Pregibon, D. (1981). Logistics regression diagnostics. Ann. Statist., 9, pp. 705-724.

[9] Rao, J.N.K., and Scott, A.J. (1984). On simple adjustments to chisquared tests with survey data: log-linear and logit models. Unpublished manuscript.

[10] Roberts, G. (1984). On chi-squared tests for logit models with cell proportions estimated from survey data. Unpublished manuscript. Carleton University.

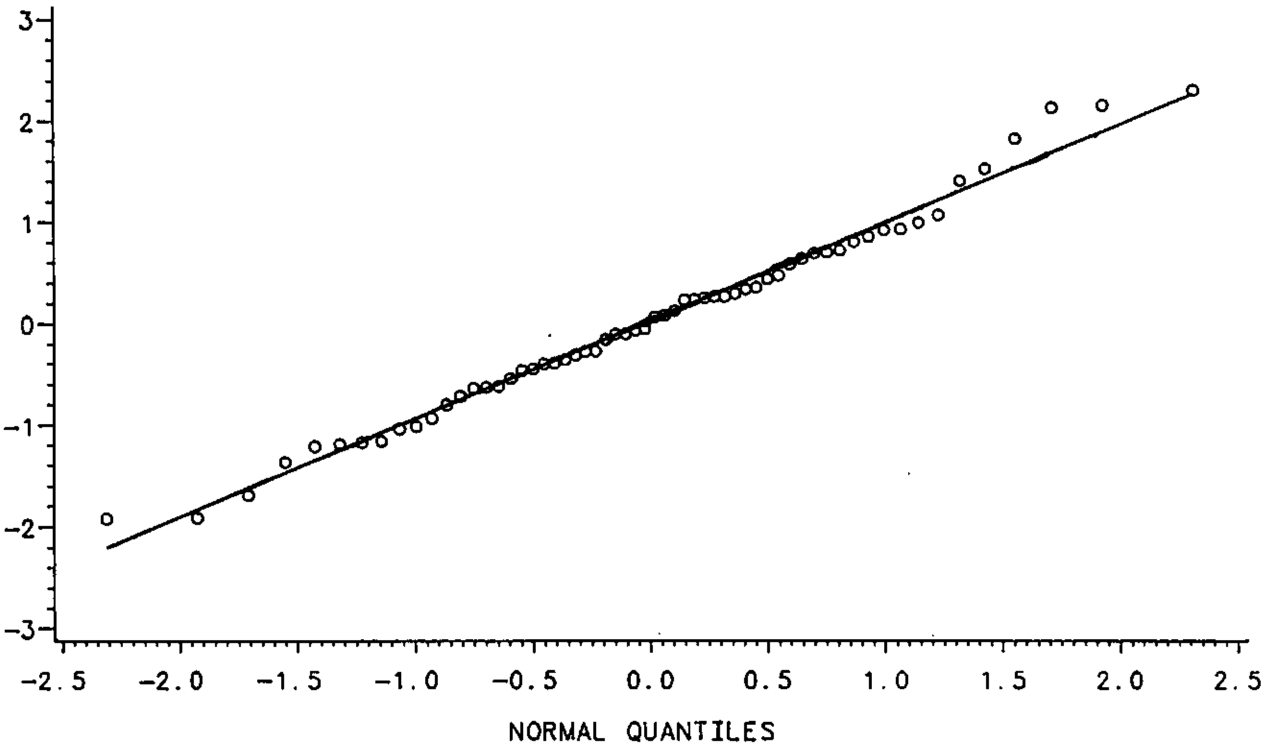

Figure 1: Normal Probability Plot of $\tilde{G}_i$

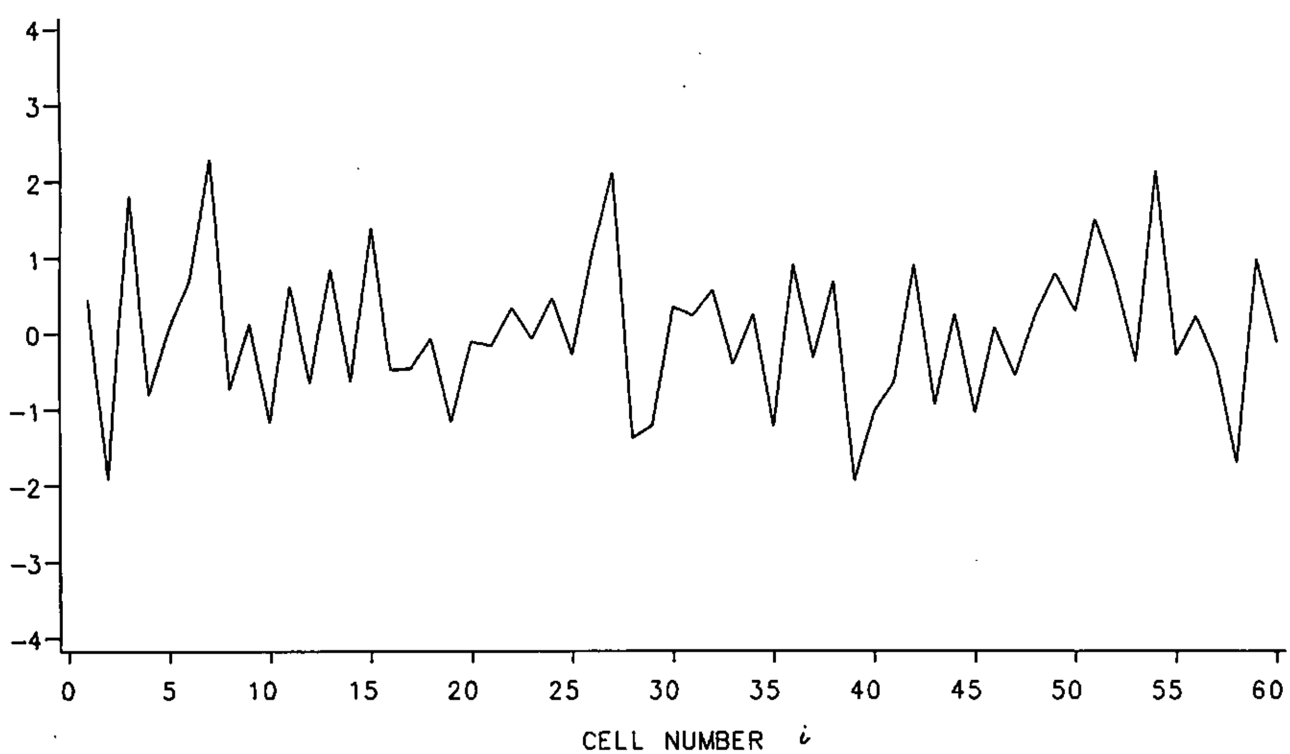

Figure 2: Index Plot of $\tilde{G}_i$

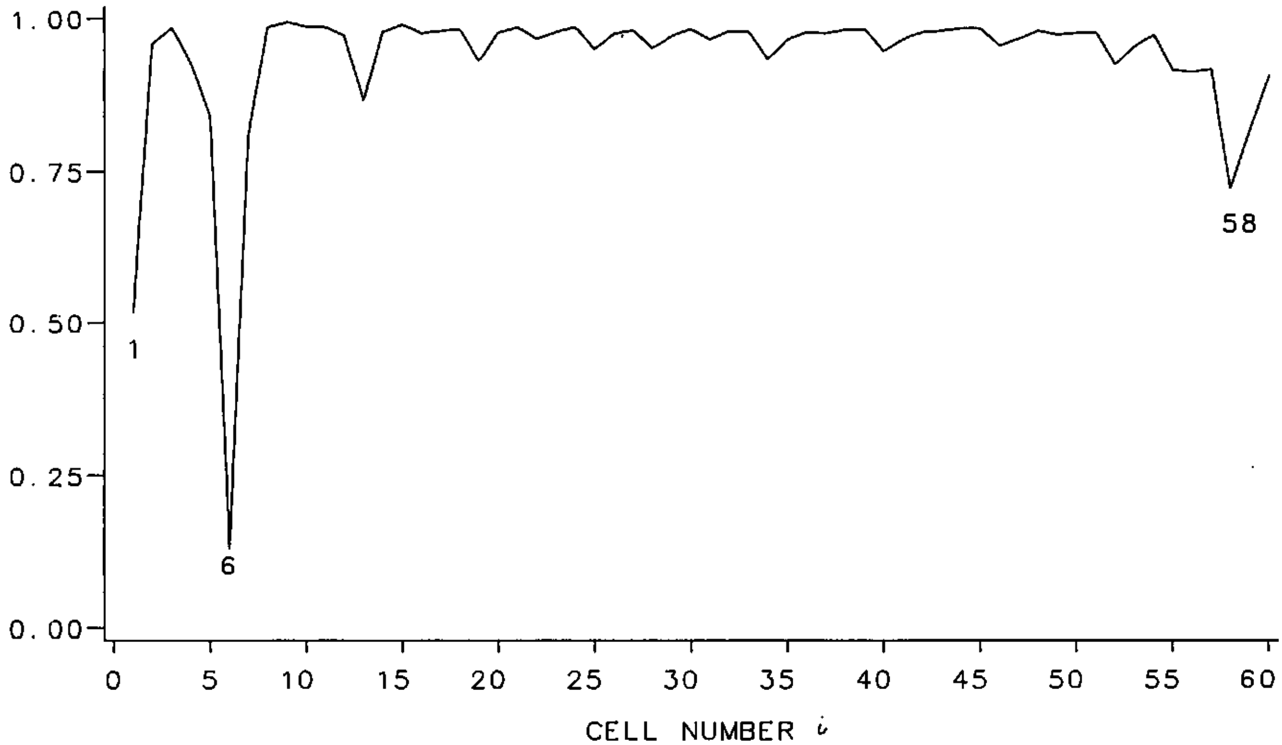

Figure 3: Index Plot of m_{ii}

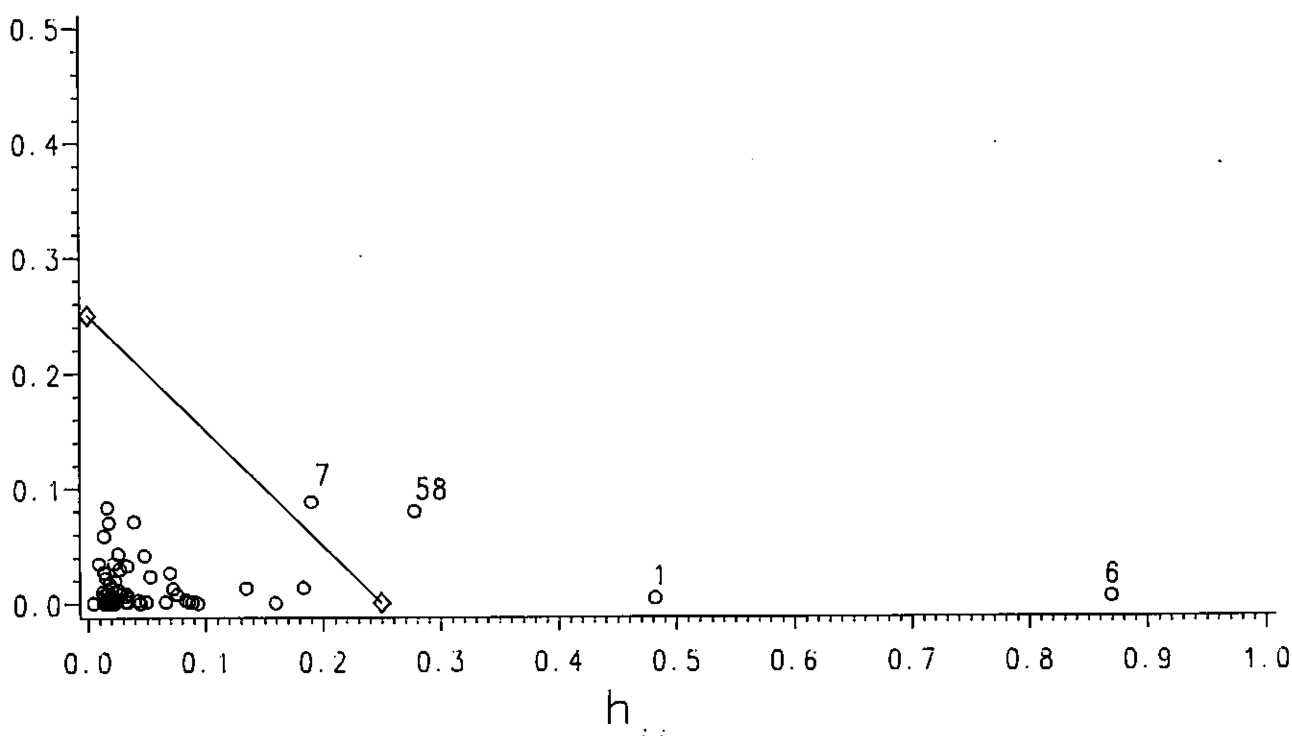

Figure 4: Scatter Plot of χ_i^2/χ^2 vs. h_{ii}

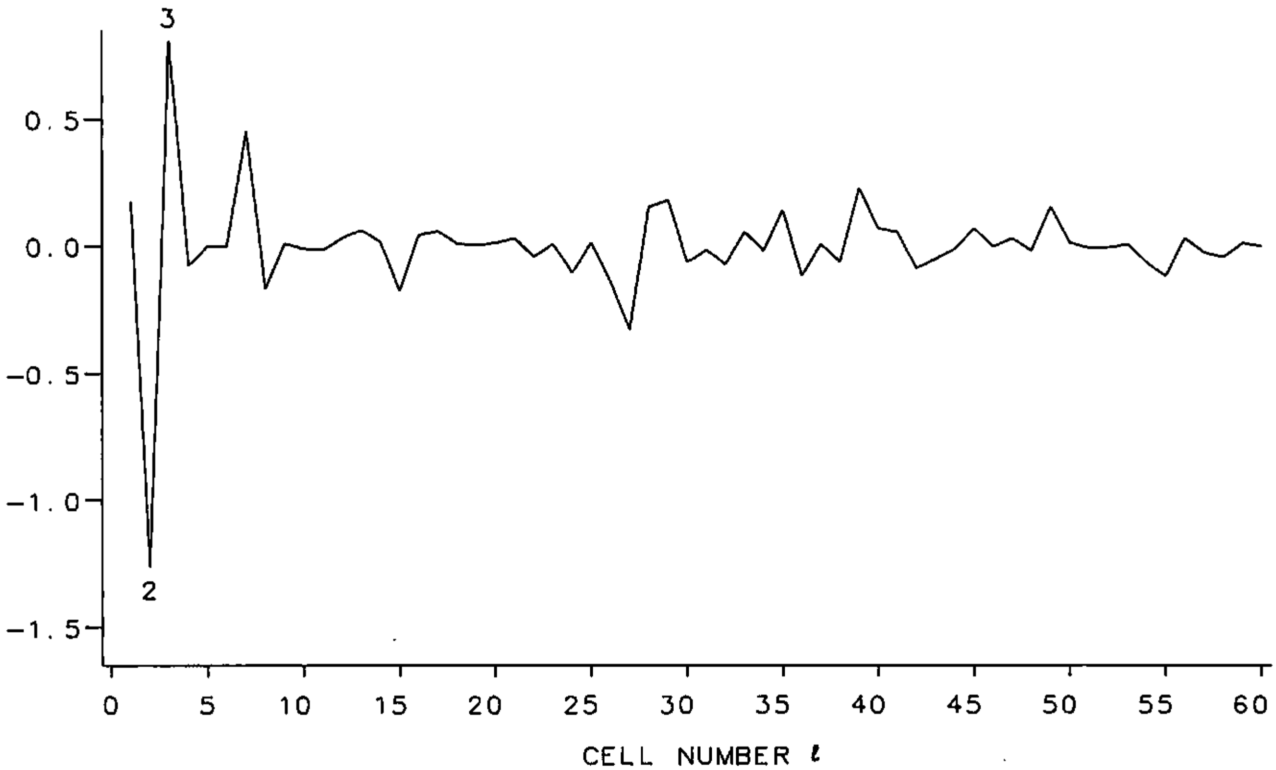

Figure 5: Index Plot of $\{\hat{\beta}_0 - \hat{\beta}_0(-\ell)\}/\text{s.e.}(\hat{\beta}_0)$

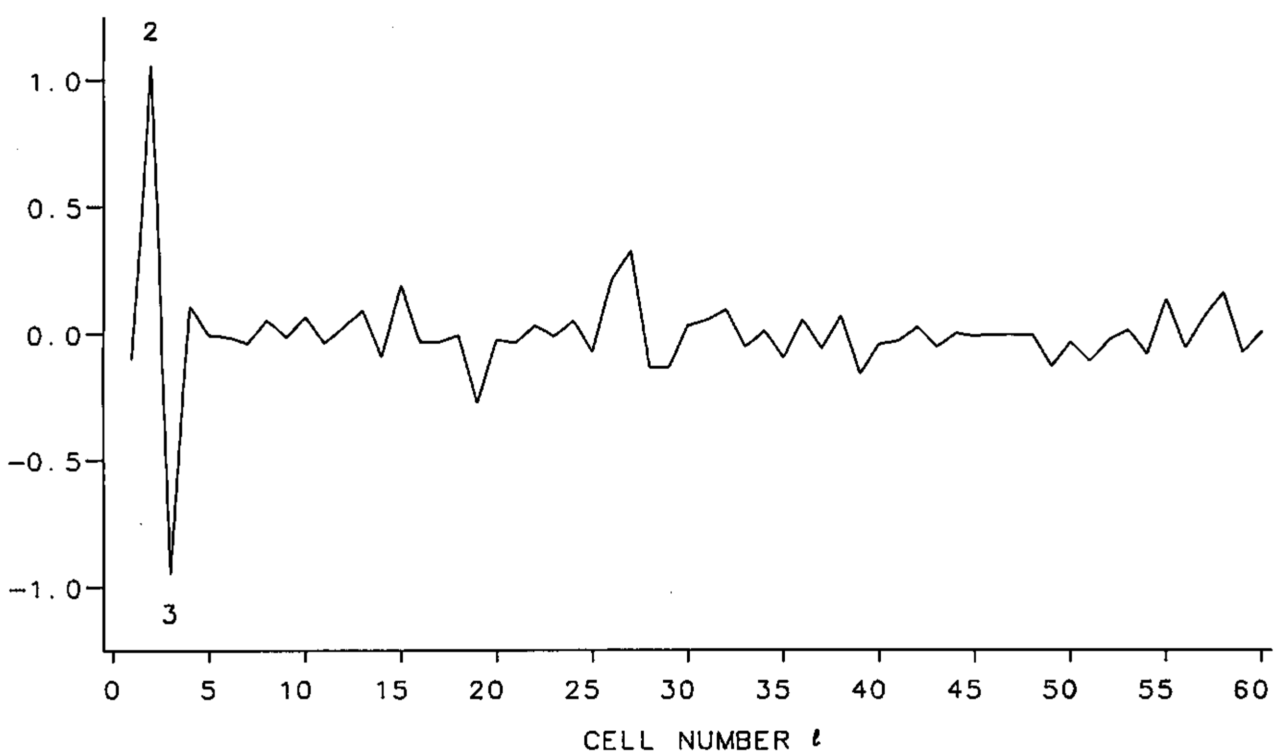

Figure 6: Index Plot of $\{\hat{\beta}_1 - \hat{\beta}_1(-\ell)\}/\text{s.e.}(\hat{\beta}_1)$

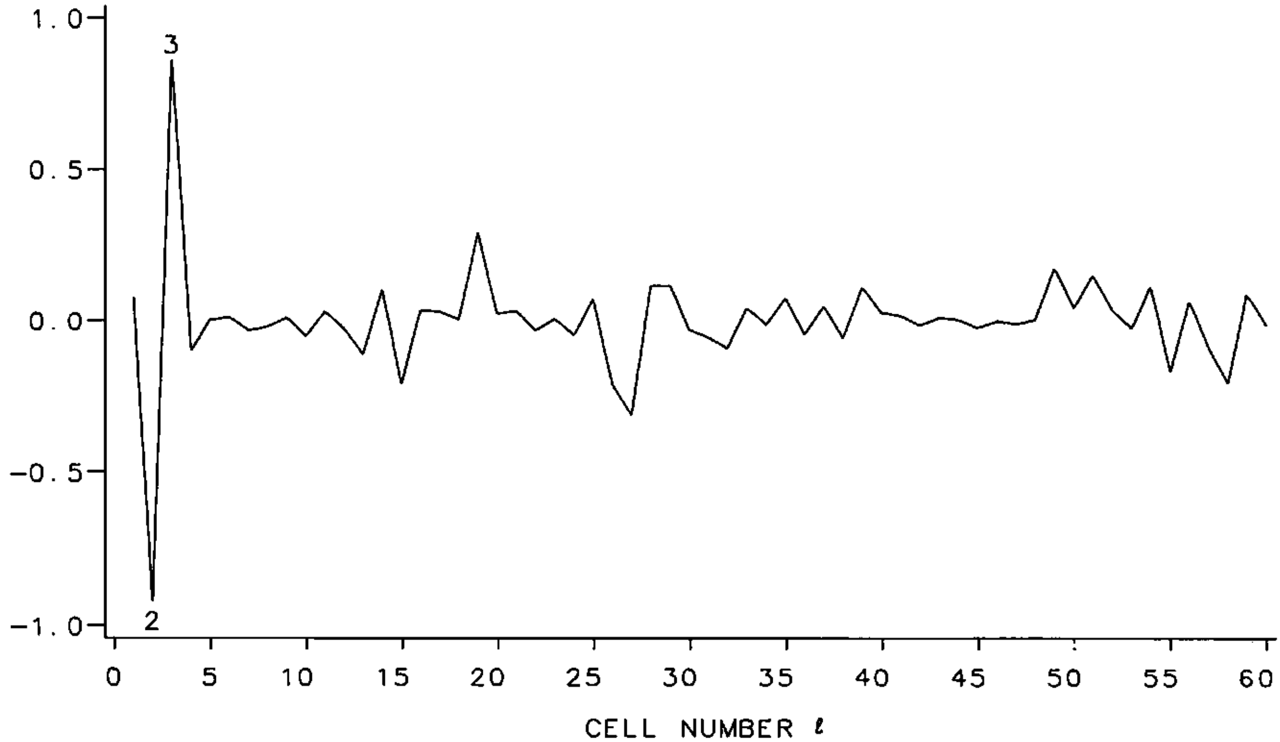

Figure 7: Index Plot of $\{\hat{\beta}_2 - \hat{\beta}_2(-\ell)\}/\mathrm{s.e.}(\hat{\beta}_2)$

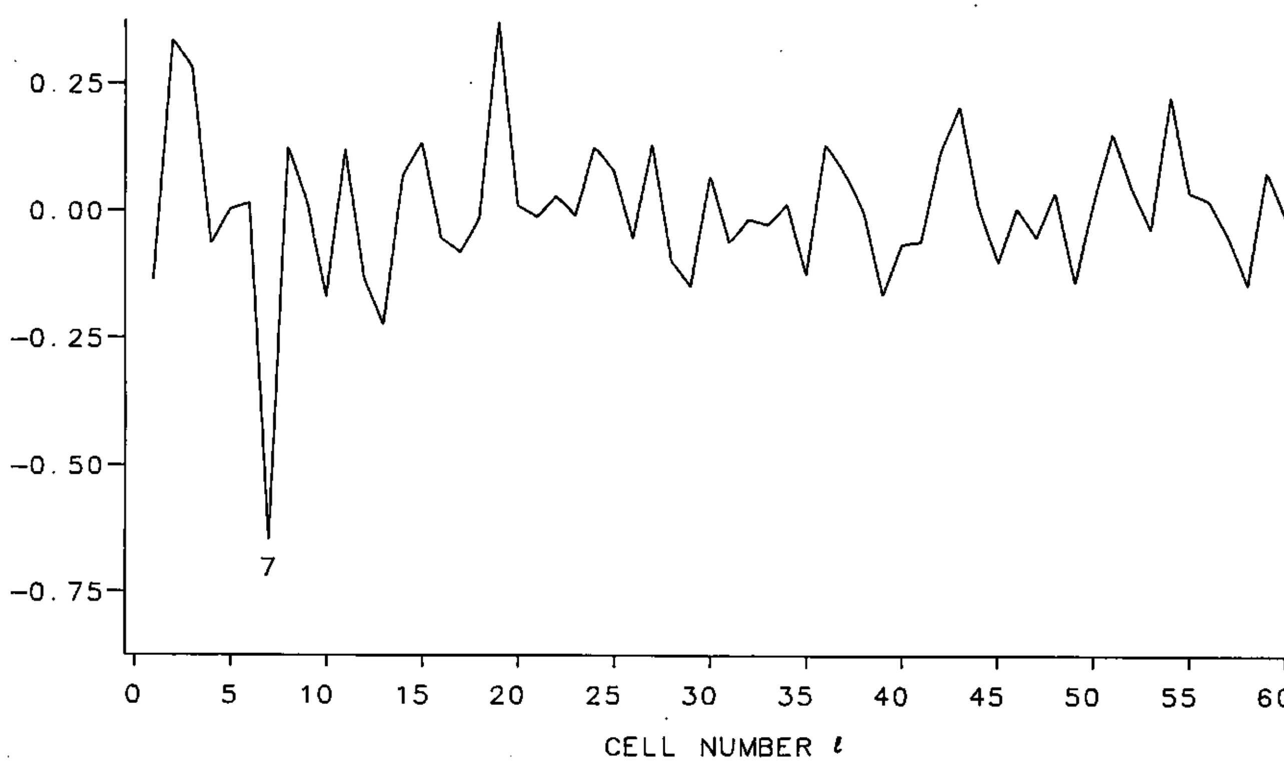

Figure 8: Index Plot of $\{\hat{\beta}_3 - \hat{\beta}_3(-\ell)\}/\mathrm{s.e.}(\hat{\beta}_3)$

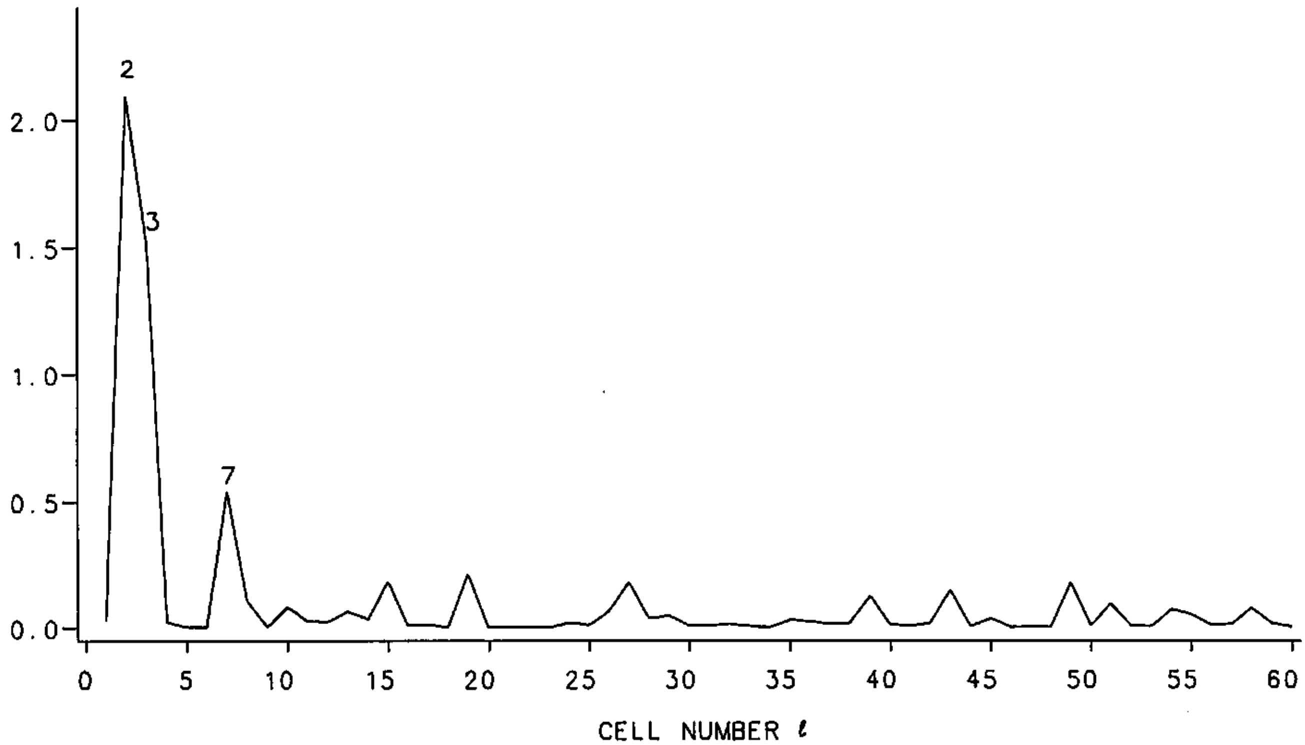

Figure 9: Index Plot of $\{G^2 - \tilde{G}^2(-\ell)\}/\hat{\hat{\delta}}$

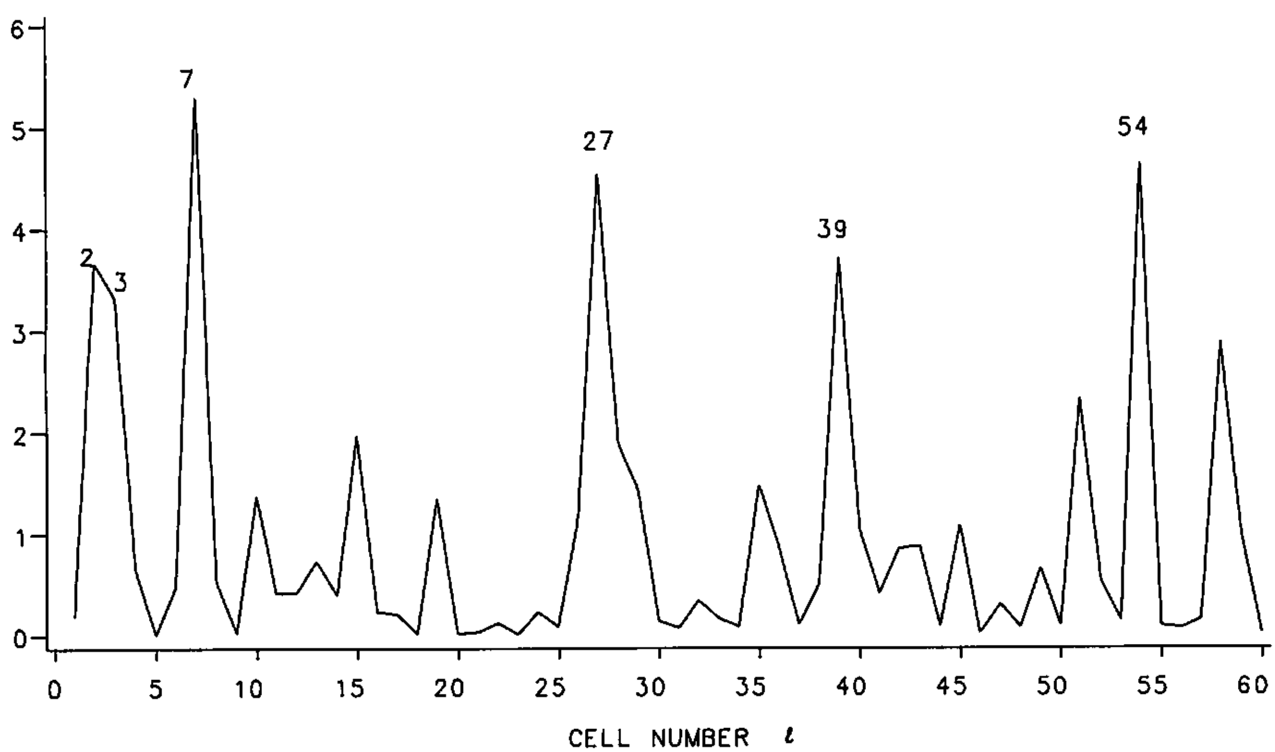

Figure 10: Index Plot of $\{G^2 - G^2(-\ell)\}/\hat{\hat{\delta}}$

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

$$\hat{X}_0 = \sum_i W_{i0} X_i. \qquad (4.1)$$

The product of the replicate weighting approach is a set of additional weights W_{ir}, $r = 1, \ldots, R$, 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 $\hat{X}_0^{(1)}, \ldots, \hat{X}_0^{(k)}$, 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, 10: 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.

LEAST SQUARES AND RELATED ANALYSES FOR COMPLEX SURVEY DESIGNS

Wayne A. Fuller[1]

1. INTRODUCTION AND MODEL

Assume that a sample of clusters of elemental units is selected from a finite population divided into L strata. The total sample of n clusters (primary sampling units) is given by

$$n = \sum_{h=1}^{L} n_h. \tag{1}$$

where $n_h \geq 2$ is the number of clusters selected in the h-th stratum. A column vector of characteristics

$$\underset{\sim}{Y}_{hij} = (Y_{hij1}, Y_{hij2}, \ldots, Y_{hijp})' \tag{2}$$

is observed for the j-th elemental unit in the i-th cluster of the h-th stratum. The vector $\underset{\sim}{Y}_{hij}$ is quite general. For example, some elements of the vector can be the powers of products of other entries. Also, one element can be, and often will be, identically equal to one. The cluster totals for the vector are defined by

$$\underset{\sim}{Y}_{hi.} = \sum_{j=1}^{m_{hi}} \underset{\sim}{Y}_{hij}. \tag{3}$$

where m_{hi} is the number of elements in the hi-th cluster.

We shall be interested in the behavior of locally continuous functions of a linear function of the vector of cluster means

[1] Wayne A. Fuller, Department of Statistics, Iowa State University.

$$\hat{\underset{\sim}{\theta}} = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{i=1}^{n_h} \underset{\sim}{Y}_{hi.}, \tag{4}$$

where W_h are fixed weights. Often the weights are

$$W_h = N_h N^{-1}, \tag{5}$$

where N_h is the number of clusters in the h-th stratum and N is the total number of clusters in the population. For the weights (5) the linear function in (4) is the usual unbiased estimator of the finite population mean per cluster. Another set of weights that often is of interest is the set of unit weights

$$W_h = n^{-1} n_h. \tag{6}$$

Our model permits us to consider functions of the mean per element. The usual estimator of the mean per element for a particular Y-variable is the ratio of the mean per cluster for the Y-variable to the mean per cluster of the number of elements. The mean number of elements per cluster is the cluster mean of a Y-variable that is identically one.

Our discussion can be easily expanded to include various forms of subsampling within clusters. Because such expansions add little to the generality of the discussion and add considerable notational complexity, we restrict our attention to single stage sampling within strata.

Our discussion rests heavily on the following central limit theorem for samples from a finite population.

Theorem 1. Let $\{\xi_r: r = 1, 2, \ldots\}$ be a sequence of stratified finite populations. Let the population in the h-th stratum of the r-th population be a random sample of size $N_{rh} \geq N_{r-1,h}$ selected from a p dimensional infinite population with absolute $2 + \delta$, where $\delta > 0$, moments bounded by $M_\delta < \infty$. Let the covariance matrix for the rh-th infinite population be $\underset{\sim}{\Sigma}_{rh}$. Let $L_r \geq L_{r-1}$ be the number of strata in the finite population and let a simple random sample of n_{rh} ($n_{rh} \geq 2$ and $n_{rh} \geq n_{r-1,h}$) units be selected in the h-th stratum. Let $f_{rh} = N_{rh}^{-1} n_{rh}$ be a triangular array such that

$$0 \leq f_{rh} < M_{fu} < 1,$$

where M_{fu} is a fixed number. Let $\underset{\sim}{Y}_{rhi.}$ be the total for the i-th cluster selected in the h-th stratum for the r-th population and let

$$\hat{\underset{\sim}{\theta}}_r = \sum_{h=1}^{L_r} W_{rh} n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \underset{\sim}{Y}_{rhi.},$$

$$\underset{\sim}{\theta}_{rf} = \sum_{h=1}^{L_r} W_{rh} N_{rh}^{-1} \sum_{i=1}^{N_{rh}} \underset{\sim}{Y}_{rhi.},$$

$$\underset{\sim}{\theta}_r = \sum_{h=1}^{L_r} W_{rh} \underset{\sim}{\mu}_{.h..},$$

where $\underset{\sim}{\theta}_{rf}$ is the finite population parameter and $\underset{\sim}{\mu}_{.h..}$ is the mean of the infinite population used to generate the h-th stratum of the finite population. Assume

$$0 < M_{SL} < \left| n_r \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \underset{\sim}{\Sigma}_{rh} \right| < M_{SU} < \infty,$$

where the M's are fixed numbers and assume that

$$n_r = \sum_{h=1}^{L_r} n_{rh} \longrightarrow \infty,$$

$$\sup_h \left[\sum_{t=1}^{L_r} W_{rt}^2 n_{rt}^{-1} \right]^{-1} W_{rh}^2 n_{rh}^{-2} \longrightarrow 0,$$

as $r \to \infty$, where W_{rh} is a triangular array of weights. Then

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\}]^{-\frac{1}{2}}(\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{r}\}]^{-\frac{1}{2}}(\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{r}) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

where

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\} = \sum_{h=1}^{L_r} W_{rh}^2(1 - f_{rh})n_{rh}^{-1}\hat{\underset{\sim}{\Sigma}}_{rh},$$

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{r}\} = \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1}\hat{\underset{\sim}{\Sigma}}_{rh},$$

$$\hat{\underset{\sim}{\Sigma}}_{rh} = (n_{rh} - 1)^{-1}\sum_{i=1}^{n_{rh}} (\underset{\sim}{Y}_{rhi.} - \bar{\underset{\sim}{Y}}_{rh..})(\underset{\sim}{Y}_{rhi.} - \bar{\underset{\sim}{Y}}_{rh..})',$$

$$\bar{\underset{\sim}{Y}}_{rhi.} = n_{rh}^{-1}\sum_{i=1}^{n_{rh}} \underset{\sim}{Y}_{rhi.}.$$

The proof of this theorem follows from Theorems 1 and 2 of Fuller (1975) and can be extended to multistage samples. Also see Krewski and Rao (1981) and Isaki and Fuller (1982).

Most of our applications are to continuous functions of $\hat{\underset{\sim}{\theta}}$.

Corollary 1. Let the assumptions of Theorem 1 hold. Let $\underset{\sim}{g}(\underset{\sim}{\theta})$ be a vector valued function of $\underset{\sim}{\theta}$, where $\underset{\sim}{g}(\underset{\sim}{\theta})$ is continuous with continuous first derivatives for $\underset{\sim}{\theta}$ in the sphere $|\underset{\sim}{\theta} - \underset{\sim}{\theta}_r| \leq \delta$ for all r, where $\delta > 0$ is fixed. Let $\underset{\sim}{G}(\underset{\sim}{\theta})$ be the nonsingular matrix of first derivatives of $\underset{\sim}{g}(\underset{\sim}{\theta})$, where the ij-th element of $\underset{\sim}{G}(\underset{\sim}{\theta})$ is

$$\frac{\partial g_i(\underset{\sim}{\theta})}{\partial \theta_j},$$

$g_i(\underset{\sim}{\theta})$ is the i-th element of $\underset{\sim}{g}(\underset{\sim}{\theta})$ and θ_j is the j-th element of $\underset{\sim}{\theta}$. Then

$$[\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_r)\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\}\underset{\sim}{G}'(\hat{\underset{\sim}{\theta}}_r)]^{-\frac{1}{2}}[\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_r) - \underset{\sim}{g}(\underset{\sim}{\theta}_{rf})] \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

$$[\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_r)\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}\underset{\sim}{G}'(\hat{\underset{\sim}{\theta}}_r)]^{-\frac{1}{2}}[\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_r) - \underset{\sim}{g}(\underset{\sim}{\theta}_r)] \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}).$$

Corollary 1 is stated for the Taylor estimator of the variance of the approximate distribution of $\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_r) - \underset{\sim}{g}(\underset{\sim}{\theta}_r)$. Suitably defined replication estimators of the variance can also be used. Replication methods include balanced replication methods (see McCarthy (1969)), jackknife methods (See Miller (1974)) and bootstrap methods (see Efron (1979, 1981)). While these methods can be adapted to the sampling situation, the adaptation is not always immediate (see Rao and Wu (1983)).

One class of continuous functions of $\hat{\underset{\sim}{\theta}}$ that deserves special attention is that obtained by using $\hat{\underset{\sim}{\theta}}$ as the dependent variable in a generalized least squares fit.

Corollary 2. Let the assumptions of Theorem 1 hold. Let $\underset{\sim}{\theta}$ satisfy

$$\underset{\sim}{\theta} = \underset{\sim}{h}(\underset{\sim}{\alpha}).$$

where $\underset{\sim}{\alpha}$ is a k-dimensional vector ($k \le p$), $\underset{\sim}{h}(\underset{\sim}{\alpha})$ is a continuous function of $\underset{\sim}{\alpha}$, with continuous first and second derivatives for all $\underset{\sim}{\alpha}$ in an open sphere containing the true $\underset{\sim}{\alpha}_r$ for all r. Let the parameter space for $\underset{\sim}{\alpha}$ be an open bounded subset of k-dimensional Euclidean space. Let $\hat{\underset{\sim}{\alpha}}_r$ be the vector that minimizes

$$[\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{h}(\underset{\sim}{\alpha}_r)]'\hat{\underset{\sim}{V}}^{-1}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}[\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{h}(\underset{\sim}{\alpha}_r)].$$

Then

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\alpha}}_r\}]^{-\frac{1}{2}}(\hat{\underset{\sim}{\alpha}}_r - \underset{\sim}{\alpha}_r) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

where

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\alpha}}_r\} = [\underset{\sim}{H}(\hat{\underset{\sim}{\alpha}}_r)\hat{\underset{\sim}{V}}^{-1}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}\underset{\sim}{H}'(\hat{\underset{\sim}{\alpha}}_r)]^{-1},$$

and $\underset{\sim}{H}(\hat{\underset{\sim}{\alpha}}_T)$ is the matrix of first derivatives of $\underset{\sim}{h}(\underset{\sim}{\alpha})$ with respect to $\underset{\sim}{\alpha}$ evaluated at $\hat{\underset{\sim}{\alpha}}$.

2. MEANS, RATIOS AND REGRESSIONS

An elementary application of Theorem 1 is the estimation of the mean per cluster and the setting of approximate confidence limits for the mean per cluster. Often the parameter of interest for the mean estimator is the finite population mean per cluster, in which case the finite population correction $(1 - f_h)$ would be included in the variance estimator.

A slightly more complex application is the estimation of the difference between the means per cluster for two domains. If we let

Y_{hij1} = observation on characteristic of interest if element hij is in domain 1
= 0 otherwise,

Y_{hij2} = observation on characteristic of interest if element hij is in domain 2
= 0 otherwise,

Y_{hij3} = 1 if element hij is in domain 1
= 0 otherwise,

Y_{hij4} = 1 if element hij is in domain 2
= 0 otherwise.

the estimated difference between the mean per element in the two domains is

$$\underset{\sim}{g}(\hat{\theta}) = g(\bar{\underset{\sim}{Y}}_{\cdots}) = \bar{Y}^{-1}_{\cdots 3}\bar{Y}_{\cdots 1} - \bar{Y}^{-1}_{\cdots 4}\bar{Y}_{\cdots 2}. \tag{7}$$

Two methods of computing the Taylor estimator of variance are often used. The first method computes the estimator of Corollary 1 directly from the matrices $\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_T)$ and $\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_T - \underset{\sim}{\theta}_T\}$ or $\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_T - \underset{\sim}{\theta}_{Tf}\}$. An algebraically identical computational procedure is to define the observations

$$\underset{\sim}{Z}(\underset{\sim}{Y}_{hi.}, \hat{\underset{\sim}{\theta}}) = \hat{\underset{\sim}{Z}}_{hi} = \underset{\sim}{G}(\hat{\underset{\sim}{\theta}})(\underset{\sim}{Y}_{hi.} - \bar{\underset{\sim}{Y}}_{h..}) \tag{8}$$

and to compute the ordinary stratified estimator of the variance of the mean per cluster for $\hat{\underset{\sim}{Z}}_{hi}$.

$$\hat{\underset{\sim}{V}}\{\hat{\bar{\underset{\sim}{Z}}}_{..}\} = \hat{\underset{\sim}{V}}\{\underset{\sim}{g}(\bar{\underset{\sim}{Y}}_{...})\}$$

$$= \sum_{h=1}^{L} W_h^2(1 - f_h)n_h^{-1}(n_h - 1)^{-1} \sum_{j=1}^{n_h}(\hat{\underset{\sim}{Z}}_{hi} - \hat{\bar{\underset{\sim}{Z}}}_{h.})(\hat{\underset{\sim}{Z}}_{hi} - \hat{\bar{\underset{\sim}{Z}}}_{h.})', \tag{9}$$

where

$$\hat{\bar{\underset{\sim}{Z}}}_{..} = \sum_{h=1}^{L} W_h \hat{\bar{\underset{\sim}{Z}}}_{h.},$$

$$\hat{\bar{\underset{\sim}{Z}}}_{h.} = n_h^{-1} \sum_{i=1}^{n_h} \hat{\underset{\sim}{Z}}_{hi}.$$

For example, the computational form (9) is used in Super Carp. See Hidiroglou et al. (1980, p. 32).

The analyst may be interested in inferences for the particular finite population sampled or for the superpopulation when working with quantities such as differences of means.

One of the more frequent analytic uses of survey data is the computation of regression equations. In fact, the difference between domain means can be expressed as a regression coefficient. Although the vector of regression coefficients is of the form $\underset{\sim}{g}(\hat{\underset{\sim}{\theta}})$ described in the previous section. it may be advantageous to partition the $\underset{\sim}{Y}$-vector of Section 1 into several parts and to give the regression coefficients explicit expressions. The regression equation can be written as

$$Y_{hij} = \underset{\sim}{X}'_{hij}\underset{\sim}{\beta} + e_{hij}, \tag{10}$$

where Y_{hij} is the dependent variable, the vector $\underset{\sim}{X}_{hij}$ is a k-dimensional vector of explanatory variables. The weighted least squares estimator of $\underset{\sim}{\beta}$ is

$$\hat{\underset{\sim}{\beta}}_W = \left[\sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} Y_{hij}. \tag{11}$$

The weights W_{hij} are permitted to be a function of hij, but we will assume that the weights are fixed in the sense that they depend only on the elemental identification. This precludes from consideration (except as an approximation) the use of weights that are a function of other elements entering the sample.

Under mild assumptions on the moments of the superpopulation generating the finite population, Theorem 1 is applicable to the estimator defined in (11). If the selection probabilities are denoted by π_{hij}, then the estimator $\hat{\underset{\sim}{\beta}}_W$ is a consistent estimator of the finite population vector

$$\underset{\sim}{B}_f = \left[\sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \pi_{hij} \underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \pi_{hij} Y_{hij}. \tag{12}$$

It follows from (12) that the estimator (11) is a consistent estimator of the finite population regression coefficient when W_{hij} is proportional to the inverse of the selection probabilities. The error in $\hat{\underset{\sim}{\beta}}_W$ as an estimator of $\underset{\sim}{B}_f$ is

$$\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{B}_f = \left[\sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} v_{hij}, \tag{13}$$

where

$$v_{hij} = Y_{hij} - \underset{\sim}{X}'_{hij} \underset{\sim}{B}_f.$$

By Theorem 1 and Corollary 1 a consistent estimator of the variance of the approximate distribution of $\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{\beta}$ is

$$\hat{V}\{\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{\beta}\} = \hat{\underset{\sim}{A}}^{-1} \hat{\underset{\sim}{G}} \hat{\underset{\sim}{A}}^{-1}. \tag{14}$$

where

$$\hat{\underset{\sim}{A}} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij} .$$

$$\hat{\underset{\sim}{G}} = (n - 1)(n - k)^{-1} \sum_{h=1}^{L} n_h (n_h - 1)^{-1} \sum_{i=1}^{n_h} \hat{\underset{\sim}{d}}_{hi.} \hat{\underset{\sim}{d}}'_{hi.} ,$$

$$\hat{\underset{\sim}{d}}_{hi.} = \sum_{j=1}^{m_{hi}} \hat{\underset{\sim}{d}}_{hij} ,$$

$$\hat{d}_{hij} = W_{hij} \underset{\sim}{X}_{hij} \hat{v}_{hij} ,$$

$$n = \sum_{h=1}^{L} \sum_{i=1}^{n_h} m_{hi} ,$$

$$\hat{v}_{hij} = Y_{hij} - \underset{\sim}{X}'_{hij} \hat{\underset{\sim}{\beta}}_{W} ,$$

and $\underset{\sim}{\beta}$ is the superpopulation analog of $\underset{\sim}{B}_f$. This particular form of the estimator of variance was suggested by Fuller (1975) and is used in Super Carp.

One of the frequently asked questions faced by survey statisticians is: "In computing the regression equation, should I use the sampling weights?" As with most such questions, the answer is "It depends." The fact that the question is asked generally means that the questionner has in mind inference for a population beyond the finite population sampled. This does not mean that the particular superpopulation is completely defined or definable. It does suggest that the questionner is postulating that the finite population is generated by a superpopulation in which some type of linear model holds. One quantification of the hypothesis that weights are not required is the superpopulation hypothesis

$$H_0 : \underset{\sim}{\theta}_{\pi} = \underset{\sim}{\theta}_{(1)} . \tag{15}$$

where the $\underset{\sim}{\theta}$'s are superpopulation analogs of (12),

$$\underset{\sim}{\theta}_{\pi} = [\sum_{h=1}^{L} \sum_{i=1}^{N_h} E_{\xi} \sum_{j=1}^{m_{hi}} \{\underset{\sim}{X}_{hij} \pi_{hij} \underset{\sim}{X}'_{hij}\}]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} E_{\xi} \{\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} \pi_{hij} Y_{hij}\} ,$$

$$\underset{\sim}{\theta}_{(1)} = [\sum_{h=1}^{L} \sum_{i=1}^{N_h} E_\xi\{\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}\underset{\sim}{X}'_{hij}\}]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} E_\xi\{\sum_{j=1}^{m_h} \underset{\sim}{X}_{hij}Y_{hij}\}, \tag{16}$$

and E_ξ denotes expectation with respect to the superpopulation. This is a testable hypothesis. It seems that, at a minimum, a test of this hypothesis should be constructed if one performs an unweighted analysis of a sample with unequal selection probabilities.

If the null hypothesis also includes the hypothesis that the estimator with unit weights is the minimum variance estimator, then the test of the hypothesis is given by the statistic

$$F^k_{n-L-2k} = k^{-1}\ \hat{\underset{\sim}{\delta}}'_2\hat{\underset{\sim}{V}}^{-1}_{22}\hat{\underset{\sim}{\delta}}_2. \tag{17}$$

where

$$(\hat{\delta}'_1,\ \hat{\delta}'_2)' = [\sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{Z}_{hij}\underset{\sim}{Z}'_{hij}]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{Z}_{hij}Y_{hij}.$$

$$\underset{\sim}{Z}'_{hij} = (\underset{\sim}{X}'_{hij},\ \underset{\sim}{X}'_{hij}W_{hij}),$$

and

$$\hat{\underset{\sim}{V}} = \begin{pmatrix} \hat{\underset{\sim}{V}}_{11} & \hat{\underset{\sim}{V}}_{12} \\ \hat{\underset{\sim}{V}}_{21} & \hat{\underset{\sim}{V}}_{22} \end{pmatrix} \tag{18}$$

is defined by (14) with $\underset{\sim}{Z}_{hij}$ replacing $\underset{\sim}{X}_{hij}$. As the notation suggests, the statistic is approximately distributed as Snedecor's F with k and n - L - 2k degrees of freedom.

Example 1. Table 1 contains observations on 37 area segments collected by the Statistical Reporting Service, U.S. Department of Agriculture in northcentral Iowa in 1978. Two determinations on the hectares of soybeans are reported. The first is obtained by personal interview in the June Enumerative Survey. The second is obtained from a classification of Landsat data based upon a classifier developed by the Statistical Reporting Service. The original objective of the study was to use the Landsat data to construct a regression estimator of the total acres. We use the data to illustrate the computation of regression statistics from survey data. The sample most nearly approximates a stratified sample with strata identified in the column headed "county". The inverse of the sampling rates is given in the weight column. The estimated regression equation for the regression of interview hectares on satellite hectares defined by estimator (11) is

$$\hat{Y} = \underset{(8.332)}{-11.845} + \underset{(0.0922)}{1.1602}X,$$

where the numbers in parentheses are the standard errors obtained from the estimated covariance matrix calculated by equation (14).

Calculations were performed using Super Carp. If the equation and standard errors are calculated using unit weights in equations (11) and (14), respectively, we have

$$\hat{Y} = \underset{(9.282)}{-3.927} + \underset{(0.0963)}{1.0850}X.$$

If we calculate the F-test suggested in equation (17), we obtain

$$F_{23}^{2} = 2.81.$$

At first glance, this test is large enough to cause to suspicion about the equality of the two coefficients. Because this sample is very small and because of the structure of the weights, the test is nearly a test between two lines, the line for county one, and the average line for the remaining counties. In this small sample the deviations from the line in county one are small. Hence, the estimated standard errors of the coefficients for the two added variables are small. This phenomenon is discussed further in Section 3. If one uses the ordinary regression F-test that assumes homogeneous error variances and ignores the stratification, one obtains

$$F_{33}^{2} = 0.68.$$

While this statistic is not distributed as Snedecor's F, it does make one feel more comfortable with the assumption that the two weighting procedures are estimating the same equation.

Table 2 contains the standard errors of regression coefficients estimated under alternative assumptions. The estimated standard errors for the intercept behave much as one might anticipate. The stratified weighted sample procedure has the smallest estimated standard error followed by the stratified unit weight procedure and the ordinary least squares procedure. Do not forget these are estimated standard errors. The two stratified procedures are consistent under the stratified model. The weighted estimator has smaller variance because the observations for stratum 1, the stratum with the largest weight, lie closer to the estimated line than do the points in other strata. The ordinary least squares estimated standard error is not consistent under the stratified model. If the sample is treated as a cluster sample of counties, the estimated standard errors for the intercept are about 30 to 40 percent larger than the corresponding values for the stratified sample.

The estimated standard errors for the slope display a different behavior. The smallest estimated standard error is associated with the unit weight cluster estimation, and the largest estimated standard error is associated with ordinary least squares. Roughly speaking, the variation of slopes among clusters is small relative to the within cluster variation. Because the weights are inversely correlated with the observed variability, the weighted estimators have smaller estimated variances. This is a small sample, but it is sufficient to demonstrate that unit weights do not always produce smaller variances than sample weights and that stratification and clustering can have rather complex effects on the estimated variances of the regression coefficients.

3. WHAT IS A LARGE SAMPLE?

Our discussion has rested on the large sample properties of estimators and of estimators of variance. If the limiting normal distribution is being used to establish confidence intervals, the size of the sample required for a good approximation depends upon the nature of the original population. For example, if the characteristic is a rare zero-one item (probability less than 0.05, say), a very large sample (more than 1,400 for a simple random sample (Cochran, 1977, p. 58)) will be required for the normal approximation. The binomial with small p is only one example of the very skewed populations often encountered in sampling practice. Measures of size such as gross sales of firms, number of employees of firms, number of animals per farm, and family income are examples of skewed populations for which large samples are required before the distribution of the mean approaches normality. On the other hand, the distribution of the mean for items such as family size may approximate the normal distribution for small (less than 100) sample sizes.

The use of the Taylor expansion is semi-nonparametric in that the approximation holds, in large samples, under very mild assumptions on the population. The large sample requirements are met if we have no isolated points in our sample space. The method may perform poorly in situations where the generating distribution and sample size are such that an observation or observations are isolated from the remaining cluster of points. We consider the problem of estimating the variance of the vector of regression coefficients used to test the effect of weighting on the coefficients in the soybean example. The original vector is

$$(1,\ X,\ XW,\ W),$$

and the hypothesis to be tested is the hypothesis that the coefficients for XW and W are zero. To illustrate the problems associated with variance estimation for the vector of coefficients for the soybean data set, we create a vector that is orthogonal in the unit weight metric. The matrix of observations on the transformed independent variables is composed of the residuals obtained in the regression of each variable, except the first, on the elements preceeding it in the original vector. Table 3 contains the transformed regression variables $(X - \bar{X},\ RWX,\ RW)$. Only a few digits have been retained to make it easier to read the table.

When we regress Y on $(1,\ X - \bar{X},\ RWX,\ RW)$ we obtain

$$\begin{array}{l} \hat{Y} = \underset{(2.24)}{95.34} + \underset{(0.093)}{1.085}(X - \bar{X}) + \underset{(0.044)}{0.093} \times 10^{-2}RWX - \underset{(0.023)}{0.015}RW, \end{array}$$

where the estimated standard errors were computed for a stratified sample with unit weights using expression (14). If the regression and standard errors are computed by ordinary least squares, we obtain

$$\hat{Y} = \underset{(3.37)}{95.34} + \underset{(0.113)}{1.085}(X - \bar{X}) + \underset{(0.086)}{0.093} \times 10^{-2}RWX - \underset{(0.034)}{0.015}RW.$$

The estimated standard error for the coefficient of RWX obtained by Taylor methods is about one half of that obtained by ordinary least squares methods. This can be explained by the data configuration.

The first observation on RWX is much larger in absolute value than any other observation. Of the total sum of squares for RWX, 67 percent is due to this observation. The Taylor approximation to the variance uses the sample variance of deviates called $\hat{\underset{\sim}{d}}_{hij}$ in (14) to estimate the variance of the statistic. The deviations from regression, denoted by $\hat{v}$, are given in the last column of Table 3. The $\hat{v}$ value for observation one is among the smaller values. The mean square for the residuals is 421. The product $(RWX)(\hat{v})$ for the first observation is -1113. This product is of the same order of magnitude as the product for observations 3, 33 and 36. Therefore, while the first observation is responsible for about 67 percent of the sum of squares of RWX, it is responsible for only about 15 percent of the sum of squares of $(RWX)(\hat{v})$. This is because $\hat{v}^2$ for the first observation is less than one tenth of the average of the squares of the other observations. Furthermore, the squared deviation for the first observation is biased downward because the method of least squares will cause the estimated plane to pass close to an observation that is separated from the other observations. Thus, if all of the observations have the same error variance, the Taylor method will produce an estimate of the variance of the coefficient for RWX that is biased downward.

Did the procedure underestimate the variance for this sample? We do not know. If we use the parametric procedure of ordinary least squares, we assign the pooled estimate of error variance to the separated observation. It is not possible to determine if this procedure is correct because our estimate of variance for the separated observation is a one degree of freedom estimator.

In this situation most people will feel more comfortable assuming that the variance for the separated point is the same as the variance of the other points rather than taking the small observed variance of the single point.

In the nonparametric world a single observation contains little information about the variability of the population that generated the observation. Furthermore, an observation separated from other observations is essentially a single observation. In the full parametric world the separated observation is in the fold because the separated observation is specified to have been created by the same generating mechanism that created the other observations. For data of the type displayed in Table 3, the answer obtained by parametric methods rests very heavily on assumptions about the error variance.

In the estimation of variances, one measure of the numerical size of the sample is the number of cluster degrees of freedom. Thus, for example, the estimated covariance matrix for a k-dimensional vector random variable is singular unless

$$\sum_{h=1}^{L} (n_h - 1) > k.$$

In setting approximate confidence intervals it seems reasonable to use Student's t distribution with degrees of freedom no greater than $\Sigma\ (n_h - 1)$. Because the variance of an estimated variance is a function of the fourth moments of the population, estimated variances are notoriously unreliable. The coefficient of variation for the squares is $2^{\frac{1}{2}}$ for the normal and considerably larger for many other common distributions.

If the error variances in the strata are unequal or if unequal weights are applied to the estimates of different strata, the variance of the variance estimator can be considerably different from that suggested by a simple calculation of error degrees of freedom. Table 4 has been constructed using the data configurations of Table 1 to illustrate these effects on the estimated variance. In the first column we assume that stratification is ineffective in that we assume each stratum variance is equal to the variance of the population. We assume the parent population to be normal so that we can give an explicit expression for the variance of the variance. In this situation stratification produces an estimated error variance for a mean with a variance that is proportional to $(26.6)^{-1}$ while a simple random sample produces a variance of the estimated variance that is proportional to 36^{-1}. The effective degrees of freedom for the stratified sample is slightly less than 27 because of the unequal sample sizes within strata. If we use the sample weights of Table 1 and the usual stratified variance estimator, the variance of the estimated variance is proportional to $(4.6)^{-1}$. This large reduction is due to the large weight for the first stratum. If the variance in the first stratum is one half of the variance in other strata, then the effective degrees of freedom for the variance estimator is 12.4. In the last column we give the effective degrees of freedom for the simple random sample if the variance of the simple random sample is twice that of the stratified sample. This illustrates the fact that stratification can reduce both the variance of the estimated mean and the variance of the estimated variance of the mean.

While we are unable to specify the number of error degrees of freedom required for our approximations, it is clear that we shall be uncomfortable with a small number of degrees of freedom, particularly with unequal weights.

The theory of Corollary 1 uses a linear approximation to the nonlinear function of the sample means to approximate the behavior of the nonlinear function. If this approximation is to perform well, the curvature of the function must be small relative to the standard error of the sample means. For example, if the function is quadratic

$$g(\bar{Y}) = \alpha_1\bar{Y} + \alpha_2\bar{Y}^2,$$

the linear approximation is

$$g(\bar{Y}) \doteq \alpha_1\mu + \alpha_2\mu^2 + (\alpha_1 + 2\alpha_2\mu)(\bar{Y} - \mu).$$

The expected value of $g(\bar{Y})$ is

$$E\{g(\bar{Y})\} = \alpha_1\mu + \alpha_2[\mu^2 + V\{\bar{Y}\}].$$

For the linear approximation to perform well we must have small $V\{\bar{Y}\}$ and/or small α_2.

In summary, to be comfortable with the use of large sample theory we require:

1. A reasonable number of observations in the sense that no observations are widely separated from the main clusters of observations. This is another way of saying that the Taylor deviates are such that the mean of the deviates is nearly normally distributed.
2. A reasonable number of effective error degrees of freedom for the estimator of variance.
3. The curvature of the nonlinear function of sample means to be small relative to the standard error of the sample means.

ACKNOWLEDGEMENTS

This research was partly supported by Research Agreement 58-319T-1-0054X with the Statistical Reporting Service of the U.S. Department of Agriculture. I thank Nancy Hasabelnaby for computations and Carol Francisco for comments.

REFERENCES

[1] Cochran, W.G. (1977). Sampling Techniques 3rd Ed. Wiley, New York.

[2] Efron, B. (1979). Bootstrap method: Another look at the jackknife. Ann. Statist. 7, pp. 1-26.

[3] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 68, pp. 589-599.

[4] Fuller, W.A. (1975). Regression analysis for sample survey. Sankhyā Series C 37, pp. 117-132.

[5] Fuller, W.A. and Hidiroglou, M.A. (1978). Regression estimation after correcting for attenuation. J. Amer. Statist. Assoc. 73, pp. 99-104.

[6] Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1980). Super Carp, Department of Statistics, Iowa State University, Ames, Iowa.

[7] Isaki, C. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. J. Amer. Statist. Assoc. 77, pp. 89-96.

[8] Kish, L. and Frankel, M.R. (1974). Inference from complex samples. J. Roy. Statist. Soc. B 36, pp. 1-22.

[9] Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist. 9, pp. 1010-1019.

[10] McCarthy, P.J. (1965). Stratified sampling and distribution-free confidence intervals for the median. J. Amer. Statist. Assoc. 60, pp. 772-783.

[11] McCarthy, P.J. (1969). Pseudo-replication: Half-samples. Rev. Int. Statist. Inst. 37, pp. 239-264.

[12] Miller, R.G., Jr. (1974). The jackknife - a review. Biometrika 61, pp. 1-15.

[13] Rao, J.N.K. and Wu, C.F.J. (1984). Bootstrap with stratified samples. Technical Report No. 19 of the Laboratory for Research in Statistics and Probability. Carleton University, Ottawa, Canada.

Table 1: Soybean Area Determined by Two Methods

County	Segment	Weight	Soybean Hectares: Interview (Y)	Soybean Hectares: Satellite (X)
1	1	502	8.09	24.75
1	2		106.03	98.10
1	3		103.60	112.50
2	1	212	6.47	43.20
2	2		63.82	80.10
3	1	188	43.50	61.65
3	2		71.43	92.70
3	3		42.49	74.25
4	1	190	105.26	98.10
4	2		76.49	99.45
4	3		174.34	152.10
5	1	134	95.67	57.60
5	2		76.57	66.15
5	3		93.48	91.80
6	1	189	37.84	34.65
6	2		131.12	97.65
6	3		124.44	116.10
7	1	172	144.15	136.35
7	2		103.60	99.45
7	3		88.59	99.90
7	4		115.58	123.30
8	1	114	99.15	85.50
8	2		124.56	121.50
8	3		110.88	77.40
8	4		109.14	102.60
8	5		143.66	133.65
9	1	193	91.05	75.15
9	2		132.33	85.95
9	3		143.14	112.05
9	4		104.13	81.90
9	5		118.57	80.55
10	1	93	102.59	117.90
10	2		29.46	39.15
10	3		69.28	72.00
10	4		99.15	99.45
10	5		143.66	155.25
10	6		94.49	85.50

Table 2: Estimated Standard Errors of Regression Coefficients Calculated by Alternative Procedures

Procedure	Estimated standard Error	
	$\hat{\beta}_0$	$\hat{\beta}_1$
Ordinary least squares	10.747	0.1116
Stratified; sample weights	8.332	0.0922
Cluster; sample weights	11.121	0.0823
Stratified; unit weights	9.282	0.0963
Cluster; unit weights	13.256	0.1071

Table 3: Data for Transformed Regression Problem

Stratum Cluster	Weight	$X - \bar{X}$	$10^{-2}RWX$	RW	$\hat{v}$
1	502	-67	-195	167	6
1	502	7	25	336	6
1	502	21	68	369	-15
2	212	-48	1	1	-37
2	212	-11	4	24	-19
3	188	-30	10	-7	-20
3	188	1	5	7	-26
3	188	-17	8	-1	-35
4	190	7	4	12	3
4	190	8	4	13	-28
4	190	61	-3	38	14
5	134	-34	28	-53	34
5	134	-25	23	-51	6
5	134	0	5	-47	-3
6	189	-57	13	-20	3
6	189	6	4	11	29
6	189	25	2	20	3
7	172	45	-9	8	1
7	172	8	3	-6	-1
7	172	8	2	-6	-16
7	172	32	-5	3	-14
8	114	-6	10	-67	8
8	114	30	-22	-66	-2
8	114	-14	18	-68	28
8	114	11	-5	-67	1
8	114	42	-32	-65	5
9	193	-16	7	4	13
9	193	-6	6	9	43
9	193	21	3	22	26
9	193	-10	6	7	19
9	193	-11	6	6	35
10	114	26	-24	-90	-21
10	114	-52	63	-84	-16
10	114	-19	26	-87	-9
10	114	8	-4	-89	-6
10	114	64	65	-93	-16
10	114	-6	12	-88	3

Table 4: Efficiency of Estimated Variance under Alternative Assumptions

Procedure	Equivalent degrees of freedom	
	$V_{SRS} = V_{st}$	$V_{SRS} = 2V_{st}$
Simple random sampling	36	9
Strat. Sa., unit weights, equal var.	26.6	26.6
Strat. Sa., unequal weights, equal var.	4.8	4.8
Strat. Sa., unequal weights, $\sigma_1^2 = 0.5\sigma^2$	13.9	13.9

SELECTED BIBLIOGRAPHY OF DATA ANALYSIS FOR COMPLEX SURVEYS[1]

[1] Bellhouse, D.R. (1982). Discussion provided for the Session on Data Analysis from Complex Designs. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 54-55.

[2] Bickel, P.J. and Freedman, D.A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. Ann. Statist., 12, pp. 470-482.

[3] Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimates from Complex Surveys. Intl. Statist. Review, 51, pp. 279-292.

[4] Binder, D.A. (1982). Non-Parametric Bayesian Models for Samples from Finite Populations. J.R. Statist. Soc. B, 44, pp. 388-393.

[5] Binder, D.A., Gratton, M., Hidiroglou, M.A., Kumar, S. and Rao, J.N.K. (1984). Analysis of Categorical Data from Surveys with Complex Designs: Some Canadian Experiences. Survey Methodology (To appear).

[6] Brier, S.E. (1980). Analysis of Contingency Tables under Cluster Sampling. Biometrika, 67, pp. 591-596.

[7] Cohen, J.E. (1976). The Distribution of the Chi-squared Statistics under Cluster Sampling from Contingency Tables. J. Amer. Statist. Ass., 71, pp. 665-670.

[8] Choi, J.W. (1981). A Further Study on the Analysis of Categorical Data from Weighted Cluster Sample Survey. Proc. Amer. Statist. Ass., Section on Survey Methods Research, pp. 15-20.

[1] Prepared by the Project Team on the Analysis of Data from Complex Surveys whose members are D. Binder, M. Gratton, M. Jeays, G. Kriger, S. Kumar, D. Paton, C. Patrick and A. van Baaren.

[9] Cowan, J. and Binder, D.A. (1978). The Effect of a Two Stage Sample Design on Tests of Independence in a 2 by 2 Table. Survey Methodology, 4, pp. 16-28.

[10] Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. J. Roy. Statist. Soc. B, 31, pp. 195-223.

[11] Fay, R.E. (1979). On Adjusting the Pearson Chi-Square Statistics for Clustered Sampling. Proc. Amer. Statist. Ass., Social Statist. Section, pp. 402-406.

[12] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part I: Descriptions and Applications of the Method. Unpublished manuscript.

[13] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part II: Asymptotic Theory. Unpublished manuscript.

[14] Fay, R. (1982). Contingency Table Analysis for Complex Survey Designs: CPLX. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 44-53.

[15] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. J. Amer. Statist. Ass., 75, pp. 261-268.

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

[17] Gross, S.T. (1980). Median Estimation in Sample Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 181-184.

[18] Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. Biometrics, 25, pp. 489-504.

[19] Hidiroglou, M.A. (1983). Approximations to the Distribution of a Sum of Weighted Chi-Square Variables. Statistics Canada, Ottawa, Ontario, Canada.

[20] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.

[21] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980). MINI CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.

[22] Hidiroglou, M.A. and Rao, J.N.K. (1983). Chi-Squared Tests for the Analysis of Three Way Contingency Tables from the Canada Health Survey. Technical report. Statistics Canada.

[23] Holt, D. and Scott, A.J. (1981). Regression Analysis using Survey Data. The Statistician, 30, pp. 169-178.

[24] Holt, D., Scott, A.J. and Ewings, P.O. (1980). Chi-Squared Tests with Survey Data. J.R. Statist. Soc. A, 143, pp. 302-320.

[25] Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. J.R. Statist. Soc. A, 143 pp. 474-487.

[26] Imrey, P.B., Koch, G.G. and Stokes, M.E. (1981). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Review. Intl. Statist. Rev., 49, pp. 265-283 (In collaboration with J.N. Darroch, D.H. Freeman, Jr. and H.D. Tolley).

[27] Imrey, P.B., Koch, G.G. and Stokes, M.E. (1982). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part II: Data Analysis. Intl. Statist. Rev., 50, pp. 35-64 (In collaboration with J.N. Darroch, D.H. Freeman, Jr. and H.D. Tolley).

[28] Imrey, P.B., Sobel, E. and Francis, M. (1980). Modeling Contingency Tables from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 212-217.

[29] Kish, L. and Frankel, M.R. (1974). Inference from Complex Sample Surveys. J. Roy. Statist. Soc. B, 36, pp. 1-37.

[30] Koch, G.G., Freeman, D.H., Jr. and Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. Intl. Statist. Review, 43, pp. 59-78.

[31] Konijn, H.S. (1962). Regression Analysis in Sample Surveys. J. Amer. Statist. Ass., 57, pp. 590-606.

[32] Landis, J.R., Lepkowski, J.M., Eklund, S.A. and Stehouwer, S.A. (1982). A Statistical Methodology for Analyzing Data from a Complex Survey. The First National Health and Nutrition Examination Survey: National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 92. Washington, D.C.

[33] Lepkowski, J.M. (1982). The Use of OSIRIS IV to Analyse Complex Sample Survey Data. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 38-43.

[34] Lepkowski, J.M., Bromberg, J. and Landis, J.R. (1981). Program for the Analysis of Multivariate Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 8-15.

[35] McCarthy, P.J. (1965). Stratified Sampling and Distribution-Free Confidence Intervals for a Median. J. Amer. Statist. Ass., 60, pp. 772-783.

[36] Nathan, G. (1969). Tests of Independence in Contingency Tables from Stratified Samples. New Developments in Survey Sampling, pp. 578-600. (N.L. Johnson, and H. Smith, eds.). Wiley: New York.

[37] Nathan, G. (1971). A Simulation Comparison of Tests for Independence in Stratified Cluster Sampling. Bull. Int. Statist. Inst., 44(2), pp. 274-280.

[38] Nathan, G. (1972). On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples. J. Amer. Statist. Ass., 67, pp. 917-920.

[39] Nathan, G. (1973). Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples. National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 53, Washington, D.C.

[40] Nathan, G. (1975). Tables of Independence in Contingency Tables from Stratified Samples. Sankhyā C, 37, pp. 77-87.

[41] Nathan, G. (1981). Notes on Inference Based on Data from Complex Sample Designs. Survey Methodology, 7, pp. 109-129.

[42] Nathan, G. and Holt, D. (1980). The Effect of Survey Design on Regression Analysis. J. Roy. Statist. Soc. B, 42, pp. 377-386.

[43] Pfefferman, D. and Nathan, G. (1977). Regression Analysis of Data from Complex Surveys. Bull. Intl. Statist. Inst., 41(3), pp. 21-42.

[44] Rao, J.N.K. (1975). Analytic Studies of Sample Survey Data. Survey Methodology (Supplementary Issue).

[45] Rao, J.N.K. (1983). Some Current Topics in Sample Survey Theory. Paper presented at Iowa State University Stat. Lab. 50 Anniversary Conference, June 1983.

[46] Rao, J.N.K. and Hidiroglou, M.A. (1981). Chi-Squared Tests for the Analysis of Categorical Data from the Canada Health Survey. Bull. Intl. Statist. Inst., 49(2), pp. 699-718.

[47] Rao, J.N.K. and Scott, A.J. (1979). Chi-Squared Tests for Analysis of Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 58-66.

[48] Rao, J.N.K. and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys - Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. J. Amer. Statist. Ass., 76, pp. 221-230.

[49] Rao, J.N.K. and Scott, A.J. (1984). On Chi-Square Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Ann. Statist., 12, pp. 46-60.

[50] Särndal, C.E. (1982). Implications of Survey Design for Generalized Regression Estimation of Linear Functions. J. of Statist. Planning and Inference, 7, pp. 155-170.

[51] Schuster, J.J. and Downing, D.J. (1976). Two-Way Contingency Tables for Complex Sampling Schemes. Biometrika, 63, pp. 271-276.

[52] Scott, A.J. and Holt, D. (1982). The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. J. Amer. Statist. Ass., 77, pp. 848-854.

[53] Scott, A.J. and Rao, J.N.K. (1981). Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data. Current Topics in Survey Sampling. (D. Krewski, R. Platek and J.N.K. Rao, eds.)

[54] Sedransk, J. and Meyer, J. (1978). Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling. J.R. Statist. Soc. B, 40, pp. 239-252.

[55] Shah, B.V. (1978). SUDAAN: Survey Data Analysis Software. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 146-151.

[56] Shah, B.V. (1981). Development of Survey Data Analysis Software. Research Triangle Institute, Research Triangle Park, North Carolina, U.S.A.

[57] Shah, B.V., Holt, M.M. and Folsom, R.E. (1977). Inference about Regression Models from Sample Survey Data. Bull. Intl. Statist. Inst., 41(3), pp. 43-57.

[58] Tepping, B.J. (1968). Variance Estimation in Complex Surveys. Proc. Amer. Statist. Ass., Social Statistics Section, pp. 11-18.

[59] Tomberlin, T.J. (1979). The Analysis of Contingency Tables of Data from Complex Samples. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 152-157.

[60] Woodruff, R.S. (1952). Confidence Intervals for Medians and Other Position Measures. J. Amer. Statist. Ass., 47, pp. 635-646.

SURVEY METHODOLOGY

December 1983 Vol. 9 No. 2

A Journal produced by Statistics Canada

C O N T E N T S

8-3200-501
Reference No.
Z - 079

ISSN: 0714-0045