BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-84/24

A LINEAR MODEL APPROACH TO THE ESTIMATION OF SURVEY REDESIGN EFFECTS

by

Edward Gbur and Charles Alexander

Bureau of the Census
Washington, D.C. 20233

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Paul P. Biemer

Report completed: August 2, 1984

Report issued: August 8, 1984

1. Introduction

Ongoing surveys are periodically redesigned to reflect changes related to the population of interest. Such sample redesigns are necessary to maintain design efficiency by taking into account known changes in the characteristics of the population and by incorporating new methodological developments in sampling. The sample redesign may or may not coincide with changes in the questionnaire or interview procedures.

Data collected during and after implementation of the redesign may be affected simultaneously by changes in the population and by the redesign itself. If so, estimates produced during this period are not directly comparable to pre-redesign estimates. As a result, the redesign must be planned and implemented in a manner which allows for the effects due to the redesign to be estimated separately from the effects due to actual changes in the population. When the redesign effects can be estimated, adjustments can be applied to the survey estimates to make them directly comparable to pre-redesign results.

In this paper a linear model approach is taken to the direct estimation of both the redesign and non-redesign related effects on estimates from ongoing surveys. The general approach presented here in terms of sample redesign can be easily adapted to questionnaire and interview redesigns. Linear models and estimation procedures for this purpose are described in general terms in Section 2. In Section 3 one possible model for the National Crime Survey sample redesign is developed. Section 4 contains numerical results relating to this model and Section 5 contains some concluding remarks.

2. A Linear Model for Survey Redesign

Let θ_{1t} represent a parameter to be estimated from the sample at time t; e.g., a population proportion. Suppose there are K_1 basic survey related factors and K_2 redesign related factors which are thought to affect the estimation of θ_{1t}. For example, if the survey design requires individuals to be interviewed for several consecutive time periods, then there may be a time in sample effect on the response; c.f., Bailar (1975) for an example. This would be classified as a basic survey related effect since it existed prior to the redesign and will persist in some form after the redesign is completed. Examples of redesign related effects include the effect of a change in sampling frames, the behavioral effect of inexperienced interviewers in new sample areas and interviewers to be terminated in outgoing sample areas, and the effect of certain administrative burdens and disruptions associated with the redesign implementation.

Let $\theta_{2t}, \ldots, \theta_{Lt}$ represent the levels of these $K_1 + K_2$ factors for time period t, $t = t_0, \ldots, t_1$. Let $\theta_t = [\theta_{1t}, \ldots, \theta_{Lt}]'$ represent the vector of parameters at time t and let $\theta = [\theta'_{t_0}, \ldots, \theta'_{t_1}]'$. The parameter vector θ contains the population parameters of interest as well as the basic survey and redesign related parameters.

Let Y_t be the vector of responses at time t. Each entry of Y_t may represent the response of an individual or a group of individuals. Let $Y = [Y_{t_0}' \ldots, Y_{t_1}']'$ and let X be the design matrix relating E(Y) and θ. Thus, we can write a linear model in matrix form as

$$Y = X\theta + e, \tag{1}$$

where e is a random vector of error terms with

$$E(e) = 0 \quad ,$$
$$Cov(e) = \Sigma_e \quad . \tag{2}$$

The model (1) can represent a fixed effects analysis of variance model, an analysis of covariance model, or a regression model. The error term e represents all sources of variation which are responsible for the deviation of the observed response from its expected value.

In some applications it may be possible to decompose e into components representing various types of sampling and/or nonsampling errors. In that case (1) would be replaced by

$$Y = X\theta + Ub + e^* \quad , \tag{3}$$

where U is the design matrix for b and b and e* are random variables with

$$\begin{aligned} E(b) &= 0 \quad , \ \mathrm{Cov}(b) = \Sigma_b \\ E(e^*) &= 0 \quad , \ \mathrm{Cov}(e^*) = \Sigma_{e^*} \end{aligned} \tag{4}$$

and we usually assume that Cov(b, e*) = 0. The components of b represent the measurable sources of error and e* represents all remaining unexplained variation.

In other applications, the independent variables in the model (1) may be measured with error, leading to an errors-in-variable model. In the following discussion we shall restrict attention to the model (1). We shall also assume that the redesign and its implementation are planned and executed in such a way that the vector θ is estimable.

Estimation in (1) can be accomplished using the method of generalized least squares (GLS); i.e., finding the value of θ which minimizes

$$S(\theta) = (Y-X\theta)'\Sigma_e^{-1} (Y-X\theta) \ . \tag{5}$$

The minimum of S(θ)occurs when

$$\hat{\theta} = (X'\Sigma_e^{-1} X)^{-1}X'\Sigma_e^{-1}Y \quad . \tag{6}$$

The covariance matrix of the GLS estimators $\hat{\theta}$ is given by

$$\Sigma_{\hat{\theta}} = (X'\Sigma_e^{-1}X)^{-1} \quad . \tag{7}$$

Although the matrix calculations in (6) and (7) are straightforward, they assume that the covariance matrix $\sum_e$ is known, at least up to a multiplicative constant. Rarely, if ever, is this the case in practice. There are several alternatives when $\sum_e$ is unknown. Among them are

(i) to replace $\sum_e$ in (6) and (7) by a consistent estimator $\hat{\sum}_e$ which is independent of $\hat{\theta}$.

(ii) to model $\sum_e$ as a function of θ, say $\sum_e = \sum_e(\theta)$, and use iteratively reweighted least squares. When θ_{1t} is a rate or proportion this alternative may be appropriate.

(iii) to model $\sum_e$ as a function of other factors (e.g., time) and use the resulting estimator in (6) and (7).

In any case, the method of GLS can be used to obtain parameter estimates and estimated standard errors.

3. A Model for the National Crime Survey

The National Crime Survey (NCS) is an ongoing address survey conducted by the Bureau of the Census for the Bureau of Justice Statistics. It utilizes a stratified multistage cluster design and rotating panels in which each panel is interviewed in six month intervals for three and one half years. Each panel is split into six groups with one group interviewed each month of the six month period. The initial interview is used to establish a reference point and is not used for estimation. At each interview individuals aged 12 and older in the sample units are questioned about all crimes which occurred in the six months preceding the month of interview. The initial interview is in person. Some subsequent interviews may be by telephone. Victimization rates for various types of personal and household crimes are produced for a variety of demographic categories. Additional information can be found in Bureau of Justice Statistics NCS reports (1983).

The two major known survey related effects are the time in sample effect and a recall lag effect associated with the time lag between the interview and the occurrence of the reported crime. The recall lag effect may be related to memory loss, to "telescoping" (misplacement of crimes into, out of, and within the reference period), or to a combination of these and other factors (Kobilarcik et al. 1983).

The NCS sample is currently being redesigned to reflect population changes measured by the 1980 Census. The phase-in of the new sample in continuing areas will begin in January 1985. After this date, new addresses entering the sample will be selected from a frame based on the 1980 Census lists, updated for new construction. Beginning in January 1986, data from incoming areas (areas which are in the new design but not in the old design) will be used in the estimation and outgoing areas will be dropped.

For purposes of modeling, only one redesign related factor encompassing all sources that may affect the estimation of crime rates will be considered. This factor will be referred to as the area type effect and will have four levels: continuing nondisrupted areas, continuing disrupted areas, outgoing areas, and incoming areas. Continuing disrupted areas usually arise from changes in PSU boundaries. The area type factor includes interviewer effects, the effect of administrative disruptions and burdens, changes in stratum and PSU definitions, and the effect of any other systematic difference between areas which fall into different categories.

For a fixed type of crime and demographic group j, let

Y_{ijstmk} = reported number of victimizations occurring in month t of a particular type of crime for the i-th sampled individual in the j-th demographic group from the k-th area type who is being interviewed for the s-th time, having a recall lag of m months; i.e., a person who is interviewed in month $t' = t+m$, where the subscript ranges are $s=1, \ldots, 6$; $t=1, \ldots, T$; $m=1, \ldots, 6$; $k=1, \ldots, 4$; $i=1, \ldots, I$ $(= I_{jstmk})$. The order of the area types k is as listed above. Let w_{ijstmk} be the weight associated with this individual. For NCS this weight is the inverse of the household's probability of selection, combined with various noninterview and post-stratification adjustments. The same weight is used for all six months of occurrence associated with a particular interview. The weighted total number of victimizations is given by

$$Y_{.jstmk} = \sum_{i=1}^{I} w_{ijstmk} Y_{ijstmk} \quad . \tag{8}$$

It can be modeled as

$$\begin{aligned} Y_{.jstmk} &= w_{.jstmk} C_{jt} + w_{.jstmk} T_{jskh} + w_{.jstmk} R_{jm} \\ &\quad + w_{.jstmk} A_{jkt'} + w_{.jstmk} RA_{jmkt'} + e_{.jstmk}, \end{aligned} \tag{9}$$

where

$$w_{.jstmk} = \sum_{i=1}^{I} w_{ijstmk} \quad ,$$

C_{jt} = "true" victimization rate for demographic category j in month of occurrence t,

T_{jskh} = effect on the rate due to interviewing individuals in the j-th category from area type k for the s-th time where the interview occurred in the h-th six month period ($h=\Phi(t, m)$),

R_{jm} = effect on the rate for the j-th category due to recalling a victimization which occurred m months prior to the interview,

$A_{jkt'}$ = effect on the rate due to interviewing individuals in the k-th area type where the interview is conducted in month $t'=t+m$,

$RA_{jmkt'}$ = effect on the rate due to the interaction between the recall lag and area type,

$e_{.jstmk}$ = the aggregate of all errors.

The parameters in the model (9) are subject to the following constraints:

$$\sum_{s=1}^{6} w_{.jstmk} T_{jskh} = 0 \qquad \text{for all } t, k, h,$$

$$\begin{aligned} \sum_{m=1}^{6} w_{\cdot j \cdot tmk} R_{jm} &= 0 && \text{for all } t, k, \\ A_{j1t'} &= 0 && \text{for all } t' , \\ RA_{jm1t'} &= 0 && \text{for all } t', m, \\ \sum_{m=1}^{6} w_{\cdot j \cdot tmk} RA_{jmkt'} &= 0 && \text{for all } t, k. \end{aligned} \tag{10}$$

As a reference point, t=1 corresponds to January 1985 and h=1 represents the six month period from January to June 1985. Since not all subscript combinations correspond to available data, when the model (9) is written in matrix form the response vector Y is reduced accordingly and only those parameters appearing in the expectation of at least one available observation are included in θ.

Although the time in sample effect T_{jskh} refers to the repeated sampling of the same individuals over time, the effects are estimated from the responses of different individuals sampled in the same month but who have differing numbers of previous interviews. This implicitly assumes that all panels exhibit approximately the same behavior.

The time in sample effect is allowed to depend on the demographic group j, the area type k, and may change with time. The use of a six month period h for the time dependence is a matter of convenience. It should also be noted that although a sample address has been included in the sample s times, the particular occupants may have been in the sample less than s times.

The recall lag constraint in (10) assumes underreporting for some lags and overreporting for others with no net effect. If the recall lag is primarily a problem of "telescoping", then the constraint may be reasonable. On the other hand, if the loss of memory of more distant events is the primary reason for the recall lag and if we are willing to assume perfect recall for the month preceding the interview, then a more reasonable constraint would be $R_{j1} = 0$. Other constraints are possible depending on the perceived nature of the recall lag effect.

The area type constraint $A_{j1t'} = 0$ means that in continuing nondisrupted areas the phase-in will have no additional effect on the victimization rate. In particular, the effect of any changes in coverage associated with the change in sampling frames is assumed to be negligible for the aggregate response in (9).

Current NCS procedures use a generalized variance function approach to calculate approximate standard error estimates for many characteristics. Empirical studies have shown that variances of crime estimates calculated using a Taylor series approximation may be approximated by a simple function of the estimated value. Thus, a single "generalized" function for the estimated variance is used for all types of crime included in the studies. Adapting this approach to our model, the variance of each response can be approximated by

$$\mathrm{var}(Y_{.jstmk}) \cong \alpha_k \, (E[Y_{.jstmk}])^2 + \beta_k \, E[Y_{.jstmk}], \qquad (11)$$

where $E(Y_{.jstmk})$ is obtained from (9) and the α_k and β_k are constants which must be estimated.

Estimation of the covariance terms in $\sum_e$ can be approached in many ways. Among the approaches are

(i) to approximate them in terms of θ using a modified generalized variance function approach,

(ii) to use known information about the nature of the effects and the survey procedures to obtain direct estimates,

(iii) to model the covariances (either linearly or nonlinearly) as a function of θ and any other factors which are thought to have an effect on them.

4. A Numerical Example

In this section an analysis using the model (9) is presented for a set of 1982 NCS data. The numerical results should be viewed only as an illustration of the proposed modeling procedure. The data consist of all reported crimes of violence (rape, robbery, and assault) which occurred during 1982 for the entire sample of persons age 12 and older. Data were collected from February 1982 through June 1983. The demographic group j consists of the entire population of persons age 12 and older.

Since the selected period does not coincide with any part of the phase-in, all areas may be classified as continuing nondisrupted areas (k=1). From the constraints (10) for $A_{jkt'}$ and $RA_{jmkt'}$ and the convention of deleting parameters which do not correspond to available data, there are no area type effects or recall lag - area type interaction terms in the model. Thus, the model (9) reduces to

$$Y_{.stm} = w_{.stm}\, C_t + w_{.stm}\, T_{sh} + w_{.stm}\, R_m + e_{.stm} \quad , \tag{12}$$

where the subscripts j and k have been dropped for notational simplicity. Let t=1 correspond to January 1982 and h=1 correspond to the six month period January - June 1982.

For simplicity in this illustrative example, the covariance matrix $\sum_e$ is assumed to be a diagonal matrix with diagonal entries given by (11) where α = -0.0000125671 and β =2355.0. The values of α and β are those used in the 1982 NCS variance estimation formulas.

The GLS estimates of the time in sample effects and recall lag effects are given in Tables 1 and 2, respectively. Following the procedure described in Bateman and Bettin (1975), the estimated values of the C_t from the model were used to calculate an estimated annual victimization rate of 33.82 per thousand with an estimated standard error of 0.62. This is comparable to the published rate of 34.3 per thousand with an estimated standard error of 0.6.

An illustration of the interpretation of the estimates in Tables 1 and 2 follows. From Table 1 the estimated effect on the victimization rate attributed to individuals interviewed for the first time (excluding bounding interviews) during the six month period from January to June 1982 is to increase the rate by approximately 0.40 victimizations per thousand. This is not statistically significant. Several estimates in Table 1, however, are significant. The estimated recall lags are interpreted similarly.

Table 1. Estimated Time in Sample Effects for Violent Crimes Occurring in 1982 Based on the Model (12).*

	January-June, 1982		July-December, 1982		January-June, 1983	
Time in Sample	Estimated Effect	Estimated Standard Error	Estimated Effect	Estimated Standard Error	Estimated Effect	Estimated Standard Error
1	0.399	0.272	0.424	0.165	0.273	0.193
2	0.571	0.275	-0.127	0.151	0.294	0.197
3	0.042	0.264	-0.209	0.147	0.013	0.186
4	-0.281	0.246	0.235	0.159	-0.072	0.187
5	-0.259	0.239	-0.434	0.140	-0.596	0.159
6	-0.472	0.228	0.111	0.156	0.088	0.187

*Entries are given as rates per thousand.

Table 2. Estimated Recall Lag Effect for Violent Crimes Occurring in 1982 Based on the Model (12)*

Recall Lag	Estimated Effect	Estimated Standard Error
1	2.290	0.148
2	0.295	0.118
3	-0.100	0.114
4	-0.509	0.104
5	-0.836	0.098
6	-1.139	0.093

*Entries are given as rates per thousand.

5. Remarks

In general the form of the model and assumptions are problem dependent and must be carefully constructed if any useful information is to be gained from its application. The NCS model (9) is relatively simple in that several interaction terms were not included. They were assumed to be negligible. These terms and other factors could be added to the model provided the parameters remain estimable. Alternatives to several of the constraints in (10) could be considered. The modeling process always allows for adjustment and revision. We expect that the NCS model (9) will also be revised and improved.

The model (9) contains only an aggregate redesign effect. To understand this effect more fully it is important to measure the various components of the area type effect through special studies and experiments. For example, special observation and record keeping for new interviewers could give additional information on the effect of new interviewers. However, the approach described in this paper is applicable even in the absence of a special redesign research program. It requires only data which will ordinarily be collected in the course of the survey.

Finally, caution must be exercised in the application of GLS to data collected from any complex sampling design. The papers by Fuller (1975) and Kish and Frankel (1974), among others, indicate the theoretical and practical difficulties which can arise. The effect of the sample design on GLS estimation in the NCS model is currently being investigated.

References

Bailar, B. A. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. JASA, 349, 23-30.

Bateman, D. and P. Bettin (1975). Standard Error Estimates for the National Crime Survey. Proceedings of the ASA Social Statistics Section, 168-177.

Fuller, W. A. (1975). Regression Analysis for Sample Survey. Sankhya C, 87, 117-132.

Kish, L. and M. R. Frankel (1974). Inference from Complex Surveys (with discussion). J. Royal Stat. Soc. B, 36, 1-37.

Kobilarcik, E. L., C. H. Alexander, R. P. Singh, and G. M. Shapiro (1983). Alternative Reference Periods for the National Crime Survey. Proceedings of the ASA Section on Survey Research Methods, 197-202.

U.S. Department of Justice (1983). Criminal Victimization in the United States, 1981. A National Crime Report. NCJ-90208. Washington, D.C.