BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR/87-14

Modeling the Effect of a Redesign on the Estimates from an Ongoing Demographic Survey

by

Edward Gbur
Statistical Research Division
Bureau of the Census
Washington, DC 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 or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, DC 20233.

Recommended by: Lawrence Ernst

Report completed: April, 1987

Report issued: May, 1987

The Census Bureau's demographic surveys are periodically redesigned for a variety of reasons. Among them are to reflect changes in the composition of the population, to improve the efficiency of the estimators through changes in the sample design or the estimation procedure, and to accommodate changes in the purpose of the survey (e.g., through changes in the survey instrument). Such redesigns are often phased-in over a time span which encompasses several data collection periods. Thus, the survey estimates produced during the phase-in are affected by the redesign as well as by actual changes in the population characteristics themselves.

The purpose of this report is to describe an approach to modeling the effect of the redesign on estimates from an ongoing survey. It consists of three parts originally written as separate reports over several years and reflects the author's attempts to set up, and then modify, the models as his experience with and information about the surveys grew. In the first part a model for the 1985-86 National Crime Survey (NCS) sample redesign is constructed. The feasibility of the analysis is demonstrated using 1982 (non-redesign) data. In Part 2 a similar model is described for the 1984-85 Current Population Survey (CPS) sample redesign. The third and final part represents an attempt to analyze the CPS redesign. A cross-sectional analysis does not prove to be adequate and a longitudinal approach is discussed but not implemented. If the obstacles to the latter approach can be overcome, it could provide a useful method for future redesigns.

The NCS analysis in Part 1 illustrates a second, and perhaps more important use of such models; namely, the potential to estimate the effect of some non-sampling errors and perhaps ultimately to remove these effects from the published estimates. Although this application is not explored in this report, it does merit further investigation.

A Linear Model Approach to the Estimation of the Redesign Effects in the National Crime Survey

by

Edward Gbur
Statistical Research Division
Bureau of the Census

March 1984

1. Introduction

The National Crime Survey (NCS) has been redesigned to reflect population changes from the 1980 Census. The final report of the CPS/NCS phase-in work group (Document No. 26 dated May 16, 1983) describes several alternative phase-in plans. Plan B in this report has been selected for use. A description of the plan and the reasons for its selection are contained in an SMD Draft Memorandum entitled "NCS Redesign: Plans for Measuring the Effect of Phase-in of New Sample Areas" (undated). A discussion of the major factors which may affect NCS estimates and several possible approaches to estimating their effect are given.

Three approaches and the necessary assumptions of each are mentioned in the draft. These can be briefly described as

(1) a comparison of the estimates obtained for the part of the population represented by the continuing sample areas with the estimates from the entire sample,

(2) a comparison of time series based forecasts with estimates from the entire sample,

(3) a linear model approach to the direct estimation of the phase-in effects.

The purpose of this report is to expand on these preliminary discussions and to construct a linear model which contains the major effects. In the framework of the proposed model, estimates of the effects and their variances can be obtained as well as the victimization levels and rates measured by NCS. This report will not consider the first two approaches listed above.

The selected phase-in plan (Plan B) is characterized by an abrupt change in noncontinuing areas. Data collection in all outgoing areas ceases after December, 1985. Bounding interviews in incoming areas are conducted from

July to December, 1985 and data collection for estimation purposes begins in January, 1986. In continuing areas the 1980 sampling frame is used to select all new rotation groups beginning with those whose bounding interviews are conducted from January to June, 1985.

The major factors which may affect NCS estimates can be broadly classified as either phase-in related or nonphase-in related. Phase-in related effects are those primarily due to the disruption caused by the redesign and should disappear after phase-in completion in June, 1988. The model presented in this report contains only one phase-in related effect, referred to as the type of area effect. Nonphase-in related effects are those which existed prior to the redesign and which will persist after completion of the phase-in. We would like to compare these effects on pre and post-redesign estimates and if possible, to study any changes in them during the phase-in period. Two nonphase-in related effects are included in the proposed model. These are referred to as the recall lag effect and the time in sample effect.

As a phase-in related effect, the type of area effect is intended to reflect changes in the survey methodology and not the population being sampled. For redesign purposes, the PSU is not always the unit of interest. The term area will be used to designate the unit of interest. It can be either a PSU or part of a PSU in those instances where the stratum and/or PSU definitions have changed with the introduction of the 1980 sampling frame. The area type effect has four levels: continuing nondisrupted areas, continuing disrupted areas, outgoing areas, and incoming areas. Examples of continuing disrupted areas are those PSUs whose boundaries have been changed to include previously nonsampled areas or to delete previously sampled areas. Under this breakdown the area type effect encompasses the effects of new interviewers and interviewers who will be fired, the effects of certain types of administrative

disruptions and burdens, and the effect of any other systematic differences between areas which fall into different categories.

Both of the nonphase-in related effects reflect characteristics associated with the sampled population's reaction to the survey procedure. The time in sample effect is well documented and need not be discussed further. The recall lag effect arises from the method of data collection; viz., the respondent is asked to describe all victimizations in the six month period preceeding the interview. Memory loss and a "telescoping effect" may be the principal components of the recall lag effect.

2. Statement of the Model and Assumptions

The model presented in this section describes the response for an entire class of individuals rather than for a single individual. The groups have been made as large as possible in terms of number of individuals represented while still allowing for the effects of all major factors on each group for which NCS estimates are regularly produced. These demographic groups could be defined by age, race, sex, and place of residence or by some smaller set of collapsed categories. The response to be modeled is given in terms of the number of victimizations rather than in terms of the number of instances of a particular type of crime.

For each sampled individual, let

Y_{ijstlk} = reported number of victimizations of a particular type of crime for the i-th sampled individual in the j-th demographic group from the k-th type of area who is being interviewed for the s-th time, reporting occurrences in month t and having a recall lag of l months,

where the subscript ranges are

$$i = 1, \ldots, I \ (= I_{jstlk}) ,$$

$$j = 1, \ldots, J ,$$

$$s = 1, \ldots, 6 ,$$

$$t = 1, \ldots, T ,$$

$$l = 1, \ldots, 6 ,$$

$$k = 1, \ldots, 4 .$$

The order of the area types k is as listed in the Introduction. As will be shown later, not all subscript combinations correspond to available data. Let w_{ijstlk} be the weight associated with Y_{ijstlk}. A model containing the major factors of interest could be constructed for each individual. However, since the vast majority of the Y_{ijstlk} will be either zero or one, it is more practical to model the weighted total number of reported victimizations; i.e.,

$$Y_{.jstlk} = \sum_{i=1}^{I} w_{ijstlk} Y_{ijstlk} . \qquad (1)$$

For a fixed demographic category j and month of victimization t, the response (1) can be modeled as

$$Y_{.jstlk} = w_{.jstlk} C_{jt} + w_{.jstlk} T_{jskh} + w_{.jstlk} R_{jl} + w_{.jstlk} A_{jkt'} + w_{.jstlk} RA_{jlkt'} + e_{.jstlk} , \qquad (2)$$

where

$$w_{.jstlk} = \sum_{i=1}^{I} w_{ijstlk} ,$$

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

T_{jskh} = effect on the rate for interviewing individuals in demographic group j from area type k for the s-th time where the interview occurred in the h-th six month period of the phase-in ($h=\phi_1(t, 1)$),

R_{jl} = effect on the rate of recalling a victimization which occurred l months prior to the interview for individuals in demographic group j,

$A_{jkt'}$ = effect on the rate of interviewing individuals in the k-th area type when the interview is conducted in month $t' = t + l$,

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

$e_{.jstlk}$ = the aggregate of all sampling errors.

The parameters in the model (2) are subject to the following constraints:

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

$$\sum_{l=1}^{6} w_{.j.tlk}\, R_{jl} = 0 \qquad \text{for all } k,$$

$$A_{jlt'} = 0 \qquad \text{for all } t' \ , \qquad (3)$$

$$RA_{jl1t'} = 0 \qquad \text{for all } l \ ,$$

$$\sum_{l=1}^{6} w_{.j.tlk}\, RA_{jlkt'} = 0 \qquad \text{for all } k \ .$$

As a reference point, t = 1 corresponds to January, 1985 and h = 1 represents the six month period from January to June, 1985.

A simple numerical example will serve to illustrate the interpretation of the terms in the proposed model. For purposes of the example we shall ignore the error term in (2). Suppose we are interested in the number of victimizations in month t for individuals in demographic group j. Suppose we are interviewing individuals for the first time (excluding bounding interviews) in incoming areas concerning victimizations which occurred three months prior to the interview. If

(i) the sampled individuals have weights which total 5000000 ($w_{.jlt34}$ = 5000000 persons),

(ii) the "true" victimization rate for this demographic group in the month of interest is 0.0300 victimizations per person ($C_{jt} = 0.0300$),

(iii) the effect on the rate due to these individuals being sampled from the first time is to increase it by 0.0020 victimizations per person ($T_{j14h} = 0.0020$),

(iv) the effect due to the sampled individuals being asked to recall an occurrence three months prior to the interview is to increase the rate by 0.0010 victimizations per person ($R_{j3} = 0.0010$),

(v) the effect due to sampling from an incoming area is to reduce the rate by 0.0050 victimizations per person ($A_{j4(t+3)} = -0.0050$),

(vi) and the effect of the interaction between the recall lag and incoming area is to reduce the rate by 0.0005 victimizations per person ($RA_{j34(t+3)} = -0.0005$),

then

$$Y_{.j1t34} = (5000000)(0.0300) + (5000000)(0.0020 + 0.0010 - 0.0050 - 0.0005)$$
$$= 150000 - 12500$$
$$= 137500 \text{ reported victimizations.}$$

Having illustrated the interpretation of the terms in the model (2), we now briefly discuss the rationale for each term and the corresponding constraints in (3).

"True" victimization rate: The estimate of C_{jt} represents a monthly estimate which, in itself, is not published. The parameter does not necessarily represent the true rate, rather it represents what the individuals in the group would be willing to tell an interviewer under "ideal" conditions. Thus, the word true is placed in quotes.

Following Bateman and Bettin (a paper entitled "Standard Error Estimation for the Natinal Crime Survey" presented at the 1975 ASA meetings in Atlanta),

the "true" quarterly and yearly rates can be expressed in terms of the C_{jt} as

$$C_{j.}^{(Q)} = \left(\sum_{t=t_0}^{t_0+2} w_{.j.t..} C_{jt}\right) / P^{(Q)} ,$$

$$C_{j.}^{(Y)} = \left(\sum_{t=t_0}^{t_0+11} w_{.j.t..} C_{jt}\right) / \bar{P}^{(.)} ,$$

respectively, where

$$P^{(Q)} = \sum_{t'=1}^{8} \omega_{t'} P_{jt'} ,$$

$P_{jt'}$ = $\frac{1}{6}$ (independent control count of the number of persons in the j-th demographic group for month of data collection t'),

$$\omega_{t'} = 1/3 \text{ if } t' = 1, 8$$

$$= 2/3 \text{ if } t' = 2, 7$$

$$= 1 \text{ if } t' = 3, 4, 5, 6 ,$$

$$\bar{P}^{(.)} = 1/4 \sum_{Q=1}^{4} P^{(Q)} .$$

The weights $\omega_{t'}$ reflect the number of months of occurrence in the quarter for which information is obtained in month of data collection t'. Hence, the monthly estimates of C_{jt} can be used to produce a set of publishable values.

Time in sample (TIS) effect: Ideally this effect would depend on the demographic group j, the type of area k, and the month of data collection $t' = t + 1$. This latter dependence would allow for changes in the TIS effect over time. If we allow the TIS effect to depend on t', then there are as many parameters T_{jsklt} as there are responses $Y_{.jstlk}$. Even with constraints on the T's of the form given in (3), we would not be able to uniquely estimate all of the parameters in the model. Therefore, the dependence on time must be modified.

Since each rotation group is divided into panels which are interviewed over a six month period, it is convenient to allow the TIS effect to depend

on the six month period h (January-June or July-December) in which the data is collected. Thus, the TIS effect is denoted by T_{jskh}. It should be noted that although the TIS effect refers to the repeated sampling of the same individuals over time, the effects are calculated from the responses of different individuals sampled in the same month but who have differing numbers of previous interviews. Hence, the use of six month periods is a matter of convenience and the actual dependence on time should be investigated further.

The constraint $\sum_s w_{.jstlk} T_{jskh} = 0$ means that within each demographic group j, area type k, and six month period h, the TIS effects can be thought of as deviations from some average level; i.e., for some numbers of interviews s, there is underreporting while for others there is overreporting, but they balance out over all values of s.

The constraint could be modified to $\sum_s w_{.jstlk} T_{jskh} = K$. If the constant K does not depend on time or on the demographic group j, then a nonzero value of K does not represent a more general constraint. This follows since we can write $T^*_{jskh} = T_{jskh} - (K/w_{.jstlk})$, replace T_{jskh} in the model by T^*_{jskh}, and add an overall mean to the right hand side of the model (2). The T^*_{jskh} sum to zero over s.

Recall lag effect: The effect of recalling a victimization which occurred l months prior to the interview is allowed to depend on the demographic group j but not on area type, month of occurrence, or number of interviews.

The constraint $\sum_l w_{.j.tlk} R_{jl} = 0$ assumes underreporting for some lags and overreporting for others with a net effect of zero. If the recall lag is primarily a problem of "telescoping" rather than a loss of memory, then the constraint appears to be reasonable. On the other hand, if loss of memory is the primary reason for the recall lag effect and if we are willing to assume perfect recall for the month preceding the month of 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.

Area type effect: The area type effect is assumed to depend on the type of area k and the month of data collection $t' = t + 1$, where

k = 1 if the area is continuing and nondisrupted,

k = 2 if the area is continuing and has been disrupted,

k = 3 if the area is outgoing,

k = 4 if the area is incoming.

Classification of the majority of the sampled areas will be clearcut. In a few cases listing a sampled area as continuing nondisrupted (k=1) or as continuing disrupted (k=2) can be somewhat arbitrary. Examples of how such cases can arise were given in the Introduction. These should be classified on a case by case basis using prespecified guidelines. If a small number of major causes of disruption in continuing areas can be identified, it may prove fruitful to break area type k = 2 into several separate factor levels.

As indicated in the Introduction, the area type effect encompasses all factors which are phase-in related; e.g., the effect of new interviewers in incoming areas and continuing disrupted areas, the effect of terminated interviewers in outgoing areas and continuing disrupted areas, and the effect of certain types of administrative problems created by the phase-in. Since many of these factors may change monthly as data is collected, the effect is allowed to depend on the month of data collection $t' = t + 1$. It is assumed that these phase-in related factors may affect each demographic category j differently.

The constraint $A_{j1t'} = 0$ for all t' means that continuing nondisrupted areas will have no additional effect on the victimization rate. In particular, it assumes that there is no effect associated with the change from the 1970 to

the 1980 sampling frame in these areas. The effects of any changes in coverage are assumed to be negligible. The reasonableness of this coverage assumption should be investigated in separate studies. The constraint $A_{j1t'} = 0$ also implies that the effects of the remaining area types in general will not cancel in the calculation of the expectation of the weighted total number of victimizations.

Recall lag-area type interaction: The inclusion of the interaction term $RA_{j1kt'}$ allows for the difference between two recall lag effects to depend on the type of area as well as on the demographic group and month of data collection. For example, lag differences for incoming areas in the early part of the phase-in may be different than those of continuing nondisrupted areas. The constraints imposed on the $RA_{j1kt'}$ are consistent with those imposed on the recall lag and area type effects.

Error term: The error terms $e_{.jst1k}$ represent all sources of variation which are responsible for the deviation of the observed response from its expected value. They are random variables having mean zero and some covariance structure. They need not be uncorrelated and homoscedastic. A more detailed discussion of the covariance structure is given in a later section.

The model (2) is relatively simple in that several first order and all higher order interactions have not been included. These interactions were assumed to be negligible. These or other factors could be added to the model but care must be taken not to exceed the available number of observations.

3. Parameter Estimation

The parameters in the model (2) subject to the constraints (3) can be estimated by the method of generalized least squares (GLS). Estimates of the variances of the GLS estimators can also be obtained. There are several possible levels of analysis; monthly, quarterly, yearly, and cumulative from

the start of the phase-in to the present. At the monthly level, each month of occurrence is analyzed separately. For analyses at the higher levels, data from all months of occurrence in the period are analyzed at the same time.

It is recommended that the analysis level be cumulative from the start of the phase-in period. In fact, it may be helpful to include data collected prior to the phase-in. Early in the phase-in this would provide estimates of nonphase-in related effects based on a larger number of observations. It would also allow for a comparison of changes in the TIS effect over a longer time period. Since the models for different demographic groups do not have any parameters in common, a combined analysis for several groups does not have an advantage over separate analyses.

For purposes of analysis, the responses for the j-th demographic group concerning occurrences in month t are collected into the column vector

$$Y_{jt} = [Y_{.j1t11}, \ldots, Y_{.j6t11}, \ldots, Y_{.j1t61}, \ldots, Y_{.j6t61}, \ldots,$$
$$Y_{.j1t14}, \ldots, Y_{.j6t14}, \ldots, Y_{.j1t64}, \ldots, Y_{.j6t64}]'.$$

The entries of Y_{jt} are grouped by area type k, recall lag l within area type, and TIS level s within recall lag.

If data are collected in some month t', then all six recall lags are available since they all occur within each interview. However, in outgoing and incoming areas (k=3,4) for a given month of occurrence t, data from all lags may not be available. For example, if t = 8 (August, 1985), incoming areas will contribute only data collected in January and February, 1986 (l=5,6) and outgoing areas will contribute only data collected from September to December, 1985 (l=1,2,3,4). Hence, the range of l values depends on k and t. Similar statements can be made for the TIS subscript s for incoming areas.

This means that for some values of t, not all entries of Y_{jt} correspond to available data. In such cases we shall reduce the length of Y_{jt} accordingly and include in the model (2) only those parameters which appear in the expectation of at least one available obervation.

Let $Y_j = [Y'_{jt_0}, \ldots, Y'_{jt_1}]'$ represent the vector of observed responses to be analyzed. For monthly level analyses $t_1 = t_0$, for quarterly level analyses $t_1 = t_0 + 2$, etc. Let

$$\begin{aligned} E(Y_j) &= X_j \theta_j , \\ Cov(Y_j) &= \Sigma_j \end{aligned} \tag{4}$$

represent the mean vector and covariance matrix of Y_j, respectively, where

X_j = the design matrix for the model (2),

θ_j = the vector of parameters in the model (2).

Then the GLS estimators of θ_j are those values of θ_j which minimize

$$S(\theta_j) = (Y_j - X_j \theta_j)' \Sigma_j^{-1} (Y_j - X_j \theta_j). \tag{5}$$

The minimum value of (5) occurs when

$$\hat{\theta}_j = (X_j' \Sigma_j^{-1} X_j)^{-1} X_j' \Sigma_j Y_j. \tag{6}$$

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

$$\Sigma(\hat{\theta}_j) = (X_j' \Sigma_j^{-1} X_j)^{-1} \tag{7}$$

Although equations (6) and (7) involve straightforward matrix manipulations, they assume that the covariance matrix Σ_j is known, at least up to a multiplicative constant. There are several alternatives for unknown Σ_j. Briefly these are

(1) to replace Σ_j in (6) by a consistent estimator $\hat{\Sigma}_j$ which is independent of $\hat{\theta}_j$. Substitution in (7) yields an estimated covariance matrix for $\hat{\theta}_j$.

(2) to model Σ_j as a function of θ_j, the mean vector, say $\Sigma_j = \Sigma_j(\theta_j)$, and use iteratively reweighted least squares. That is, replace Σ_j by an identity matrix, obtain an initial estimate $\tilde{\theta}_j$ of θ_j from (6), estimate $\Sigma_j(\theta_j)$ by $\tilde{\Sigma}_j(\tilde{\theta}_j)$, re-estimate θ_j from (6) with $\tilde{\Sigma}_j$ in place of Σ_j, and iterate until some convergence criterion is met.

(3) to model Σ_j as a function of some other factors and use the resulting estimator in place of Σ_j in (6).

The form of $E(Y_j)$ and Σ_j and the use of the alternative procedures for unknown Σ_j are discussed in later sections. In any case, we can obtain parameter estimates and estimated standard errors, and by appealing to large sample theory, approximate tests of significance.

4. Expectations of Monthly Totals

In this section we investigate the expected value of the monthly total number of victimizations. These calculations will provide additional insight into the model and will show that these totals are not unbiased estimators of the corresponding "true" total number of victimizations during the phase-in period under the model (2). However, unbiased estimators of the "true" totals can be obtained from the method of generalized least squares applied to the model (2).

For a fixed month of occurrence t and demographic category j, the total number of victimizations for a particular type of crime is given by

$$Y_{.j.t..} = \sum_k \sum_l \sum_s Y_{.jstlk} \quad ,$$

where the ranges of summation ultimately depend on t and their order cannot be interchanged in general (An example was given in the previous section.).

The mean of $Y_{.j.t..}$ is given by

$$E(Y_{.j.t..}) = \sum_k \sum_l \sum_s E(Y_{.jstlk})$$

$$= (\sum_k \sum_l \sum_s w_{.jstlk})\, C_{jt} + \sum_k \sum_l (\sum_s w_{.jstlk}\, T_{jskh})$$

$$+ \sum_k (\sum_l \sum_s w_{.jstlk}\, R_{jl}) + \sum_k \sum_l \sum_s w_{.jstlk}\, A_{jkt'}$$

$$+ \sum_k (\sum_l \sum_s w_{.jstlk}\, RA_{jlkt'}) , \qquad (8)$$

where $t' = t + 1$ is the month of data collection,

$h = \phi_1(t,l)$

$= [(t+l-1)/6] + 1$,

with $[\,\cdot\,]$ denoting the greatest integer function.

To simplify the general expression in equation (8), the following cases need to be considered.

Case 1: $1 \leq t \leq 6$ (occurrences prior to July, 1985)

In this case no data are available from incoming areas and complete data are available from the remaining types of areas.

Case 2: $6 < t < 12$ (occurrences from July to November, 1985)

In this case partial data are available from incoming and outgoing areas and complete data are available for both types of continuing areas.

Case 3: $12 \leq t < 42$ (occurrences from December, 1985 to the end of the phase-in)

In this case no data are available from outgoing areas, partial data are available from incoming areas, and complete data are available from both types of continuing areas.

In the calculations which follow a dot in place of a subscript indicates a summation over the entire range of the subscript. We shall allow an abuse of this notation when we write the monthly total $Y_{.j.t..}$.

Case 1: $1 \leq t \leq 6$

For each t the following combinations of subscripts correspond to available data:

$k = 1, 2, 3,$

$l = 1, \ldots, 6,$

$s = 1, \ldots, 6.$

From (8) and the constraints (3), we have

$$E(Y_{.j.t..}) = \left(\sum_{k=1}^{3} w_{.j.t.k}\right) C_{jt} + \sum_{k=1}^{3} \sum_{l=1}^{6} \left(\sum_{s=1}^{6} w_{.jstlk}\, T_{jskh}\right)$$

$$+ \sum_{k=1}^{3} \left(\sum_{l=1}^{6} w_{.j.tlk}\, R_{jl}\right) + \sum_{k=2}^{3} \sum_{l=1}^{6} w_{.j.tlk}\, A_{jk(t+l)}$$

$$+ \sum_{k=2}^{3} \left(\sum_{l=1}^{6} w_{.j.tlk}\, RA_{jlk(t+l)}\right)$$

$$= \left(\sum_{k=1}^{3} w_{.j.t.k}\right) C_{jt} + \sum_{k=2}^{3} \left(\sum_{l=1}^{6} w_{.j.tlk}\, A_{jk(t+l)}\right) . \qquad (9)$$

The second term in equation (9) represents the effects of the continuing disrupted areas and the outgoing areas on the expected total number of victimizations.

Case 2: $6 < t < 12$

For each t the following combinations of subscripts correspond to available data:

$k = 1, 2$, $l = 1, \ldots, 6$, $s = 1, \ldots, 6$,

$k = 3$, $l = 1, \ldots, 12-t$, $s = 1, \ldots, 6$,

$k = 4$, $l = 13-t, \ldots, 6$, $s = 1$ $(h = 3)$.

From (8) and the constraints (3), we have

$$E(Y_{.j.t..}) = \left(\sum_{k=1}^{2} w_{.j.t.k} + \sum_{l=1}^{12-t} w_{.j.t13} + \sum_{l=13-t}^{6} w_{.jltl4}\right) C_{jt}$$

$$+ \left[\sum_{k=1}^{3} \sum_{l} \left(\sum_{s=1}^{6} w_{.jstlk} T_{jskh} \right) + \sum_{l=13-t}^{6} w_{.jltl4} T_{jl43} \right]$$

$$+ \left[\sum_{k=1}^{2} \left(\sum_{l=1}^{6} w_{.j.tlk} R_{jl} \right) + \sum_{l=1}^{12-t} w_{.j.tl3} R_{jl} \right.$$

$$\left. + \sum_{l=13-t}^{6} w_{.jltl4} R_{jl} \right] + \left[\sum_{l=1}^{6} w_{.j.tl2} A_{j2(t+l)} \right.$$

$$\left. + \sum_{l=1}^{12-t} w_{.j.tl3} A_{j3(t+l)} + \sum_{l=13-t}^{6} w_{.jltl4} A_{j4(t+l)} \right]$$

$$+ \left[\sum_{l=1}^{6} w_{.j.tl2} RA_{jl2(t+l)} + \sum_{l=1}^{12-t} w_{.j.tl3} RA_{jl3(t+l)} \right.$$

$$\left. + \sum_{l=13-t}^{6} w_{.jltl4} RA_{jl4(t+l)} \right]$$

$$= \eta_1 C_{jt} + \sum_{l=13-t}^{6} w_{.jltl4} T_{jl43} + \sum_{k=3}^{4} \left(\sum_{l} \sum_{s} w_{.jstlk} R_{jl} \right)$$

$$+ \sum_{k=2}^{4} \left(\sum_{l} \sum_{s} w_{.jstlk} A_{jk(t+l)} \right) + \sum_{k=3}^{4} \left(\sum_{l} \sum_{s} w_{.jstlk} RA_{jlk(t+l)} \right), \quad (10)$$

where $\eta_1 = \sum_{k=1}^{2} w_{.j.t.k} + \sum_{l=1}^{12-t} w_{.j.tl3} + \sum_{l=13-t}^{6} w_{.jltl4}$.

The actual expressions for the additional terms on the right hand side of (10) are not as important as are their existence. The importance of (10) is that the expected total depends on the TIS effect for incoming areas, the recall lag effect for outgoing and incoming areas, the area type effect for all areas except continuing non-disrupted, and the recall lag-area type interaction for outgoing and incoming areas. The presence of the TIS effect is due to the disturbance of the rotation group pattern in incoming areas; i.e., all sampled individuals are having their first interview for data collection purposes. The presence of the recall lag effect is due to the

cessation of data collection in outgoing areas and the beginning of data collection in incoming areas. The survey related phase-in effects are represented by the area type term.

Case 3: $12 \leq t < 42$

For each t the following combination of subscripts correspond to available data:

$$k = 1, 2 \ , \ l = 1, \ldots, 6 \ , \ s = 1, \ldots, 6 \ ,$$

$$k = 4 \ , \ l = 1, \ldots, l^* \ , \ s = 1, \ldots, s^* \ ,$$

$$k = 4 \ , \ l = l^* + 1, \ldots, 6 \ , \ s = 1, \ldots, s^* + 1,$$

where

s^* = maximum value of s for $l = 1$

$= [\, t/6 \,] - 1$,

l^* = number of l in the same six month period (h) as $l = 1$

$= 6\,([\, t/6 \,] + 1) - t$,

and we ignore the second set of subscript combinations for $k = 4$ when $l^* = 6$. The need for two sets of subscript combinations for $k = 4$ arises from data collection in two consecutive six month periods h which have different ranges of TIS subscripts for incoming areas.

From (8) and the constraints (3), we have

$$E(Y_{.j.t..}) = \Big(\sum_{k=1}^{2} w_{.j.t.k} + \sum_{l=1}^{l^*}\sum_{s=1}^{s^*} w_{.jstl4} + \sum_{l=l^*+1}^{6}\sum_{s=1}^{s^*+1} w_{.jstl4}\Big) C_{jt}$$

$$+ \Big[\sum_{l=1}^{l^*}\sum_{s=1}^{s^*} w_{.jstl4}\, T_{jst4h} + \sum_{l=l^*+1}^{6}\sum_{s=1}^{s^*+1} w_{.jstl4}\, T_{jst4h'}\Big]$$

$$+ \sum_{l=l^*+1}^{6} w_{.j(s^*+1)tl4}\, R_{jl} + \Big[\sum_{l=1}^{6} w_{.j.tl2}\, A_{j2(t+l)}$$

$$+ \sum_{l=1}^{l^*}\sum_{s=1}^{s^*} w_{.jstl4}\, A_{j4(t+l)} + \sum_{l=l^*+1}^{6}\sum_{s=1}^{s^*+1} w_{.jstl4}\, A_{j4(t+l)}\Big]$$

$$+ \left[\sum_{l=1}^{l^*} \sum_{s=1}^{s^*} w_{\cdot jstl4} \, RA_{jl4(t+1)} + \sum_{l=l^*+1}^{6} \sum_{s=1}^{s^*+1} w_{\cdot jstl4} \, RA_{jl4(t+1)} \right] , \tag{11}$$

where h′ is used for emphasis to indicate a different h value than h in the pervious summand.

As in Case 2, the actual expressions in (11) involving the TIS, recall lag, area type, recall lag-area type interaction are not as important as is their presence in the expression.

5. The Covariance Structure

The discussion of GLS estimation in Section 3 did not deal explicitly with the structure of the covariance matrix $\sum_j$ of the response vector Y_j. In this section, one possible structure is described. A generalized variance function approach similar to that currently employed in calculating estimated standard errors of NCS estimates is taken. This approach will not lead to the exact form of $\sum_j$; it will only yield approximations. The deviations of these approximations from the exact forms should be investigated.

The demographic group j and type of crime are fixed throughout this section. For notational convenience, write the covariance matrix of Y_j in partitioned form

$$\sum_j = \begin{bmatrix} \sum_{jt_0} & & & * \\ \sum_{j(t_0+1)t_0} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \sum_{jt_1t_0} & \cdots & \sum_{jt_1(t_1-1)} & \sum_{jt_1} \end{bmatrix} , \tag{12}$$

where $\sum_{jt} = \mathrm{Cov}(Y_{jt})$,

$\sum_{jtr} = \mathrm{Cov}(Y_{jt}, Y_{jr})$ for $t \neq r$.

We shall adopt the convention that only rows of $\sum_j$ corresponding to available data $Y_{.jstlk}$ for month of occurrence t are included in $\sum_j$.

Nonzero covariance terms in $\sum_j$ arise from two major sources of correlation: the correlation between different individuals within the same sampled area (PSU or part of a PSU) and the correlation arising from multiple contributions from the same individual, either from more than one interview or from information for more than one month of occurrence within a single interview. The latter source applies only to analyses at levels higher than the monthly level. The former source includes the effect of the same interviewer collecting data from several individuals and the effect of the sampled area's characteristics; e.g., individuals in a "higher risk" area tend to have above average numbers of victimizations.

Although only one PSU is selected per stratum there will be instances in which there may be more than one sampled area per stratum. For example, the boundaries of a selected PSU may have been redefined in such a way that it now consists of a continuing disrupted area (k=2) and an incoming area (k=4). Since the number of such cases is relatively small, we shall assume that their contribution to the covariance terms is negligible; i.e., assume that $\mathrm{Cov}(Y_{.jstlk}, Y_{.js'rl'k'}) = 0$ for $k \neq k'$. This assumption, together with with the ordering of the entries of each subvector Y_{jt} of Y_t means that each block in expression (12) for $\sum_j$ will be block diagonal; i.e.,

$$\Sigma_{jt} = \begin{bmatrix} \Sigma_{jt}^{(1)} & & & 0 \\ & \Sigma_{jt}^{(2)} & & \\ & & \Sigma_{jt}^{(3)} & \\ 0 & & & \Sigma_{jt}^{(4)} \end{bmatrix}$$

$$= \text{Blk diag} \left[\Sigma_{jt}^{(k)} ; \quad k = 1, \ldots, 4 \right]$$

and

$$\Sigma_{jtr} = \text{Blk diag} \left[\Sigma_{jtr}^{(k)} ; k = 1, \ldots, 4 \right] \quad , t \neq r.$$

The generalized variance function approach assumes that the variance of an estimator X of a total has the form

$$\text{Var}(X) \cong \alpha (E[X])^2 + \beta E[X], \tag{13}$$

where α and β are unknown constants. The coefficients α and β are estimated from a set of statistics whose variances have the same general form. The advantage of the generalized variance function approach is that it relates the variance of the estimator to its mean. In terms of model (2), under this approach the second method (iteratively reweighted least squares) in Section 4 can be used to obtain parameter estimates and estimated standard errors.

Applying (13), the variance of $Y_{.jstlk}$ has the form

$$\text{Var}(Y_{.jstkj}) \cong \alpha_k (E[Y_{.jstlk}])^2 + \beta_k E[Y_{.jstlk}], \tag{14}$$

where, from model (2),

$$E[Y_{.jstlk}] = w_{.jstlk} (C_{jt} + T_{jskh} + R_{jl} + A_{jkt'} + RA_{jlkt'}).$$

Hence, the covariance matrix Σ_j will be a function of all of the entries of θ_j. If we are willing to assume that the "true" victimization rate C_{jt} is "large" relative to the combined effects of the remaining factors, then (14) reduces to

$$\text{Var}(Y_{.jstlk}) \cong (\alpha_k w_{.jstlk}^2) C_{jt}^2 + (\beta_k w_{.jstlk}) C_{jt}. \tag{15}$$

Although this reduced form may have intuitive appeal, the estimation procedure is not simplified by using (15) in place of (14) in $\sum_j$. Thus, the more general form (14) is recommended.

The coefficients α_k and β_k in (14) are assumed to be dependent on the area type k. This allows the variance function to be affected by the phase-in. This may not be necessary and in fact, data limitations may make it necessary to use the same coefficients throughout. This needs to be investigated further.

Since each block in $\sum_j$ is block diagonal only covariances of the form $\mathrm{Cov}(Y_{.jstlk}, Y_{.js'rl'k})$ need to be considered. There are several possibilities for these terms. Among them are:

(1) Assume that $\mathrm{Cov}(Y_{.jstlk}, Y_{.js'rl'k}) = 0$ for all s, s', t, r, l, l'.

(2) Develop models for these covariances in terms of the entries of θ_j. To do this, define the relative covariance (correlation) for two estimators X and Y as

$$V_{XY} = \frac{\mathrm{Cov}(X, Y)}{E[X]\,E[Y]} .$$

As in the generalized variance function approach, for a set of pairs of statistics whose covariances have similar forms, model the relative covariance as a function of E[X] and E[Y], say

$$V_{XY} \cong \Psi(E[X], E[Y]),$$

so that

$$\mathrm{Cov}(X, Y) = \Psi(E[X], E[Y]) \cdot E[X]\,E[Y]$$

is given in terms of the parameters which determine the means of X and Y.

Regardless of whether (1), (2), or some other approach is taken, the nondiagonal entries of $\sum_j$ need to be expressed as functions of θ_j or as

known constants. Then the method of iteratively reweighted least squares can be applied to obtain parameter estimates and estimated standard errors.

6. A Numerical Example

In this section an analysis using the model (2) is presented for a set of 1982 NCS data. The data consist of all reported crimes of violence (rape, robbery, and assult) which occurred during 1982 for the entire population of persons age 12 and older. Data was collected from February, 1982 through June, 1983. The data was obtained from tapes prepared by Paul Wakim (SRD) from the NCS incident files.

Since the selected period does not coincide with any part of the phase-in only continuing nondisrupted areas (k=1) appear in the sample. From the constraints (3) for A_{jkt}' and RA_{jlkt}' 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 (2) reduces to

$$Y_{.stl} = w_{.stl}\, C_t + w_{.stl}\, T_{sh} + w_{.stl}\, R_l + e_{.stl} \quad , \tag{16}$$

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.

From (16), the parameter vector θ is the thirty-two entry vector

$$\theta = [C_1, \ldots, C_{12}, T_{11}, \ldots, T_{51}, T_{12}, \ldots, T_{52},$$
$$T_{13}, \ldots, T_{53}, R_1, \ldots, R_5]' \ .$$

The parameters T_{61}, T_{62}, T_{63}, and R_6 are not included since

$$T_{6h} = - \sum_{s=1}^{5} w_{.stl}\, T_{sh} \,/\, w_{.6tl} \qquad \text{for all } t, h,$$

$$R_6 = - \sum_{l=1}^{5} w_{..tl} R_l / w_{..t6} \qquad \text{for all } t.$$

The response vector Y contains the 432 observed weighted totals and the design matrix X is given by

$$X = W^* X^* \quad ,$$

where

$$W^* = \text{diag } [w_{.stl} \; ; \; s=1, \ldots, 6, \; l=1, \ldots, 6, \; t=1, \ldots, 12] \; ,$$

X^* = incidence matrix for θ.

For simplicity assume that

$$\sum = \text{Cov}(Y)$$

$$= \text{diag } [\text{var}(Y_{.stl})] \; ,$$

where

$$\text{var}(Y_{.stl}) \cong \alpha \, (E[Y_{.stl}])^2 + \beta \, E[Y_{.stl}] \qquad \text{for all } s, t, l,$$

with $\alpha = - 0.0000125671$ and $\beta = 2355.0$. The values of α and β are those used in the 1982 NCS variance estimation formulas.

The method of iteratively reweighted least squares (c.f., Section 3) was applied to the data and model (16). Calculations were carried out using a combination of Minitab and Fortran programs. The analysis of variance table is given in Table 1. Although the distributions of the F-statistics are only approximate, the results are clear-cut. The monthly rates C_t are nonzero; there is a highly significant recall lag effect; there is a significant time in sample effect. The TIS effect is not as striking as the recall lag effect which may be an order of magnitude larger. In fact, for the first six month data collection period the TIS effect is not significant at the 10 percent level.

Table 1. Analysis of Variance Table for Violent Crimes Occurring in 1982 Based on the Model (16).

Source of Variation	Sum of Squares	df	Mean Square	F-ratio	p-value
Monthly rate	2364.656	12	197.055	196.66	<0.0001
Time in sample	56.817	15	3.787	3.78	<0.001
Jan-June, 1982	6.068	5	1.214	1.21	>0.10
July-Dec, 1982	24.511	5	4.902	4.89	<0.001
Jan-June, 1983	26.238	5	5.248	5.24	<0.001
Recall lag	325.514	5	65.103	64.97	<0.0001
Error	400.740	400	1.002		
Total (Uncorrected)	3147.727	432			

The estimated monthly rates and their estimated standard errors are given in Table 2 and the corresponding information for quarterly and yearly rates is listed in Table 3. The estimates in Table 3 were obtained using the formulas given in Section 2. Since the actual weights $w_{..t.}$ can only be obtained from the NCS complete victimization file, independent control counts were used in place of the $w_{..t.}$ in these expressions. In practice the actual weights would be used. The published yearly rate for 1982 is included for comparison purposes.

Table 2. Monthly Violent Crime Rates for 1982 Based on the Model (16).

Month	Estimated Rate	Estimated Standard Error
1*	2.623**	0.167**
2	2.573	0.162
3	2.702	0.172
4	2.815	0.179
5	2.890	0.174
6	3.046	0.187
7	3.372	0.195
8	3.217	0.189
9	2.827	0.177
10	2.796	0.179
11	2.752	0.172
12	2.772	0.169

* t=1 corresponds to January, 1982.
** Entries are given as rates per thousand.

Table 3. Quarterly and Yearly Violent Crime Rates for 1982.

Time Period	Estimated Rate	Estimated Standard Error
(a) Quarter		
1	7.866*	0.294*
2	8.726	0.315
3	9.388	0.326
4	8.291	0.305
(b) Year		
From model (16)	34.274	0.655
Published	34.3	0.6

*Entries are given as rates per thousand.

As expected, the published yearly rate and the estimated rate from the model (16) do not differ significantly. However, such close agreement will not necessarily occur during the phase-in since, in addition to area type effects, the TIS and recall lag patterns will be disrupted.

Residual plots and other diagnostics in the analysis indicated several candidates for outliers. The analysis was rerun with the observations corresponding to the five largest positive standardized residuals and the only large negative standardized residual deleted. The large negative residual corresponded to the only observed weighted total equal to zero. Although there were minor changes in various F-ratios and parameter estimates, no significant changes occurred in the results. The yearly estimated rate is 33.822 victimizations per thousand with an estimated standard error of 0.616 victimizations per thousand.

The estimated time in sample and recall lag effects are given in Tables 4 and 5, respectively. For example, from Table 4 the estimated effect on the rate attributed to individuals interviewed for the first time (excluding the bounding interview) during the six month period from January to June, 1982

is to increase the rate by approximately 0.4 victimizations per thousand. This is not statistically significant ($t \cong 1.47$). In general, there are no striking patterns in Table 4. The estimated recall lag effects in Table 5 are interpreted similarly. The pattern in Table 5 is as expected; viz., there is a tendency to overreport victimization levels in the immediate past and to underreport for the distant past.

Table 4. Estimated Time in Sample Effects for Violent Crimes Occurring in 1982 Based on the Model (16).

	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 5. Estimated Recall Lag Effect for Violent Crimes Occurring in 1982 Based on the Model (16)

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.

A Linear Model Approach to the Estimation of the Redesign Effects in the Current Population Survey

by

Edward Gbur

Statistical Research Division

May, 1984

1. Introduction

The Current Population Survey (CPS) has been redesigned to reflect population changes from the 1980 Census and to improve the efficiency of estimators at the state level. The final report of the CPS/NCS phase-in work group (Document No. 26, dated May 16, 1983) discusses several alternative phase-in plans. A modified version of Plan L described in this report has been selected for use. The most recent version of this plan, Plan R', is described in an SMD Memorandum entitled "CPS Redesign: Change in Dummy Assignments in Phase-In Plan R'" (ID# K-10, dated January 11, 1984). The major factors which may affect CPS estimates are described in the CPS/NCS final report and in an SMD Memorandum entitled "Plans to Measure the Effect of Phase-In on CPS Redesign" (ID# K-13, dated January 17, 1984). The latter paper also describes one approach to measuring the effects of these factors.

The purpose of this report is to provide an alternative approach to the problem of measuring the phase-in effects. A linear model containing the major factors of interest is proposed. In the context of the proposed model, estimators of these effects, as well as the rates and levels measured by CPS, and their standard errors can be obtained.

The selected phase-in plan (Plan R') is characterized by a gradual phase-in in continuing areas and, in contrast, an abrupt change in noncontinuing areas. In continuing areas the 1980 sampling frame is used to select all rotation groups introduced after March, 1984. The first 1980 based group is A48-5. The phase-in ends in continuing areas after June, 1985. In outgoing areas the last rotation group to be interviewed for the full set of eight months is A48-3. Groups A48-4 through A49-3 are interviewed only for the first set of four months and then replaced by groups selected from incoming areas. These incoming replacement groups are interviewed for four consecutive months corresponding to the

interview times in the rotation pattern of the groups being replaced. Sampling in outgoing areas ceases after May, 1985. Rotation groups introduced after October, 1984 (beginning with A49-4) in incoming areas receive the full set of eight interviews. The full set of month in sample times for incoming areas is not available until November, 1985, at which point the phase-in is complete in all areas.

The major factors which may affect CPS estimates can be broadly classified as either phase-in related or nonphase-in related. Phase-in related effects are those primarily due to the disruption caused by the redesign and should disappear after phase-in completion in November, 1985. The model presented in this report contains one phase-in related effect, referred to as the type of area effect. Nonphase-in related effects are those which existed prior to the redesign and which will persist after completion of the phase-in. The nonphase-in related effect included in the proposed model is referred to as the month in sample effect.

As a phase-in related effect, the type of area effect is intended to reflect changes in survey methodology and not in the population being sampled. For redesign purposes, the term area is used to designate the unit of interest. It can be either a PSU or part of a PSU in those instances where the stratum and/or PSU definitions have changed with the introduction of the 1980 sampling frame. The area type effect has four levels: continuing nondisrupted areas, continuing disrupted areas, outgoing areas, and incoming areas. Examples of continuing disrupted areas arise from PSUs whose boundaries have been changed to include previously nonsampled areas or to drop previously sampled areas and from continuing PSUs in which sample sizes are expected to increase or decrease dramatically (c.f., SMD Memorandum ID# K-8, "CPS Phase-In: Change in Sample Size for Continuing Areas"). Under this breakdown the area type

effect encompasses the effects of new interviewers and interviewers who will be fired, the effects of certain types of administrative disruptions and burdens, and the effect of any other systematic differences between areas which fall into different categories.

The levels for the area type effect chosen in this report differ from those considered in SMD Memorandum ID# K-13, where continuing areas are divided according to their self-representing, nonself-representing classification. They also differ from the phase-in regions used to specify weighting procedures as described in an SMD Memorandum entitled "CPS Phase-In Specificatons for Assigning Base Weights and SR/NSR Status" (ID# K-7, dated December 20, 1983). The selection of levels of the area type factor is discussed in more detail in the following section.

The nonphase-in related effect reflects characteristics associated with the sampled population's reaction to the survey procedure. The month in sample effect is well documented (e.g., B.A. Bailar (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. JASA, 70, 23-30) and will not be discussed further.

2. Statement of the Unemployment Model and Assumptions

The model presented in this section describes the response for an entire group of individuals rather than a single individual. The groups can be defined by age, race, sex, and geographic variables or by a smaller set of collapsed categories. The response to be modeled is given in terms of a weighted total rather than a rate.

For each sampled individual, let

Y_{ijksmt} = 0-1 unemployment response for the i-th sampled individual in the j-th demographic group from the k-th type of area who is interviewed for the m-th time within the s-th four month period for the individual's rotation group concerning the individual's status in month t,

where the subscript ranges are

$i = 1, \ldots, I \quad (= I_{jksmt})$

$j = 1, \ldots, J$

$k = 1, 2, 3, 4$

$s = 1, 2$

$m = 1, 2, 3, 4$

$t = 1, \ldots, T$,

and

$Y_{ijksmt} = 1$ if the individual is unemployed,

$= 0$ if the individual is employed or not in in the civilian labor force (CLF).

The ordering of the area types k corresponds to that given in the Introduction. As will be shown later, not all subscript combinations correspond to available data.

Let w_{ijksmt} be the weight associated with the i-th sampled individual. The weighted total number of unemployed represented by the part of the sample corresponding to an observable subscript combination (j,k,s,m,t) is given by

$$Y_{.jksmt} = \sum_{i=1}^{I} w_{ijksmt} Y_{ijksmt} \quad . \tag{1}$$

The 0-1 responses Y_{ijksmt} could be modeled directly using techniques associated with binary response regression models; e.g., probit or logit analysis. This approach will not be considered here. Instead, the weighted total $Y_{.jksmt}$ defined in equation (1) will be used as the response variable in the model.

For a fixed demographic category j, the response $Y_{.jksmt}$ can be modeled as

$$Y_{.jksmt} = C_{.jksmt} R_{jt} + C_{.jksmt} A_{jkt} + C_{.jksmt} S_{jst} + C_{.jksmt} M_{jmt} + C_{.jksmt} SM_{jsmt} + C_{.jksmt} AS_{jkst} + C_{.jksmt} AM_{jkmt} + e_{.jksmt} , \quad (2)$$

where

$C_{.jksmt}$ = the number of individuals in the j-th demographic group in the CLF in month t represented by the (k,s,m)-th part of the sample,

R_{jt} = true unemployment rate for the j-th demographic group in month t.

A_{jkt} = effect on the rate due to interviewing individuals from the j-th demograhpic group in areas of type k concerning month t,

S_{jst} = effect on the rate due to interviewing individuals from the j-th demographic group in the s-th four month period for that rotation group concerning month t,

M_{jmt} = effect on the rate due to interviewing individuals from the j-th demographic group in the m-th month of a four month period concerning month t,

SM_{jsmt} = effect on the rate due to the period-month within period interaction,

AS_{jkst} = effect on the rate due to the area type-month within period interaction,

$e_{.jksmt}$ = the aggregate of all sampling errors.

The parameters in the model (2) are subject to the following constraints:

$$A_{j1t} = 0 \qquad \text{for all t,}$$

$$\sum_{s=1}^{2} C_{.j.s.t} S_{jst} = 0 \qquad \text{for all t,}$$

$$\sum_{m=1}^{4} C_{.j..mt} M_{jmt} = 0 \qquad \text{for all t,}$$

$$\sum_{s=1}^{2} C_{.j.smt}\, SM_{jsmt} = 0 \qquad \text{for all } m, t,$$

$$\sum_{m=1}^{4} C_{.j.smt}\, SM_{jsmt} = 0 \qquad \text{for all } s, t, \tag{3}$$

$$AS_{j1st} = 0 \qquad \text{for all } s, t,$$

$$\sum_{s=1}^{2} C_{.jks.t}\, AS_{jkst} = 0 \qquad \text{for all } k, t,$$

$$AM_{j1mt} = 0 \qquad \text{for all } m, t,$$

$$\sum_{m=1}^{4} C_{.jk.mt}\, AM_{jkmt} = 0 \qquad \text{for all } k, t.$$

As a reference point for the development of the model (2), let t = 1 correspond to January, 1984. It should prove useful to include data collected for an even more extensive period prior to the beginning of the phase-in. This would enable us to study the behavior of the terms which represent the month in sample effect before the phase-in and to compare the changes in them during the phase-in with pre and post phase-in levels. The use of data prior to January, 1984 would depend on its availability.

The three factor interaction between area type, period, and month within period was not included in the model (2) so that the error variance could be estimated and approximate tests of significance could be carried out.

The coefficients $C_{.jksmt}$ in the model (2) are unknown so that (2) does not represent a linear model. Moreover, the information in the responses $Y_{.jksmt}$ by themselves will not provide a means of estimating the $C_{.jksmt}$. One solution to this problem would be to replace the $C_{.jksmt}$ by estimates and develop estimation and inference procedures for the model (2) conditional on their estimated values. A natural source of estimates of the $C_{.jksmt}$ is

the CLF information collected from the same individuals in the CPS sample used to estimate the parameters in the unemployment model (2). A discussion of the estimation of these coefficients is given in Section 4. Difficulties with this approach and some alternatives are described in Section 5.

Let $\hat{c}_{.jksmt}$ denote the estimated value of $C_{.jksmt}$. Then the model (2) can be rewritten with the estimated coefficients as

$$\begin{aligned} Y_{.jksmt} = {} & \hat{c}_{.jksmt} R_{jt} + \hat{c}_{.jksmt} A_{jkt} + \hat{c}_{.jksmt} S_{jst} \\ & + \hat{c}_{.jksmt} M_{jmt} + \hat{c}_{.jksmt} SM_{jsmt} + \hat{c}_{.jksmt} AS_{jkst} \\ & + \hat{c}_{.jksmt} AM_{jkmt} + e_{.jksmt} \cdot \end{aligned} \qquad (2')$$

The constraints (3) are modified similarly and will be referred to as equation (3').

A simple numerical example will serve to illustrate the interpretation of the terms in the proposed model. For purposes of the example we shall ignore the error term in the model (2'). Suppose we are interested in the number of unemployed individuals from a particular demographic group j in month t. Consider individuals interviewed in incoming areas (k=4) for the second month (m=2) within their first set of four months in the survey (s=1). If

(i) the sampled individuals have weights which total 1000000 ($\hat{c}_{.j412t} = 1000000$),

(ii) the true unemployment rate for this group in the month of interest is 9.2% ($R_{jt} = 0.092$),

(iii) the effect on the unemployment rate due to sampling from an incoming area in month t is to reduce it by 0.3% ($A_{j4t} = -0.003$),

(iv) the effect on the rate due to sampling for the first four month period for this group in month t is to increase it by 0.1% ($S_{j1t} = 0.001$),

(v) the effect on the rate due to sampling for the second consecutive month for this group in month t is to increase the rate by 0.08% ($M_{j2t} = 0.0008$),

(vi) the effect on the rate due to the period-month within period interaction for this group in month t is to increase it by 0.01% ($SM_{j12t} = 0.0001$),

(vii) the effect on the rate due to the area type-period interaction for this group in month t is to decrease it by 0.04% ($AS_{j41t} = -0.0004$),

(viii) the effect on the rate due to the area type-month within period interaction for this group in month t is to reduce it by 0.02% ($AM_{j42t} = -0.0002$),

then the reported number of unemployed is

$$Y_{.j412t} = (1000000)(0.092) + (1000000)(-0.003 + 0.001 + 0.0008 + 0.0001 - 0.0004 - 0.0002) = 92000 - 1700 = 90300.$$

This reported value is approximately a 1.8% underestimate of the true level.

Having illustrated the interpretation of the terms in the model (2'), we now briefly discuss the rationale for various model terms and the corresponding constraints in (3').

Unemployment rate: The estimates of the monthly unemployment rates R_{jt} can be used for publication. They can be thought of as having been adjusted for time in sample and phase-in related effects. Furthermore, annual averages can be obtained from the estimated R_{jt} which would provide an alternative to that described in Technical Report 40 (pp. 64-65).

Since the estimates of the R_{jt} from the unemployment model (2') utilize data from all previous months, they are directly comparable to the composite estimators currently in use for CPS regardless of whether or not

we are in a phase-in period. A comparison of the estimators from the model (2′) and the composite estimators will be the subject of future research.

Area type effect: The area type effect depends on the demographic group j, type of area k, and month of interest t, where

k = 1 if the area is continuing and nondisrupted,

k = 2 if the area is continuing and disrupted,

k = 3 if the area is outgoing,

k = 4 if the area is incoming.

Classification of the majority of the sampled area will be clear-cut. In a few cases listing a sampled area as continuing nondisrupted (k=1) or as continuing disrupted (k=2) can be somewhat arbitrary. Examples of how such cases arise were given in the Introduction. These should be classified on a case by case basis using prespecified guidelines.

The area type effect encompasses all factors which are phase-in related; e.g., the effect of new interviewers in incoming areas and continuing disrupted areas having large increases in sample size, the effect of terminated interviewers in outgoing areas and continuing disrupted areas having large decreases in sample sizes, and the effect of certain types of administrative problems created by the phase-in. Since many of these factors may change monthly the effect is allowed to depend on the month of interest t. They may affect different demographic groups in different ways; thus the dependence of j.

The constraint $A_{j1t} = 0$ for all t means that the effect of the phase-in on rates in continuing nondisrupted areas is negligible. In particular, it assumes that there is no effect associated with the change from the 1970 to the 1980 sampling frame in these areas. Any changes in coverage are assumed to be negligible at the aggregate level found in the model (2′). The

reasonableness of this coverage assumption should be investigated in separate studies. The constraint $A_{j1t} = 0$ also implies that the effects of the remaining area types in general will not cancel in the calculation of the expectation of the weighted total.

The phase-in regions found in SMD Memorandum ID# K-7 are defined in terms of the frame used to select the sample. Since it has been assumed that the frame effect is negligible compared to interviewer and administrative effects, the use of phase-in regions instead of area types in defining the levels of this factor does not appear to have any advantages. The weights in the model (2') are assumed to have been calculated from the appropriate base weights as described in the memorandum. The breakdown of continuing areas into the types listed above enables us to assume that certain sampled areas are not seriously affected by the phase-in. The classification of continuing areas as selfrepresenting, nonselfrepresenting does not appear to be important for the aggregate responses modeled in (2').

Month in sample effect: The month in sample (MIS) effect is the non-phase-in related factor in the model (2'). It is represented by the terms S_{jst}, M_{jmt}, and SM_{jsmt}. The use of two main effects and their interaction rather than a single term more closely reflects the effect of the 4-8-4 sampling pattern employed by CPS. The parameter S_{jst} represents the effect of sampling the same individuals in two consecutive years while the M_{jmt} represents the efffect of sampling the same individuals for four consecutive months each year. An SMD Memorandum from Robert Tegels entitled "CPS Redesign: Estimates and Bias Indices Calculated for the Rotation of Sample Redesign Project" (dated January 9, 1984) gives indirect evidence that the MIS effect follows the same general form in each four month period. The evidence is only indirect since the bias indices generally used to measure the MIS effect do not correspond

exactly to the MIS effect in model (2'), which is additive rather than multiplicative. However, they are useful in indicating the nature of the dependence on time in sample.

The period effect S_{jst} and the month within period effect M_{jmt} are dependent on the demographic group j and month of interest t but are assumed to be independent of the area type k. Any differences in MIS effects for different area types is accounted for by the area type-period and area type-month within period interactions in the model (2'). The period-month within period interaction SM_{jsmt} allows for differences between months within a period to differ for the two periods. For example, the difference between the first and second months in the first four month period may be different, perhaps larger, than the corresponding difference for the second period. Such differences may, in part, reflect a "carry over effect" from the first to second period in the respondents' reaction to being recycled into the survey after a rest period. If the interactions are large, the benefits of the 4-8-4 sampling pattern should be weighed against the disadvantages of a sample composed of two dissimilar groups where the disparity is related to the survey design.

The constraint $\sum_{m} \hat{c}_{.j..mt} M_{jmt} = 0$ means that for each demographic group j and month t, the month within period effects can be thought of as deviations from some average level; i.e., there is underreporting for some months within the period and overreporting for others but they balance out over all months within the period. A similar interpretation holds for $\sum_{s} \hat{c}_{.j.s.t} S_{jst} = 0$. The constraints on the interaction SM_{jsmt} are consistent with those imposed on the period and month within period effects.

An alternative approach to modeling the MIS effect would be to use the month in sample m as a covariate and replace the model (2′) by an analysis of covariance type model of the form

$$Y_{.jksmt} = \hat{c}_{.jksmt} R_{jt} + \hat{c}_{.jksmt} A_{jkt} + \hat{c}_{.jksmt} \phi_{jkst}(m) + e_{.jksmt} , \tag{4}$$

where $\phi_{jkst}(m)$ is some function of the month in sample m (m = 1, 2, 3, 4). The data from Tegels' memorandum for average bias indices by month in sample (Table 4 therein) indicates that a quadratic function

$$\phi_{jkst}(m) = \alpha_{1jkst}(m - \bar{m}) + \alpha_{2jkst}(m - \bar{m})^2 \tag{5}$$

for each four month period provides a reasonable description of the relationship.

However, the analysis of covariance model (4) with the quadratic covariate function (5) suffers from two major drawbacks. First, the model requires a quadratic function to be fit to four data points. If common coefficients were used for several sets of (j,k,s,t), a lack of fit test for the quadratic function could not be followed by higher degree polynomial model fits if it were deemed necessary.

The second, and perhaps more damaging drawback is that at crucial points in the phase-in outgoing and incoming areas do not have a full set of four values of m. For example, in November, 1984 all useable data from incoming areas is collected from individuals being interviewed for the first time; i.e., only data corresponding to s=1, m=1 is available. Without modifying the model (4) to handle these cases, the parameters cannot be uniquely estimated.

In light of these difficulties, the MIS effect will be modeled as in (2′) and analysis of covariance type models will not be pursued further.

In the analysis of variance for the model (2') it is possible to decompose the sums of squares associated with the month within period effect and the period-month within period interaction into orthogonal contrasts representing linear and quadratic trends. Significant contrasts would indicate the nature of the dependence on month within period.

Area type-month in sample interactions: The area type-month in sample interaction is represented in the model (2') by the terms AS_{jkst} and AM_{jkmt}. The three factor interaction ASM_{jksmt} is assumed to be zero so that estimators of the covariance structure of the responses can be obtained. The included interaction terms allow the MIS effect to change from one area type to another as well as from one demographic group j and month of interest t to another. The constraints correspond to those imposed on the main effects.

There is an implicit assumption in all methods of measuring the MIS effect that all rotation groups exhibit approximately the same behavior toward the survey. Although the MIS factor refers to the effect of the repeated sampling of the same individuals over time, the effects are calculated from the responses of different individuals sampled at the same point in time but who have differing numbers of previous interviews. Thus the necessity of the assumption. At certain critical points in the phase-in (such as in the last month of sampling in outgoing areas) this may not be a valid assumption across area types. Violations should manifest themselves as large interactions for the corresponding months t.

Error term: The error terms $e_{.jksmt}$ represent all sources of variation not included in the model (2') which are responsible for the deviation of the observed response from its expected value. They are random variables having mean zero and some covariance structure. They need not be uncorrelated and homoscedastic.

3. Parameter Estimation

The parameters in the model (2′) subject to the constraints (3′) can be estimated by the method of generalized least squares (GLS). Estimates of the standard errors of the GLS estimators can also be obtained.

For purposes of analysis for a particular demographic group j, all data from the first month of available data (t=1) up to and including the current month (t=T) is utilized. The responses $Y_{.jksmt}$ for month t are collected into a column vector

$$Y_{jt} = [\ Y_{.j111t}, \ldots, Y_{.j114t}, Y_{.j121t}, \ldots, Y_{.j124t}, \ldots, Y_{.j421t}, \ldots, Y_{.j424t}\ ]'.$$

The entries of Y_{jt} are grouped by area type k, period s within area type, and month m within period. Let the entire vector of responses be denoted by

$$Y_j = [\ Y_{j1}', \ldots, Y_{jT}'\ ]'.$$

Since the models for different demographic groups j have no parameters in common, there is no advantage to combining distinct demographic groups into a single analysis.

For certain months t not all entries of Y_{jt} will correspond to available data. For example, for May, 1985 (t=17 if we use January, 1984 as a reference point) in outgoing areas (k=3) the only available data is collected from individuals being interviewed for the eighth time (s=2, m=4) and in incoming areas (k=4) all individuals, including the outgoing area's rotation group replacements, are being interviewed at some point in the first set of four months (s=1, m=1, 2, 3, 4). In such cases we shall reduce the length of Y_{jt} and Y_j accordingly and include in the model (2′) only those parameters which appear in the expectation of at least one available observation. For the May, 1985 example this affects only $AS_{j31,17}$, $AS_{j42,17}$, and $AM_{j3m,17}$, m = 1, 2, 3.

Let $E(Y_j) = X_j \theta_j$ and $Cov(Y_j) = \sum_j$ represent the mean vector and covariance matrix of the reduced Y_j, respectively, where

X_j = the design matrix for the model (2′),

θ_j = the reduced parameter vector for the model (2′).

Then the GLS estimators of θ_j are those values of θ_j which minimize

$$S(\theta_j) = (Y_j - X_j \theta_j)' \sum_j^{-1} (Y_j - X_j \theta_j) \quad . \tag{6}$$

The minimum of (6) occurs when

$$\hat{\theta}_j = (X_j' \sum_j^{-1} X_j)^{-1} X_j' \sum_j^{-1} Y_j \quad . \tag{7}$$

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

$$\sum(\hat{\theta}_j) = (X_j' \sum_j^{-1} X_j)^{-1} \quad . \tag{8}$$

The calculations in equations (7) and (8) assume that $\sum_j$ is known, at least up to a multiplicative constant, or is replaced by a consistent estimator $\hat{\sum}_j$.

Technical Report 40 (pp. 93-94) describes formulas involving the estimated level $E(Y_j) = X_j \theta_j$ which provide approximate variance estimates for $E(Y_j)$. Similar methods could be used to develop approximate estimates of the off-diagonal entries of $\sum_j$ in terms of the entries of the vector $E(Y_j)$. These approximate estimates of $\sum_j$ can then be used in an iteratively reweighted least squares procedure to estimate θ_j and $\sum(\hat{\theta}_j)$. That is, replace $\sum_j$ in (6) by an identity matrix and obtain an initial estimate $\tilde{\theta}_j$ of θ_j from (7). Then calculate $\sum_j = \sum_j(\tilde{\theta}_j)$ and use it in equations (6)-(8) in place of $\sum_j$ to obtain $\hat{\theta}_j$ and $\sum(\hat{\theta}_j)$.

A more detailed discussion of the iteratively reweighted least squares procedure and methods of determining approximate formulas for the entries of $\sum_j$ can be found in a companion report on the National Crime Survey (NCS) redesign. The companion report also includes a discussion of several possible covariance structures and the assumptions associated with each structure. Since the discussions are virtually unchanged for CPS, they are not repeated here.

Under the model (2′) and the constraints (3′) the observed weighted monthly totals $Y_{.j...t} = \sum_k \sum_s \sum_m c_{.jksmt}\, Y_{.jksmt}$ are biased estimators of the corresponding true totals $C_{.j...t}\, R_{jt}$ for months t during the phase-in period. However, unbiased estimators of the rates R_{jt}, as well as estimated standard errors, can be obtained from the method of generalized least squares. In addition, it provides estimates and estimated standard errors of the MIS effects and phase-in related effects represented in the area type factor. Verification that $Y_{.j...t}$ is a biased estimator is similar to that provided in the companion NCS report and is not included here.

4. A Model for the Civilian Labor Force Level

The unemployment model (2) in Section 2 uses the CLF levels as weights. Since the true levels are unknown, they are replaced by estimates obtained from the survey, resulting in model (2′). The observed CLF levels could be used as the estimated coefficients in the model (2′). However, these observed levels are subject to both phase-in and nonphase-in related effects. As an alternative, estimates of these coefficients can be obtained from GLS applied to a linear model for CLF levels which is similar to the unemployment model (2). There are two advantages to a linear model approach; viz., the estimates of the coefficients have been adjusted for phase-in and nonphase-in related effects and the remaining model parameters are of interest in themselves.

For each sampled individual, let

$Z_{ijksmt} = 1$ if the individual is in the CLF,

$= 0$ if the individual is not in the CLF,

where the subscripts (i,j,k,s,m,t) have the same interpretation as in the definition of Y_{ijksmt}. Let w_{ijksmt} be the weight associated with the i-th sampled individual.

The weighted total number of individuals in the CLF represented by the part of the sample corresponding to an observable subscript combination (j,k,s,m,t) is given by

$$Z_{.jksmt} = \sum_{i=1}^{I} w_{ijksmt} Z_{ijksmt} \quad . \tag{9}$$

The weighted CLF totals $Z_{.jksmt}$ defined in (9) can be modeled as

$$\begin{aligned} Z_{.jksmt} = {} & w_{.jksmt} P_{jt} + w_{.jksmt} A_{jkt} + w_{.jksmt} S_{jst} + w_{.jksmt} M_{jmt} \\ & + w_{.jksmt} SM_{jsmt} + w_{.jksmt} AS_{jkst} \\ & + w_{.jksmt} AM_{jkmt} + e_{.jksmt} \quad , \end{aligned} \tag{10}$$

where

$$w_{.jksmt} = \sum_{i=1}^{I} w_{ijksmt} \quad ,$$

P_{jt} = true proportion of the j-th demographic group in the noninstitutionalized civilian population which belongs to the CLF in month t ($C_{.jksmt} = w_{.jksmt} P_{jt}$) ,

and the terms A_{jkt}, S_{jst}, M_{jmt}, SM_{jsmt}, AS_{jkst}, AM_{jkmt}, and $e_{.jksmt}$ have the same interpretation as they did in the model (3). The P_{jt} are sometimes referred to as participation rates.

The use of the area type and MIS factors notation in the model (10) actually constitutes an abuse of notation. For example, the area type effect in the model (10) represents the effect of the same phase-in related factors as the A_{jkt} in the model (2) but for a different response variable. Thus, the numerical estimates in the two models will generally differ. Since this duplicate notation should not cause any confusion, we shall continue to use it in both models.

The constraints on the parameters in the model (10) are the same as those given for the model (2) in equation (3) with the model weights $C_{.jksmt}$ replaced by the survey weights $w_{.jksmt}$.

Parameter estimation and inference in the model (10) proceeds in a straightforward fashion using the method of GLS as described in Section 3. The same type of covariance structure should be used for both models. GLS estimates of the P_{jt}, say $\hat{P}_{jt}$, are used to calculate the estimated levels $\hat{c}_{.jksmt} = w_{.jksmt}\,\hat{P}_{jt}$ required in the model (2'). Estimated standard errors of the $\hat{C}_{.jksmt}$ can also be obtained.

5. Problems and Alternatives

The weights $\hat{c}_{.jksmt}$ in the unemployment model (2') are estimated from the survey since the corresponding weights are unknown. All estimation and inference in the model (2') as described in Section 3 is conditional on the observed values. The use of estimated levels as weights is a potential source of difficulty. The possible difficulties are related to the fact that

(1) the weights $\hat{C}_{.jlsmt}$ are random variables and hence, are subject to variation. That is, the weights used in the model (2') are generally not equal to corresponding true CLF levels,

(2) the weights $\hat{C}_{.jksmt}$ are probably correlated with the response $Y_{.jksmt}$ since they are functions of $Z_{.jksmt}$ which are measured on the same individuals as the $Y_{.jksmt}$. That is, the terms in the expression for the mean response are correlated with the response and hence, with the error term.

For these reasons, the use of the estimated levels and their effect on the estimates of the parameters in the model (2') should be investigated further.

As an alternative to modeling the observed totals, the observed unemployment rate could be modeled. Let

$$p_{.jksmt} = \frac{Y_{.jksmt}}{Z_{.j...t}} \tag{11}$$

and

$$p_{.j\ldots t} = \sum_k \sum_s \sum_m p_{.jksmt} \quad , \tag{12}$$

where $Z_{.j\ldots t} = \sum_k \sum_s \sum_m Z_{.jksmt}$ is the observed total CLF for the j-th demographic group in month t. The proportion $p_{.j\ldots t}$ in (12) is the observed unemployment rate and $p_{.jksmt}$ in (11) is the contribution to the rate from the part of the sample defined by the subscript triple (k,s,m). The $p_{.jkmst}$ can be used in forming the response variable in a model similar to the model (3). They are preferable to $Y_{.jksmt} / Z_{.jksmt}$, the proportion of the j-th demographic group represented by the (k,s,m)-th part of the sample who are unemployed in month t. These latter proportions have the disadvantage that they do not sum to the overall proportion $p_{.j\ldots t}$.

Let $T(p_{.jksmt})$ be a transformation of $p_{.jksmt}$. The transformed proportion can be modeled as

$$T(p_{.jksmt}) = R_{jt} + A_{jkt} + S_{jst} + M_{jmt} + SM_{jsmt} + AS_{jkst} + AM_{jkmt} + e_{.jksmt} \quad , \tag{13}$$

where the terms in (13) have the same interpretation as in the model (2'). Three transformations are considered; others are possible.

(1) The identity transformation: $T(p_{.jksmt}) = p_{.jksmt}$.

Under the identity transformation, (13) becomes an additive model in the observed proportions. Two difficulties occur; the estimated unemployment rate $\hat{R}_{jt}$ and the predicted proportions $\hat{p}_{.jksmt}$ need not be between zero and one. For these reasons, use of the identity transformation is not advised.

(2) The logarithmic transformation: $T(p_{.jksmt}) = \ln(p_{.jksmt})$.

The logarithmic transformation is often used because it is thought of as converting a multiplicative model for the $p_{.jksmt}$ into an additive model

for the $\ln(p_{.jksmt})$. There are several difficulties associated with this transformation. First, if R^*_{jt} is the unemployment rate and R_{jt} in the model (13) is the transformed rate, then $R^*_{jt} = \exp(R_{jt})$ and the estimate of R^*_{jt} is guaranteed to be positive but will not necessarily be less than one. Thus, it is possible (but highly unlikely) that the estimated unemployment rate will be greater than 100 percent. Second, the estimated rates $\hat{R}^*_{jt} = \exp(\hat{R}_{jt})$ are biased estimators even though the $\hat{R}_{jt}$ are unbiased estimators of the transformed rated.

The remaining difficulty is one associated with any transformation. The transformation may create interactions which do not exist in the original scale or may eliminate interactions which exist in the original scale (c.f., H. Scheffe, The Analysis of Variance, Chapter 10). The effect of interactions on the original scale is of direct interest in understanding the overall effect of the phase-in on the estimates produced by CPS as well as in understanding the time in sample effect.

(3) The logit transformation: $T(p_{.jksmt}) = \ln[p_{.jksmt} / (1 - p_{.jksmt})]$.

The logit transformation has the advantage that the estimated unemployment rate will be between zero and one. However, it possesses the remaining drawbacks mentioned for the logarithmic transformation and, in addition, it does not have the intuitively appealing feature of converting a multiplicative model into an additive one.

If the model (13) is fit using GLS for either a logarithmic or logit transformation, then any observed $p_{.jksmt}$ which are zero and, for the logit transformation any observed $p_{.jksmt}$ which are one, make the response $T(p_{.jksmt})$ undefined and must be handled separately. Although $p_{.jksmt} = 1$ is highly unlikely, $p_{.jksmt} = 0$ can reasonably be expected to occur. The use of "working T's" (c.f., Finney, Statistical Methods in Biological Assay) provides one

solution to the problem. A more difficult, but much less likely problem arises if $Z_{.jksmt} = 0$. In such cases $Y_{.jksmt}$ is also zero and $p_{.jksmt}$ is undefined. A method of dealing with this problem must be specified. In contrast, models (2′) and (10) are not generally affected by zero responses; in particular, a zero response in (10) will not produce a corresponding estimated value of zero for $\hat{c}_{.jksmt}$.

An Analysis of the Effect of the Sample Redesign of the Current Population Survey on Unemployment Estimates

by

Edward Gbur
Statistical Research Division

April 1987

1. Introduction

The Current Population Survey (CPS) was redesigned to reflect population changes from the 1980 Census and to improve the efficiency of the estimators at the state level. The phase-in of the new design took place from March 1984 to November 1985. A brief description of the phase-in plan and related references is given in Gbur (1984). The report also describes an alternative approach to the problem of measuring the phase-in effects on CPS unemployment statistics. The approach is based on a linear model for the weighted total number of unemployed in a demographic group as a function of phase-in and nonphase-in related effects.

Since Gbur (1984) was written, data collected as a part of SMD's monitoring of the phase-in has been obtained from Sid Schwartz and Debbie Fenstermaker (SMD). By slightly modifying the definitions of the levels of the area type effect in the model, SMD's data could be used in conjunction with the proposed model if additional assumptions were made or if a cross-sectional rather than longitudinal analysis were carried out. A longitudinal analysis along the lines orginally described in Gbur (1984) would have been preferred. However, it was not possible to obtain an estimate of the correlation structure of the response variables during the phase-in period since only aggregate level data was available.

The purpose of this report is to describe the modifications to the linear model approach and to present the results of a cross-sectional analysis.

2. A Model for Unemployment

The model developed in Gbur (1984) contains, in addition to a parameter representing the true unemployment rate, an effect due to the rotation sampling pattern and an effect due to the

redesign phase-in. The latter was referred to as the area type effect and had four levels: continuing nondisrupted, continuing disrupted, incoming, and outgoing. To accommodate the use of SMD's data in the analysis summarized in Section 3, the two continuing type area categories were collapsed into a single category labeled as continuing area. The 4-8-4 rotation pattern in CPS was modeled by two factors; a two level factor representing the "year" or set of four months in sample and a four level factor representing the position of the month within the set of four months. The interaction of these two factors and an area type - set of months and area type - position of month within set interactions rounded out the terms in the model.

Since the response variable was the weighted number of unemployed and the parameters representing the various effects were given as rates (to avoid significant results due solely to differences in the sizes of the groups represented by particular parameters), each term in the model was expressed as a product of the appropriate rate parameter and a count of the individuals in the civilian labor force (CLF) represented by the corresponding part of the population. These latter coefficients were unknown and needed to be estimated. The CLF model used to obtain the coefficients was similar to the unemployment model and would have to be fitted to data from the same sample as the unemployment model would. Thus, the proposed approach involved two stages of model fitting on the same data, the second depending on predicted values from the first. The unanswered theoretical questions involved in this two stage approach has led us to consider models for the transformed proportion (probability) of unemployed. The particular transformation chosen was the logit transformation.

For a fixed demographic group and month t, let

p_{ijk} = proportion of the demographic group in the population represented by the sample from the k^{th} area type, j^{th} month within the i^{th} four month period who are unemployed in month t,

for the subscript ranges

$i = 1,2$

$j = 1,2,3,4$

$k = 1,2,3$ (continuing, outgoing, incoming).

As illustrated in Gbur (1984), for a given month t, not all subscript combinations (i,j,k) correspond to available data. The logit of p_{ijk} is defined as the natural logarithm of the odds ratio; i.e.,

$$\mu_{ijk} = \ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right) . \qquad (1)$$

A brief description motivating the modeling of logits as a function of a set of independent variables is given in Appendix 1.

The logit model corresponding to the model for the unemployment level in Gbur (1984, Section 2) is given by

$$\mu_{ijk} = R + S_i + M_j + SM_{ij} + A_k + SA_{ik} + MA_{jk} , \qquad (2)$$

where

R = mean of the logits for all (i,j) combinations from area type 1 (continuing areas),

A_k = effect on the logit due to sampling individuals from from area type k,

S_i = effect on the logit due to sampling individuals from the i^{th} four month period in the rotation pattern,

M_j = effect on the logit due to sampling individuals for the j^{th} month within a four month period in the rotation pattern,

SM_{ij} = effect on the logit due to the period - month within period interaction,

SA_{ik} = effect on the logit due to the area type - period interaction,

MA_{jk} = effect on the logit due to the area type - month within period interaction.

As before, the parameters in the model (2) are subject to the following contraints:

$$A_1 = 0 , \qquad S_1 + S_2 = 0 , \qquad \sum_{j=1}^{4} M_j = 0 ,$$

$$SM_{1j} + SM_{2j} = 0 \quad \text{for all } j , \qquad \sum_{j=1}^{4} SM_{ij} = 0 \quad \text{for all } i, \qquad (3)$$

$$SA_{i1} = 0 \quad \text{for all } i , \qquad SA_{1k} + SA_{2k} = 0 \quad \text{for all } k>1 ,$$

$$MA_{j1} = 0 \quad \text{for all } j , \qquad \sum_{j=1}^{4} MA_{jk} = 0 \quad \text{for all } k>1 .$$

Except for the change in the levels of the area type factor discussed above, comments concerning each term in the model (2) and the constraints (3) are essentially unchanged from those found in Gbur (1984) when references to "rates" are changed to "logits". Expressions and interpretations for each of the effects in the model in terms of the logits μ_{ijk} are given in Appendix 2.

If we assume independent binomial sampling in each cell, the logit model (2) or any hierarchical submodel can be fitted using maximum likelihood techniques. The statistical packages BMDP and SPSS both allow for parameter estimation in the model subject to the constraints (3). By fitting a sequence of nested models each term in (2) can be tested for significance using likelihood ratio statistics which are asymptotically distributed as chi-square random variables.

However, the estimated proportions used to calculate the observed logits in our model are based on weighted totals obtained from survey weights derived from the complex sample design and other adjustments. It has been shown (cf., Kumar and

Rao, 1984) that under these conditions, likelihood ratio tests based on standard methods for binominal sampling have inflated significance levels. Hence, corrective action must be taken to ensure a valid analysis.

Asymptotically valid methods have been developed using Wald type statistics (e.g., Koch et al. 1975). Such methods require individual data records or an estimate of the entire convariance matrix of the estimated cell proportions. In our application, neither of these were available for the entire phase-in period; only weighted cell totals suitable for secondary analysis were available. For such situations approximations to the asymptotic distributions of the likelihood ratio statistics have been proposed (e.g., Binder et al. 1984). The exact form of the approximation varies with the model and the necessary theory has been developed for many of the common models. For log-linear models which admit a direct solution of the likelihood equations under multinomial sampling, only knowledge of the individual cell design effects is used. Unfortunately, logit models are not generally in this class. Recently, Rao and Scott (1987) have obtained simple upper bounds on the approximation for logit models in terms of the cell design effects and certain generalized design effects which do not depend on any hypothesis. An outline of the necessary adjustments is given in Appendix 3.

3. A Cross-Sectional Logit Analysis

An initial cross-sectional analysis was carried out using the modified SMD data. Logit models of the form (2) were fit to each month's data for total unemployment. At this point it became apparent that for some months prior to the redesign and in the early stages of the redesign where the effect would presumably be minimal, there were differences in the proportion unemployed due to the area type (continuing versus outgoing). Further investigation revealed that, in fact, there are inherent differences between continuing and outgoing areas. The latter

tend to be rural or small metropolitan areas whereas the former are generally the more populated metropolitan areas. Thus, the effect of the redesign and the inherent differences between continuing and noncontinuing areas are confounded in the area type factor in the model.

Despite this confounding it may still be possible to obtain some information about the effect of the redesign. Suppose that a pattern can be found for the difference between the effect of the continuing and noncontinuing areas on the proportion unemployed in the months prior to the redesign and during its initial phase. Then these differences could be used as a baseline for comparison of the area type effects during the remainder of the redesign. If we assume that the effect of the inherent differences between the two area types is approximately constant over a period extending from immediately before the redesign until after its completion, then any deviation from the baseline pattern would be attributed to the effect of the redesign. This approach was investigated using the data from January through October, 1984 to establish the baseline.

For each of the ten months a sequence of nested logit models was fit to the data for total unemployed using the SPSS-X procedure LOGLINEAR. Selected portions of the output were saved and used as input into a FORTRAN program which computed a complex sample design correction to the χ^2 goodness of fit and test of a significant effect statistics. A diagonal covariance matrix was used in the correction factor calculation with variances obtained from a generalized variance function having the same coefficients as were used for the published CPS data during this period. The assumption of uncorrelated estimates within each month is reasonable since the rotation groups represent (at least approximately) independent subsamples of the CPS sample.

The p-values for the adjusted χ^2 tests are given in Table 1. A large p-value indicates that the effect of the corresponding factor on the logit of the proportion unemployed is not significantly different from zero. Since the existence of a

rotation group bias has been firmly established on the proportion scale and the logit transformation will not eliminate this effect, it has been partially forced into the model by automatically including the term M_j in the model for each month. Assuming the hierarchy principle (if an interaction has been included in a model, then all lower order terms involving the factors in the interaction must also be incuded), the resulting logit models are given in Table 2.

Table 1. p-values for the Corrected χ^2 Tests for Siginificant Effects in the Logit Model (2) for Total Unemployed

	Month (1984)									
Effect	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.
S × A	0.58	0.05	0.07	0.02	*	0.31	0.83	0.79	*	0.83
M × A	0.95	0.94	0.68	0.67	0.32	0.94	0.77	0.85	0.43	0.20
A	0.04	*	0.01	*	0.02	0.16	0.54	0.59	0.95	0.41
S × M	0.05	0.52	0.08	0.17	0.17	0.70	0.78	0.30	0.61	0.09
S	0.43	0.47	0.76	0.95	0.70	0.58	0.57	0.75	0.21	0.17

* p-value is less than 0.005.

From Tables 1 and 2 it is clear that a number of different models are appropriate, depending on the month in the baseline period. Moreover, the area type effect is not significant in every month and the form of the rotation sampling effect varies over time. These results are not totally unexpected. In past studies of the rotation group bias in CPS (e.g., Bailar (1975) and Bailar (1979)), estimates of the bias indices have been based on averages over a number of months of data. In addition, Bailar's (1979) numerical work suggests that the bias may be a function of time. Given this evidence and the results in Tables 1 and 2, a cross-sectional approach to providing a baseline for area type comparisons did not produce the desired results.

Table 2. Fitted Logit Models for Total Unemployed

Month (1984)	Model
January	$\hat{\mu}_{ijk} = \hat{R} + \hat{S}_i + \hat{M}_j + \hat{SM}_{ij} + \hat{A}_k$
February	$\hat{\mu}_{ijk} = \hat{R} + \hat{S}_i + \hat{M}_j + \hat{A}_k + \hat{SA}_{ik}$
March	$\hat{\mu}_{ijk} = \hat{R} + \hat{S}_i + \hat{M}_j + \hat{SM}_{ij} + \hat{A}_k + \hat{SA}_{ik}$
April	$\hat{\mu}_{ijk} = \hat{R} + \hat{S}_i + \hat{M}_j + \hat{A}_k + \hat{SA}_{ik}$
May	$\hat{\mu}_{ijk} = \hat{R} + \hat{S}_i + \hat{M}_j + \hat{A}_k + \hat{SA}_{ik}$
June	$\hat{\mu}_{ijk} = \hat{R} + \hat{M}_j$
July	$\hat{\mu}_{ijk} = \hat{R} + \hat{M}_j$
August	$\hat{\mu}_{ijk} = \hat{R} + \hat{M}_j$
September	$\hat{\mu}_{ijk} = \hat{R} + \hat{M}_j$
October	$\hat{\mu}_{ijk} = \hat{R} + \hat{S}_i + \hat{M}_j + \hat{SM}_{ij}$

4. Comments and Conclusions

The cross-sectional analysis described in the previous section avoided the need for an estimate of the correlation structure required for a longitudinal analysis. However, it failed to provide even the baseline estimates necessary to unravel the confounding of the redesign effect and the effect of the inherent differences between continuing and noncontinuing areas. The next logical step is the construction and fitting of a longitudinal model for the logit of the proportion unemployed.

Two major obstacles to a longitudinal approach are the need for an estimate of the correlation structure at the detailed level specified in the model and the need to account for (potential) trend and seasonal effects. Successful resolution of the latter may also remove any time dependency in the rotation group effects and allow for a stable baseline pattern for the area type effect. Huang and Ernst (1981) contains estimates of

the correlation structure at the rotation group level for CPS. However, to make use of these estimates we would have to assume that (i) the correlation structure is unchanged over time, (ii) it is approximately the same at the more detailed level of the model, and (iii) it is not affected by the redesign. Despite these problems, a longitudinal approach should be attempted.

References

Agresti, A. (1984). **Analysis of Ordinal Categorical Data.** New York, John Wiley and Sons.

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. **JASA** 70, 23-30.

Bailar, B.A. (1979). Rotation sample biases and their effects on estimates of change. **Proc. Inter. Stat. Inst.** 48(2), 385-407.

Binder, D.A., M. Gratton, M.A. Hidiroglou, S. Kumar, and J.N.K. Rao (1984). Analysis of categorical data from surveys with complex designs: Some Canadian experiences. **Survey Meth.** 10, 141-156.

Gbur, E. (1984). A linear model approach to the estimation of the redesign effects in the Current Population Survey. Internal Census Bureau report.

Huang, E. and L. Ernst (1981). Comparison of an alternative estimator to the current composite estimator in CPS. **Proc. of Section on Survey Research Methods**, ASA. 303-308.

Jensen, D.R. and H. Solomon (1972). A Gaussian approximation to the distribution of a definite quadratic form. **JASA** 67, 898-902.

Johnson, N.L. and S. Kotz (1968). Tables of the distribution of positive definite quadratic forms in central normal variables. **Sankhya B** 30, 303-314.

Koch, G.G., D.H. Freeman and J.L. Freeman (1975). Strategies in the multivariate analysis of data from complex surveys. **Inter. Stat. Rev.** 43, 59-78.

Kumar, S. and J.N.K. Rao (1984). Logistic regression analysis of Labour Force Survey data. **Survey Meth.** 10, 62-81.

Rao, J.N.K. and A.J. Scott (1987). On simple adjustments to chi-square tests with sample survey data. **Annals of Stat.** 15, 385-397.

Roberts, G. (1984). On chi-square tests for logit models with cell proportions estimated from survey data. Unpublished manuscript.

Appendix 1

The development of logit models in this appendix follows Agresti (1984, Chapter 6). Let y be a binary random variable taking on values 0 and 1 with probability 1-p and p, respectively. Then

$$E[y] = P(y=1) = p .$$

Suppose that p is a function of a vector of explanatory variables x. A linear relationship of the form

$$p(x) = \beta_0 + \sum_{i=1}^{I} \beta_i x_i$$

is generally considered inappropriate since predicted values of p can be outside the interval [0,1] unless the ranges of the x_i are restricted. Curvilinear relationships between p and x are usually considered more appropriate.

For a single explanatory variable (I = 1) a sigmoidal curve is a natural shape for monotone relationships between p and x. One of many functions which has this general shape is the logistic function

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = [1 + e^{-(\beta_0 + \beta_1 x)}]^{-1} . \qquad (A1)$$

Under the logistic assumption, the odds ratio for the response y = 1 to the response y = 0 is

$$p(x) / (1 - p(x)) = e^{\beta_0 + \beta_1 x} \qquad (A2)$$

so that the logarithm of the odds ratio, or logit, is given by

$$\text{logit}(p(x)) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x \quad , \tag{A3}$$

where ln represents the natural logarithm.

In the simple logit model (A3), if β_1 is positive, then the logit of p(x) is an increasing function of x. Equivalently, the odds ratio increases multiplicatively by $\exp(\beta_1)$ for every unit increase in x. On the original probability scale ,

$$p(x+1) = \frac{1 + e^{\beta_0 + \beta_1 x}}{e^{-\beta_1} + e^{\beta_0 + \beta_1 x}} \cdot p(x) \quad ,$$

where the multiplier on the right hand side is greater than one since $\beta_1 > 0$ implies $\exp(-\beta_1) < 1$. Hence, on the probability scale a unit increase in x leads to an increase in p(x) which depends nonlinearly on the value of x itself.

The independent variables in the logit model can be either continuous or discrete. Letting the x's be 0-1 indicator variables yields the analysis of variance-like model which is used in this report.

The preceding development can be generalized to the case where y is expressed in terms of sampling weights and adjustment factors. In this case, p(x) is thought of as the proportion of individuals in the population with the characteristic of interest and the specified set of x values. It is this extension which is utilized in this report.

Appendix 2

In this appendix we develop expressions for each term in the logit model (2) subject to the constraints (3) given in Section 2 as a function of the logits μ_{ijk} . We begin by observing that for

k=1 (continuing areas)

$$\mu_{ij1} = R + S_i + M_j + SM_{ij} \qquad i = 1,2, \quad j = 1,2,3,4 \qquad \text{(A4)}$$

and for k = 2, 3 (outgoing and incoming areas, respectively) ,

$$\begin{aligned} \mu_{ijk} &= R + S_i + M_j + SM_{ij} + A_k + SA_{ik} + MA_{jk} \\ &= \mu_{ij1} + A_k + SA_{ik} + MA_{jk} , \end{aligned} \qquad \text{(A5)}$$

for appropriate (i,j) .

The simplest expressions in terms of the μ_{ijk} occur in the first stage of the phase-in from April through October 1984. During this period, the new frame was introduced in continuing areas, all outgoing areas were sampled, and no data for estimation purposes was collected from incoming areas (k=3). In this case, from (A4) and the constraints (3) we obtain

$$R = \frac{1}{8}\sum_i \sum_j \mu_{ij1} = \mu_{..1} , \qquad \text{(A6)}$$

where a dot as a subscript indicates on average over that subscript. Substituting for R from (A6), fixing one subscript and summing over the other in (A4) yields

$$S_i = \mu_{i.1} - \mu_{..1} , \qquad \text{(A7)}$$

$$M_j = \mu_{.j1} - \mu_{..1} . \qquad \text{(A8)}$$

Substituting (A6) - (A8) into (A4) yields

$$SM_{ij} = \mu_{ij1} - \mu_{i.1} - \mu_{.j1} + \mu_{..1} . \qquad \text{(A9)}$$

The expressions in (A6) - (A9) correspond to the usual analysis of variance type formulas for the parameters in a two factor model with interaction if the third subscript is ignored. Thus, for example, R is the overall mean of the logits for continuing areas and S_i is the deviation of the mean over the

months within the i^{th} period from the continuing areas from the overall mean of the logits for the continuing areas.

Summing (A5) over i and j and solving for A_2 gives

$$A_2 = \mu_{..2} - \mu_{..1} \cdot \tag{A10}$$

Again from (A5) we have

$$\sum_i \mu_{ij2} = 2R + 2M_j + 2A_2 + 2MA_{j2} \quad ,$$

which, upon substituting, rearranging, and simplifying, yields

$$\begin{aligned} MA_{j2} &= (\mu_{.j2} - \mu_{..2}) - (\mu_{.j1} - \mu_{..1}) \\ &= (\mu_{.j2} - \mu_{.j1}) - (\mu_{..2} - \mu_{..1}) \\ &= (\mu_{.j2} - \mu_{.j1}) - A_2 \quad . \end{aligned} \tag{A11}$$

Similarly,

$$\begin{aligned} SA_{i2} &= (\mu_{i.2} - \mu_{..2}) - (\mu_{i.1} - \mu_{..1}) \\ &= (\mu_{i.2} - \mu_{i.1}) - A_2 \quad . \end{aligned} \tag{A12}$$

From (A10), the outgoing area type effect A_2 is the difference between the means of the logits for outgoing and continuing areas. Hence, it measures the effect of the outgoing areas relative to those which are retained in the sample. The month within period - area type interaction in (A11) can be interpreted in two ways. First, it is the difference in the deviations of the mean logits for the j^{th} month within period from the overall mean from the outgoing and continuing areas. Alternatively, it is the difference in mean logits for the j^{th} month within period for the two area types, adjusted for the area type effect A_2 . The period - area type interaction SA_{i2} can be interpreted analogously.

Finally, substituting (A6) - (A8) and (A10) - (A12) into (A5) and solving for SM_{ij} yields

$$SM_{ij} = \mu_{ij2} - \mu_{i.2} - \mu_{.j2} + \mu_{..2} , \qquad (A13)$$

which is of the same form as (A9) except that it is based on the logits for the outgoing areas rather than the continuing areas. This result is as expected since the lack of a three way period - month within period - area type interaction in the model means that the two way interaction between period and month within period is the same for all area types. From a different perspective, if a three way interaction SMA_{ijk} and the corresponding constraints were included in the model, then it can be shown that

$$SMA_{ij2} = (\mu_{ij2} - \mu_{i.2} - \mu_{.j2} + \mu_{..2}) - (\mu_{ij1} - \mu_{i.1} - \mu_{.j1} + \mu_{..1}) , \qquad (A14)$$

which is the difference between (A13) and (A9). In a similar manner, alternative expressions for SA_{i2} and MA_{j2} can be shown to be equivalent by rearranging the terms on th right hand side of (A14).

For the remaining months of the phase-in there are missing (i,j) cell combinations for both outgoing and incoming areas (k=2,3). As a result, the sums in (A10) - (A12) are generally not taken over the full range and the corresponding expressions, although conceptually the same, do not simplify to those given above.

Appendix 3

In this appendix the likelihood ratio statistics are developed for both the goodness of fit and significance of an effect hypotheses in the logit model. The general approach used to modify the likelihood ratio statistics to account for the complex sample design is outlined and then applied to the logit

model.

To establish notation, assume that there are I cells in the model and let a plus sign as a subscript indicate a sum over that subscript. Let

N_i = population size of cell i (unweighted),

$N_+ = \sum_i N_i$ = population size,

N_{i1} = number in cell i with the characteristic of interest,

$N_{+1} = \sum_i N_{i1}$ = number in the population in the characteristic of interest,

n_i = sample size for cell i (unweighted),

$n_+ = \sum_i n_i$ = overall sample size,

$\pi_i = \frac{N_{i1}}{N_i}$ = population proportion in cell i with the characteristic,

y_i = sample number in cell i with the characteristic (unweighted),

$q_i = \frac{y_i}{n_i}$ = sample proportion estimate of π_i (unweighted),

$\hat{p}_i = \frac{\hat{N}_{i1}}{\hat{N}_i}$ = weighted sample proportion estimate of π_i (survey estimates of π_i),

$\hat{w}_i = \frac{\hat{N}_i}{\hat{N}_+}$ = weighted sample proportion estimates of $\frac{N_i}{N_+}$.

Let

$$\pi_i = f_i(\beta)$$

$$= e^{x_i\beta} / (1 + e^{x_i\beta}) \quad , \qquad i=1,\ldots,I$$

so that

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = x_i\beta \quad ,$$

where x_i is the i^{th} row of the model design matrix X and β is the K dimensional parameter vector. In matrix form the model is given by

$$\mu = \text{logit}(\pi) = X\beta \quad . \tag{A15}$$

If we assume independent binomial sampling in each cell, then the likelihood function is

$$L(\beta) = \prod_{i=1}^{I} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

and the log-likelihood is

$$\ln L(\beta) = c + \sum_i [y_i \ln(\pi_i) + (n_i - y_i) \ln(1-\pi_i)]$$

$$= c + \sum_i n_i [q_i \ln(\pi_i) + (1-q_i) \ln(1-\pi_i)] \tag{A16}$$

or in terms of β ,

$$\ln L(\beta) = c + \sum_i n_i \left[q_i \ln\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right) + (1-q_i) \ln\left(\frac{1}{1+e^{x_i\beta}}\right)\right]$$

$$= c + \sum_i n_i [q_i x_i\beta - \ln(1+e^{x_i\beta})] \quad . \tag{A17}$$

Differentiation with respect to β_j yields

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{I} n_i \left[q_i x_{ij} - \left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right) x_{ij} \right]$$

$$= \sum_i n_i x_{ij} (q_i - \pi_i) \quad , \qquad j=1,\ldots K \quad .$$

Setting the partial derivatives equal to zero and writing the

resulting system of likelihood equations in matrix form yields

$$X'y - X'(n\hat{\pi}) = 0 \ , \tag{A18}$$

where $(n\hat{\pi}) = [n_1\hat{\pi}_1,\ldots,n_I\hat{\pi}_I]'$. Solution of the likelihood equations (A18) leads to the fitted model

$$\hat{\mu} = \text{logit}(\hat{\pi}) = X\hat{\beta} \ .$$

To test the significance of a particular effect in the model (A15), for convenience rearrange the order of the entries in β (and the corresponding columns of X) so that the parameters associated with the effect being tested are the last entries in the vector β. Then let $\beta = [\beta_1'\beta_2']$ and partition $X = [X_1 X_2]$ correspondingly. The model (A15) becomes

$$\mu = X_1\beta_1 + X_2\beta_2 \ . \tag{A19}$$

The likelihood ratio test statistic for H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$ is given by

$$\begin{aligned} G^2(2|1) &= 2\ln[L(\hat{\beta})/L(\hat{\hat{\beta}})] \\ &= 2n_+\sum_i(\frac{n_i}{n_+})\{q_i \ln[\hat{\pi}_i/\hat{\hat{\pi}}_i] \\ &\quad + (1-q_i)\ln[(1-\hat{\pi}_i)/(1-\hat{\hat{\pi}}_i)]\} \ , \end{aligned} \tag{A20}$$

where $\hat{\beta}$ and $\hat{\hat{\beta}}$ are the maximum likelihood estimators obtained from (A18) under H_1 and H_0, respectively.

In general, when q_i and n_i/n_+ are not available, they will be replaced by the survey estimates $\hat{p}_i$ and $\hat{w}_i$, respectively. This assumes that the unweighted proportion in cell i with the characteristic is approximately the same as the weighted estimate and that the distribution of sample units over the cells is approximately the same as the distribution of final person weights over the cells. Making these substitutions in (A20)

yields

$$G^2(2|1) \cong 2n_+ \sum_i \hat{w}_i \{p_i \ln[\hat{\pi}_i/\hat{\hat{\pi}}_i] + (1-\hat{p}_i) \ln[(1-\hat{\pi}_i)/(1-\hat{\hat{\pi}}_i)]\} \quad . \tag{A21}$$

Under the hierarchy assumption, expression (A21) for G^2 is equivalent to

$$G^2(2|1) \cong 2n_+ \sum_i \hat{w}_i \{\hat{\pi}_i \ln[\hat{\pi}_i/\hat{\hat{\pi}}_i] + (1-\hat{\pi}_i) \ln[(1-\hat{\pi}_i)/(1-\hat{\hat{\pi}}_i)]\} ,$$

which is the more familar form of the likelihood ratio statistic.

For the goodness of fit likelihood ratio test we have $\hat{p}_i$ in place of $\hat{\pi}_i$ and, denoting the maximum likelihood estimates under H_0: "model fits" by $\hat{\pi}_i$, $\hat{\pi}_i$ in place of $\hat{\hat{\pi}}_i$, the statistic is

$$G^2 \cong 2n_+ \sum_i \hat{w}_i \{\hat{p}_i \ln[\hat{p}_i/\hat{\pi}_i] + (1-\hat{p}_i)\ln[(1-\hat{p}_i)/(1-\hat{\pi}_i)]\} \quad . \tag{A22}$$

Under independent binomial sampling and the appropriate null hypothesis, both of the likelihood ratio statistics (A21) and (A22) are asymptotically chi-square with the appropriate degrees of freedom. However, for general sample designs, the likelihood ratio statistics are asymptotically distributed as $\sum \delta_i U_i$, where the U_i are independent $\chi^2_{(1)}$ and the δ_i are the eigenvalues of certain "design effect" matrices. The basic approach for obtaining these matrices and the resulting modification of the statistics is sketched below.

The usual goodness of fit statistic, Wald type statistics, and the likelihood ratio statistic for testing both the goodness of fit and significance of an effect hypotheses have been shown to be asymptotically equivalent. Hence, any of them can be used to derive their common asymptotic distribution. The most convenient is usually a Wald statistic.

The basic computation relies on the following theorem (e.g., Graybill, 1976, Section 4.4).

Theorem: If $X \sim MVN(0,V)$ and $Q = X'AX$, where A is a symmetric matrix, then $Q \sim \sum w_i U_i$, where U_i are independent $\chi^2_{(1)}$ and w_i are eigenvalues of the matrix AV.

The theorem is usually applied to an appropriate Wald statistic W (Q in the theorem) constructed from the maximum likelihood estimators of the parameter being tested in the null hypothesis and the inverse of their estimated covariance matrix under multinomial sampling. If the vector of maximum likelihood estimators is asymptotically normal with a sample design dependent covariance matrix (V in the theorem), then the asymptotic distribution of W follows. Note that the "design effect" matrix (AV in the theorem) is a generalization of the univariate design effect concept; i.e., the ratio of the variance of the estimator under the given sample design to that under simple random sampling.

For the logit model (A19), the Wald statistic for testing H_0: $\beta_2 = 0$ is given by (cf., Binder et al., 1984 or Kumar and Rao, 1984)

$$W = n_+ \hat{\beta}_2' (\tilde{X}_2' \hat{D}_{\pi(1-\pi)} \tilde{X}_2) \hat{\beta}_2 , \qquad (A23)$$

where $\hat{\beta}_2$ is the maximum likelihood estimator under the general model and

$$\hat{D}_{\pi(1-\pi)} = \text{diag}[\hat{\pi}_i(1-\hat{\pi}_i)] ,$$

$$\tilde{X}_2 = [I - X_1(X_1' \hat{D}_{\pi(1-\pi)} X_1)^{-1} X_1' \hat{D}_{\pi(1-\pi)}] X_2 .$$

Thus, the generalized design effects matrix is given by

$$M_{\beta_2=0} = (\tilde{X}_2' \hat{D}_{\pi(1-\pi)} \tilde{X}_2)^{-1} (\tilde{X}_2' \hat{V} \tilde{X}_2) , \qquad (A24)$$

where $\hat{V}$ is the estimated covariance matrix of $\hat{\pi}$ under the given sample design.

For the goodness of fit hypothesis, the likelihood ratio statistic G^2 is asymptotically equivalent to the goodness of fit

statistic

$$W = n_+ \sum_{i=1}^{I} \hat{w}_i(\hat{p}_i - \hat{\pi}_i)^2/[\hat{\pi}_i(1-\hat{\pi}_i)] \ , \tag{A25}$$

where, as in (A18), $\hat{\pi}_i$ is the maximum likelihood estimator of π_i under H_0. The generalized design effects matrix is given by

$$\begin{aligned} M_{gof} &= \frac{1}{\hat{N}_+} [(\hat{D}_{N_+} \tilde{X} \hat{D}_{N_+}^{-1})' \hat{D}_{N_+\pi(1-\pi)}(\hat{D}_{N_+} \tilde{X} \hat{D}_{N_+})^{-1}]^{-1} \\ &\quad \cdot [(\hat{D}_{N_+} \tilde{X} \hat{D}_{N_+}^{-1})' \hat{V}(\hat{D}_{N_+} \tilde{X} \hat{D}_{N_+}^{-1})] \\ &= \hat{D}_w \tilde{X} \hat{D}^{-1}_{N_+\pi(1-\pi)}(\hat{D}_{N_+} \hat{V} \hat{D}_{N_+}) \tilde{X} \hat{D}_{N_+}^{-1} \ , \end{aligned} \tag{A26}$$

where

$$\hat{D}_{N_+} = \text{diag}[\hat{N}_i] \ ,$$

$$\hat{D}_w = \text{diag}[\hat{w}_i] \ ,$$

$$\hat{D}_{N_+\pi(1-\pi)} = \text{diag}[\hat{N}_i \hat{\pi}_i(1-\hat{\pi}_i)] \ ,$$

$$\tilde{X} = I - X(X' \hat{D}_{N_+\pi(1-\pi)} X)^{-1} X' \hat{D}_{N_+\pi(1-\pi)} \ ,$$

and $\hat{V}$ is the estimated covariance matrix of $\hat{p}$ under the given complex sample design.

Percentile points of the distribution of $\sum \delta_i U_i$ under H_0 are generally not available. Johnson and Kotz (1968) and Jensen and Solomon (1972) contain selected percentiles for up to five summands. Approximate values can be obtained using the usual chi-square tables if we approximate the distribution of $\sum \delta_i U_i$ by a scalar multiple of a chi-square, say $a\chi^2_{(b)}$, where a and b are obtained by matching moments of the two distributions. If only the means are equated, then it can be shown that

$$\begin{aligned} a &= \frac{1}{d} \sum_{i=1}^{d} \delta_i \ , \\ b &= d \ , \end{aligned} \tag{A27}$$

where d is the appropriate number of degrees of freedom. If both the means and variances are equated, then

$$a = \left(\sum_{i=1}^{d} \delta_i^2\right)/\left(\sum_{i=1}^{d} \delta_i\right) ,$$

$$b = \left(\sum_{i=1}^{d} \delta_i\right)^2/\left(\sum_{i=1}^{d} \delta_i^2\right) . \qquad \text{(A28)}$$

Since

$$\sum_{i=1}^{d} \delta_i = \text{trace}(M) ,$$

$$\sum_{i=1}^{d} \delta_i^2 = \text{trace}(M^2) ,$$

where M is the appropriate design effects matrix, the individual δ_i's need not be calculated to obtain values for a and b in (A27) and (A28).

Roberts (1984) obtained the following expressions for $\sum \delta_i$ under (A27). For the goodness of fit hypothesis,

$$\sum_{i=1}^{d} \delta_i = n_+ \sum_{i=1}^{I} \hat{w}_i \hat{V}_{ii}(r)/[\hat{\pi}_i(1-\hat{\pi}_i)] , \qquad \text{(A29)}$$

where $\hat{V}_{ii}(r) = \text{var}(\hat{p}_i - \pi_i)$ is the i^{th} diagonal element of the matrix $A\hat{V}A'$ with

$$A = I - \hat{D}_{\pi(1-\pi)}X(X'\hat{D}_{w\pi(1-\pi)}X)^{-1} X'\hat{D}_w .$$

For testing H_0: $\beta_2 \neq 0$,

$$\sum_{i=1}^{d} \delta_i = n_+ \sum_{i=1}^{I} \hat{w}_i \tilde{V}_{ii}(r)/[\hat{\hat{\pi}}_i(1-\hat{\hat{\pi}}_i)] , \qquad \text{(A30)}$$

where $\tilde{V}_{ii}(r) = \text{var}(\hat{\pi}_i - \hat{\hat{\pi}}_i)$ is the i^{th} diagonal element of the matrix $\tilde{V} = \hat{D}_{\pi(1-\pi)}\tilde{X}_2 B\, \tilde{X}_2'\, \hat{D}_{\pi(1-\pi)}$ with

$$B = (\tilde{X}_2'\, \hat{D}_{w\pi(1-\pi)}\tilde{X}_2)^{-1}\, \tilde{X}_2\hat{D}_w\hat{V}\, \hat{D}_w(\tilde{X}_2'\, \hat{D}_{w\pi(1-\pi)}\tilde{X}_2)^{-1} .$$

In summary, hypothesis testing for goodness of fit and significance of an effect in the logit model (A15) is modified to account for the complex sample design by calculating the usual likelihood ratio statistic (A22) or (A21) and comparing the observed value to critical values of a multiple of a chi-square, where the multiplier is obtained from (A29) or (A30).