Comparison of Alternative Model-Based Estimators for the Proportion of Housing Units Victimized by Crime During a Year

Charles H. Alexander, U.S. Census Bureau

I. INTRODUCTION AND SUMMARY

This paper considers the problem of a random sample, for most of whose elements two random variables are observed. In some cases only one of the two variables is observed. The goal is to make certain inferences about the "population", without knowing for sure what reasonably may be assumed about the similarity of the completely observed and partially observed cases.

Specifically, a sample of housing units receive two interviews covering consecutive six-month time periods. The purpose is to estimate the proportion of units in the population which experienced some particular type of crime during the twelve-month period covered by the two interviews. The following results, given as proportions, were obtained for the crime of burglary from the U.S. National Crime Survey for interviews during 1980.

TABLE 1
Reported Burglary in Second Interview

		No	Yes	
Reported Burglary in 1st Interview	No	.931499	.032916	.964415
	Yes	.031947	.003639	.035586
		.963446	.036555	1.000001

Let "Group A" refer to this group of units with two completed interviews. The entries in this table will be denoted by $p_A(i,j)$; $i=0,1$, $j=0,1$; where i pertains to the first interview and j to the second. The corresponding population proportions will be denoted by $\pi_A(i,j)$. Based on these Group A housing units, for which two interviews were completed, it would be estimated that $1-p_A(0,0)=.068501$, or 6.9% of the units in the population were "touched by burglary" during a twelve-month period.

However, for some housing units in the sample, one of the two interviews may be missing. The interest is then in what would have been reported on the other interview. For example, the following results were obtained for those units (Group C) whose first interview was missing.

TABLE 2
Reported Burglary in Second Interview

		No	Yes	
Reported Burglary in 1st Interview	No	$p_C(0,0)$	$p_C(0,1)$	$p_C(0,\cdot)$
	Yes	$p_C(1,0)$	$p_C(1,1)$	$p_C(1,\cdot)$
		.959426	.040574	1.00000

The problem is to estimate $1-\pi_C(0,0)$ using information from Group A, but keeping in mind that Groups A and C may have different characteristics. Group B (those with missing second interview) is treated similarly.

This paper compares several approaches to this problem. Section II describes four alternative estimators which may be calculated using the known values shown in the above tables, based on alternative assumptions about the relationships between the groups. Section III describes an alternative approach, which stems from the work of Eddy, et al (1981). Data are collected on the proportion of housing units which report crimes in each three-month quarter of the twelve-month period. The idea is to assume a specific family of probability distributions for the joint distribution of random variables which indicate victimization in each quarter. It is assumed that the distributions for Groups A, B, and C are members of the same family, but may have different parameters. Maximum likelihood estimates of the "touched-by-crime" probability for the missing data groups are calculated under the assumed model. This approach not only uses more information, but permits a test of the model.

These model-based estimators can be very sensitive to serious lack of fit in the assumed model (Alexander and Roebuck (1983)). Consequently it is important to test the fit of these models. In Sections IV, V, and VI, three of the models are tested for Group A and, to the extent possible, for Group C. In Section VII, the models are used to study the behavior of the simpler estimators of Section II, assuming that one of the models applies. Section VIII contains a discussion of the results.

The outline of this work was presented in Alexander and Roebuck (1983). The data for testing the fit of the various models have now been obtained from the NCS Public Use File at the University of Michigan. The models proposed earlier do not fit the data in certain important respects. This appears to be because the models have not taken into account certain response error patterns which are known to be present in the NCS survey data. This calls for further work. However, the results of Section VII indicate that some of the estimators in Section II may eventually prove to be slightly superior to the present published estimator.

The problem has special features as it applies to the NCS. The proportions in Table 1 are for a combination of six different twelve-month periods, with the first interviews taking place in the months of January to June 1980. (For example, the group with first interviews in January was also interviewed in July and the two interviews covered the period July 1979- June 1980.) As with the usual NCS estimated crime rates, the estimated proportions are calculated using weights equal to the unit's inverse probability of selection, with various other adjustments. For these estimates, Group A consisted of about 43,000 units; Groups B and C contained about 11,000 and 13,000 respectively. Each of the latter groups contains about 2500 refusals or other noninterviews of eligible units. Of the remaining cases, about 8500 in Group B are "outgoing" units which are getting their last scheduled interview and are about to be replaced by roughly 8500 "first bounded interview" units new to the survey, which are included in Group C. The remaining 2000 units in Group C are primarily those which missed an interview due to an experiment which ended in early 1980.

It is likely that Groups B and C differ from Group A and from each other. The new housing units getting their first bounded interview tend to report crime at a somewhat higher rate than the average unit, and units which are in for their final visit have a slightly lower rate. Units which have a refusal on one interview may have different characteristics than the population as a whole. In addition, some units in Group C, which were vacant or refused interview at the previous visit, will have a reference period which is not bounded by a previous interview, so that they may report extra crimes which occurred before the six months of interest. Such unbounded interviews may also occur in Group A, when a household moves from a sample address and a new household moves in, but there are probably more unbounded units in Group C.

For simplicity of exposition, membership in Group A, B, or C is described as a fixed characteristic of each unit in the population. It may be more realistic to think of the group as a random characteristic, so that, for example, $\pi_A(i,j)$ would be viewed as a conditional probability given membership in Group A. The approach taken in this paper is consistent with this view.

This paper ignores units which are noninterviews both times and also those which are interviewed one time, but at the other visit are found to be vacant or destroyed. These units must be taken into account in the NCS estimation, but present a different kind of missing data problem than the one considered here.

The NCS is sponsored by the U.S. Bureau of Justice Statistics (BJS) and conducted by the U.S. Bureau of the Census. A public use file is maintained at the University of Michigan. The author wishes to thank Dr. Christopher Innes of the National Criminal Justice Data Archives at the University of Michigan and Marshall DeBerry of the Bureau of Justice Statistics (formerly at the Census Bureau) for their able assistance in obtaining the data in this paper from the NCS Public Use File.

II. FOUR ESTIMATORS FOR $\pi_C(0,0)$

Let $Z_1 = 0$ if no burglary on the first interview
$\quad\quad 1$ otherwise

Let Z_2 be defined similarly for the second interview.

Then $\pi_A(z_1,z_2)$ is the joint discrete probability function for Z_1 and Z_2, given that one is dealing with a case in Group A.

The estimators will be defined as they apply to Group C. The application to Group B will be obvious. Some omitted details are contained in Alexander and Roebuck (1983).

1. Present Touched by Crime Estimator

$$E1 = \frac{1 - p_A(0,0)}{p_A(\cdot,1)} \cdot p_C(\cdot,1) = .07603$$

E1 uses $\frac{1 - p_A(0,0)}{p_A(\cdot,1)}$ to approximate

$$\frac{1 - \pi_C(0,0)}{\pi_C(\cdot,1)}.$$

It is consistent under the assumptions that

$$(2.1) \quad \frac{P_C\{Z_1=1\}}{P_C\{Z_2=1\}} = \frac{P_A\{Z_1=1\}}{P_A\{Z_2=1\}} \quad \text{and}$$

$$(2.2) \quad P_C\{Z_1=0 | Z_2=1\} = P_A\{Z_1=0 | Z_2=1\}$$

2. Griffin's Estimator

$$E2 = p_C(\cdot,1) + p_C(\cdot,0) \cdot \frac{p_A(1,0)}{p_A(\cdot,0)} = .07239$$

This estimator was suggested by Griffin (1981) and is based on the assumption that

$$(2.3) \quad P_A\{Z_1=1 | Z_2=0\} = P_A\{Z_1=1 | Z_2=0\},$$

in which case it is consistent. Unlike E1, E2 is guaranteed to lie between zero and one.

Conditions (2.2) and (2.3) seem to be quite strong. In particular, in the special case that Z_1 and Z_2 are independent, then either condition implies that $P_A\{Z_1=1\} = P_C\{Z_1=1\}$, i.e., that Groups A and C have the same victimization probability for the first interview.

3. Equal Correlation Estimator

Define $T_A = \dfrac{p_A(1,\cdot)}{p_A(\cdot,1)}$ and

let r_A be the sample correlation between Z_1 and Z_2 for Group A. Then define

$$E3 = p_C(\cdot,1)\cdot(1+T_Ap_C(\cdot,0)-r_A(T_Ap_C(\cdot,0)\cdot(1-T_Ap_C(\cdot,1)))^{.5}).$$

This estimator is consistent if (2.1) holds and the population correlations between Z_1 and Z_2 are the same for Groups A and C.

For the data in Section I, T_A = .97349 and r_A = .06725, so E3 = .07589

It is not necessarily the case that E3 is between zero and one. Partly because of this, a similar estimator more appropriate to dichotomous random variables will be considered.

4. Equal Odds Ratio Estimator

Assume that the ratio of the odds that $Z_1 = 1$ relative to the odds that $Z_2 = 1$ is the same for Groups A and C, i.e., that

$$(2.4)\quad D_A = \frac{\pi_A(1,\cdot)/\pi_A(0,\cdot)}{\pi_A(\cdot,1)/\pi_A(\cdot,0)} = \frac{\pi_C(1,\cdot)/\pi_C(0,\cdot)}{\pi_C(\cdot,1)/\pi_C(\cdot,0)} = D_C$$

Assume also that the following odds ratios are equal in the population:

$$(2.5)\quad OR_A = \frac{\pi_A(0,0)/\pi_A(1,1)}{\pi_A(1,0)/\pi_A(0,1)} = \frac{\pi_C(0,0)/\pi_C(1,1)}{\pi_C(1,0)/\pi_C(0,1)} = OR_C$$

Substituting the known sample quantities for Group A for the corresponding unknown quantities for Group C, it is possible to solve for an estimate of $\pi_C(0,0)$. For $OR_A=1$, the resulting equation is quadratic, whose solution in the interval (0,1) is

$$E4 = 1 - \frac{B-(B^2-4OR_A(OR_A-1)p_C(\cdot,0)\cdot)^{.5}C}{2(OR_A-1)}$$

where $C = 1 - \dfrac{D_Ap_C(\cdot,1)/p_C(\cdot,0)}{1 + D_Ap_C(\cdot,1)p_C(\cdot,0)}$

estimates $\pi_C(0,\cdot)$ and

$B = 1 + (p_C(\cdot,0) + C)\cdot(OR_A - 1)$.

In this example, D_A = .97251, OR_A = 3.22350, and E4 = .07565.

If OR = 1, then E4 = $1 - C\cdot p_C(\cdot,0)$. It can be shown that $0 < E4 < 1$, provided that $0 < OR_C < \infty$.

The difficulty in choosing among these estimators is that the assumptions cannot directly be tested; the observations for Group C are incomplete. An indirect approach would be to partition the Group A sample into various demographic groups based on income, urban/rural location, race of householder, etc. If for some given type of crime, OR_A were about the same as OR_C for all these groups, this would lend some credence to (2.5), and similarly for the other assumptions. The data for such an analysis have not yet been tabulated. Further, the partial respondents may differ from complete repondents in unknown ways. An alternative approach to choosing among these estimators is discussed in Section VII.

III. ESTIMATORS BASED ON MORE DETAILED MODELS

Let X_1, X_2, X_3, X_4 be zero-one random variables denoting whether a given randomly selected housing unit reported a crime in each of the four consecutive three-month periods (quarters) covered by the unit's two interviews. For example

X_1 = 1 ; if the unit reported a burglary in the first three months covered by the first interview
0 ; otherwise.

For units in Group A, all four random variables are observed. For those in Group C only X_3 and X_4 are observed.

The discrete joint probability function (pf) will be denoted by $f(x_1,x_2,x_3,x_4|\theta_A)$ for Group A and analogously for Group C. The form of the function is assumed to be the same for both groups, but the parameters may differ. The pf of X_3 and X_4 for Group C will be written as $f(x_3,x_4|\theta_C)$. The empirical probability function (epf) for Group A will be denoted by $f^*(x_1,x_2,x_3,x_4)$.

For each of the following models, a specific form for f will be chosen. A maximum likelihood estimator (mle) will be found for θ_C based on the epf $f^*(x_3,x_4)$. The mle for the touched-by-crime probability is then $1 - f(0,0,0,0|\theta^*)$, where θ^* is the mle for θ_C. Additional details are presented in Alexander and Roebuck (1983).

5. Independent Bernoulli Model

$$(3.1)\quad f(x_1,x_2,x_3,x_4|p) = p^{\Sigma x}(1-p)^{4-\Sigma x}$$

The parameter p represents the probability that the selected unit is victimized in a given quarter.

$$E5 = 1 - f(0,0,0,0|p^*) = 1 - (1-p^*)^4,$$

where the mle p^* is the estimated expected value of $(X_3+X_4)/2$ for Group C, i.e.,

$$p^* = .5(f^*(0,1)+f^*(1,0)) + f^*(1,1).$$

6. Markov model

This model assumes that the probability of a victimization in a given quarter depends only on whether there was a victimization in the previous quarter. These probabilities are:

$$P\{X_i = 1|\ X_{i-1} = 0\} = P0$$

$$P\{X_i = 1|\ X_{i-1} = 1\} = P1, \text{ for } i=2,3,4.$$

Assuming that $P\{X_i = 1\} = P$ for all i, then

$$P\{X_i = 1\} = P = P0\cdot(1 - P) + P1\cdot P, \text{ so}$$

$$P = P0/(1+P0-P1).$$

The pf $f(x_1,x_2,x_3,x_4|P0,P1)$ is easily determined, and leads to the estimator

$$E6 = 1 - (1 - P^*)\cdot(1 - P0^*)^3.$$

The mles are calculated as follows.

$$(6.1)\quad \begin{aligned} P0^* &= (f^*(1,0)+f^*(0,1))/(f^*(1,0)+f^*(0,1)+2f^*(0,0)) \\ P1^* &= (f^*(1,1)+ P0^*\cdot f^*(1,1))/(f^*(1,1)+P0^*\cdot(1-f^*(1,1))). \end{aligned}$$

7. Beta-binomial Model

For the selected HU in Group A, the joint distribution of X_1,X_2,X_3,X_4 is assumed to be the same independent Bernoulli distribution defined above, except that p is now assumed to be a random variable having a beta distribution with parameters α_A and β_A. The likelihood function is obtained by taking the expectation of (3.1) with respect to p.
For Group C

$$P\{X_3=x_3,X_4=x_4|p\}=p^{x_3+x_4}\cdot(1-p)^{2-(x_3+x_4)},$$

where p has a beta distribution with parameters α_C and β_C.

For either Group A or Group C the mles for this model must be obtained by numerical maximization of the likelihood function. This model is similar to Model 5 for any given housing unit, but allows different units to have different victimization probabilities.

Besides using more information than the simpler estimators of Section II, these more detailed models have the advantage that the model can be tested for fit, both for Group A and for the available data from Group C. The main practical drawback occurs when no closed form expression is available for the mles.

It is, of course, essential to examine the fit of the models before using the estimator based on the model. Alexander and Roebuck (1983) give examples illustrating that use of the wrong models (among others, using E5 or E6 when the data come from the beta-binomial model) can lead to substantially worse results than the simpler estimators E1, E2, and E3.

IV. FIT OF THE MODELS FOR THE COMPLETE DATA

Table 4 shows the actual epf $f^*(x_1,x_2,x_3,x_4)$ for the crime of burglary (with an approximate 95% confidence interval), along with the maximum likelihood estimates of the pf for models 5, 6, and 7. Several discrepancies are apparent between the epf and the models. In all these models, the events 0001 and 1000 have the same probability, as do 0100 and 0010. However, in the epf, the event "0001 or 0100" occurs about 50% more frequently than "1000 or 0010". This difference is probably due to a well known NCS "recency effect" (see Kobilarcik, et al (1983)), the effect that a greater proportion of crimes are reported during the three months immediately preceding the interview than in the earliest three months of the reference period. This is presumably due to some form of response error.

All three models miss the pattern for the events with $\Sigma x_i = 2$. The events 0011 and 1100 each have much higher actual frequency than does 0110. All the models assign roughly equal probabilities to these three events.

There are other apparent discrepancies for Model 6. For example, the events 0111 and 1110 have a combined frequency of .000235, slightly higher than the observed .000270 for the union of the events 1011 and 1101. For the estimated pf for model 6, the corresponding probabilities are .000112 and .000036. This difference does not have an obvious explanation in terms of the recency effect, but it may not be valid because of the large standard errors on these estimates.

It is apparent that these models fail to describe important features of the epf. It is necessary to make an adjustment for the recency effect and perhaps to examine additional models.

In spite of these discrepancies, Model 7 gives a very close approximation to the observed "touched by burglary" proportion (.0685). This will be discussed further in Sections V-VII.

To summarize the lack of fit of the three models, a version of the chi-square goodness of fit statistic (times a constant) has been included in Table 7, namely,

$$X^2 = \Sigma\ ((O-E)^2/E),$$

where O represents the observed proportion in a "cell" and E represents the "expected" proportion calculated under the model, using the mle values of the parameters. (The case when O = E = 0 for model 5 is replaced by zero.) The X^2 values for Models 6 and 7 have been expressed as a percentage of the value for Model 5. Model 5 is used as a baseline, because the other two Models may be viewed as generalizations of Model 5 and thus can be expected to have better fit. Model 5 is a special case of Model 6, with P0 = P1. It can be shown that Model 5 is a limiting case of Model 7, with α and β approaching infinity keeping $\alpha/(\alpha+\beta) = p$, where p is a constant between zero and one.

V. THE PROBLEM OF TESTING THE MODEL FOR FIT

This section considers more carefully the question of testing the fit of the data to the hypothesized distribution. A distinction now will be made between the hypothesized family of pfs $f(x_1,x_2,x_3,x_4 \mid \theta)$ and the true but unknown family $h(x_1,x_2,x_3,x_4 \mid \tau)$. The assumption that "the same model fits Groups A and C" then means that there is some unknown distribution $h(x_1,x_2,x_3,x_4 \mid \tau_A)$ which is the true pf for Group A and, if τ_A is replaced by τ_B or τ_C, is the pf for Groups B and C respectively. The existence of such a family h is assumed throughout our discussion.

It is unfortunately not enough to show that the hypothesized family f fits the data for Group A. Even if it is true that for some value θ_A, $f(x_1,x_2,x_3,x_4 \mid \theta_A) = h(x_1,x_2,x_3,x_4 \mid \tau_A)$ for all x_1,x_2,x_3,x_4, this does not imply that the hypothesized family (f) describes the other Groups. What really needs to be demonstrated is that

(5.1) for every parameter value τ, there exists a value θ such that $f(x_1,x_2,x_3,x_4 \mid \theta) = h(x_1,x_2,x_3,x_4 \mid \tau)$, for all x_1,x_2,x_3,x_4.

(Of course, it would be sufficient for this to be true only for $\tau = \tau_A, \tau_B, \tau_C$, but since τ_B and τ_C are unknown, the more general proposition must be addressed.)

The "test" of such a sweeping proposition cannot be purely statistical. One approach would be to test statistically whether the family f fits the data for Group A and then to consider the extent to which model f corresponds to a plausible explanation of the phenomenon of interest. As has been seen, the models considered above show substantial lack of fit for Group A. In addition, each model fails to describe some well documented features of NCS crimes. Because of the NCS recency effect and the known seasonality of crime, it is not to be expected that $P(X_i=1)$ is the same for all i, as all these models require. Additionally, the usual published NCS victimization statistics show that the probability of victimization varies dramatically depending on the urban/rural status of the housing unit, the ages of the occupants, etc. Models 5 and 6 assume that different units have identical probabilities. (Model 7 allows this probability to vary.) There are undoubtedly situations in which the occurrence of a crime at a given housing unit affects behavior (of victim or offender) in such a way as to change the probability of victimization in subsequent quarters, although there is little evidence regarding the extent of this effect for the NCS. Models 5 and 7 allow no such dependence for a given housing unit, although Model 6 does. Thus the present models fail according to this approach. It does not seem likely that, even with better models, a simple model can be justified a priori as a complete explanation of the distribution of crime.

Another approach to this problem corresponds to assuming that for different types of crime (or for crime rates for different demographic groups), the pf is also described by $h(x_1,x_2,x_3,x_4 \mid \tau)$, where τ depends on the type of crime and the demographic group. Under this assumption, if the hypothetical family (f) fits well to h for a wide variety of crimes and demographic groups, then this would tend to support the assertion that more generally (5.1) is true, so that the family f would fit Groups B and C.

Theoretically, in order for the model to yield a consistent estimator of $h(0,0,0,0 \mid \tau)$ from the Group A data, it is not essential that the assumed family f fit for all values of x_1,x_2,x_3,x_4. To see this, let $T(x_1,x_2,x_3,x_4)$ be a sufficient statistic for θ under the hypothesized family f. Let the possible values of T be denoted by $t_0,t_1,\ldots,t_K$, where $K < 15$, letting $t_0 = T(0,0,0,0)$. Let

$$f_T(t_k \mid \theta)=\Sigma f(x_1,x_2,x_3,x_4 \mid \theta) \text{ and}$$
$$h_T(t_k \mid \tau) = \Sigma h(x_1,x_2,x_3,x_4 \mid \tau),$$

where the summations range over all values of x_1,x_2,x_3,x_4 such that $t_k=T(x_1,x_2,x_3,x_4)$. Assume that the function f satisfies the necessary regularity conditions for the mle θ^* to exist and be a consistent estimator of θ, if the familiy f were the true model. Let f_T also satisfy these conditions. Assume also that if θ' is any consistent estimator of θ, then $f(0,0,0,0 \mid \theta')$ is a consistent estimator of $f(0,0,0,0 \mid \theta)$. (This is true for all the families f discussed in this paper.)

PROPOSITION: If for every parameter value τ, there exists a parameter value θ such that

(i) $f_T(t_k \mid \theta) = h_T(t_k \mid \tau)$,
for $k=0,1,\ldots,K$ and
(ii) $f(0,0,0,0 \mid \theta) = h(0,0,0,0 \mid \tau)$,

then whatever the true value of τ, $f(0,0,0,0 \mid \theta^*)$ is a consistent estimator of $h(0,0,0,0 \mid \tau)$. (If $T(x_1,x_2,x_3,x_4) = t_0$ only when $x_1=x_2=x_3=x_4=0$, then condition (ii) is redundant.)

Proof: Condition (i) implies that for some value θ_A, $f_T(t \mid \theta_A)$ is the true pf for $T(x_1,x_2,x_3,x_4)$. Then the mle θ^* calculated based on the epf of $T(X_1,X_2,X_3,X_4)$ is a consistent estimator of θ, viewed as a parameter of f_T. Since the mle depends only on the distribution of the sufficient statistic, this same value θ^* is the mle calculated as a parameter of f based on the epf of X_1,X_2,X_3,X_4. Thus θ^*, calculated as a parameter of f, is a consistent estimator of θ, even though the true pf is h, not f. Therefore $f(0,0,0,0 \mid \theta^*)$ is a consistent estimator of $f(0,0,0,0 \mid \theta) = h(0,0,0,0 \mid \tau)$.QED

Note that it is not necessarily the case that $f(0,0,0,0 \mid \theta^*)$ is a maximum likelihood estimator of $h(0,0,0,0 \mid \tau)$; it must further be assumed that T is also sufficient for the family h.

This proposition may explain why Model 7 gave a good estimate for the "touched by burglary" rate. The statistic Σx_i is a minimal sufficient statistic for (α,β) under Model 7. The fit of the model to the empirical distribution of this statistic is comparatively good, and condition (ii) applies. By contrast, the same statistic is a minimal sufficient statistic for Model 5, but Table 5 shows that the fit of the likelihood function under this model to the epf of the statistic is relatively poor.

A minimal sufficient statistic for P0 and P1 under model 6 is given by

S(0000)=1	S(0110)=6
S(1000)=S(0001)=2	S(1001)=7
S(0100)=S(0010)=3	S(1110)=S(0111)=8
S(1100)=S(0011)=4	S(1101)=S(1011)=9
S(0101)=S(1010)=5	S(1111)=10

Table 6 shows the distribution of this statistic. Note that the major discrepancies involve S=4,5,6,7,8, the values for which $\Sigma x_i = 2$.

The practical utility of the above proposition is limited. Even though the beta-binomial model exhibits fairly good fit for the distribution of a minimal sufficient statistic, the lack of fit for the complete distribution is disturbing. It suggests that the model does not accurately reflect the phenomenon being measured. In addition it is difficult to test assumption (ii). The fact that the mle for $f(0,0,0,0 \mid \theta)$ is close to $f^*(0,0,0,0)$ does not necessarily support (ii); see the discussion of Model 6 in the next Section.

VI. FIT TO THE INCOMPLETE DATA

For Group C, the epf of X_3 and X_4 are given in Table 7, along with the mles of the pfs for the three models. Again there is evidence of a lack of fit, for the events 01 and 10, due to the recency effect.

Note that Model 6 fits nearly perfectly the distribution for 00 and 11 in Table 7. It is easy to show that nearly perfect fit holds in this case regardless of the observed values, so that this is not really a test of fit. Indeed, substituting the expressions for P0* and P1* from (6.1) into the pf for Model 6, one obtains as the mle for $f(0,0)$,

$$f^*(0,0)/(1+(f^*(1,1))^2/(1-f^*(1,1)),$$

which is very close to $f^*(0,0)$ if $f^*(1,1)$ is small. Similarly the mle for $f(1,1)$ is

$$f^*(1,1)/(1-f^*(1,1)),$$

which is close to $f^*(1,1)$ if $f^*(1,1)$ is small. Something similar may happen with Model 7; however, analysis of the pf is much more difficult.

A result similar to the Proposition in the last section could be proved for Group C, with $T(x_3,x_4)$ being a sufficient statistic based only on the observations of X_3 and X_4. The problem in this case is that condition (ii), which is exactly as before, seems to be impossible to check, since $f^*(0,0,0,0)$ is not observed for Group C.

VII. COMPARISON OF THE SIMPLE ESTIMATORS UNDER THREE MODELS

One way of comparing the estimators of Section II is to see how well they would perform if the population fit one of the models in Section III. Trying the simple estimators under a variety of models gives some idea of their robustness. However, since these models clearly need (at least) a correction for known response error, this analysis will be of more interest when better-fitting models are found.

Based on the parameter estimates for Groups A and C, if the populations exactly fit the models, then the populations would have the following characteristics, corresponding to the tables in Section I. Note that the full table for Group C would not be observed.

MODEL 5

Group A			Group C		
.927249	.035688	.962938	.917443	.040390	.957832
.035688	.001374	.037062	.040390	.001778	.042168
.962938	.037062	1.000000	.957832	.042168	1.000000

MODEL 6

Group A			Group C		
.929590	.034208	.963798	.922031	.037393	.959424
.034208	.001993	.036202	.037393	.003183	.040576
.963798	.036202	1.000000	.959424	.040576	1.000000

MODEL 7

Group A			Group C		
.931515	.032199	.963715	.925912	.033506	.959418
.032199	.004086	.036285	.033506	.007075	.040482
.963715	.036285	1.000000	.959418	.040582	1.000000

Using the given results for Group A and the observable value $p_c(\cdot,1)$, the four estimators from Section I can be calculated and compared to the actual value $1-\pi_c(0,0)$ which is implied by the model with the assumed parameters. The results (with % error = 100(E-Actual)/Actual) are as follows.

TABLE 3
Estimated "touched by burglary" proportions
Assuming different models for the population

Model	"Actual"	E1	E2	E3	E4
#5	.0826	.0828(0.3%)	.0777(-5.9%)	.0826(0.0%)	.0825(-0.0%)
#6	.0780	.0789(1.2%)	.0746(-4.3%)	.0787(1.0%)	.0787(0.9%)
#7	.0741	.0766(3.4%)	.0726(-2.0%)	.0764(3.2%)	.0761(2.8%)

For these three models for burglary, estimators E3 and E4 do slightly better than the present published estimator E1, and do substantially better than E2 for Models 5 and 6, but worse for Model 7. This analysis needs to be repeated for other crimes and, especially, for models which give a better fit to the Group A data, to make it sufficiently conclusive to warrant changing the form of the published estimator.

VIII. DISCUSSION OF THE RESULTS

Model 7 appears to fit the best of the three models considered. However, the main conclusion of this paper is that none of the proposed model fits without a modification to take into account the recency effect. In a different kind of analysis, the lack of fit of Model 6 was also observed in Griffin (1983).

The next step in the research is to attempt to develop models which do not require $P(X_i = 1)$ to be the same for $i = 1,2,3,4$. An additional model from Alexander and Roebuck (1983), the "independent with additional victimization" model, has not been considered for reasons of space. This model also requires $P(X_i = 1)$ to be constant, but may be easy to modify to eliminate this restriction.

The immediate goal of this research is to find a model which fits well for all the crime categories of interest, and to use this model to select one of the six closed-form estimators (E1 - E4 and the Group C mles for Models 5 and 6). It is probably desirable to apply the missing-data adjustment separately for different subgroups of the sample. For this purpose, it will be necessary to repeat the analysis for different subgroups.

The results of Table 3 are of some immediate interest. The present estimator (E1) has a relatively small bias under the three models. (The bias is only for the incomplete cases, which in any given month are at most about one-fourth of the sample.) Thus there is no strong reason to replace E1 by the mle under any of the selected models, since E1 does fairly well under the assumptions upon which such an alternative estimator would be based.

The data in this paper should be viewed as preliminary. The numerical likelihood calculations need further scrutiny, especially for Model 8, whose maximum likelihood appears to be along a "ridge" in the function. (The mles were calculated in UNIVAC single precision arithmetic, using the IMSL subroutine ZXMWD. The maximum was checked using single precision on an IBM personal computer, by inspecting the likelihood at a grid of parameter values.)

The "Actual" values in Table 4 were calculated using the NCS "final" design-based weights. The main effect of these weights is due to a correction for instances of subsampling in the field, and to a "post-stratification" adjustment bringing the weighted age-race-sex distribution of the full sample into agreement with independent estimates for the population. However, the weights also include noninterview adjustments which are not appropriate for application to Group A in our present problem. It was not possible to reweight the Group A cases separately for this analysis. This is felt to make little difference to the results; indeed, almost identical results were obtained using unweighted results for burglary.

The approximate standard errors in Table 4 are calculated using a design effect appropriate to the usual NCS estimates. The appropriateness of this design effect for this purpose has not been tested.

TABLE 4
Fit of the Models to the Data
for Group A

$X_1X_2X_3X_4$	Actual (95% CI)	Model 5	Model 6	Model 7
0000	.931499 ±.0036	.927249	.929592	.931515
0001	.020016 ±.0020	.017676	.016950	.015639
0010	.011872 ±.0015	.017676	.016283	.015639
0011	.001028 ±.0005	.000337	.000976	.000921
0100	.017552 ±.0019	.017676	.016283	.015639
0101	.001173 ±.0005	.000337	.000297	.000921
0110	.000551 ±.0003	.000337	.000937	.000921
0111	.000169 ±.0002	.000006	.000056	.000096
1000	.013242 ±.0016	.017676	.016950	.015639
1001	.000819 ±.0004	.000337	.000309	.000921
1010	.000591 ±.0003	.000337	.000297	.000921
1011	.000119 ±.0002	.000006	.000018	.000096
1100	.001152 ±.0005	.000337	.000976	.000921
1101	.000151 ±.0002	.000006	.000018	.000096
1110	.000066 ±.0001	.000006	.000056	.000096
1111	.000000	.000000	.000003	.000015
X^2 for model:		.020492	.008356	.003269
As % of Model 5:		100%	40.8%	16.0%

Tables 4, 5, and 6 are based on the following values for the mles for Group A.
Model 5: p = .018706
Model 6: P0= .017907, P1= .056625
Model 7: α= .422101, β= 22.1416

TABLE 5
Fit of the Models to the Distribution
of Σx_i for Group A

Σx_i	Actual	Model 5	Model 6	Model 7
0	.931499	.927249	.929592	.931515
1	.062682	.070704	.066466	.062556
2	.005315	.002022	.003792	.005526
3	.000505	.000024	.000148	.000384
4	.000000	.000000	.000003	.000015

TABLE 6
Fit of the models to the Distribution
of the Minimal Sufficient Statistic S(for Model 6)

S	Actual	Model 5	Model 6	Model 7
1	.931499	.927249	.929592	.931515
2	.033258	.035352	.033900	.031278
3	.029424	.035352	.032566	.031278
4	.002181	.000674	.001952	.001842
5	.001764	.000674	.000594	.001842
6	.000551	.000337	.000937	.000921
7	.000819	.000337	.000309	.000921
8	.000235	.000012	.000112	.000192
9	.000270	.000012	.000036	.000292
10	.000000	.000000	.000003	.000015

TABLE 7
Fit of the models to the distribution
of X_3 and X_4 for Group C

X_3X_4	Actual	Model 5	Model 6	Model 7
00	.959426	.957832	.959424	.959418
10	.015512	.020857	.019262	.019247
01	.023012	.020857	.019262	.019247
11	.002049	.000454	.002053	.002094

Table 7 is based on the following values of the mles for Group C.
Model 5: p = .021311
Model 6: P0 = .019681, P1 = .096326
Model 7: α = .250632, β = 11.4937

References

(1) Eddy, W.F., S.E. Fienberg, and D.L. Griffin (1981) "Some First Attempts at Estimating Victimization Prevalence in a Rotating Panel Survey", Technical Report No. 220, Department of Statistics, Carnegie-Mellon University, September 1981.

(2) Alexander, Charles H. and Michael J. Roebuck (1983) "National Crime Survey, Empirical and Model-Based Estimators for the Proportion of Households Victimized in a Year," Proceedings of the American Statistical Association Survey Research Methods Section.

(3) Griffin, D.L. (1981) "Discussion of Several Estimators of the Proportion of Households Victimized in a Year," Working Memorandum NCS-3, Department of Statistics, Carnegie-Mellon University, June 1981.

(4) Griffin, D.L.(1983) "Estimation of Victimization Prevalence using Data from the National Crime Survey," unpublished Phd. dissertation, Carnegie-Mellon University, August 1983.

(5) Kobilarcik, Edward L., Charles H. Alexander, Rajendra P. Singh, and Gary M. Shapiro, (1983) "Alternative Reference Periods for the National Crime Survey," Proceedings of the American Statistical Association Survey Research Methods Section.