EFFECTS OF ATTRITION IN THE NATIONAL CRIME VICTIMIZATION SURVEY

Lynn M. R. Ybarra, Sharon L. Lohr, Arizona State University
Lynn Ybarra, Department of Mathematics, Arizona State University, Tempe AZ 85287-1804

Key Words: Imputation, Longitudinal surveys, Missing data, Nonresponse

1 INTRODUCTION

Many crime victims experience multiple victimizations over time. Estimating the rate of repeat victimization from a longitudinal survey such as the U.S. National Crime Victimization Survey (NCVS), however, is challenging because individuals often have missing data for some of the interviews. Households that move are more likely to have experienced a victimization than are households that remain in one location (Saphire, 1984; Lohr and Sun, 1998; Dugan, 1999); by using data only from households that complete all interviews, repeat victimization rates are likely to be underestimated. Stasny (1990), Conaway (1993), and Lohr and Sun (1998) found that the estimated number of households victimized at least once was greater when the information from households missing some interviews was included in the analysis. Since repeat victimization is rare, estimates of repeat victimization are probably more affected by missing data than are estimates of overall cross-sectional victimization rates.

In this paper, we use data from the 1996-98 NCVS to explore potential effects of missing data and to estimate repeat victimization rates for violent crime. In contrast to much of the earlier work, we examine repeat victimizations among individuals rather than households. We introduce two algorithms for estimating repeat victimization rates, using logistic models to impute values for individuals who have partial data. These models are applied to estimate rates of repeat victimization for violent crime, and to explore sensitivity of estimates to assumptions.

2 GROSS FLOWS AND ATTRITION MODELS

A gross flow matrix is an $m \times m$ contingency table showing the transitions from each state of an m-class categorical variable to other states for successive time periods. In the following 2×2 gross flow matrix, x_{NN} is the number of persons who are not victimized in either time 1 or time 2, x_{NV} persons who are not victimized in time 1 but are victimized in time 2, etc.

		Time 2	
		Nonvictim	Victim
Time 1	Nonvictim	x_{NN}	x_{NV}
	Victim	x_{VN}	x_{VV}

When individuals are missing interviews in either time period 1 or time period 2, we do not observe data for the table above. Instead, we observe:

		Time 2		
		Nonvictim	Victim	Missing
Time 1	Nonvictim	y_{NN}	y_{NV}	y_{NM}
	Victim	y_{VN}	y_{VV}	y_{VM}
	Missing	y_{MN}	y_{MV}	

The estimate $y_{NV}/(y_{NN} + y_{NV})$ for the conditional probability of victimization given victim status in time 1 will be biased if the missing values y_{NM} are more or less likely to be victims at time 2. The estimate $y_{NV}/(y_{NN} + y_{NV} + y_{NM})$ will be too small, and the estimate $(y_{NN} + y_{NM})/(y_{NN} + y_{NV} + y_{NM})$ will be too large. These estimates would result if all observations were assigned to the nonvictim or victim cells, respectively, with possible misclassification of some of the observations.

Our goal is to develop models that reduce the misclassification that would occur if all persons missing an interview were assigned to nonvictim status or all were assigned to victim status for that interview. We assume that all persons who provide responses for an interview are classified correctly. Thus we are not dealing with misclassification caused by response error, as described by Chua and Fuller (1987) and Singh and Rao (1995). Those authors used re-interview data and data from other sources to adjust for misclassification caused by erroneous responses from survey participants. In reality, respondents in crime surveys do sometimes misreport their victimization experiences, but this misreporting is difficult to detect without independent sources of information. In this paper, we concentrate on errors attributable to missing data.

Our primary interest is in grows flows from year to year. In the NCVS, a year comprises two interviews, each covering victimization experiences in a six-month period, so it is possible for a survey participant to provide information for one six-month period but not the other. If such a person reported a violent victimization, he or she would be correctly classified as a victim, regardless of experiences in the period for which the person were a nonrespondent. If a person reported no violent victimizations during one interview and was nonrespondent for the other interview, assigning the person to nonvictim status may result in misclassification.

Consider a population with N persons, of which n are sampled. Let $Y_{i1}, Y_{i2}, \ldots, Y_{it}$ represent the responses

for person i (i =1, ..., N) at times 1 through t. We assume that Y_{ij} is an ordinal variable taking on values 1, ..., m. We are interested in two discrete time periods, A_1 = {1, ..., a} and A_2 = {a+1, ..., t}. Let Z_{ik} = max{Y_{ij}, j $\in$ A_k} for k = 1, 2. Since many of the responses are not observed, let R_{ij}, j =1, ..., t be an indicator variable where R_{ij} = 1 if person i contributes data at interview j and R_{ij} = 0 if person i is missing at interview j. If R_{ij} = 0, then Y_{ij} is not observed; depending on the values of the observed responses for person i, Z_{ik} may or may not be missing. In our application, a = 2, and Y_{ij} can take on values of either 0 (nonvictim) or 1 (victim). If Y_{i1} = 1 and R_{i2} = 0, then Z_{i1} = 1; however, if Y_{i1} = 0 and R_{i2} = 0, then Z_{i1} is missing. If a value of 0 were imputed for every missing Z_{i1} and Z_{i2}, some persons would be misclassified.

Let $B \subset A_1$, and consider interviews $\{j, j \in B\}$ for person i. Define the indicator variable M_{iB} to be 1 if $\max_{j \in B} Y_{ij} > Y_{ik}$ for all $k \in A_1 \cap B^c$; and 0 otherwise; M_{iB} indicates potential misclassification.

For the NCVS, suppose B = {2}. If person i is a victim at interview 1 so that Y_{i1} = 1, then M_{iB} = 0. If Y_{i1} = 0, then M_{iB} = 0 if Y_{i2} = 0 and M_{iB} = 1 if Y_{i2} = 1. For a record with complete data, we can assign a value to M_{iB}; this value tells whether the record would have been misclassified if missing the observations in B and assigned to a category based only on observations in $A_1 \cap B^c$. We define M_{iB} for $B \subset A_2$ in similar fashion.

The goal is to predict $P(Z_1 = z_1, Z_2 = z_2 | X)$, where X represents explanatory variables. With no covariates, this would give the estimated probabilities for a gross flow matrix. In our application, however, many of the z's are missing. To explore the effect of missing data on the predictions, we derive models for imputing the missing values. Pfeffermann et al. (1998) incorporated time dependence of the responses into the model through Markov models. We incorporate the intraperson correlation in victimization by using z_2 as a covariate in imputing z_1, and vice versa.

For $B \subset A_1$, define

$$\eta_{iB} = P(M_{iB} = 1 | X_i, Z_{i2}) \quad (1)$$

Similarly, for $B \subset A_2$, let

$$\varphi_{iB} = P(M_{iB} = 1 | X_i, Z_{i1}) \quad (2)$$

Separate models are used because the effect of covariates or status may differ in the two time periods. In our application, η_{iB} and φ_{iB} are the conditional probabilities of being misclassified given the victimization status in the other time period and subject-specific covariates. We model η_{iB} and φ_{iB} using logistic regression, with the models

$$\text{logit}(\eta_{iB}) = Z_{i2}\beta_B + X_i^T \gamma_B \quad (3)$$

$$\text{logit}(\varphi_{iB}) = Z_{i1}\beta_B + X_i^T \gamma_B \quad (4)$$

We fit the models two ways to explore sensitivity of our inferences to the imputations. The first method used only records with complete data for time period A_1 in logistic regression (3) and only records with complete data for time period A_2 in logistic regression (4). The second initially set the missing observations to 0 and then included all cases in both logistic regressions. The following algorithms give the steps used in computing the estimates for the two methods.

Algorithm 1

Step 0. Set Y_{ij} = 0 whenever R_{ij} = 0. Calculate Z_{ij} for all persons using the imputed values for Y_{ij}. Using only persons with complete records for time period A_k, calculate M_{iB} for all subsets B of A_k, k = 1,2.

Step 1. Estimate the parameters of the models in (3) and (4). The logistic regressions will only use cases for which M_{iB} was calculated in Step 0.

Step 2. Use the parameter estimates from Step 2 to impute values for Y_{ij} when R_{ij} = 0.

Step 3. Recalculate Z_{ij} for all persons using the new imputed values for Y_{ij}.

Repeat steps 1 through 3 until the parameters in models (3) and (4) converge.

Algorithm 2

Step 0. Set Y_{ij} = 0 whenever R_{ij} = 0. Calculate Z_{ij} for all persons using the imputed values for Y_{ij}. Using all persons (including the imputed values for Y_{ij}) calculate M_{iB} for all subsets B of A_1 and A_2.

Step 1. Estimate the parameters of the models in (3) and (4). The logistic regressions will use all cases.

Step 2. Use the parameter estimates from Step 2 to impute values for Y_{ij} when R_{ij} = 0.

Step 3. Using all persons and the new imputed values for Y_{ij} recalculate Z_{ij} and M_{iB} for all subsets B of A_1 and A_2.

Repeat steps 1 through 3 until the parameters in models (3) and (4) converge.

If few records have missing data, algorithms 1 and 2 will give similar estimates for the misclassification model parameters and for the gross flow matrices; the complete records largely determine the logistic coefficients. With more missing data, the two algorithms are expected to give different estimates and provide an indication of how much the estimates in the gross flow matrices depend on the assumptions about the missing data. We construct an initial table by assigning zeros to missing interviews. Algorithm 2, since it uses the initial imputed zeros in the iteration, is expected to move fewer persons from the initial table cells than is algorithm 1. If the models adopted in (3)

and (4) reflect the true classification mechanism, the initial assignment of missing values to cells should make little difference.

Both algorithms are iterative procedures, so estimates of standard errors from the statistical software used to perform the logistic regression will be incorrect. The jackknife will give consistent estimates of the variances of the logistic regression coefficients and the gross flow matrix entries under the assumption of uniform response rates within each stratum (Rao, 1996). Jackknife standard errors, however, only reflect the uncertainty due to sampling error; they do not incorporate possible effects of misspecifying the non-response model.

3 1996-98 NCVS LONGITUDINAL FILE

The NCVS is a stratified multistage cluster survey with a rotating panel design. Selected households are interviewed every six months for three and a half years. During an interview, every household member aged 12 and over is asked about his or her victimization experiences in the previous six months. The first interview is used for bounding purposes; interviews 2 through 7 are released in the public use data sets. The data codebook (U.S. Department of Justice, 2000) gives a detailed description of the NCVS study design.

To maximize the available information, we constructed two data sets from the 1996-98 annual files. The first consisted of data from interviews 2 through 5 for persons whose second interview was scheduled between January 1996 and June 1997. The second data set consisted of interviews 4 through 7 for persons whose fourth interview occurred between January 1996 and June 1997. For space reasons, we present only the results from the data set using interviews 2 through 5.

We matched persons from the annual files using the linkage variables provided by the Census Bureau, plus race and gender. The final data sets contained one record for each person in the survey, with information on the number of violent victimizations, their severity, victim actions, and other variables of interest for each interview period. This increases the amount of missing data. Table 1 displays the missing data patterns in the data. There are, for example, 8284 persons who completed interview 2 but not interviews 3, 4, or 5.

The first two interviews constitute the first year and the last two interviews constitute the second year. Table 2 contains observed gross flow matrices, using person weights. A person has victim status (V) if he or she reported at least one violent victimization during the year. Status N indicates the person was present for both interviews and reported no violent victimizations. Persons with missing data and no observed victimizations for the year were placed in status .N, N., or .., depending on whether they were missing the first interview, the second interview, or both interviews.

Table 1: Attrition Patterns

		Number of Persons
Interviews Completed	2	8284
	3	3092
	4	2815
	5	6319
	23	4318
	24	381
	25	268
	34	1269
	35	321
	45	3604
	234	3568
	235	1004
	245	1001
	345	3582
	2345	26470
	TOTAL	66296

The missing data in Table 2 may be viewed as a misclassification problem. If we ignored persons with pattern N., .N, or .., we would expect biased estimates since most of these cases represent nonvictims. If we assigned persons with pattern N. or .N to the nonvictim, N, cell, we would be misclassifying some of them and underestimating the victimization and repeated victimization rates. Persons who experienced a victimization but did not report it due to missing the interview would be misclassified as nonvictims.

Table 2: Observed Gross Flow Matrix

(in 100,000 persons)

	Year 2					
Year 1	V	N	N.	.N	..	SUM
V	3.17	18.66	4.76	1.08	14.99	42.66
N	10.89	608.30	83.42	22.63	97.99	823.23
N.	0.71	21.84	8.13	6.15	189.90	226.72
.N	3.37	86.83	31.16	7.78	77.27	206.42
..	13.31	90.59	68.75	160.10		332.75
SUM	31.45	826.22	196.22	197.74	380.16	1631.78

4 NCVS GROSS FLOW ESTIMATES

We applied the algorithms using the 1996-1998 longitudinal NCVS data sets described in Section 3. In the notation of the model, Y_{ij} is 1 if person i reported a violent victimization during interview j and 0 otherwise. The four time periods, $t = 4$, are divided into two years; $A_1 = \{2,3\}$ and $A_2 = \{4,5\}$. The goal is to examine the probability of being misclassified during times 2 and 3 as a result of missing one or both of these interviews and also of being misclassified during times

4 and 5 as a result of missing one or both of these interviews. We view these probabilities as a function of a person's victimization status during the other time period and of person-specific covariates. Table 3 gives the covariates used in the models. All covariates except status23 and status45 refer to levels at first completed interview. Other models with different covariates are given in Tobin (1999); these produced similar results.

Table 3: Explanatory Variables

Variable	Levels
Age	1 if age 25 of older
	0 if between ages 12 and 24
Gender	1 if female
	0 if male
Marital Status	1 if married
	0 otherwise
Move	1 if moved in previous five years
	0 otherwise
Home	1 if home owned or being bought
	0 otherwise
Status 2,3	1 if victimized in interview 2 or 3
	0 otherwise
Status 4,5	1 if victimized in interview 4 or 5
	0 otherwise

Tables 4 and 5 give the adjusted gross flow matrices for the two algorithms in 100,000 persons, and Table 6 gives the logistic regression coefficients. All models were fit using weights; unweighted analyses gave similar results.

Table 4: Gross Flow Matrices for all Violent Crime Victims from Algorithm 1

		Year 2 Interviews 4 & 5		
		V	N	SUM
Year 1 Interviews 2 & 3	V	13.10	63.43	76.54
	N	44.75	1511.00	1555.75
	SUM	57.85	1574.43	1632.29

Table 5: Gross Flow Matrices for all Violent Crime Victims from Algorithm 2

		Year 2 Interviews 4 & 5		
		V	N	SUM
Year 1 Interviews 2 & 3	V	8.97	54.23	63.20
	N	37.95	1531.00	1568.95
	SUM	46.92	1585.23	1632.15

Below are conditional probabilities calculated from the gross flow matrices:

	Algorithm 1	Algorithm 2
P (V year 2 | V year 1)	0.171	0.142
P (V year 2 | N year 1)	0.029	0.024

The difference in the estimates of P (V year 2 | V year 1) from the two algorithms indicates the sensitivity to the initial assumptions about the missing data. The estimates from algorithm 2 are similar to estimates that use only complete records. Below CR indicates estimates calculated from the tables where only complete records were used. All N indicates estimates calculated by assuming all missing interviews were nonvictims.

	CR	all N
P (V year 2 | V year 1)	0.145	0.080
P (V year 2 | N year 1)	0.018	0.018

Assuming that all missing interviews represent nonvictims severely underestimates the repeat victimization rate. The estimates from algorithm 2, thought to be conservative, are similar to the CR estimates. The estimates from algorithm 1 are larger than the estimates from algorithm 2, indicating that algorithm 1 moves more people from the nonvictim cells to the victim cells. This is largely because the intercepts in models predicting η_{23} and φ_{45} are more negative for algorithm 2, which means the baseline rate of misclassification is smaller in algorithm 2 and fewer people will move. This was expected since many cases are missing both interviews for a given year, i.e. both 2 and 3 or both 4 and 5. Algorithm 1 would not use such people in the logistic regression models whereas algorithm 2 would assume both interviews were nonvictims and would use the people in all models.

Algorithm 1 generally gives larger coefficients for the victimization status variables than algorithm 2. The coefficients for age, gender, and marital status are statistically significant at the 0.05 level in all models whereas the coefficients for move and home are significant in only some models.

In all models, the variables with the largest effect are those related to status in the opposite time period. These are all statistically significant at the 0.01 level, indicating that persons who experience a victimization in one time period are much more likely to be erroneously classified as nonvictims based on partial information. Using estimates from algorithm 1, a person who is missing both interviews 2 and 3 is about 6.5 times more likely to be misclassified if the person was victimized in interviews 4 and 5 than if the person was not victimized. Algorithm 2 reports this rate to be about 5.2.

Table 6: Coefficients for Logistic Regression Models

		Coefficients						
		Int.	Age	Gender	Marital Status	Move	Home	Status 4,5
η_2	Alg. 1	-3.53	-0.72	-0.49	-0.77	0.43	0.02	1.59
	Alg. 2	-3.55	-0.56	-0.45	-0.84	0.38	-0.06	1.85
η_3	Alg. 1	-4.36	-0.49	-0.35	-0.47	0.81	0.09	0.77
	Alg. 2	-4.01	-0.51	-0.27	-0.55	0.71	-0.18	0.97
η_{23}	Alg. 1	-1.49	-0.83	-0.44	-0.73	0.02	-0.78	1.88
	Alg. 2	-2.45	-0.63	-0.42	-0.77	0.29	-0.28	1.65
φ_4	Alg. 1	-3.90	-0.57	-0.23	-0.46	0.01	-0.18	1.62
	Alg. 2	-4.01	-0.46	-0.23	-0.49	-0.02	-0.17	1.73
φ_5	Alg. 1	-4.28	-0.45	-0.50	-0.30	0.55	-0.29	0.91
	Alg. 2	-3.71	-0.38	-0.23	-0.51	-0.27	-0.54	0.36
φ_{45}	Alg. 1	-1.83	-0.67	-0.27	-0.50	-0.39	-0.98	1.84
	Alg. 2	-2.77	-0.43	-0.24	-0.54	-0.16	-0.52	1.65

5 CONCLUSIONS

Estimates of repeat violent victimization from the NCVS depend heavily on the form of the model used to impute values for missing data. Algorithm 2, initially imputing nonvictim status to missing observations, gave results similar to gross flow tables constructed from complete records only. Algorithm 1, which used only the complete records for the imputation models, resulted in substantially higher estimates of repeat victimization rates.

One aspect of the models presented here is that the imputation depends in part on the victimization status in the other time period. Since we do not know all of the factors relating to victimization, inclusion of status in the other time period served as a proxy for unobserved variables that might better predict repeat victimization. As with all models for imputation, model checking is possible only if we are able to obtain data for the nonrespondents, because we do not know the true nonresponse mechanism. Our models, though, indicate that the attrition in the NCVS may have a large effect on estimates of repeat violent victimizations.

ACKNOWLEDGEMENTS

This research was supported by a grant from the U.S. Bureau of Justice Statistics. We are grateful to Marshall DeBerry for invaluable help and advice in constructing the longitudinal data set.

REFERENCES

Chua, T.C. and Fuller, W.A. (1987). A model for multinomial response error applied to labour flows. *J. Amer. Statist. Assoc.*, 82, 46-51.

Conaway, M.R. (1993). Non-ignorable non-response models for time-ordered categorical variables. *Applied Statistics*, 42, 105-115.

Dugan, L. (1999). The effect of criminal victimization on a household's moving decision. *Criminology*, 37, 903-930.

Fienberg, S.E. (1980). The measurement of crime victimization: prospects for panel analysis of a panel survey. *The Statistician*, 29, 313-350.

Lohr, S. and Sun, S. (1998). Probability of victimization over time: results from the U.S. National Crime Victimization Survey. *Proceedings of Statistics Canada Symposium 98, Longitudinal Analysis for Complex Surveys*, 163-168.

Pfeffermann, D., Skinner, C., and Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *J. Roy. Statist. Soc., Ser. A* 161:13-32.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *J. Amer. Statist. Assoc.* 91:499-506.

Saphire, D. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey.* New York: Springer-Verlag.

Singh, A.C. and Rao, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *J. Amer. Statist. Assoc.*, 90, 478-488.

Stasny, E.A. (1990). Symmetry in flows among reported victimization classifications with nonrandom nonresponse. *Survey Methodology*, 16, 305-330.

Tobin, L.M.R. (1999). Time-in-sample and attrition effects on misclassification and rates of violent crimes in the National Crime Victimization Survey. Unpublished M.S. thesis, Arizona State University.

U.S. Department of Justice, Bureau of Justice Statistics (2000). National Crime Victimization Survey, 1992-1998 [Computer file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census. 8th ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor].