Paper 269-25

Correcting for Missing Discrete Responses in Business Surveys

Lawrence C. Marsh, University of Notre Dame, South Bend, IN 46617-1226

ABSTRACT

Businesses often need an accurate profile of their customers in order to better serve them, improve products and make advertising more effective. Unfortunately, customers do not always completely fill out the survey forms, especially those product registration cards. The forms often have multiple choice questions and customers may leave some questions blank. This paper presents a maximum likelihood method of correcting for the biased sample selection that occurs when the dependent variable is an incomplete categorical response where the data are not missing at random. The paper includes a test of the missing-at-random hypothesis.

INTRODUCTION

Marsh and Wells (1996) and Baxter and Marsh (1996) attempted to correct for missing categorical dependent variable values using a two-step procedure. In a footnote they provided the likelihood function for a single-step approach. Marsh and Wells (1996) applied their technique to the employment status of public employees while Baxter and Marsh (1996) analyzed a survey of businesses in Minneapolis/ Saint Paul. These authors tested the missing-at-random (MAR) hypothesis and found the data to be systematically missing (in other words, not-missing-at-random (NMAR)).

Lee and Marsh (1998) used the single step maximum likelihood estimator (MLE) which they again applied to the data on the employment status of public employees. They also found the missing data to be NMAR.

Marsh (1999) compared the maximum likelihood method developed by Lee and Marsh (1998) with a maximum entropy method and a sample augmentation method (SAM) along with several variations of the MLE.

The purpose of this paper is to present the MLE method of correcting for the distortions that arise when missing data from business surveys are ignored and/or not adequately accounted for in multinomial logit models. The basic problem is that systematically missing data will produce biased and inconsistent estimates of the underlying joint and marginal probabilities if not handled properly.

The next section will discuss the theory behind the MLE method, which will be referred to as Restore-MLE. The traditional approach to this problem is to ignore the missing data by dropping the observations containing the missing categorical responses and estimating a multinomial discrete response model using only the complete observations. This traditional approach will be referred to as Partial-MLE.

Three discrete outcome variables with no missing data were obtained from the Panel Study of Income Dynamics (PSID). The three variables were home ownership status, employment status and marital status. These three dependent variables were first estimated by MLE without any missing data. This estimator was designated Full-MLE. After designating a substantial number of observations as having missing discrete responses, the Partial-MLE and Restore-MLE methods were compared in their ability to retrieve the correct joint probabilities for the dependent responses in all 9343 observations.

MISSING RESPONSE MODEL

Typically when confronted with some missing data values for the dependent variable, researchers will just drop the observations with missing data and estimate their model under the assumption that the data are missing at random. When this assumption does not hold, the resulting estimators are generally biased and inconsistent due to biased sample selection. This problem has been well documented, originally by such well known authors as Rubin (1976), Cosslett (1981), Manski and McFadden (1981), and Kmenta and Balestra (1986), and in a more recent context by Manski (1994), and Heckman, Ichimura, Smith and Todd (1998) among others.

Lien and Reardon (1990) examine the case of missing independent variable values in limited dependent variable models. Their results do not apply when the limited dependent variable itself has missing values. Bhat (1994) looks only at ordered categorical dependent variable responses and uses a bivariate normal with one variate representing the latent degree-of-missingness and the other for the latent variable underlying the ordered response. Fitzmaurice, Laird and Zahner (1996) used one logit equation as the selection equation and another for the latent variable underlying a *binary* response with some missing values. Hausman, Abrevaya and Scott-Morton (1998) considered misclassified discrete-response dependent variables but not missing value ones.

Unlike the earlier research mentioned above, the current paper will consider the case of categorical dependent variable responses that are unordered and have some missing values.

Three examples will be used to clarify the Restore-MLE method. The dependent variable for the primary example will be OWNRENT which takes on three values: OWN for home ownership, RENT for renting, and OTHER for neither OWN nor RENT. The other two examples will be introduced later.

One way of thinking about the biased sample selection problem that is implied by discrete response data that are not missing-at-random is depicted in Table 1. Table 1 shows a joint bivariate probability distribution between home ownership status (Own, Rent or Other) and selectivity status (Observed or Missing).

Information is directly available only for the joint probabilities in the first row, P_{o1}, P_{o2}, P_{o3} and for the marginal probability for Missing, P_m. It is well known that consistent maximum likelihood estimation of P_{o1}, P_{o2} and P_{o3} (e.g. using only the complete observations) will yield consistent estimates of the marginal probabilities P_1, P_2 and P_3 only when the missing data are truly missing-at-random.

Table 1: Joint Probability Distribution

	Own	Rent	Other	row
Observed	P_{o1}	P_{o2}	P_{o3}	P_o
Missing	P_{m1}	P_{m2}	P_{m3}	P_m
column	P_1	P_2	P_3	1.0

In an attempt to estimate these marginal probabilities consistently Table 2 introduces a simple structure for the missing joint probabilities such that $P_{m1}= C_1P_{o1}$, $P_{m2}= C_2P_{o2}$ and $P_{m3}= C_3P_{o3}$.

Table 2: Joint PDF - Simple Structure

	Own	Rent	Other	row
Observed	P_{o1}	P_{o2}	P_{o3}	P_o
Missing	C_1P_{o1}	C_2P_{o2}	C_3P_{o3}	P_m
column	P_1	P_2	P_3	1.0

If the missing data are truly missing-at-random then C_1, C_2, and C_3 will each be equal to the ratio obtained by dividing the number of missing cases by the number of complete cases. Therefore, the null and alternative hypotheses for testing whether the data are truly missing-at-random may be expressed as follows:

H_O: $C_1 = C_2 = C_3$
H_a: H_O not true

In addition to the three "legitimate" outcomes there is the fourth outcome called "missing".

The problem of estimating C_1, C_2, and C_3 was addressed by Marsh and Wells (1996) and Baxter and Marsh (1996). These papers proposed a two-step procedure using PROC CATMOD in the first step on the categorical dependent variable which included a category for the "missing" response. The second step was to use PROC REG to regress the vector of missing response probabilities onto the response probabilities of the legitimate (.i.e. nonmissing) categories.

Under the assumption of n independent observations, the likelihood function is given as follows:[1]

$$L = \prod_{i=1}^{n} \left[\ P_{o1i}^{y_{1i}} \ P_{o2i}^{y_{2i}} \ P_{o3i}^{y_{3i}} \ P_{mi}^{y_{mi}} \right]$$

where y_{1i}, y_{2i}, y_{3i} and y_{mi} are so-called dummy or binary variables that equal one when their outcome occurs and zero otherwise. Since a marginal probability is just the sum of the corresponding joint probabilities, Table 2 allows for the substitution of:

$$P_{mi} \ = \ C_1 P_{o1i} + C_2 P_{o2i} + C_3 P_{o3i}$$

in the likelihood function to obtain:

$$L = \prod_{i=1}^{n} \left[P_{o1i}^{y_{1i}} \ P_{o2i}^{y_{2i}} \ P_{o3i}^{y_{3i}} \left(C_1 P_{o1i} + C_2 P_{o2i} + C_3 P_{o3i} \right)^{y_{mi}} \right]$$

The two key aspects of this research are this formulation of the likelihood function and the expression for the joint probabilities which are defined as:

$$P_{oji} = \frac{exp(x_{ji}{}'\beta_j)}{\left(1 + C_j\right) \sum_{k=1}^{3} exp(x_{ki}{}'\beta_k)} \qquad \text{for } j = 1, 2, 3$$

and

$$P_{mji} = \frac{C_j \ exp(x_{ji}{}'\beta_j)}{\left(1 + C_j\right) \sum_{k=1}^{3} exp(x_{ki}{}'\beta_k)} \qquad \text{for } j = 1, 2, 3$$

where $\mathbf{x}_{ji}$ is a vector of independent variables, β_j is a vector of parameters and j = 1, 2, 3. The coefficient vector for the first category, β_1, is set equal to zero for normalization. The two corresponding marginal *row* probabilities are:

$$P_{oi} \ = \ \sum_{j=1}^{3} P_{oji} \ = \ \sum_{j=1}^{3} \left[\frac{\exp(x'_{ji}\beta_j)}{(1 + C_j) \sum_{k=1}^{3} \exp(x'_{ki}\beta_k)} \right]$$

and

$$P_{mi} \ = \ \sum_{j=1}^{3} P_{mji} \ = \ \sum_{j=1}^{3} \left[\frac{C_j \ \exp(x'_{ji}\beta_j)}{(1 + C_j) \sum_{k=1}^{3} \exp(x'_{ki}\beta_k)} \right]$$

The three marginal *column* probabilities are:

$$P_{ji} \ = \ P_{oji} + P_{mji} \ = \ \frac{(1 + C_j)\exp(x'_{ji}\beta_j)}{(1 + C_j) \sum_{k=1}^{3} \exp(x'_{ki}\beta_k)}$$

$$= \ \frac{\exp(x'_{ji}\beta_j)}{\sum_{k=1}^{3} \exp(x'_{ki}\beta_k)}$$

for j = 1, 2, 3.

THREE SIMULATED EXAMPLES

It is not unusual for a significant number of observations from business surveys to be missing values for some key variables of interest. This is particularly true of those ubiquitous product registration cards. However, in order to evaluate how well an estimator performs in restoring the missing data, it is useful to have a data set that is complete, and then simulate one or more patterns for the missing responses. See Marsh (1999) for five such missing data patterns. Consequently, in order to simulate missing data in business surveys a complete data set was extracted from the Panel Study of Income Dynamics (PSID). The PSID extract data set consisted of 9493 observations on nine socioeconomic variables that are typical of the kind of variables often

[1] See Marsh and Wells (1995) for a discussion of this likelihood function in the context of joint decision making where only the joint outcome is observed and individual decisions are "missing" (e.g. married couples, corporate boards, courtroom juries, Federal Reserve Board, etc.).

obtained from product registration cards. Three of the variables will serve as dependent variables in three missing response examples.

These three dependent variables are OWNRENT for home ownership status, EMPLOY for employment status and MARITAL for marital status.

The same set of independent variables was used for each of these three examples. The independent or explanatory variables were MALE, a dummy variable indicating male gender, EDUC, representing years of formal education, AGE, specifying the person's age in years, FAMINC, recording the family income (in thousands of dollars), and AGECHILD, for the age of the youngest child in the family. OWNRENT which was already discussed above, was broken down into three possible responses: a dummy variable OWN indicating home ownership, a dummy variable RENT indicating status as a renter, and OTHER to cover other situations.

To establish the baseline model, a multinomial logit model was estimated using the full data set before any values were designated as missing. This "true" model was designated Full-MLE and provides the correct or "target" model.

Once this target Full-MLE model is estimated, certain values of the OWNRENT variable can then be set to missing to pretend that their true values are not known. This will provide the target coefficients and target probabilities that the Partial-MLE and Restore-MLE estimators will attempt to determine. In order to test the missing-at-random hypothesis, Restore-MLE was rerun as CRestore-MLE with the missing-at-random restriction: $C_1 = C_2 = C_3$.

Table 3a: Home Ownership Status, OWNRENT (default group is OWN)

		Partial-MLE		Full-MLE		Restore-MLE		CRestore-MLE	
	Variable	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*
RENT	Const	4.3231	0.0001	4.2398	0.0001	4.2398	0.0001	5.8666	0.0001
	Male	-0.7106	0.0001	-0.6806	0.0001	-0.6806	0.0001	-0.8131	0.0001
	Educ	-0.0355	0.0008	-0.0434	0.0001	-0.0434	0.0001	-0.0620	0.0001
	Age	-0.0488	0.0001	-0.0485	0.0001	-0.0485	0.0001	-0.0469	0.0001
	Faminc	-0.0372	0.0001	-0.0370	0.0001	-0.0370	0.0001	-0.0290	0.0001
	Agechild	-0.0339	0.0001	-0.0376	0.0001	-0.0376	0.0001	-0.0259	0.0026
OTHER	Const	2.4952	0.0001	2.6006	0.0001	2.6006	0.0001	4.2169	0.0001
	Male	-0.3506	0.0114	-0.4146	0.0001	-0.4146	0.0001	-0.5747	0.0001
	Educ	0.0265	0.2357	0.0048	0.7832	0.0048	0.7832	-0.0166	0.4113
	Age	-0.0549	0.0001	-0.0549	0.0001	-0.0549	0.0001	-0.0533	0.0001
	Faminc	-0.0759	0.0001	-0.0737	0.0001	-0.0737	0.0001	-0.0631	0.0001
	Agechild	-0.0990	0.0001	-0.0990	0.0001	-0.0990	0.0001	-0.0849	0.0001
	d_1=					2.0784	0.0001	0.8533	0.0001
	d_2=					0.0000	1.0000	0.8533	0.0001
	d_3=					0.0000	1.0000	0.8533	0.0001
	Loglike =	-3959.6	n=5493	-6699.3	n=9493	-9080.0	n=9493	-10054.6	n=9493

In order to simulate a difficult missing-response problem, 4000 of the 9493 observations were designated as have missing values for the OWNRENT variable. If these missing values were distributed over the three possible outcomes in proportion to their relative frequencies, this would approximate a missing-at-random situation.

However, since the objective is to attempt to restore the true coefficients and probabilities when there is substantial sample selection bias, the 4000 missing values for this simulation experiment were all drawn from the first category (OWN). Since there were 4926 home owners in the original data set, this left only 926 values indicating home owner status.

In contrast, the values of OWNRENT for the 4037 renters and the 530 individuals who neither owned nor rented property were left untouched.

The most important evidence revealed in Table 3a is the fact that the Restore-MLE estimator is able to fully recover the original coefficients from Full-MLE in spite of having to deal with 4000 missing values for OWNRENT. The coefficients d_1, d_2 and d_3 are the square roots of C_1, C_2 and C_3 respectively. Restore-MLE correctly detected the fact that the correct values for d_2 and d_3 are zero since all the missing observations were taken from the first category.

Moreover, some distortions were evident in the coefficients from Partial-MLE since it not only did not have the full data to work with but was also using data that exhibited strong sample selection bias.

However, by directly forcing the assumption of missing-at-random on the model by imposing the restriction $C_1 = C_2 = C_3$, the CRestore-MLE produced even greater distortions in the coefficient values.

The inappropriateness of the missing-at-random assumption is further seen in the Likelihood Ratio Test which produced a chi-square test statistic value of 1949.2 strongly rejected that null hypothesis.

Table 3b: Estimation of Employment Status, EMPLOY (default group is JOB)

		Partial-MLE		**Full-MLE**		**Restore-MLE**		**CRestore-MLE**	
	Variable	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*
N	Const	0.2115	0.4752	-0.0456	0.8477	-0.0456	0.8477	0.8653	0.0011
O	Male	-0.1624	0.1342	-0.0169	0.8496	-0.0169	0.8496	-0.0594	0.5600
J	Educ	-0.0559	0.0013	-0.0521	0.0002	-0.0521	0.0002	-0.0562	0.0005
O	Age	-0.0044	0.3206	-0.0025	0.4573	-0.0025	0.4573	-0.0019	0.6088
B	Faminc	-0.0489	0.0001	-0.0481	0.0001	-0.0481	0.0001	-0.0425	0.0001
	Agechild	0.0074	0.4943	0.0103	0.2243	0.0103	0.2243	0.0192	0.0449
R	Const	-12.1870	0.0001	-11.7265	0.0001	-11.7265	0.0001	-10.1365	0.0001
E	Male	0.6064	0.0001	0.5158	0.0001	0.5158	0.0001	0.3777	0.0033
T	Educ	-0.0070	0.7100	-0.0054	0.6872	-0.0054	0.6872	-0.0169	0.2977
I	Age	0.1988	0.0001	0.1915	0.0001	0.1915	0.0001	0.1845	0.0001
R	Faminc	-0.0222	0.0001	-0.0181	0.0001	-0.0181	0.0001	-0.0170	0.0001
E	Agechild	-0.0237	0.1609	-0.0307	0.0134	-0.0307	0.0134	-0.0266	0.0514
O	Const	-0.2522	0.3634	-0.3105	0.1579	-0.3105	0.1579	0.5940	0.0208
T	Male	-1.0458	0.0001	-1.0931	0.0001	-1.0931	0.0001	-1.2003	0.0001
H	Educ	-0.0979	0.0001	-0.0992	0.0001	-0.0992	0.0001	-0.1080	0.0001
E	Age	0.0442	0.0001	0.0460	0.0001	0.0460	0.0001	0.0470	0.0001
R	Faminc	-0.0748	0.0001	-0.0758	0.0001	-0.0758	0.0001	-0.0654	0.0001
	Agechild	0.0041	0.6905	-0.0015	0.8585	-0.0015	0.8585	0.0065	0.5158
	d_1=					1.3453	0.0001	0.8533	0.0001
	d_2=					0.0000	1.0000	0.8533	0.0001
	d_3=					0.0000	1.0000	0.8533	0.0001
	d_4=					0.0000	1.0000	0.8533	0.0001
	Loglike=	-3550.4	n=5493	-5959.8	n=9493	-10002.6	n=9493	-10968.2	n=9493

A second example examines employment status using the variable EMPLOY with dummy variables JOB, indicating someone working for money, NOJOB, for someone seeking work, RETIRE, for people who are retired, and OTHER, for all others not in the labor force. Again, the important result is the fact that the Restore-MLE method is able to exactly reproduce the true Full-MLE results. Here again, the Restore-MLE has successfully detected that all of the missing values were from the first group, JOB, and that none came from any of the other three groups. Thus, the values of d_2, d_3 and d_4 are appropriately all zero. The Partial-MLE estimator does not find the correct coefficient values nor does the CRestore-MLE estimator. Moreover, the Likelihood Ratio Test yields a chi-square test statistic value of 1931.4

Table 3c: Estimation of Marital Status, MARITAL (default group is MARR)

		Partial-MLE		Full-MLE		Restore-MLE		CRestore-MLE	
	Variable	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*	*Estimate*	*P-Value*
N	Const	9.0778	0.0001	9.2353	0.0001	9.6796	0.0001	10.0169	0.0001
E	Male	-7.6135	0.0001	-7.9532	0.0001	-8.3909	0.0001	-7.2195	0.0001
V	Educ	0.1004	0.0001	0.1071	0.0001	0.1069	0.0001	0.0769	0.0001
E	Age	-0.0742	0.0001	-0.0707	0.0001	-0.0707	0.0001	-0.0745	0.0001
R	Faminc	-0.0454	0.0001	-0.0493	0.0001	-0.0493	0.0001	-0.0412	0.0001
	Agechild	-0.2821	0.0001	-0.2932	0.0001	-0.2979	0.0001	-0.2471	0.0001
W	Const	0.7907	0.2311	1.4023	0.0085	1.8515	0.0063	1.8654	0.0090
I	Male	-8.6563	0.0001	-9.1817	0.0001	-9.6230	0.0001	-8.4010	0.0001
D	Educ	0.0071	0.7428	-0.0003	0.9837	-0.0003	0.9861	-0.0170	0.3674
O	Age	0.0943	0.0001	0.0906	0.0001	0.0906	0.0001	0.0911	0.0001
W	Faminc	-0.0293	0.0001	-0.0258	0.0001	-0.0259	0.0001	-0.0187	0.0001
	Agechild	-0.1773	0.0001	-0.1751	0.0001	-0.1805	0.0001	-0.1343	0.0001
D	Const	5.4996	0.0001	5.7916	0.0001	6.2399	0.0001	6.3814	0.0001
I	Male	-7.7152	0.0001	-8.0411	0.0001	-8.4794	0.0001	-7.3057	0.0001
V	Educ	0.0696	0.0002	0.0791	0.0001	0.0787	0.0001	0.0583	0.0007
O	Age	0.0006	0.8719	-0.0001	0.9594	-0.0001	0.9731	-0.0006	0.8567
R	Faminc	-0.0240	0.0001	-0.0290	0.0001	-0.0289	0.0001	-0.0226	0.0001
C	Agechild	-0.1941	0.0001	-0.1916	0.0001	-0.1969	0.0001	-0.1514	0.0001
S	Const	6.8971	0.0001	7.2487	0.0001	7.6946	0.0001	7.8736	0.0001
E	Male	-7.7252	0.0001	-8.1642	0.0001	-8.6008	0.0001	-7.4344	0.0001
P	Educ	-0.0197	0.3345	-0.0138	0.4162	-0.0139	0.4141	-0.0351	0.0666
A	Age	-0.0150	0.0010	-0.0174	0.0001	-0.0174	0.0001	-0.0187	0.0001
R	Faminc	-0.0279	0.0001	-0.0340	0.0001	-0.0341	0.0001	-0.0272	0.0001
	Agechild	-0.1976	0.0001	-0.1969	0.0001	-0.2020	0.0001	-0.1565	0.0001
	d_1=					1.7123	0.0001	0.8533	0.0001
	d_2=					0.0000	1.0000	0.8533	0.0001
	d_3=					0.0000	1.0000	0.8533	0.0001
	d_4=					-0.0593	0.0265	0.8533	0.0001
	Loglike=	-4220.0	n=5493	-6792.5	n=9493	-9831.7	n=9493	-12064.9	n=9493

which again strongly rejects the missing-at-random null hypothesis.

The third and final example to demonstrate the effectiveness of the Restore-MLE method uses the dependent variable, MARITAL, indicating the person's marital status. MARITAL is represented by five dummy variables beginning with MARR which means married and living together (not separated), NEVER, for never married, WIDOW, for widowed, DIVORC, for divorced, and SEPAR, for separated.

In this case the Restore-MLE estimated coefficients and those of the Full-MLE method are close but do not match exactly. The reason for this may be the somewhat ambiguous nature of the SEPAR variable. Some people are separated because they are having difficulty in their marriage and are, in effect, in an intermediate stage between marriage, MARR, and divorce, DIVORC.

However, some others who are not currently living with their spouse may be in the armed forces or in the hospital so their state of separation is, in some sense, qualitatively different than that of the first group of separated couples. In some sense this last group probably should have really been assigned to MARR.

The d_4 value of –0.0593 corresponds to a C_4 value of 0.0035 indicating that some of the missing responses which were really from MARR have been mistakenly assigned to SEPAR by the Restore-MLE estimator.

This demonstrates the importance of working with categories that are distinct and clearly defined as well as conceptually coherent and meaningful. When people see choices as ambiguous or poorly grouped, then the Restore-

Table 4: True Relative Frequency vs. Estimated Probabilities

	FIRST	SECOND	THIRD	FOURTH	FIFTH
OWNRENT	**OWN**	**RENT**	**OTHER**		
Full Rel-Freq	0.5189	0.4253	0.0558		
Full MLE	0.5189	0.4253	0.0558		
Restore MLE	0.5189	0.4253	0.0558		
Partial MLE	0.4935	0.4499	0.0566		
CRestore MLE	0.2344	0.6875	0.0781		
EMPLOY	**JOB**	**NOJOB**	**RETIRE**	**OTHER**	
Full Rel-Freq	0.6542	0.0836	0.1484	0.1138	
Full MLE	0.6542	0.0836	0.1484	0.1138	
Restore MLE	0.6542	0.0836	0.1484	0.1138	
Partial MLE	0.6539	0.0849	0.1460	0.1152	
CRestore MLE	0.5059	0.1529	0.1769	0.1647	
MARITAL	**MARR**	**NEVER**	**WIDOW**	**DIVORC**	**SEPAR**
Full Rel-Freq	0.5649	0.1632	0.0986	0.1088	0.0645
Full MLE	0.5649	0.1632	0.0986	0.1088	0.0645
Restore MLE	0.5645	0.1632	0.0986	0.1092	0.0645
Partial MLE	0.5520	0.1673	0.0959	0.1115	0.0733
CRestore MLE	0.4153	0.2240	0.1163	0.1560	0.0885

MLE estimator will see them as ambiguous or poorly grouped as well. Here again the Partial-MLE and CRestore-MLE estimators have difficulty matching the Full-MLE coefficient values. The Likelihood Ratio Test yields a chi-square test statistic value of 4466.4 which even more emphatically rejects the missing-at-random null.

Table 4 presents the relative frequencies from the original data set as well as the estimated marginal probabilities from the Full-MLE, Restore-MLE, Partial-MLE and CRestore-MLE methods.

Since Full-MLE is simply estimating a standard multinomial logit model using the original data before any dependent variable values were set to missing, it is no surprise (although it is certainly reassuring) that the Full-MLE marginal probabilities match the relative frequencies exactly.

More importantly, in both the OWNRENT and EMPLOY examples, the Restore-MLE estimator produces exactly the same average marginal probability values as the Full-MLE and relative frequencies. In the MARITAL example, however, the Restore-MLE probabilities are slightly off the mark in the first and fourth categories.

In contrast, the Partial-MLE and CRestore-MLE estimators do not do so well. In particular, these two estimators consistently underestimate the probabilities from the first category in each of the three examples. This is not surprising since all of the missing dependent variable values came from the first category in each of the three examples.

The Newton-Raphson algorithm (NRR) was run in SAS 6.12 (TSO45) on a UNIX[2] computer (SUN Enterprise 4000) under SUN OS 5.6 for all three of these simulations.

[2] I am grateful to James J. Ostlund for providing and managing this "Stats" computer, and to Kevin D. Brunson for consulting assistance.

Table 5: Mean Squared Errors, CPU Time, Iterations, LogLikelihood

	MSE	CPU time	Iterations	LogLike
OWNRENT				
Full MLE	0.3137	2.2 min.	24	-6699.3
Partial MLE	0.3138	1.3 min.	24	-3959.6
CRestore MLE	0.1165	2.1 min.	13	-10054.6
Restore MLE	0.0357	2.5 min.	13	-9080.0
EMPLOY				
Full MLE	0.2330	4.8 min.	27	-5959.8
Partial MLE	0.2329	2.7 min.	27	-3550.4
CRestore MLE	0.0585	4.4 min.	15	-10968.2
Restore MLE	0.0260	5.6 min.	16	-10002.5
MARITAL				
Full MLE	0.1869	7.6 min.	30	-6792.5
Partial MLE	0.1871	4.1 min.	27	-4220.0
CRestore MLE	0.0466	8.6 min.	16	-12064.9
Restore MLE	0.0125	14.3 min.	40	-9831.7

The CPU time for these simulations varied from 1.3 minutes for the Partial-MLE estimator in the OWNRENT example to 14.3 minutes for the Restore-MLE estimator in the MARITAL example.

In each of these examples the Partial-MLE took the least amount of time while the Restore-MLE took the most time. The Full-MLE took more time than the Partial-MLE because, although they had the same number of parameters to estimate, the Full-MLE had to process more observations.

In general the CRestore-MLE and the Restore-MLE took the fewest number of iterations, except for the MARITAL example where the Restore-MLE took considerably more iterations than the others, possibly related to the difficulties discussed earlier for MARITAL status. The most important statistics in Table 5 are the mean squared errors which give the average squared deviations of the estimated probabilities from the actual, observed binary (0,1) outcomes. The Full-MLE and Partial-MLE estimators had the highest mean squared errors which is not surprising given that they have fewer parameters, and, therefore, less flexibility than the other two estimators. The CRestore-MLE estimator had one additional parameter and somewhat smaller mean squared errors than the Full-MSE and Partial-MLE estimators. The Restore-MLE had the smallest mean squared errors of all which, again, was not surprising since it has the most parameters.

Finally, the last column in Table 5 reports the loglikelihood scores which were used to calculate the Likelihood Ratio Test statistics discussed earlier. The missing-at-random null was rejected in all three examples.

CONCLUSION

This research has demonstrated the ability of the Restore-MLE estimator to recover the coefficients and probabilities of the true model with data typically found in business surveys of consumers. These results show that simply ignoring the observations with missing values is not optimal. Restore-MLE merges the typical estimation step and prediction step into one step with improved results as shown throughout this paper. Moreover, recovering the underlying coefficients and probability structure is not difficult using the fairly simple SAS program provided in the Appendix.

REFERENCES

Baxter, S.S. and L.C.Marsh (1996), "Testing for and Correcting for Missing Data in Survey Responses", *Proceedings of the Midwest SAS Users Group*, 7, 165-174.

Bhat, C.R. (1994), "Imputing a Continuous Income Variable from Grouped and Missing Income Observations," *Economics Letters*, 46, 311-319.

Cosslett, S.R. (1981), "Efficient Estimation of Discrete-Choice Models," in *Structural Analysis of Discrete Data with Econometric Applications*, Chapter 2, edited by C.F.Manski and D.McFadden.

Fitzmaurice, G.M., N.M.Laird, and G.E.Zahner (1996), "Multivariate Logistic Models for Incomplete Binary Responses,"*Journal of American Statistical Association*, 91, 99-108.

Hausman, J.A., J.Abrevaya and F.M.Scott-Morton (1998), "Misclassification of the Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics*, 87, 239-269.

Heckman, J.J., H.Ichimura, J.Smith and P.Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.

Lee, B.J. and L.C. Marsh (1998), "Nested Logit Analysis of Missing Response Observations," *Applied Economics Letters*, 5, 751-755.

Lien, D. and D.Rearden (1990), "Missing Measurements in Discrete Response Models,"*Economics Letters*, 32, 231-235.

Manski, C.F. (1994), "The Selection Problem," in *Advances in Econometrics*, Cambridge University Press, 143-170.

Manski, C.F. and D.McFadden (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, Chapter 1, edited by C.F.Manski and D.McFadden.

Marsh, L.C. (1999), "Alternative Approaches to Correcting for Missing Categorical Dependent Variable Responses", working paper, University of Notre Dame.

Marsh, L.C. and K.L.Wells (1995), "Karnaugh Maps, Interaction Effects, and Creating Composite Dummy Variables for Regression Analysis in SAS Software". 1995 SAS Conference Proceedings, *SUGI*, 20, 1194-1203.

Marsh, L.C. and K.L.Wells (1996), "An Analysis of Changes in the Public Employment Using a New Method of Accounting for Missing Data on Job Loss Status", working paper, University of Notre Dame.

Rubin, D.B. (1976), "Inference and Missing Data," *Biometrika*, 63(3), 581-592.

CONTACT INFORMATION

Lawrence C. Marsh
Department of Economics
University of Notre Dame
302 E. Pokagon Street
South Bend, IN 46617-1226
Phone: (219) 287-0458
Fax: 9219) 287-2173
E-mail: Lawrence.C.Marsh.1@nd.edu
Web1: www.nd.edu/~lmarsh
Web2: www.nd.edu/~meg
Notre Dame SAS Homepage:
www.nd.edu/~meg/NDSAS.html

Trademarks

APPENDIX: SAS® PROGRAM THAT CORRECTS FOR MISSING DATA

```
options nocenter linesize=80 pagesize=max;              * Home Ownership Example;
libname survey 'PSID';
data one;  set survey.SUGI2000; * Questions: call Larry Marsh at (219) 287-0458;
INDEX=RANUNI(123456789);            FAMINC=FAMINC/1000;
keep OWNRENT OWN  RENT  OTHER  MALE EDUC AGE FAMINC AGECHILD INDEX;
PROC SORT data=one out=one;     by OWNRENT INDEX;
data one;   set one;   n+1;   *create MISSING=1 for missing responses and MISSING=0 otherwise;
if n le 4000  then OWNIT=0;   else OWNIT=OWN;          * simulate 4000 MISSING responses;
if n le 4000  then MISSING=1;  else MISSING=0;  *drop these two lines when data are really missing;

PROC NLP data=one maxit=99999  maxfu=99999  cov=2 gtol=0 tech=nrr outest=betas  PCOV;
b11=0;     b21=0;     b31=0;     b71=0;     b91=0;     b121=0;
parms  d1=1,
d2=1,b12=-3.1665,b22=-4.0821,b32=0.0138,b72=0.0174,b92=0.000034,b122=0.1969,
d3=1,b13=2.0921,b23=-0.1055,b33=0.1209,b73=-0.0533,b93=-0.00002,b123=-0.0963;

XB1=b11+b21*MALE+b31*EDUC+b71*AGE+b91*FAMINC+b121*AGECHILD;
XB2=b12+b22*MALE+b32*EDUC+b72*AGE+b92*FAMINC+b122*AGECHILD;
XB3=b13+b23*MALE+b33*EDUC+b73*AGE+b93*FAMINC+b123*AGECHILD;

          C1=d1*d1;                       C2=d2*d2;                      C3=d3*d3;

     EXPXB1=exp(XB1);                EXPXB2=exp(XB2);               EXPXB3=exp(XB3);

sumex= EXPXB1+ EXPXB2+ EXPXB3;
sumcexc=(C1*  EXPXB1/(1+C1))+(C2*  EXPXB2/(1+C2))+(C3*  EXPXB3/(1+C3));

max loglike;    *please reference Lawrence C. Marsh, SUGI 25 Proceedings (2000);

loglike=MISSING*log(sumcexc)+OWNIT*(XB1-log(1+C1))+RENT*(XB2-log(1+C2))
+OTHER*(XB3-log(1+C3))-log(sumex);

data betas; set betas; if _TYPE_='PARMS';
PROC PRINT  data=betas;

data one;  if _N_=1 then set betas; set one;    XB1=0;
XB2=b12+b22*MALE+b32*EDUC+b72*AGE+b92*FAMINC+b122*AGECHILD;
XB3=b13+b23*MALE+b33*EDUC+b73*AGE+b93*FAMINC+b123*AGECHILD;
C1=d1*d1;   C2=d2*d2;   C3=d3*d3;   EXPXB1=exp(XB1);  EXPXB2=exp(XB2);  EXPXB3=exp(XB3);
sumex= EXPXB1+ EXPXB2+ EXPXB3;
p1= EXPXB1/sumex;      p2= EXPXB2/sumex;       p3= EXPXB3/sumex;
pm1 = C1* p1/(1+C1);    pm2 = C2*p2/(1+C2);      pm3 =  C3*p3/(1+C3);

PROC MEANS;   var p1 p2 p3 pm1 pm2 pm3;     /* PROC PRINT; var p1 p2 p3 pm1 pm2 pm3;*/
```