OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 62, 2 (2000)
0305-9049

SAMPLE SELECTION BIAS CORRECTION FOR MISSING RESPONSE OBSERVATIONS

Byung-Joo Lee and Lawrence C. Marsh[†]

I. INTRODUCTION

Missing data can pose a significant problem when the goal is to understand the behavior of all subjects rather than just the ones with complete data. Often researchers simply have ignored missing data with the hope the data are missing at random (see Rubin (1976), Kmenta and Balestra (1986)). Unfortunately, data frequently are missing in a nonignorable manner, and that can lead to substantial distortions when only complete data are used. Guttman and Menzefricke (1983) have estimated models with the complete data and then used the estimated models to impute the missing data. Bhat (1994) uses a bivariate normal error structure where one dimension measures willingness to report income and the other dimension is a latent measure of the log of income. His analysis is strictly for ordered response data such as income groups. It does not provide the flexibility needed for dealing with general response categories. Fitzmaurice, Laird and Zahner (1996) propose some multivariate logistic models for dealing with missing binary response data. Their approach is basically the same as Bhat (1994) and other researchers in that they estimate a logit for non-response and another logit for the mean of the responses. Thus, they provide a likelihood function based on conditional distributions for handling systematically missing data. If data are missing systematically, this practice can lead to a sample selection problem. Systematically missing data pose a related but somewhat different variation on the sample-selection methods developed by Manski (1994) and Heckman *et al*. (1998).

Often survey data include a number of incomplete or missing responses to a dependent variable of interest. To correct the sample selection bias problem, a common approach is to specify a selection equation and estimate the structural parameters conditional on selected (or observed) responses using a joint density or a product of conditional and joint density functions. However, except the normal distributional assumption of error terms, the conditional density and/or joint density functions are very hard to obtain,

[†]We appreciate the comments and suggestions of the editor and an anonymous referee, which greatly enhance the quality of our paper. However, we are solely responsible for all remaining errors.

and the maximum likelihood estimation of the structural parameters is very complicated. In this paper, we propose a simpler solution to correct the sample selection bias problem of missing response observations in categorical dependent variables and show that this estimator is consistent for the structural parameter. We develop a maximum likelihood function to accommodate the joint probability structure implicit in the missing responses.

In Section II we briefly summarize the standard procedure for sample selection bias models and discuss possible problems with typical selection bias models. In Section III we propose a simpler procedure that is consistent with the standard procedure to recover the joint probability distribution. The proposed procedure differs from the standard procedure in two ways. First, it assumes that those failing to respond fit into one of the response categories. Second, the likelihood of non-response is estimated as a function of the dependent variable categories rather than as a direct function of the set of explanatory variables. The proposed procedure can use the nested multinomial logit model. Section IV applies this estimation method to the problem of determining the correct distribution of job-loss status with data from the Panel Study of Income Dynamics (PSID) for the years 1977-1991. Finally, Section V provides the summary and conclusions of our analysis and suggestions for future research.

II. SAMPLE SELECTION BIAS CORRECTION

When survey data contain missing responses, the probabilities for the responses consist of two sets: One set of probabilities corresponds to the observed responses and a second set corresponds to the unobserved (missing) responses. If we ignore missing response observations, we will encounter sample selection bias problems. Manski and McFadden (1981) and Cosslett (1981) provide extensive discussion about discrete choice models and their estimation. Lien and Rearden (1990) investigate the missing data problem in various discrete response models. Consider the following multinomial response model with selection criteria:

$$y_{tj}^{*} = x_{tj}'\beta_j + \varepsilon_{tj},\ t = 1, 2, \ldots, T \text{ and } j = 1, 2, \ldots, J \qquad (1)$$

where y_{tj}^{*} is a latent variable but the response variable y_{tj} is observed where $y_{tj} = 1$ and $y_{tk} = 0$ for all $k \neq j$ if the t^{th} individual belongs to the j^{th} category. Individual response variable y_{tj} is observed if the following selection criteria is met:

$$s_t^{*} = x_t'\gamma + \delta_t \text{ and } s_t^{*} \geq 0.$$

Define $s_t = 1(s_t^{*} \geq 0)$, where $1(\lambda)$ is an indicator function which returns to one if λ is true and zero otherwise. In other words, we can observe the t^{th} individual response only if s/he chooses to respond to a particular question. This is a typical sample selection bias problem. To estimate the structural

parameter β consistently, we need to use the joint density function of (y_{tj}^*, s_t) or the product of conditional density of y_{tj}^* and the marginal density of s_t. The log likelihood function for the selection problem is as follows:

$$\ln L(\beta, \gamma) = \sum_{t=1}^{T} \left\{ s_t \sum_{j=1}^{J} y_{tj} (\ln \Pr(y_{tj} = 1 | s_t = 1, \beta, \gamma) + \ln \Pr(s_t = 1 | \gamma)). \right.$$

$$\left. + (1 - s_t) \ln \Pr(s_t = 0 | \gamma) \right\} \tag{2}$$

To estimate this log-likelihood function, we need to derive the conditional density of y_{tj}^* and the marginal density of s_t. Typically we assume that the selection equation has a probit or logit model structure. We can also assume that the errors in the response equation have a normal distribution or a multinomial logit probability distribution. However, for practical purposes, when we have a multinomial response model, it is much simpler to assume that y_{tj}^* has multinomial logit probabilities because of the closed form expression of the probability of each alternative. If we assume that $(\varepsilon_{tj}, \delta_t)$ has an extreme value distribution, then the conditional probability of y_{tj} given s_t or the joint probability of (y_{tj}, s_t) is not very easy to derive. To avoid this complication, we can assume that $(\varepsilon_{tj}, \delta_t)$ has a multivariate normal distribution, and the conditional distribution of ε_{tj} given δ_t is also normal. However, in this case, the multinomial response probabilities involve multiple integral evaluations and the maximum likelihood estimation becomes very time consuming and almost impractical for the higher dimension response models even with modern high-speed computing power. In the next section, we will propose a simpler technique to estimate the structural parameters when we have a selection bias problem.

III. THE JOINT MULTINOMIAL PROBABILITY DISTRIBUTION

In this section we will start with the same multinomial response model as in equation (1).

$$y_{tj}^* = x_{tj}' \beta_j + \varepsilon_{tj},\ t = 1, 2, \ldots, T \text{ and } j = 1, 2, \ldots, J \tag{3}$$

All notations have exactly same meaning as in the previous section. For simplicity, assume that ε_{tj} has a multinomial logit distribution. In typical survey data, when there are J possible responses, each individual chooses either one of J categories or chooses not to respond at all. Then those who did not respond to a particular question become missing responses for our analytical purposes. As we discussed in the previous sections, these missing responses cause a sample selection bias problem if we do not properly take this into consideration for estimation. In this section, we propose a new approach to handling missing responses that is different from the standard approach to sample selection bias correction.

In survey data, we observe that each individual responds to either one of J categories or does not respond (missing response). When an individual does not respond, but if s/he would have responded, s/he would have belonged to any one of J categories. We will treat the missing responses from this perspective. This is a key assumption that is needed to avoid the possibility that non-response may be driven by some missing option. Our method is appropriate when each of the non-respondents truly belongs to one of the specified categories. Our method is not appropriate when non-respondents fail to respond because they don't fit into any of the offered categories. In other words, we do not claim to be able to correct for poorly designed surveys or inappropriately worded survey questions, or even for those cases where the best possible survey cannot adequately account for all relevant responses. The following table will clarify this concept. For simplicity, assume that $J = 4$ categories.

Table 1 shows the joint probabilities of the observed responses and the marginal probability of the missing response, where P_{tjo} is the joint probability of the t^{th} individual responding to the j^{th} category which is observable, and P_{tm} is the probability of a missing response. We can decompose the marginal probability of missing as following. When the t^{th} individual did not respond to the question, but if s/he would have responded, s/he could have belonged to any one of four categories. The missing probability, P_{tm}, is decomposed into four joint probabilities and is considered as the marginal probability of the four joint probabilities as in Table 2.

If the missing observations are missing at random, then for each category, the missing joint probability is proportional to the observed joint probability (i.e. the observed probabilities are independent of missing probabilities.) If this is the case, there is no sample selection bias problem, and we can

TABLE 1
Probability Distribution of Four Categories

Category	*1*	*2*	*3*	*4*	*Missing*	*Total*
Probability	P_{t1o}	P_{t2o}	P_{t3o}	P_{t4o}	P_{tm}	1.0

TABLE 2
Joint Probability Distribution of Observed and Missing Responses

Category	*1*	*2*	*3*	*4*	*Marginal Prob.*
Observed	P_{t1o}	P_{t2o}	P_{t3o}	P_{t4o}	P_{to}
Missing	P_{t1m}	P_{t2m}	P_{t3m}	P_{t4o}	P_{tm}
Marginal Prob.	P_{t1}	P_{t2}	P_{t3}	P_{t4}	1.0

estimate the structural parameters consistently by using only the observed responses. However, if missing observations are not missing at random, we will have a selection bias problem if we use only observed data. In this case, the ratio between observed and missing probabilities for each category will be different. To facilitate this idea, we assume that each category, whether observed or missing, has a probability structure of its own and each category has its own proportional weight. This is shown in Table 3.

After suppressing the individual t subscript, we can see the missing marginal probability as the sum of the properly weighted observed probabilities as follows:

$$\alpha_1 P_{1o} + \alpha_2 P_{2o} + \alpha_3 P_{3o} + \alpha_4 P_{4o} = P_m \tag{4}$$

If the missing observations are missing at random regardless of category (i.e., if they are independent), then the weights, $\alpha_j' s$, $j = 1, 2, 3, 4$ are equal. However, if the missing observations are not missing at random, then each category has different weight and it must be estimated differently for each of the four alternatives. Therefore, the testable hypothesis of missing-at-random may be expressed as:

$$H_o\colon \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$

$$H_A\colon not\, H_o \tag{5}$$

The rejection of this null hypothesis suggests that the missing responses are not missing at random, and there is a sample selection bias problem. If the missing data are deleted, the remaining data will not provide consistent estimates for the structural parameters.

From equation (3), we assume that ε_{tj} has a multinomial logit distribution. This means that the marginal probability (P_{tj}) of each category j has a multinomial logit probability. However, what we observe from data is the joint probability of each observed category and the marginal probability of missing observations. To derive the joint probability of each observed category, from Table 3 we can see that

$$(1 + \alpha_j) P_{tjo} = P_{tj},$$

where

TABLE 3
Joint Probability Distribution in Terms of Observed Probabilities

Category	*1*	*2*	*3*	*4*	*Marginal Prob.*
Observed	P_{t1o}	P_{t2o}	P_{t3o}	P_{t4o}	P_{to}
Missing	$\alpha_1 P_{t1o}$	$\alpha_2 P_{t2o}$	$\alpha_3 P_{t3o}$	$\alpha_4 P_{t4}$	P_{tm}
Marginal Prob.	P_{t1}	P_{t2}	P_{t3}	P_{t4}	1.0

$$P_{tj} = \frac{e^{x'_{tj}\beta_j}}{\sum_{i=1}^{J} e^{x'_{ti}\beta_i}}$$

is the multinomial logit marginal probability of the j^{th} alternative for the t^{th} individual. The observed (joint) probability of j^{th} category is then

$$P_{tjo} = \frac{e^{x'_{tj}\beta_j}}{(1+\alpha_j)\sum_{i=1}^{J} e^{x'_{ti}\beta_i}}.$$

The marginal probability of observed is

$$P_{to} = \sum_{j=1}^{J} P_{tjo} = \sum_{j=1}^{J} \frac{e^{x'_{tj}\beta_j}}{(1+\alpha_j)\sum_{i=1}^{J} e^{x'_{ti}\beta_i}},$$

and the marginal probability of missing is

$$P_{tm} = \sum_{j=1}^{J} \alpha_j \cdot P_{tjo} = \sum_{j=1}^{J} \alpha_j \cdot \frac{e^{x'_{tj}\beta_j}}{(1+\alpha_j)\sum_{i=1}^{J} e^{x'_{ti}\beta_i}}.$$

Thus, we express the missing probability as the weighted sum of observed joint probabilities. By using these probabilities, we can directly estimate the model parameters. This approach is not only simpler, because only the structural parameters need to be estimated (no need to estimate the selection equation parameters), but it also makes the joint probability calculation straightforward. As we have seen in the previous section, the joint probability or the conditional probability derivation could be either very complicated or not practical under the standard sample selection bias correction procedure.

We can write the log-likelihood function for our multinomial response model with missing response observations as following:

$$\begin{aligned} \ln L(\beta, \alpha) &= \sum_{t=1}^{T} \left\{ \left(\sum_{j=1}^{J} y_{tj} \cdot \ln P_{tjo} \right) + y_{tm} \cdot \ln P_{tm} \right\} \\ &= \sum_{t=1}^{T} \left\{ \left(\sum_{j=1}^{J} y_{tj} \cdot \ln P_{tjo} \right) + y_{tm} \cdot \ln \left(\sum_{j=1}^{J} \alpha_j \cdot P_{tjo} \right) \right\} \end{aligned} \quad (6)$$

where $y_{tj} = 1$ if the t^{th} individual belongs to the j^{th} category and 0

otherwise and $y_{tm} = 1$ if the t^{th} individual response is missing and zero otherwise. P_{tjo} is the t^{th} individual *observed joint probability* defined above. The log-likelihood function in equation (6) is equivalent to the log-likelihood function of equation (2) that corrects the sample selection bias. The first part of the log-likelihood function is the observed joint probability of each category and the second part is the marginal probability of missing observations expressed as the weighted sum of the observed joint probabilities. Therefore, the estimates are consistent for the structural model parameters. To contrast the proposed procedure and the standard procedure, we provide the simple example in Appendix I.

The next section will use this log-likelihood function to estimate the empirical model of the public sector job-loss status. These data contain a significant number of missing responses.

IV. ESTIMATION OF WORK STATUS WITH MISSING RESPONSES

We analyzed public sector employment status using the 1991 wave of the Panel Study of Income Dynamics (PSID), which covers everyone who worked in the public sector anytime between 1977 and 1991. The data set includes 2117 observation with 268 missing responses for the job status variable.[1] Job status categories include: 1) stayed on the job (work), 2) quit the job (quit), 3) dismissed from the job (dismiss), 4) left the job for retirement or health related reasons (other), 5) left the job but for unknown reasons (unknown), 6) person's job status is completely missing (missing). The sixth response is due solely to the missing nature of the data. The explanatory variables used to predict job status are two job skill dummy variables for three job skill groups (control group: Unskilled worker), two education level dummy variables for three education groups (control group: Less than High School), job duration, real wage, state unemployment rate, and years of job experience since 18 years old. Data were reduced by 350 observations due to the unobservability of one or more values of these explanatory variables.[2] Therefore, the final analysis contains 1767 observations with six categories including missing observations. Table 4 shows the frequency distribution of job status of total available observations.

Table 5 reports descriptive statistics for the explanatory variables used to determine the job status of public sector employment.

[1]We are grateful to Karin L. Wells for providing us with this particular PSID extract which she developed from the PSID data for the years 1977 through 1991.

[2]We assume that the explanatory variables are missing at random. This is the exogenous sampling procedure explained by Manski and McFadden (1981). At the referee's suggestion, we tested whether this assumption is justified. The following is the frequency distribution table before deleting 350 observations due to one or more missing explanatory variable values. To see whether each category is similarly represented before and after deleting observations by missing explanatory variables, the testable null hypothesis is: H_o: $\hat{P}_i = P_i$ *vs.* H_1: *not* H_o, where $\hat{P}_i$ is the proportion of each category from Table 4, and P_i is the proportion from Table F.1.

continued overleaf

TABLE 4
Frequency Distribution of Six Categories of Job Status

Work	*Quit*	*Dismiss*	*Other*	*Unknown*	*Missing*	*Total*
368	227	109	172	688	203	1767
(0.2083)	(0.1285)	(0.0617)	(0.0973)	(0.3894)	(0.1149)	(1.0)

TABLE 5
Descriptive Statistics of Explanatory Variables

Variables	*Description*	*Mean*	*Std. Dev.*
Occu1	Skilled worker	0.2716	0.4449
Occu2	Semi-skilled worker	0.4776	0.4996
Hschool	High School Diploma	0.3724	0.4836
College	College Graduate	0.3780	0.4850
Duration	Job Duration (in years)	3.3662	3.5818
Rwage	Real Wage (per hour in 1983 $)	8.0867	4.2618
Unemp	State Unemployment rate (%)	7.1781	2.0723
Yrs18	Job Experience since 18	16.1975	12.0404

We estimated three different models. Model 1 completely ignores the observations with missing responses. This model incorrectly assumes that the conditional distribution (instead of marginal distribution) of ε_{tj} has a multinomial distribution. Model 2 includes missing responses but assumes missing response is random and no sample selection bias exists (constant α). Model 3 also includes missing responses, but this model allows the possibility of non-random missing responses, and, therefore, we need to correct for sample selection bias to obtain consistent estimators of structural

TABLE F.1
Frequency Distribution of Six Categories of Job Status (Entire Sample)

Category	*Work*	*Quit*	*Dismiss*	*Other*	*Unknown*	*Missing*	*Total*
Frequency	410	257	130	215	837	268	2117
Proportion	(0.1937)	(0.1214)	(0.0614)	(0.1016)	(0.3954)	(0.1266)	(1.0)

To compare this frequency distribution with Table 4, we used the Pearson's goodness-of-fit test, which has a chi-square distribution with 5 degrees of freedom. Chi-square test statistics is calculated as: $\chi^2_{(5)} = \sum_{i=1}^{6}((Y_i - n \cdot P_i)^2)/n \cdot P_i = 0.1427$, where Y_i is the frequency of each category from Table 4, n is 1767, the sample size after deleting missing explanatory observations, and P_i is the entire sample proportion of each category from Table F.1. The critical value of the 1% one-sided χ^2 distribution with 5 degrees of freedom is 15.09. Therefore, we do not reject the null hypothesis.

model (α's vary).[3] The following are the three log-likelihood functions that correspond to each of these three models.

For Model 1, we deleted the 203 missing response observations and estimated the parameters using only the 1564 observed response observations using equation (7).

$$\ln L_1(\beta) = \sum_{t=1}^{T} \sum_{j=1}^{5} y_{tj} \cdot \ln \left(\frac{\exp(x'_{tj}\beta_j)}{\sum_{i=1}^{5} \exp(x'_{ti}\beta_i)} \right) \tag{7}$$

It is clear that Model 1 estimates conditional (on observed response) probabilities assuming the missing observations are independent of categories. This will be an inconsistent parameter estimator if the missing observations cause sample selection bias. In fact we need to estimate the joint probabilities.

In Model 2, we included the missing response observations, but we assume that they are missing at random, and there is no sample selection bias problem. Therefore, we estimated equation (8) using the constant α.

$$\ln L_2(\beta, \alpha) = \sum_{t=1}^{T} \left\{ \sum_{j=1}^{5} y_{tj} \cdot \ln \left(\frac{e^{x'_{tj}\beta_j}}{(1+\alpha) \sum_{i=1}^{5} e^{x'_{ti}\beta_i}} \right) + y_{tm} \cdot \ln \left(\sum_{j=1}^{5} \alpha \cdot \frac{e^{x'_{tj}\beta_j}}{(1+\alpha) \sum_{i=1}^{5} e^{x'_{ti}\beta_i}} \right) \right\} \tag{8}$$

This likelihood function estimates the joint probability of each category, but it still assumes that the observations are missing randomly, independent of categories. If missing observations are independent of categories, then there is no sample selection bias problem whether we include the missing observations or completely disregard those observations for analysis. In this case, we expect the parameter estimates from Model 2 and Model 1 to be very similar.

Model 3 is the most comprehensive one. This model includes all observations. We allowed for the possibility that the missing response observations are not randomly missing. This model estimates equation (9).

[3]Since the α_js are used to calculate the probabilities as we can see in Table 3, they must be positive. However, the maximum likelihood estimation of α_j is unrestricted and some data produce negative values of α_j. To avoid this problem, we actually estimated d_j where $\alpha_j = d_j^2$ to ensure that α_j will be positive.

$$\ln L_3(\beta, \alpha) = \sum_{t=1}^{T} \left\{ \sum_{j=1}^{5} y_{tj} \cdot \ln \left(\frac{e^{x'_{tj}\beta_j}}{(1+\alpha_j) \sum_{i=1}^{5} e^{x'_{tj}\beta_j}} \right) + y_{tm} \cdot \ln \left(\sum_{j=1}^{5} \alpha_j \cdot \frac{e^{x'_{tj}\beta_j}}{(1+\alpha_j) \sum_{i=1}^{5} e^{x'_{ti}\beta_i}} \right) \right\} \quad (9)$$

This likelihood function is equivalent to equation (6), and this will correct the sample selection bias problem, if any.

Table 6 reports parameter estimates of each model along with standard errors and related statistics.

There are four sets of parameter estimates. The first set is estimates for the Quit category; the second, the Dismiss category; the third, the Other category and the last one, the Unknown category. Parameters of the first category, Work, are all set to zero as normalization.

As we can see from Table 6, the estimates from Model 1 and Model 2 are almost identical. These results are intuitively expected because Model 1 deletes missing responses (implicitly assumes they are randomly missing) while Model 2 explicitly assumes that the missing responses are randomly missing. The estimate of α from Model 2 gives constant weight ($0.1298 = 0.3603^2$) to each category for the randomly missing response variables.[4,5] Model 3 gives all different weights for each category. While the parameter estimates from all three models are qualitatively comparable, the job skills parameters (Occu1, Occu2) in Model 3 are somewhat different from the parameters of the other two models.[6] This may be due to the under-representation of the Quit category in the missing at random models as discussed later.

We tested the null hypothesis in equation (5) by log-likelihood ratio test and the null hypothesis of missing at random is rejected at the 1 percent significance level. The test statistic is, $\chi^2_{(4)} = 2 \cdot (-2369.05 - (-2382.06)) = 26.02$, where the 1 percent one-sided critical value is 13.28. We also recalculated the marginal probability distribution of each category based on

[4]This estimate ($\hat{\alpha}$) is equivalent to the following estimation, $\tilde{\alpha}$. Let five categories be 1, 2, ..., 5. We have the following relationship: $P_1 + P_2 + P_3 + P_4 + P_5 + \alpha \cdot (P_1 + P_2 + P_3 + P_4 + P_5) = 1$. Then, $(1+\alpha) \cdot (1 - P_m) = 1$ and $\tilde{\alpha} = 1/(1 - \hat{P}_m) - 1$, where $\hat{P}_m = Prob(missing) = 0.1149$ from Table 7. Therefore, $\tilde{\alpha} = 0.1298 = \hat{\alpha} (\equiv \hat{d}_1^2 = 0.3603^2$ *from Model 2 in Table 6*).

[5]Probability of missing observations in the sample is $\alpha/(1+\alpha)$. When the estimate of α is 0.1298, the probability of missing observations is $0.1298/(1 + 0.1298) = 0.1149$. This number corresponds to the probability of missing from Table 7.

[6]Equations (7) and (8) have the same first order conditions with respect to β. In equation (8), α and β are separable. For more details, see Appendix II.

TABLE 6
Maximum Likelihood Estimates

Variable	Model 1 *Estimates*	*Std. Err.*	Model 2 *Estimates*	*Std. Err.*	Model 3 *Estimates*	*Std. Err.*
Const	0.0341	0.0475	0.0342	0.0608	−0.2282	0.0195
Occu1	0.8094	0.0421	0.8094	0.0652	0.5783	0.0194
Occu2	−0.2188	0.0416	−0.2192	0.0635	−0.0508	0.0201
Hschool	−0.5055	0.0418	−0.5056	0.0639	−0.5107	0.0193
College	−0.8757	0.0419	−0.8756	0.0644	−0.9442	0.0193
Duration	−0.3619	0.0273	−0.3620	0.0324	−0.3130	0.0160
Rwage	−0.0293	0.0188	−0.0293	0.0217	−0.0122	0.0132
Unemp	0.2963	0.0214	0.2963	0.0258	0.3142	0.0136
Yrs18	−0.0200	0.0091	−0.0200	0.0101	−0.0257	0.0068
Const	−0.0118	0.0544	−0.0099	0.0361	0.0503	0.0204
Occu1	0.2755	0.0429	0.2758	0.0677	0.1678	0.0195
Occu2	−0.1591	0.0425	−0.1596	0.0650	−0.1806	0.0195
Hschool	−0.7424	0.0424	−0.7430	0.0659	−0.8189	0.0194
College	−1.6293	0.0426	−1.6298	0.0669	−1.6374	0.0195
Duration	−0.3900	0.0349	−0.3901	0.0456	−0.4051	0.0184
Rwage	−0.0888	0.0255	−0.0888	0.0316	−0.0872	0.0157
Unemp	0.3444	0.0241	0.3443	0.0296	0.3534	0.0149
Yrs18	−0.0335	0.0119	−0.0335	0.0128	−0.0348	0.0094
Const	−1.8981	0.0429	−1.8971	0.0682	−1.9329	0.0195
Occu1	1.4130	0.0424	1.4134	0.0662	1.3169	0.0194
Occu2	0.0789	0.0430	0.0788	0.0692	0.0876	0.0201
Hschool	−0.3562	0.0421	−0.3565	0.0648	−0.4080	0.0194
College	−1.1882	0.0423	−1.1885	0.0659	−1.2070	0.0194
Duration	−0.3182	0.0258	−0.3182	0.0300	−0.3156	0.0164

TABLE 6
Continued

Rwage	−0.1045	0.0206	−0.1046	0.0240	−0.0983	0.0141
	Model 1		Model 2		Model 3	
Variable	*Estimates*	*Std. Err.*	*Estimates*	*Std. Err.*	*Estimates*	*Std. Err.*
Unemp	0.3651	0.0231	0.3651	0.0276	0.3753	0.0148
Yrs18	0.0580	0.0076	0.0580	0.0083	0.0561	0.0062
Const	0.4584	0.0427	0.4589	0.0670	0.5127	0.0194
Occu1	0.6324	0.0417	0.6322	0.0638	0.5430	0.0194
Occu2	0.0322	0.0585	0.0318	0.0582	0.0292	0.0220
Hschool	−0.5098	0.0407	−0.5099	0.0605	−0.5864	0.0192
College	−1.5622	0.0412	−1.5620	0.0624	−1.5991	0.0193
Duration	−0.5569	0.0262	−0.5569	0.0308	−0.5737	0.0165
Rwage	0.0495	0.0159	0.0494	0.0187	0.0570	0.0113
Unemp	0.3190	0.0196	0.3189	0.0237	0.3242	0.0126
Yrs18	0.0161	0.0070	0.0161	0.0077	0.0143	0.0059
d_1			0.3603	0.0132	0.2022	0.0186
d_2					0.8717	0.0180
d_3					0.0001	0.0429
d_4					−0.0000	0.0650
d_5	$\alpha_j = d_j^2$	*(see fn3)*			0.2999	0.0192
Log-Lik.	−1751.93	n = 1564	−2382.06	n = 1767	−2369.05	n = 1767

the α_j estimates. If we ignore the missing response observations for data analysis, we are working with the corresponding conditional (on observed) probability distribution based on the assumption that the missing data are missing at random. To focus on the observed probability distribution, we calculated the predicted observed probability distribution for each category and reported it in Table 7.

From Table 7, Model 1 and Model 2 show a very similar predicted probability distribution for each category, which is very close to the conditional frequency distribution that ignores the missing data. This is not surprising since these two models assume that all missing responses are missing at random and therefore assume no sample selection bias problem. However, we rejected the null hypothesis of missing at random, and Model 3 shows that the Quit and Unknown categories are underrepresented due to the non-response (missing response) observations. The Quit category is severely underestimated while all other categories are slightly overestimated under the missing at random assumption (Model 1 and Model 2). This suggests the possibility that workers who quit their jobs were more reluctant to report the job status than workers in other categories. If we simply ignore missing response observations, we obviously encounter selection bias problem and thus inconsistent structural parameter estimates.

V. SUMMARY AND CONCLUSIONS

We have introduced a new method of correcting the joint probability distribution of multinomial responses (both observed and unobserved) for systematically missing data. We proposed a simpler method to construct the log-likelihood function and applied this to data from the PSID on job loss status, and found missing responses to be systematically missing. If we ignore missing response observations for analysis, the structural parameter estimators are inconsistent. In the job loss status analysis, we have discov-

TABLE 7
Predicted Probability Distribution of Each Category

	Work	*Quit*	*Dismiss*	*Other*	*Unknown*	*Missing*	*Total*[7]
Frequency (from Table 4)							
	0.2083	0.1285	0.0617	0.0973	0.3894	0.1149	1.0001
Conditional on Observed							
	0.2353	0.1452	0.0697	0.1099	0.4399	—	1.0000
Model 1	0.2345	0.1473	0.0709	0.1092	0.4379	—	0.9998
Model 2	0.2346	0.1472	0.0708	0.1092	0.4379	—	0.9997
Model 3	0.2168	0.2261	0.0617	0.0973	0.4244	—	1.0263

[7]Due to the rounding error, total may not be equal to one.

ered that people who quit their job appear to be recorded as missing more often than those in the other four categories. While this analysis does not provide any explanation for this phenomenon, it does appear to be an interesting outcome that might be worth investigating in future research. We hope that this technique for recovering the underlying probability distribution in the presence of missing data will produce equally interesting results in other areas of study.

APPENDIX I: COMPARISON OF TWO METHODS

We will consider a simple categorical response model. As explained already, to make the standard procedure work, we assume that there are only two response categories and error terms are normally distributed.

$$y_t^* = x_t'\beta + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{A.1.1}$$

We observe response category $y_t = 1$ if $x_t'\beta + \varepsilon_t \geqslant 0$ and category 0 if $x_t'\beta + \varepsilon_t < 0$. Each individual will respond to category 1, category 0 or does not respond to the question (missing response).

The standard procedure assumes that the respondent will go through a two stage thought process. First he will decide whether to respond or not, and if he decides to respond, then he will choose either category 1 or category 0. Therefore, we need to introduce the selection equation.

$$s_t^* = x_t'\gamma + \delta_t, \quad \delta_t \sim N(0, \sigma_\delta^2) \tag{A.1.2}$$

Each individual will respond if his selection equation crosses a threshold value. He will respond if $x_t'\gamma + \delta_t \geqslant 0$ and will not respond if $x_t'\gamma + \delta_t < 0$. $(\varepsilon_t, \delta_t)$ has a bivariate normal distribution with covariance $\rho \cdot \sigma_\varepsilon \sigma_\delta$, where ρ is a correlation coefficient. To estimate the model parameters (β, γ), we have the following log-likelihood function.

$$\begin{aligned} \ln L(\beta, \gamma) = \sum_{t=1}^{T} \Bigg\{ s_t \cdot \Bigg(y_t \cdot \ln\Bigg(\int_{-x_t'\beta}^{\infty} \int_{-x_t'\gamma}^{\infty} q(\varepsilon_t, \delta_t) d\delta_t d\varepsilon_t \Bigg) \\ + (1 - y_t) \cdot \ln\Bigg(\int_{-\infty}^{-x_t'\beta} \int_{-x_t'\gamma}^{\infty} q(\varepsilon_t, \delta_t) d\delta_t d\varepsilon_t \Bigg) \Bigg) \\ + (1 - s_t) \cdot \ln\Bigg(1 - \Phi\Bigg(\frac{x_t'\gamma}{\sigma_\delta} \Bigg) \Bigg) \Bigg\} \end{aligned} \tag{A.1.3}$$

where $q(\varepsilon_t, \delta_t)$ is a bivariate normal density function and $\Phi(\cdot)$ is a standard normal cumulative density function.

The proposed method does not require the selection equation to handle missing response observations. The new method assigns the response

probability based on the different category. From Table 3, we can calculate the joint probability of the observed for each category. The joint probabilities of the observed for each category are:

$$P_{t1o} \equiv \Pr(y_t = 1) = \frac{1}{1+\alpha_1}\left(\Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right) \tag{A.1.4}$$

$$P_{t0o} \equiv \Pr(y_t = 0) = \frac{1}{1+\alpha_0}\left(1 - \Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right) \tag{A.1.5}$$

The probability of the missing responses is:

$$\begin{aligned}&\alpha_1 \cdot \Pr(y_t = 1) + \alpha_0 \cdot \Pr(y_t = 0)\\ &= \alpha_1 \cdot \frac{1}{1+\alpha_1}\left(\Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right) + \alpha_0 \cdot \frac{1}{1+\alpha_0}\left(1 - \Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right)\end{aligned} \tag{A.1.6}$$

Therefore, the log-likelihood function is:

$$\begin{aligned}\ln L(\beta, \alpha) = \sum_{t=1}^{T}\Bigg\{ & y_t \cdot \ln\left(\frac{1}{1+\alpha_1}\left(\Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right)\right)\\ & + (1 - y_t) \cdot \ln\left(\frac{1}{1+\alpha_0}\left(1 - \Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right)\right)\\ & + y_{tm} \cdot \ln\left(\frac{\alpha_1}{1+\alpha_1}\left(\Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right) + \frac{\alpha_0}{1+\alpha_0}\left(1 - \Phi\left(\frac{x_t'\beta}{\sigma_\varepsilon}\right)\right)\right)\Bigg\}\end{aligned} \tag{A.1.7}$$

By comparing these two log-likelihood functions for this simple model with only two categories, it is clear that the proposed method provides a much simpler log-likelihood function. Even for this simple model, the standard method requires evaluating a bivariate normal density function while the proposed method only requires calculating a univariate standard normal density function.

APPENDIX II: LOG-LIKELIHOOD FUNCTIONS

For Model 1, the log-likelihood function is:

$$\ln L_1(\beta) = \sum_{t=1}^{T}\sum_{j=1}^{5} y_{tj} \cdot \left(x_{tj}'\beta - \ln\left(\sum_{i=1}^{5} e^{x_{ti}'\beta_i}\right)\right) \tag{A.2.1}$$

The first order condition is:

$$\frac{\partial \ln L_1(\beta)}{\partial \beta_j} = \sum_{t=1}^{T} y_{tj} \cdot \left(x_{tj} - \frac{x_{tj} \cdot e^{x'_{ti}\beta_j}}{\sum_{i=1}^{5} e^{x'_{tj}\beta}} \right) = \sum_{y=1}^{T} y_{tj} \cdot \left(1 - \frac{e^{x'_{tj}\beta_j}}{\sum_{i=1}^{5} e^{x'_{ti}\beta}} \right) x_{tj} \tag{A.2.2}$$

For Model 2, the log-likelihood function is:

$$\begin{aligned} \ln L_2(\beta, \alpha) = \sum_{t=1}^{T} \Bigg\{ & \sum_{j=1}^{5} y_{tj} \cdot \left(x'_{tj}\beta_j - \ln(1+\alpha) - \ln\left(\sum_{i=1}^{5} e^{x'_{ij}\beta_i} \right) \right) \\ & + y_{tm} \cdot \ln \left(\alpha \cdot \frac{\sum_{i=1}^{5} e^{x'_{ij}\beta_i}}{(1+\alpha) \sum_{i=1}^{5} e^{x'_{ij}\beta_i}} \right) \Bigg\} \\ = \sum_{t=1}^{T} \Bigg\{ & \sum_{j=1}^{5} y_{tj} \cdot \left(x'_{tj}\beta_j - \ln(1+\alpha) - \ln\left(\sum_{i=1}^{5} e^{x'_{ti}\beta_i} \right) \right) \\ & + y_{tm} \cdot (\ln(\alpha) - \ln(1+\alpha)) \Bigg\} \end{aligned} \tag{A.2.3}$$

Two sets of parameters (β, α) are separable, and the first order conditions are:

$$\frac{\partial \ln L_2(\beta, \alpha)}{\partial \beta_j} = \sum_{t=1}^{T} y_{tj} \cdot \left(x_{tj} - \frac{x_{tj} \cdot e^{x'_{tj}\beta_j}}{\sum_{i=1}^{5} e^{x'_{ti}\beta_i}} \right) = \sum_{t=1}^{T} y_{tj} \cdot \left(1 - \frac{e^{x'_{tj}\beta_j}}{\sum_{i=1}^{5} e^{x'_{ti}\beta_i}} \right) x_{tj} \tag{A.2.4}$$

$$\frac{\partial \ln L_2(\beta, \alpha)}{\partial \alpha} = \sum_{t=1}^{T} \left\{ \sum_{j=1}^{5} y_{tj} \cdot \left(\frac{-1}{1+\alpha} \right) + y_{tm} \cdot \left(\frac{1}{\alpha} - \frac{1}{1+\alpha} \right) \right\} \tag{A.2.5}$$

As we can see from here, the first order conditions for β_j from Model 1 and Model 2 are identical and estimates of the β_j are almost identical as we can see from Table 6.

For Model 3, we have a slightly more complicated log-likelihood function:

$$\ln L_3(\beta, \alpha) = \sum_{t=1}^{T} \left\{ \sum_{j=1}^{5} y_{tj} \cdot \left(x'_{tj}\beta_j - \ln(1+\alpha_j) - \ln\left(\sum_{i=1}^{5} e^{x'_{ti}\beta_i} \right) \right) \right.$$

$$\left. + \; y_{tm} \cdot \ln\left(\sum_{j=1}^{5} \alpha_j \frac{e^{x'_{tj}\beta_j}}{(1+\alpha_j) \sum_{i=1}^{5} e^{x'_{ij}\beta_i}} \right) \right\}$$

$$= \sum_{t=1}^{T} \left\{ \sum_{j=1}^{5} y_{tj} \cdot \left(x'_{tj}\beta_j - \ln(1+\alpha_j) - \ln\left(\sum_{i=1}^{5} e^{x'_{ti}\beta_i} \right) \right) \right.$$

$$\left. + \; y_{tm} \cdot \left(\ln\left(\sum_{j=1}^{5} \frac{\alpha_j \cdot e^{x'_{tj}\beta_j}}{1+\alpha_j} \right) - \ln\left(\sum_{i=1}^{5} e^{x'_{tj}\beta_i} \right) \right) \right\} \quad \text{(A.2.6)}$$

The first order conditions are:

$$\frac{\partial \ln L_3(\beta, \alpha)}{\partial \beta_j} = \sum_{t=1}^{T} y_{tj} \cdot \left(x_{tj} - \frac{x_{tj} \cdot e^{x'_{ji}\beta_j}}{\sum_{i=1}^{5} e^{x'_{ti}\beta_i}} \right)$$

$$+ \; y_{tm} \cdot \left(\frac{x_{tj} \cdot \frac{\alpha_j \cdot e^{x'_{tj}\beta_j}}{1+\alpha_j}}{\sum_{i=1}^{5} \frac{\alpha_i \cdot e^{x'_{ti}\beta_i}}{1+\alpha_i}} - \frac{x_{tj} \cdot e^{x'_{tj}\beta_j}}{\sum_{i=1}^{5} e^{x'_{ti}\beta_i}} \right) \quad \text{(A.2.7)}$$

$$\frac{\partial \ln L_3(\beta, \alpha)}{\partial \alpha_j} = \sum_{t=1}^{T} y_{tj} \cdot \left(\frac{-1}{1+\alpha_j} \right) + y_{tm} \cdot \left(\frac{\frac{1}{(1+\alpha_j)^2} e^{x'_{tj}\beta_j}}{\sum_{i=1}^{5} \frac{\alpha_i}{1+\alpha_i} e^{x'_{ti}\beta_i}} \right) \quad \text{(A.2.8)}$$

We can see that the first part of equation (A.2.7) is the same as equations (A.2.2) and (A.2.4). At least for the logit probability structure this provides us with a direct comparison between the missing at random models and our proposed model.

University of Notre Dame, Indiana

Date of Receipt of Final Manuscript: January 1999

REFERENCES

Bhat, C. R., (1994), "Imputing a Continuous Income Variable from Grouped and Missing Income Observations," *Economics Letters* Vol. 46, pp 311–19.

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

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

Guttman, I. and Menzefricke, U. (1983), "Bayesian Inference in Multivariate Regression with Missing Observations on the Response Variables," *Journal of Business and Economics Statistics*, Vol. 1, pp 239–48.

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

Kmenta, J. and Balestra, P. (1986), "Missing Measurements in a Regression Problem with No Auxiliary Relations," in *Advances in Econometrics*, JAI Press.

Lien, D. and Rearden, D. (1990), "Missing Measurements in Discrete Response Models," *Economics Letters*, Vol. 32, pp 231–35.

Manski, C. F., (1994), "The Selection Problem," in *Advances in Econometrics*, Cambridge University Press, pp 143–70.

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

Rubin, D. B., (1976), "Inference and Missing Data," *Biometrika*, Vol. 63, pp 581–92.