Using Multiple Imputation in a Customer Satisfaction Survey

Stephen Ash, Bureau of the Census, Washington D.C. 20233

Keywords: ECM algorithm, Data Augmentation Bayesian Iterative Proportional Fitting.

1. INTRODUCTION

This paper will describe how we applied missing data methods for univariate statistics and loglinear models. We used the results of the loglinear models to describe the associations between the questions of a customer satisfaction survey. The "customers" of our survey were the persons who called to the Census 2000 Telelphone Questionnaire Assistance (TQA) operation.

2. BACKGROUND

This section describes the Census 2000 Inbound Telephone Questionnaire Assistance (TQA) operation, and then it describes the customer satisfaction data associated with this operation.

2.1. Census 2000 Telephone Questionnaire Assistance

Census 2000 implemented an extensive Inbound TQA operation to support calls in English, Spanish, and other languages. The goals of the Census 2000 Inbound TQA operation included:

< providing the public with convenient access to general Census 2000 information
< providing help in completing census forms
< fielding requests for replacement forms, and
< collecting short form data from callers

This operation included a toll-free number and an Interactive Voice Response (IVR) system that handled a large number of calls concurrently. The components of the IVR system included an automated touch-tone menu and a voice recognition option for callers using a rotary telephone. Most callers were initially routed to the IVR system. The IVR system then provided callers with menu options which applied to their reason for calling, e.g., requests for questionnaires. At different points within the IVR system, a caller could also be transferred to an agent by request.

The system was designed to handle 11 million calls, but received slightly over 6 million during the period of March 3, 2000 to June 30, 2000.

2.2. Customer Satisfaction Survey of the TQA

To determine if callers were satisfied with the service provided by TQA, we conducted a customer satisfaction survey using an automated touch-tone instrument. When a caller, sampled for the customer satisfaction survey finished receiving assistance, they were routed to the customer satisfaction survey.

For the customer satisfaction survey, we partitioned persons who called for assistance into two types based on their experience: (1) "IVR-only" or callers who used the IVR system and did not speak with an agent and (2) "IVR and Agent" or callers who used the IVR system and also spoke with an agent.

The questions of the survey referred to aspects of the call which could directly affect a caller's experience and thereby their satisfaction. The responses to the survey are all on a 7-point Likert-type scale, where one is the least favorable response and 7 is the most favorable response for a given question.

At the beginning of the call, an automated system selected the 1-in-160 sample from all those persons who called for assistance. In the middle of the data collection period, we learned that we were not getting the number of sample cases we expected, so we increased the sampling fraction to 1-in-80. The data collection period started on March 3, 2000 and ended June 30, 2000. We changed the sampling fraction on March 23, 2000. The analysis of this paper uses the weighted counts. Table 1 provides a summary of the resultant sample.

Summary of Sample Results (Table 1)

	Fully observed		Total sample size		% w/ missing data
	Unwgt	Wgt	Unwgt	Wgt	
IVR-only	2,448	2,781	3,045	3,441	19.2
IVR and Agent	607	773	1,248	1,833	42.2

At this point we noticed that any multivariate analysis would exclude a substantial portion of our sample when analyzing either type of experience. We would exclude all sample cases without a complete response for all of the questions of the survey because for most kinds of multivariate analysis, e.g., loglinear models, we can only use the observations that have responded to every question.

For the IVR and Agent sample the resultant sample size was too small and the percent of item missing data was too large to apply methods for missing data. However for the IVR-only sample, this was not the case and we decided to complete our analysis using both 1) the usual analysis methods that ignore the missing data; and 2) multivariate tools for missing data. The remainder of the paper discusses only the analysis of the IVR-only sample. The questions of the IVR-only customer satisfaction survey include:

1:[Menu] *An automated menu system answered your call today and gave you a list of options. Once you made your first menu selection, rate how well the information that followed fit your expectation for that selection, with 7 being exactly what you expected and 1 being not at all what you expected.*

2:[Navigate] *Rate how easy it was to move through the automated menu system with 7 being very easy and 1 being not at all easy.*

3:[Issue] *Thinking of the main reason you called today, rate the effectiveness of the automated system in handling that particular issue with 7 being very effective and 1 being not at all effective.*

4:[Participate] *Rate how much the information you received today will help you participate in Census 2000, with 7 being very helpful and 1 being not at all helpful.*

5:[Satisfaction] *Rate your overall satisfaction with your call today to the Census 2000 Assistance Center with 7 being very satisfied and 1 being not at all satisfied.*

Although we asked callers to respond on a 7-point Likert-type scale for each of the five questions, we realized that the resultant saturated multinomial table had 7^5 = 16,807 cells. We certainly did not have enough sample to estimate these 16,807 cells, so we collapsed the 7-point scale to a 2-point scale which resulted in a more manageable 2^5 = 32 cells. For the collapsing, we mapped the responses 1, 2, 3, and 4 to 1 and responses 5, 6 and 7 to 2.

3. METHODS

This section describes the missing-data models and measures that we used to examine how satisfied callers were with the TQA Inbound operation. We can group the models in the following categories:

Methods for the saturated model
< EM Algorithm
< Data Augmentation (DA)

Methods for loglinear models
< ECM Algorithm
< Data Augmentation Bayesian Iterative Proportional Fitting (DABIPF)

For our application of these missing data methods, we require the following assumptions:

(1) We assume that the observations are independent and identically distributed (i.i.d.) or approximately i.i.d. As previously noted, this was not the case. We sampled the observations under two different sampling fractions, where one sampling fraction was twice as large as the other. Since we selected a systematic sample with both sampling fractions, we assume that the weighted counts came from a large population and thereby represent population approximately i.i.d.

(2) We assume that the missingness is Missing At Random (MAR) in the sense of Rubin (1987). This assumption says that we believe "the missing values behave like a random sample of all values within subclasses defined by observed data" (Schafer 1997).

3.1. What do the data look like?

Before we describe the models and measures further, we'll describe some notation borrowed from Schafer (1997). We assume that the data has n observations and p variables (for us, the five questions of the survey) and we represent this by x, a $n \times p$ data matrix. Since all of our variables of interest are categorical, we'll also note that for our application each x_j takes possible values 1 or 2. We'll assume that the sample size n is fixed, therefore x has a multinomial distribution with parameter 2.

In this paper, the *fully observed* data refers to the set of observations for which we have a response for every variable. The *missing data* will refer to the set of all observations – the observations with some item nonresponse and those with a response for every variable.

3.2. How did we estimate the parameters of the saturated model?

The following section outlines the missing-data tools we used to estimate the parameters of the saturated model.

EM Algorithm

We used the EM algorithm (Dempster, Laird and Rubin 1977) to find solutions to the observed-data likelihood of the saturated model. We need data augmentation because the EM algorithm does not provide an estimate of the variance of 2. So why did we bother with the EM algorithm? It is easy to implement, contributes to the understanding of DA and we can use it to check the DA estimates of 2. Therefore we can describe customer satisfaction by estimating the parameter 2.

Data Augmentation

Since the EM algorithm does not provide variance estimates for our parameters of interest, we created m=25 different sets of imputed data. We used these m different datasets to estimate different functions of the parameter *2* and their associated variances as suggested by Rubin (1987).

To create the multiple imputations, Tanner and Wong (1987) suggest using data augmentation (DA). We calculated estimates of the five marginal proportions and their associated variance from our *m* multiple imputations we derived from DA for the saturated model.

We also estimated the measure of association gamma for all possible pairs of the five questions of our customer satisfaction survey. Gamma has a range of -1 to 1, where a value of zero indicates low association between the questions of interest and a value of 1 or -1 indicates high association. We calculated gamma and it's associated asymptotic variance as described by Goodman and Kruskal (1963).

3.3. How did we use loglinear models?

This section describes how we estimated cell counts using loglinear models which we can use to describe the associations of the variables.

ECM Algorithm

We used the ECM algorithm (Meng & Rubin 1991) instead of the EM algorithm because the M-step is not simple for loglinear models. A full M-step of the EM algorithm would require several iterations of Iterative Proportional Fitting (IPF). That means EM would altogether require many iterations within iterations.

The ECM algorithm circumvents this problem by replacing the M-step of the EM algorithm with a CM-step or Conditional Maximization. In general, this CM-step is a series of conditional maximization steps that are not equivalent to the M-step, but allow the ECM algorithm to converge using these much simpler steps. For categorical data, ECM is simple because it uses one iteration of IPF for each of the sufficient configurations.

Data Augumentation Bayesian IPF

Data Augmentation Bayesian Iterative Proportional Fitting or DABIPF (Schafer 1997) is the application to DA for loglinear models. It uses part of the Bayesian IPF algorithm as suggested by Gelman et. al. (1995) to supplement the DA for the saturated model.

We iterated the following steps of DABIPF to create each of our *m*=25 multiple imputations:

For the I-Step:

We "imputed" *x* as we did for DA of the saturated model -- we randomly allocated the missing data totals to the fully observed totals according to a multinomial distribution with parameter *2* from the previous P-step.

For the P-step:

To generate random draws from our posterior *2 / x*, we completed the following steps of the general IPF algorithm, except we replace the known marginals with random draws from a gamma distribution that has parameters associated with the expected cell counts and the parameter of the conjugate Dirlecht.

As summarized by Schafer (1997) , we calculated across each of the sufficient configurations *j*, each cell proportion as

$$q_{ij}^{(t+1)} = \frac{g_{j+}}{g_{++}}\left(\frac{q_{ij}^{(t)}}{q_{j+}^{(t)}}\right)$$

where we randomly generated g_{j+} from a gamma distribution, i.e., $g_{j+} \sim gamma\left(a'_{j+}\right)$ with the parameter $a'_{j+} = \sum_i \left(a_{ij} + x_{ij}\right)$ and $g_{++} = \sum_i g_{j+}$.

3.4. How did we examine the fit of the models?

We examined the fit of the models derived by EM, DA and ECM with the deviance for missing data as described by Fuchs (1982)

$$G_{miss}^2 = 2\sum_{s=1}^{S}\sum_{y\in Y} x_y^{(s)} \log\frac{x_y^{(s)}}{n_s b_y^{(s)}} - G_0^2$$

where $b_y^{(s)}$ is defined as in Schafer (1997) as

$$b_y^{(s)} = \sum_{M_s(y)\in M_s} q_y$$

For DABIPF, we examined the fit by calculating the following mean deviance for the *m* multiple imputations as

$$\overline{G}_{miss}^2 = 2\sum_{s=1}^{S}\sum_{y\in Y} x_y^{(s)} \log\frac{x_y^{(s)}}{n_s \overline{b}_y^{(s)}} - G_0^2$$

where we calculate $\overline{b}_y^{(s)}$ as the sum of the mean cell proportions from the *m* multiple imputations

$$\overline{b}_y^{(s)} = \sum_{M_s(y)\in M_s} \overline{q}_y = \sum_{M_s(y)\in M_s} (1/m)\sum_m q_{y,i}$$

What is interesting to note about this deviance, is that we cannot estimate the deviance as the mean of the deviances from the *m* multiple imputations. Each deviance calculated for an individual multiple imputation will be greater than the ECM estimate of the deviance -- the solution which minimizes the deviance. Therefore the mean of the deviances will always be greater than the ECM estimate. However, the mean of the cell proportions over the multiple imputations does converge to the set of cell proportions which minimize the deviance.

4. LIMITS

There were several operational problems with the data collection of the customer satisfaction survey. Taken together the reader should be wary of the representativeness of the resultant sample.

(1) Three of the 22 call center sites were not included in the sample universe for technical reasons.

(2) We had to stop interviewing for the customer satisfaction survey during the peak calling period of March 13 and 14 to complete some re-programming.

(3) Some unidentifiable portion of both types of callers received the wrong set of questions. Some IVR-only callers received the IVR and Agent questions and some IVR and Agent callers received the IVR-only questions.

(4) We thought that the system would automatically transfer sampled calls from the agent to the customer satisfaction survey. This was not the case. The agent had to manually transfer the sampled callers to the customer satisfaction survey at the end of each call. Because this was not the original understanding, we had to implement new procedures in the beginning of production.

(5) Other programming errors resulted in what we believe to be numerous lost calls.

5. RESULTS

5.1. Which model best describes the data?

From Table 2 we determined that Model 2 or (124, 13, 15, 23, 25, 34, 35, 45) is the model which has the best fit with the data for all three estimation methods. Adding the three-factor effect 145 to Model 2 does not significantly reduce in the deviance, e.g., for the fully observed data $G^2_{\text{model 2}} - G^2_{\text{model 1}} = 2.5804 < 2.706 = c^2_{a=0.10,df=1}$. Also, note that all of the models in this set of candidate models contain all possible two-way associations -- all five variables are significantly associated to each other.

So how do our missing data estimates compare? The estimates from ECM and DABIPF agree with each other from model to model as expected. The missing data estimates are also very similar to the fully observed estimates, suggesting that the observations from the fully observed and missing data are not very different.

What does it mean that our final model has the three-way effects 124, or Menu|Navigate|Participate? As Whittaker (1990) explains, the three-way effects represents the interaction of the two-way interactions, i.e., the three-way effect is significant if any of the two-way effects within the three-way effect are not constant within levels of the two-way effects. Often three-way or higher order effects are significant when there is a strong two-way effect present. For our 124 effect, we will see that it does not include the strongest two-way effect, i.e., Satisfied|Participate, but it does include the second strongest two-way effect Menu|Participate.

Deviances of models (Table 2)

	Model[1]	df	Fully Observed $\left(G^2\right)$	ECM $\left(G^2_{miss}\right)$	DABIPF $\left(\overline{G}^2_{miss}\right)$
1	(124, 13, 145, 23, 25, 34, 35, 45)	14	16.59	16.64	16.78
< 2	(124, 13, 15, 23, 25, 34, 35, 45)	15	19.18	18.02	18.28
3	(12, 13, 145, 23, 24, 25, 34, 35)	15	27.39	30.55	30.32
4	(12, 13, 14, 15, 23, 245, 34, 35)	15	27.95	29.61	29.74
5	(12, 13, 14, 15, 23, 24, 25, 34, 35, 45)	16	32.33	34.23	33.96

We also add that intuitively the parts of the three-way effect Menu|Navigate|Participate are certainly related. The callers ability to navigate and use the menu system contributed to their self-reported participation.

For the calculation of the ECM estimates we used 30 iterations. For the DABIPF estimates, we used 100 burn-in iterations for each of our 25 multiple imputations. Also, we calculated G^2_{miss} for ECM and $\overline{G}^2_{miss}$ for DABIPF using the value G^2_0= 363.29 for the deviance for the saturated model. We found G^2_0 using the EM algorithm.

5.2. Comparison of univariate estimates

Next, we calculated the simple marginal proportion for each question of the survey. For each question this value represents the proportion of respondents that answered most favorable, e.g., it was easy or very easy to move through the automated menu system. Table 3 summarizes these estimates from which we note five comments:

[1] We specify models by the associations in the given model, for example, 13 is the two-way effect Menu|Issue and 124 is the three-way effect, Menu|Navigate|Participate.

(1) Persons using the Census 2000 IVR system answered most favorably to each of the five questions we asked.

(2) The missing data estimates of the marginal proportions are approximately 3-4 percent lower in magnitude than the fully observed for all estimates. This means that if our model is appropriate, then the fully observed estimates are biased upward.

(3) As expected, the ECM and DABIPF estimates agree in general from model to model. The different models do define different estimated proportions, but the differences are smaller than the standard errors for most of the estimates.

Estimates of the marginal proportions as percentages (Table 3)

Model	Satisfaction	Menu	Navigate	Issue	Participate
			Fully Observed		
	78.41	80.09	86.62	78.51	79.53
	(0.78)	(0.76)	(0.64)	(0.78)	(0.77)
			EM Algorithm		
	74.58	76.08	83.96	75.47	75.94
			Data Augmentation		
	76.16	79.29	85.72	77.32	78.40
	(1.17)	(0.94)	(0.87)	(1.01)	(1.10)
			ECM Algorithm		
1	74.71	76.20	84.13	75.60	76.07
2	74.71	76.20	84.13	75.60	76.16
3	74.70	76.21	84.14	75.61	76.16
4	74.71	76.21	84.14	75.61	76.19
5	74.72	76.21	84.14	75.61	76.30
			DA Bayesian IPF		
1	74.45	75.83	85.09	74.84	76.58
	(1.00)	(1.05)	(0.94)	(1.03)	(1.00)
2	75.41	76.33	83.88	75.95	76.86
	(0.98)	(0.93)	(0.86)	(1.03)	(1.00)
3	75.01	77.04	84.34	75.88	76.38
	(1.04)	(1.04)	(0.85)	(1.15)	(1.00)
4	73.90	77.04	84.80	75.35	75.92
	(1.17)	(1.21)	(0.95)	(1.23)	(1.12)
5	73.27	75.42	83.93	75.25	75.82
	(1.15)	(1.05)	(0.95)	(1.06)	(1.03)

(4) Also the ECM and DABIPF estimates derived with their reduced set of parameters are similar to the EM and DA estimates for the saturated model.

(5) Overall, the estimates of the marginal proportions variances for DA and DABIPF are roughly 1.5 to 3 times greater than the fully observed estimates. The values in parentheses are the estimates of the standard errors of the marginal proportions.

5.3. Comparison of marginal associations

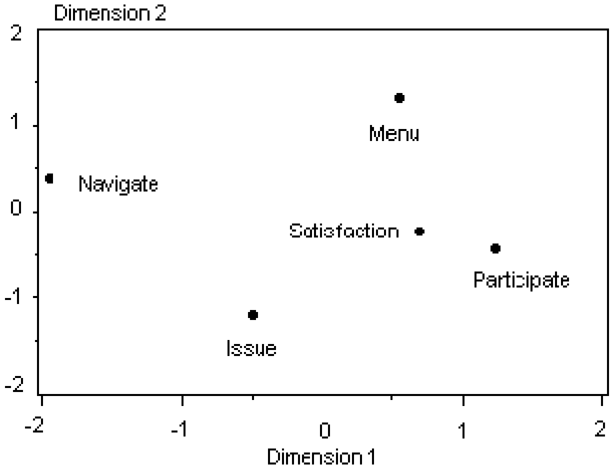
Because all of the two-way associations are significant we decided that it would be interesting to further examine the associations relative to each other. Table 4 presents the estimated values of gamma we calculated using DA.

Estimated values of Gamma (Table 4)

	Satisfaction	Menu	Navigate	Issue	Participate
Satisfaction	--				
Menu	0.931 (0.016)	--			
Navigate	0.900 (0.026)	0.913 (0.023)	--		
Issue	0.930 (0.017)	0.923 (0.017)	0.925 (0.021)	--	
Participate	0.978 (0.006)	0.951 (0.013)	0.915 (0.023)	0.945 (0.014)	--

The values in parentheses are the estimates of the standard errors of gamma calculated using DA.

Graph of Gamma using MDS (Figure 1)



An easy way to examine the values of gamma relative to each other is to plot the values using Multidimensional Scaling (MDS). In Figure 1, the associations are represented by the relative distances between points on the graph. The highly associated variables are the points closest to each other on the graph and the least associated variables are the points furthest from each other on the graph. See also Borg and Groenen (1997) for a good reference on MDS.

Figure 1 additionally shows how Participate and Satisfaction (or Participate|Satisfaction) have the strongest relative association.

6. CONCLUSIONS

Because respondents answered favorably to all of our questions about their satisfaction, we conclude that there was a high level of satisfaction with the assistance that the TQA IVR-only operation provided.

There is strong evidence that all of the questions of the survey are strongly associated. This is demonstrated by the model we selected and the high estimated values for gamma, our measure of association. The model we selected, using both the fully observed data and ECM model for missing data, included all two-way interactions. Also the smallest value for the measure of association gamma was 0.900, which indicates a strong association.

All of the questions of our customer satisfaction survey may be associated because callers responded with respect to their overall experience, and not entirely considering each question separately.

There is not much evidence for us to conclude that the missing data was much different than the fully observed data. The estimates derived from the fully observed data and the missing data did not differ greatly. We saw that the estimated marginal proportions differed in magnitude by 3-4 percent. We also selected the same model using both methods – meaning that the same associations are represented. From the small amount of disagreement for all estimates we conclude that the sample with item nonresponse was not much different from those of the fully observed sample.

The missing data methods did however provide insight to the estimates of variance. We saw how much more variable the missing data estimates were as compared to the fully observed estimates. The increase in variation may not be important for our example where the fully observed and missing data estimates did not greatly differ. It does become important in those applications when the fully observed and missing data estimates do differ. In those instances we would have to choose between a biased estimate with a small variance and an unbiased estimate with a large variance.

We also note that all of our conclusions about the population's customer satisfaction are tempered by the limitations of the analysis.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

References

Borg, Ingwer and Groenen, Patrick (1997) *Modern Multidimensional Scaling: Theory and Applications*, Springer-Verlag, New York, NY.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) "Maximum likelihood estimation from incomplete data via the EM Algorithm" (with discussion) Journal of the Royal Statistical Society Series B, *39*, 1-38.

Fuchs, Camil (1982) "Maximum likelihood estimation and model selection in contingency tables with missing data", Journal of the American Statistical Association, *77*, 270-278.

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. (1995) *Bayesian Data Analysis*, Chapman & Hall.

Goodman, Leo A., and Kruskal, William H., (1963) "Measures of Association for Cross Classifications III: Approximate Sampling Theory", Journal of the American Statistical Association, *58*, 310-364.

Meng, Xio-Li and Rubin, Donald B. (1991) "IPF for Contingency Tables with Missing Data via the ECM Algorithm", American Statistical Association Proceedings of the Statistical Commuting Section, 244-247.

Rubin, Donald B. (1987) *Multiple Imputation*, John Wiley & Sons, Inc.

Schafer, Joseph L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC.

Tanner, Martin A., and Wong, W. H. (1987) "The calculation of posterior distributions by data augmentation" (with discussion), Journal of the American Statistical Association, *82*, 528-550.

U.S. Bureau of Census (2001) "Census 2000: Telephone Questionnaire Assistance (TQA), Program Master Plan", from Teresa Angueira to Distribution List, August 14, 2001.

Whittaker, Joe (1990) *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons.