

Partial Fully Efficient Fractional Imputation for Incomplete Contingency Tables with Covariates

Shin-Soo Kang¹, Kenneth J. Koehler², and Michael D. Larsen²

Department of Information and Statistics, KwanDong University, Kangwon-Do, 210-701, South Korea¹

Department of Statistics, Iowa State University, Ames, Iowa 50011-1210, U.S.A.², larsen@iastate.edu

Abstract

This article concerns the analysis of contingency tables when values of some categorical variables are missing. It is assumed that covariates (continuous or otherwise) that are predictive of the categorical variables are available and observed. The approach is one of imputation or filling-in the missing values. It is imagined that one wants to release completed tables to users so that they can conduct their chosen analyses without the complication of missing data. At the same time it is desirable that users of the released data be able to accurately estimate uncertainty associated with their estimates. One existing option for accomplishing the two goals is a multiple imputation approach. This article presents a new fractional imputation method to produce completed tables and estimate uncertainty. Based on the observed categorical variables and covariates one estimates logistic regressions to predict cell membership for the partially classified observations. A completed table with fractional counts based on predicted cell membership probabilities is created and analyzed to produce estimates. A jackknife method is used to estimate standard errors of estimates. One disadvantage is that the method must be run for each new jackknife sample in order to produce a variance estimate. The primary advantage is that it appears to improve efficiency of estimates. In limited simulations, when covariates contain information about the categorical variables, the new method provides more efficient estimates of a log-odds ratio than either multiple imputation based on data augmentation or complete case analysis.

Keywords: Complete Case Analysis; Fractional Imputation; Missing at random; Multiple Imputation; Nonresponse.

1. Introduction

In the analysis of contingency tables, it may happen that one or more of the categorical variables is not observed for some respondents, but there is covariate information that can be used to predict the missing information. The predictions, conditional on the covariates and the observed categorical variables, then can be used to fill-in or impute categorical values for the missing responses. This article proposes and studies a procedure that is an analog to fully efficient fractional imputation (FEFI; Kim and Fuller 2004) for the cell mean model and for two-way tables without covariates (Kang, Koehler, and Larsen 2006, 2007a, 2007b).

One simple approach when some categorical variables are missing is to discard cases with missing values and conduct complete-case (CC) analysis, in which case standard computer programs for the analysis of complete tables with covariates should be readily applicable. If a statistical model, such as a multinomial model or a Poisson model is assumed for the counts in the table, then one can perform maximum likelihood estimation (MLE) given the observed information. Computation of estimates likely would proceed with an iterative algorithm, such as the EM algorithms (Dempster, Laird, and Rubin 1977) and its variants (McLachlan and Krishnan 1996). If a prior distribution is proposed for unknown model parameters, then a Bayesian analysis using all observed values can be conducted; see, for example, Schafer (1997; chapters 7-9). As an approximation to the Bayesian approach, one can use multiple imputation (Rubin 1978, 1987; see also Schafer 1997).

Section 2 defines notation and the fractional imputation method. We restrict attention to the case in which missingness is confined to at most two categorical variables. Section 3 discusses variance estimation. Section 4 presents a simulation study. Section 5 is a summary and discussion. Methods for when more than two categorical variables are missing values are discussed in this section.

2. PFEFI with Logistic Regression

2.1 Notation

Let $X = (X_1, X_2, \dots, X_q)$ be q categorical variables that define a q -dimensional contingency table. Let $Z = (Z_1, Z_2, \dots, Z_p)$ be a set of covariates. Let $\pi_{ab\dots q}$ be the probability of being in cell (a, b, \dots, q) of the table; that is, the probability of being in the cell defined by $X_1 = x_a, X_2 = x_b, \dots, X_q = x_q$, where x_i is the i^{th} possible value for variable $X_i, i = 1, \dots, q$. If $q = 2$, then p_{iab} is the cell probability for being in row a and column b in a two-way table defined by variables X_1 and X_2 . In a two-way table, let θ be the log odds ratio for membership in cells at two levels of variable X_2 for two levels of variable X_1 . For example, if a and a' are two levels of X_1 and b and b' are two levels of X_2 , then for the corresponding log odds ratio

$$\theta = \log \frac{\pi_{ab}/\pi_{ab'}}{\pi_{a'b}/\pi_{a'b'}} = \log \frac{\pi_{ab}\pi_{a'b'}}{\pi_{a'b}\pi_{a'b}}.$$

In a two-by-two table, there is only one log odds ratio, $\theta = \log \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$. In a multiway table, θ is defined similarly for any two variables and pairs of values with all values of other variables being held constant. Suppose that the first two variables are contrasted and the others are held constant. Then

$$\theta = \log \frac{\pi_{abc\dots q}\pi_{a'b'c'\dots q}}{\pi_{ab'c'\dots q}\pi_{a'bc'\dots q}}.$$

Log odds involving other pairs of variables are defined similarly.

2.2 Estimating Conditional Probabilities of Cell Membership

Suppose that some values of X_1 and X_2 are missing. If there are $q > 2$ dimensions to the contingency table, then without loss of generality suppose that the first two variables have missing values. Let the set of completely observed X -variables be $X_{\text{obs}} = (X_3, X_4, \dots, X_q)$. Values can be missing on only X_1 , on only X_2 , or on both variables. Let $\underline{z} = (z_1, z_2, \dots, z_p)$ represent observed values for the covariates. Let $\underline{x} = (x_3, x_4, \dots, x_q)$ be observed categorical values.

The probability that $X_1 = a$ given the $X_2 = b$, $X_{\text{obs}} = \underline{x}$, and $Z = \underline{z}$ is $\pi_{a|b,\underline{x},\underline{z}} = P(X_1 = a | X_2 = b, X_{\text{obs}} = \underline{x}, Z = \underline{z})$. If X_1 has two levels, then this probability function can be estimated using logistic regression based on cases with both X_1 and X_2 observed. If X_1 has more than two levels, then polychotomous logistic regression must be used. If X_1 is observed but X_2 is missing, then probabilities and the estimation procedures are analogous. The probability that $X_2 = b$ given that $X_1 = a$, $X_{\text{obs}} = \underline{x}$, and $Z = \underline{z}$ is $\pi_{b|a,\underline{x},\underline{z}} = P(X_2 = b | X_1 = a, X_{\text{obs}} = \underline{x}, Z = \underline{z})$. It can be estimated using logistic regression or polychotomous logistic regression as appropriate.

If both X_1 and X_2 are missing, then the situation is more complicated. Feasible cell probabilities (there are AB feasible cell probabilities in a $A \times B$ table given values of X_{obs}) are determined by all cases with more information observed. That is, they are estimated based on all available partial information as in Kang, Koehler, and Larsen (2006, 2007a,b). Thus

$$\begin{aligned} \pi_{a,b|\underline{x},\underline{z}} &= P(X_1 = a, X_2 = b | X_{\text{obs}} = \underline{x}, Z = \underline{z}) \quad (1) \\ &= \pi_{a,b|(X_1, X_2)\text{observed},\underline{x},\underline{z}} + \\ &\quad \pi_{a|X_2\text{missing},\underline{x},\underline{z}}\pi_{b|a,\underline{x},\underline{z}} + \\ &\quad \pi_{b|X_1\text{missing},\underline{x},\underline{z}}\pi_{a|b,\underline{x},\underline{z}} \end{aligned}$$

For example, if X_1 and X_2 are both binary variables, then a case with values for both variables missing could receive as a donor potentially eight types of cases: four with both X_1 and X_2 observed, two with X_1 observed but X_2 missing, and two with X_1 missing but X_2 observed. If the donor has partial information,

then there is a probability that that donor would receive a donor from a particular cell. Thus for a two-by-two table it will be necessary to estimate up to eleven logistic regressions to produce fractional imputation weights. Two logistic regression could be needed if X_1 has missing cases. Two could be needed if X_2 has missing cases. The other seven could be needed if X_1 and X_2 are simultaneously missing. For an $A \times B$ table there would potentially be $AB + A + B - 1 + A(B - 1) + B(A - 1) = 3AB - 1$ logistic regressions: $A(B - 1)$ for missing X_1 's, $B(A - 1)$ for missing X_2 's, and $AB + A + B - 1$ for missing pairs (X_1, X_2) .

2.3 The Fractional Imputation Procedure and Estimator

The estimates of $\pi_{a|b,\underline{x},\underline{z}}$, $\pi_{b|a,\underline{x},\underline{z}}$, and $\pi_{a,b|\underline{x},\underline{z}}$ from the previous section are used as weights in the fractional imputation procedure. Table 1 shows an example of estimated fractional imputation weights associated with cases in a 2×2 table. If there are two categorical variables ($q = 2$) and Z is empty, then weights correspond to those in Kang, Koehler, and Larsen (2006, 2007a,b).

Table 1: Example of fractional imputation weights for 2×2 table with missing values.

Subset of cases	Count	Donor values		Weight		
		X_1	X_2	X_1	X_2	
S_1	n_1	1	1	–	–	1
S_2	n_2	1	0	–	–	1
S_3	n_3	0	1	–	–	1
S_4	n_4	0	0	–	–	1
S_5	n_5	1	?	–	1	$\hat{\pi}_{b=1 a=1,\underline{x},\underline{z}}$
				–	0	$\hat{\pi}_{b=0 a=1,\underline{x},\underline{z}}$
S_6	n_6	0	?	–	1	$\hat{\pi}_{b=1 a=0,\underline{x},\underline{z}}$
				–	0	$\hat{\pi}_{b=0 a=0,\underline{x},\underline{z}}$
S_7	n_7	?	1	1	–	$\hat{\pi}_{a=1 b=1,\underline{x},\underline{z}}$
				0	–	$\hat{\pi}_{a=0 b=1,\underline{x},\underline{z}}$
S_8	n_8	?	0	1	–	$\hat{\pi}_{a=1 b=0,\underline{x},\underline{z}}$
				0	–	$\hat{\pi}_{a=0 b=0,\underline{x},\underline{z}}$
S_9	n_9	?	?	1	1	$\hat{\pi}_{1,1 \underline{x},\underline{z}}$
				1	0	$\hat{\pi}_{1,0 \underline{x},\underline{z}}$
				0	1	$\hat{\pi}_{0,1 \underline{x},\underline{z}}$
				0	0	$\hat{\pi}_{0,0 \underline{x},\underline{z}}$

The weights are used in fractional imputation to produce fractionally imputed cell counts. In a 2×2 table the FI cell counts are given in table 2. Note that the estimated probabilities that are being added together in the sum vary with \underline{z} and \underline{x} and are not necessarily the same for all cases in a particular subset of the responses. The generalization of this to a $A \times B$ table is straight

forward. The FI weighted count in cell (a, b) is

$$\hat{n}_{ab} = n_{ab} + \sum_{i \in S: X_1=a, X_2=?} \hat{\pi}_{b|a, \underline{x}_i, \underline{z}_i} + \sum_{i \in S: X_2=b, X_1=?} \hat{\pi}_{a|b, \underline{x}_i, \underline{z}_i} + \sum_{i \in S: X_1=?, X_2=?} \hat{\pi}_{ab|\underline{x}_i, \underline{z}_i}. \quad (2)$$

Table 2: FI weighted cell counts for a 2×2 table. Response sets are defined in Table 1.

(X_1, X_2)	Weighted cell count
(1,1)	$n_1 + \sum_{i \in S_5} \hat{\pi}_{b=1 a=1, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_7} \hat{\pi}_{a=1 b=1, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_9} \hat{\pi}_{1,1 \underline{x}_i, \underline{z}_i}$
(1,0)	$n_2 + \sum_{i \in S_5} \hat{\pi}_{b=0 a=1, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_8} \hat{\pi}_{a=1 b=0, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_9} \hat{\pi}_{1,0 \underline{x}_i, \underline{z}_i}$
(0,1)	$n_3 + \sum_{i \in S_6} \hat{\pi}_{b=1 a=0, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_7} \hat{\pi}_{a=0 b=1, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_9} \hat{\pi}_{0,1 \underline{x}_i, \underline{z}_i}$
(0,0)	$n_4 + \sum_{i \in S_6} \hat{\pi}_{b=1 a=0, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_8} \hat{\pi}_{a=0 b=0, \underline{x}_i, \underline{z}_i} + \sum_{i \in S_9} \hat{\pi}_{0,0 \underline{x}_i, \underline{z}_i}$

The FI estimator of the log odds ratio is based on the FI weighted cell counts in Table 2 for a 2×2 table and formula 2 for a $A \times B$ table. For example, if a and a' are two levels of X_1 and b and b' are two levels of X_2 , then for the corresponding log odds ratio the estimate of θ is

$$\hat{\theta} = \log \frac{\hat{n}_{ab}/\hat{n}_{ab'}}{\hat{n}_{a'b}/\hat{n}_{a'b'}} = \log \frac{\hat{n}_{ab}\hat{n}_{a'b'}}{\hat{n}_{a'b}\hat{n}_{ab}}.$$

2.4 Fully Efficient Fractional Imputation (FEFI) and Partial FEFI

Fractional imputation methods that use all possible donors or all possible imputations for missing values are called fully efficient fractional imputation (FEFI) methods, because they use all available information to estimate the conditional distribution of a missing value (Kim and Fuller 2004). As such, they should have small imputation variance in the sense of Kalton and Kish (1984). The methods described so far in the paper therefore are instances of FEFI.

A large complication for the method in terms of the amount of estimation that is required occurs for the cases that are missing both X_1 and X_2 . In simulations in this work, partial FEFI (PFEFI) is used. In PFEFI, the cases with both variables missing are imputed directly into the cells defined by crossing levels of X_1 and X_2 . That is, $\pi_{a,b|\underline{x}, \underline{z}}$ is estimated directly via polychotomous logistic regression instead of through 1. In the case that values of X_1 and X_2 are missing completely at random with possibly heterogeneous response rates for the two variables results using PFEFI should not be substantially different from those using FEFI. It could be important to use FEFI as opposed to

PFEFI in other circumstances. Future work will provide detailed explanations.

3. Variance Estimation

Variance estimation for estimates of the log odds ratio is described in this section. In the case of no missing data or complete case analysis, the variance estimates for a log odds ratio estimator from a complete contingency table usually are given as $1/n_{ab} + 1/n_{ab'} + 1/n_{a'b} + 1/n_{a'b'}$. In the case of maximum likelihood estimation under a model, large sample maximum likelihood theory and the delta method for nonlinear transformations can be used to produce asymptotic formulas for variances. If maximum likelihood estimates are produced using the EM algorithm or one of its extensions, then the SEM algorithm (Meng and Rubin 1991) or a suitable variation (McLachlan and Krishnan 1996) plus the delta method can be used to produce variance estimates without evaluating the second derivative of the log likelihood.

Bayesian estimation would base its assessment of uncertainty on the posterior distribution of the parameter, θ , given the observed data. Variance estimation when multiple imputation is used is described in Section 3.2 below. Multiple imputation has the advantage that once multiple sets of imputations are created it is possible to run any number of complete data analyses on the replicate versions of the completed data.

For the fractional imputation method described in this article the jackknife resampling method is considered. The jackknife should produce consistent variance estimates for $\hat{\theta}$. Its primary disadvantage, however, is a large increase in the amount of computing required. In the fractional imputation method presented here, even in the case of a 2×2 table, several logistic regression functions must be estimated. For each jackknife replicate in principle it is necessary to estimate the same logistic regressions. Future work will consider alternatives. Section 3.1 discusses the jackknife procedure for this version of fractional imputation. Alternatives such as generalized cross validation and delete-d jackknife (with a random sample of jackknife samples) procedures might be considered to reduce computational burden (Shao and Tu 1995, Efron and Tibshirani 1993).

3.1 Jackknife Variance Estimation

The simple jackknife for the estimate of the log-odds ratio, θ , based on complete data is as follows. Let $\hat{\theta}$ be a consistent estimator of θ based on a sample S of n independent observations. Let $S^{(-j)}$ be the jackknife sample of size $n - 1$ obtained by dropping the j^{th} observation from the original sample S . Let $\hat{\theta}^{(-j)}$ be the estimate of θ based on $S^{(-j)}$. The set of n jackknife

estimates is $(\hat{\theta}^{(-1)}, \dots, \hat{\theta}^{(-n)})$. The jackknife estimator of θ is $\hat{\theta}_{\text{jack}} = \hat{\theta} + (n-1)(\bar{\hat{\theta}} - \hat{\theta})$, where $\bar{\hat{\theta}} = \frac{1}{n} \sum_{j=1}^n \hat{\theta}^{(-j)}$. The variance estimate, \hat{V}_{jack} , for the variance of $\hat{\theta}$ (or of $\hat{\theta}_{\text{jack}}$) is

$$\hat{V}_{\text{jack}} = \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}^{(-j)} - \bar{\hat{\theta}})^2. \quad (3)$$

Suppose some observations of the original sample S are incomplete on categorical variables. We can impute the missing data by the fractional imputation procedure as described in section 2. This produces the FI estimate of θ . In order to apply the jackknife procedure to estimate the variance of the point estimator in this case, the following three steps are repeated n times to produce the estimates $(\hat{\theta}^{(-1)}, \dots, \hat{\theta}^{(-n)})$.

1. Delete the j^{th} observation from S with incomplete some observations, yielding the sample $S^{(-j)}$.
2. Fill in the missing data in $S^{(-j)}$ by applying the fractional imputation procedure introduced in section 2, yielding $\hat{S}^{(-j)}$, an imputed version of $S^{(-j)}$.
3. Compute $\hat{\theta}^{(-j)}$ based on $\hat{S}^{(-j)}$.

Once the three steps are completed n times, use equation (3) to produce a variance estimate for $\hat{\theta}$. The jackknife variance estimation procedure should also be applicable to PFEFI.

The jackknife variance estimation procedure should produce a consistent estimator for the variance of $\hat{\theta}$. First, the estimator $\hat{\theta}$ is a smooth function of four cell counts. Second, the imputed cell counts are sums of weights. Third, the weights are computed based on logistic regressions or are products of terms produced through logistic regressions. In any case, the weights are smooth functions of probabilities produced through logistic regressions. Let $\pi(\beta)$ be the function of logistic regression parameters, β , based on one of the logistic regressions from Section 2. The jackknife method could be used to produce a consistent estimate of the variance of the estimated logistic regression parameters, $\hat{\beta}$. The following would be a consistent estimate of $V(\hat{\beta})$, the variance of $\hat{\beta}$:

$$V(\hat{\beta}) \cong \frac{n-1}{n} \sum_{j=1}^n (\hat{\beta}^{(-j)} - \hat{\beta})(\hat{\beta}^{(-j)} - \hat{\beta})'. \quad (4)$$

See Shao and Tu (1995, sections 2.1 and 8.2).

Using Taylor linearization applied to $\pi(\beta)$ as a function of β (see, Fuller 1996, theorem 5.1.7),

$$\pi(\hat{\beta}^{(-j)}) = \pi(\hat{\beta}) + \left(\frac{\partial \pi}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right)' (\hat{\beta}^{(-j)} - \hat{\beta}) + o_p\left(\frac{1}{n}\right).$$

Then the jackknife variance estimator defined by $V_{\text{jack}} = \frac{n-1}{n} \sum_{j=1}^n [\pi(\hat{\beta}^{(-j)}) - \pi(\hat{\beta})]^2$ is asymptotically equivalent to

$$\begin{aligned} V_{\text{jack}} &\cong \frac{n-1}{n} \sum_{j=1}^n \left[\left(\frac{\partial \pi}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right)' (\hat{\beta}^{(-j)} - \hat{\beta}) \right]^2 \quad (5) \\ &= \left(\frac{\partial \pi}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right)' \sum_{j=1}^n \frac{n-1}{n} (\hat{\beta}^{(-j)} - \hat{\beta})(\hat{\beta}^{(-j)} - \hat{\beta})' \\ &\quad \left(\frac{\partial \pi}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right) \\ &\cong \left(\frac{\partial \pi}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right)' V(\hat{\beta}) \left(\frac{\partial \pi}{\partial \beta} \Big|_{\beta=\hat{\beta}} \right). \end{aligned}$$

Thus V_{jack} is a consistent estimate of the variance of the estimated probabilities $\pi_{a|b}$, $\pi_{b|a}$, and π_{ab} from Section 2. Consequently, the jackknife procedure should produce consistent estimates of the variance of the estimated log odds ratio.

3.2 MI Variance Estimation

Multiple imputation methods create M versions of completed tables. From each table estimates of a log odds ratio and the standard error of the estimates are produced. The MI estimate of the log odds ratio is then the average of the M estimates of the log odds ratio. The variance estimate for the MI estimate is produced using standard formulas for combining estimates of the variability within one completed-table analysis and the variability between estimates from the M tables. See Rubin (1987), Schafer (1997), or Little and Rubin (2002) for formulas.

4. Simulation

4.1 Simulation Design

Let X_1 and X_2 denote two categorical response variables that have 0 or 1 binary values. Let Z_1 and Z_2 have a bivariate normal distribution with mean vector μ_d , $d = 1, 2, 3, 4$, which depends on (X_1, X_2) , and variance-covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. Let μ_1 be a conditional mean vector given $X_1 = 1, X_2 = 1$, μ_2 be for $X_1 = 1, X_2 = 0$, μ_3 be for $X_1 = 0, X_2 = 1$ and μ_4 be for $X_1 = 0, X_2 = 0$.

The values of (X_1, X_2) are generated from the multinomial distribution with $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.3, 0.2, 0.2, 0.3)$. The true log-odds ratio, $\log\left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right)$, is 0.8109.

Three types of data sets are generated with different associations, R^2 , between (X_1, X_2) and (Z_1, Z_2) by varying μ_d as stated in Table 3. One thousand data sets are generated per each type. The

association R^2 was empirically computed by the following formula (6) from the generated data sets;

$$R^2 = 1 - \frac{|S_{XX} - S_{XZ}S_{ZZ}^{-1}S_{ZX}|}{|S_{XX}|}, \quad (6)$$

where S is the sample variance-covariance matrix of X_1, X_2, Z_1, Z_2 such that $S = \begin{pmatrix} S_{XX} & S_{XZ} \\ S_{ZX} & S_{ZZ} \end{pmatrix}$.

Table 3: Three types of data sets.

Data Type	μ_1'	μ_2'	μ_3'	μ_4'	R^2
1	(10,11)	(10,11)	(10,11)	(10,11)	0.00
2	(10,11)	(11,12)	(11,12)	(12,13)	0.31
3	(10,11)	(11,12)	(12,13)	(13,14)	0.53

The generated sample size per each data set is 200. Variables X_1 and X_2 are missing completely at random (MCAR) with probability 0.2.

Multiple imputed data sets for the categorical variables are generated by data augmentation with a Dirichlet prior, using 'dataDepPrior' and 'daCgm' functions in S-Plus (2001; see also Schafer 1997). The program parameter "nPriorObs" indicates the number of 'prior observations' that are used in computations. The default prior distribution assumes independence between the variables, or a log odds ratio of 0.00.

4.2 Simulation Results for Log Odds Ratio

The new method, which is partial fully efficient fractional imputation (PFEFI) with logistic regression, is compared with multiple imputation (MI), and complete case analysis (CC). Results are presented assuming no missing data for comparison (Full).

Table 4 contains the averages of 1000 log-odds point estimates; values in parentheses are standard deviations of the estimates. Table 4 shows that the new method and CC have similar performances which are close to the results of fully observed data set. The new method tends to have smaller standard deviations than CC when R^2 is increased. As the number of prior observations are increased, MI estimates are moved toward zero and are less variable.

Table 5 contains the averages of 1000 standard errors of log-odds ratio; the values in parenthesis are standard deviations of 1000 standard errors of log-odds. The values in Table 5 shows that the new method with jackknife and CC have less biased estimates of standard errors than MI. The values in parenthesis in Table 4 are close to the values in Table 5. The new method with jackknife is more efficient for estimating standard errors than CC when

R^2 is increased. MI has the biggest standard deviations of 1000 standard errors of log-odds ratio.

4.3 Simulation Results for a Cell Probability

The FEFI and PFEFI methods also can be used to produce estimates of cell probabilities. For a 2×2 table, π_{11} denotes the probability that a case has $X_1 = 1$ and $X_2 = 1$. The values in Table 6 and Table 7 show that all three methods provide essentially unbiased estimates for the cell probability and standard errors. The standard errors of the estimates differ across methods. Complete-case analysis provides the estimate of π_{11} with the largest variance. The new method, PFEFI with jackknife, tends to provide smaller standard errors of the cell proportion than MI in most cases.

5. Summary and Discussion

Fully efficient fractional imputation (FEFI) for missing categorical responses when covariate information is available has been defined. Estimation of a log odds ratio has been presented. Variance estimation using a jackknife procedure has been described, justified, and illustrated. An alternative to FEFI, called partial FEFI (PFEFI), should be appropriate for independent sampling of cases from a population and missing completely at random.

Simulation results suggest that the new method has the best performance of alternatives considered for estimating cell probabilities. Although the imputation method improves the estimation of individual cell probabilities relative to complete-case analysis, it is not always true that it can improve the estimation of a measure of association between the two variables. When the covariates have less information about the categorical variables, the imputation methods, the new method and MI can not provide more information on association between two categorical variables. In this case, the complete case analysis is good enough to estimate the log-odds ratio. If the correlation between categorical variables and covariates is higher, the new method provides better estimates of log-odds ratio.

Although the missingness of X_1 and X_2 are missing completely at random (MCAR) in the simulation study, FEFI and PFEFI should also work under MAR when the missingness of X_1 and X_2 depends on the covariates Z .

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0532413. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the

views of the National Science Foundation.

Table 4: Point estimation of log-odds ratio. Standard deviations of estimates are in parentheses.

Data Type (R^2)	1 (0.00)	2 (0.31)	3 (0.53)	
New	0.833 (0.3740)	0.824 (0.3592)	0.825 (0.3342)	
MI	nPriorObs=0	0.843 (0.3853)	0.830 (0.3704)	0.830 (0.3408)
	nPriorObs=5	0.829 (0.3773)	0.821 (0.3673)	0.825 (0.3399)
	nPriorObs=10	0.817 (0.3755)	0.816 (0.3594)	0.820 (0.3373)
CC	0.833 (0.3733)	0.826 (0.3774)	0.823 (0.3644)	
Full	0.835 (0.2921)	0.818 (0.2961)	0.821 (0.2901)	

References

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm (with comments)." *JRSS-B*, 39, 1-37.
- Efron, B., and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
- Fuller, W. A. (1996). *Introduction to Statistical Time Series*, New York: Wiley.
- Kalton, G., and Kish, L. (1984). "Some efficient random imputation methods." *Communications in Statistics A*, 13, 1919-1939.
- Kang, S. S., Koehler, K. J., and Larsen, M. D. (2007a). "Fractional imputation and tests of independence for incomplete two-way contingency tables." *Computational Statistics and Data Analysis*, submitted.
- Kang, S. S., Koehler, K. J., and Larsen, M. D. (2007b). "Fractional imputation for incomplete two-way contingency tables." *Communications in Statistics – Theory and Methods*, submitted.
- Kang, S. S., Koehler, K. J., and Larsen, M. D. (2006). "Tests of independence with incomplete contingency tables using likelihood functions." *2006 Proceedings of the Survey Research Methods Section, American Statistical Association*. [CD-ROM]. Alexandria, VA: American Statistical Association.
- Kim, J. K., and Fuller, W. A. (2004). "Fractional hot deck imputation." *Biometrika*, 91, 559-589.
- Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data.*, 2nd edition J. Wiley & Sons, New York.
- McLachlan, G. J., and Krishnan, T. (1996). *The EM Algorithm and Extensions*. Wiley-Interscience.
- Meng, X. L., and Rubin, D. B. (1991). "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm." *Journal of the American Statistical Association*, 86, 899-909.
- Rubin, D. B. (1978). "Multiple imputation in sample surveys - A phenomenological Bayesian approach to nonresponse." *Proceedings of the Survey Research Methods Section, American Statistical Association 1978*, 20-34.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall / CRC Press.
- Shao, J., and Tu, D. (1995). *The jackknife and bootstrap*. Springer-Verlag, New York, Inc.
- S-Plus 6.1 Manual: Analyzing Data with Missing Values in S-Plus (2001). Insightful Corporation. Seattle, Washington.

Table 5: Estimation of S.E. (log-odds ratio). Standard deviations of estimates are in parentheses.

Data Type (R^2)	1 (0.00)	2 (0.31)	3 (0.53)
New with Jackknife	0.3742 (0.0147)	0.3568 (0.0139)	0.3431 (0.0120)
nPriorObs=0	0.3698 (0.0352)	0.3552 (0.0301)	0.3387 (0.0228)
MI	0.3707 (0.0348)	0.3523 (0.0292)	0.3395 (0.0235)
nPriorObs=10	0.3680 (0.0346)	0.3534 (0.0285)	0.3381 (0.0228)
CC	0.3661 (0.0133)	0.3662 (0.0138)	0.3662 (0.0125)
Full	0.2913 (0.0050)	0.2911 (0.0050)	0.2912 (0.0048)

Table 6: Point estimation of π_{11} . Standard deviations of estimates are in parentheses.

Data Type (R^2)	1 (0.00)	2 (0.31)	3 (0.53)
New	0.3019 (0.03680)	0.2996 (0.03748)	0.3004 (0.03515)
nPriorObs=0	0.3021 (0.03705)	0.2997 (0.03789)	0.3003 (0.03547)
MI	0.3012 (0.03663)	0.2993 (0.03763)	0.3001 (0.03526)
nPriorObs=10	0.3005 (0.03664)	0.2988 (0.03743)	0.2997 (0.03509)
CC	0.3022 (0.04016)	0.2993 (0.04163)	0.3003 (0.04141)
Full	0.3026 (0.03162)	0.2993 (0.03368)	0.2998 (0.03206)

Table 7: Estimation of S.E. ($\hat{\pi}_{11}$). Standard deviations of estimates are in parentheses.

Data Type (R^2)	1 (0.00)	2 (0.31)	3 (0.53)
New with Jackknife	0.0374 (0.00138)	0.0363 (0.00142)	0.0355 (0.00132)
nPriorObs=0	0.0371 (0.00246)	0.0362 (0.00220)	0.0353 (0.00183)
MI	0.0371 (0.00239)	0.0361 (0.00220)	0.0354 (0.00193)
nPriorObs=10	0.0370 (0.00248)	0.0362 (0.00218)	0.0353 (0.00184)
CC	0.0405 (0.00189)	0.0403 (0.00200)	0.0404 (0.00196)
Full	0.0324 (0.00097)	0.0323 (0.00105)	0.0323 (0.00100)