Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 1, pp. 185 ~ 194

Estimation of Log-Odds Ratios for Incomplete 2×2 Tables with Covariates using FEFI

Shin-Soo Kang[1] · Je-Min Bae[2]

Abstract

The information of covariates are available to do fully efficient fractional imputation(FEFI). The new method, FEFI with logistic regression is proposed to construct complete contingency tables. Jackknife method is used to get a standard errors of log-odds ratio from the completed table by the new method. Simulation results, when covariates have more information about categorical variables, reveal that the new method provides more efficient estimates of log-odds ratio than either multiple imputation(MI) based on data augmentation or complete case analysis.

Keywords : Complete Case Analysis, Fractional Imputation, Multiple Imputation, Wald Statistic

1. Introduction

In the analysis of 2×2 contingency tables, it may happen that one of the binary responses is not observed for some respondents, but there is covariate information that can be used to impute the missing responses. Although using imputation in the analysis of missing data has been studied for a long time, the analysis of incomplete contingency tables with covariates has not received sufficient attention. One simple approach, known as complete-case(CC) analysis, discards the missing data ignoring covariates information. An alternative approach involves constructing a complete table, in which all cases are completed classified, by imputing information for the missing row or column classification. Multiple

1) Professor, Department of Management and Information, Kwandong University, Kangnung, 210-701, Korea
E-mail: sskang@kd.ac.kr
2) Associate Professor, Department of Computer Education, Kwandong University Kangnung, 210-701, Korea
E-mail: gemini@kd.ac.kr

imputation, proposed by Rubin (1978), provides a way analyzing completely classified tables.

The incomplete data with one binary response variable and covariates has been studied widely. Fitzmaurice et al. (1994) described a likelihood-based method to estimate logistic regression coefficients for incomplete binary response and proposed a consistent estimator of the asymptotic variance-covariance matrix of the estimators using only the first derivatives of log-likelihood function. When the data has a binary outcome variable with incompletely observed categorical covariates, Vach and Schumacher (1993) estimated logistic regression coefficients using likelihood based approach.

Let X_1, X_2 be two categorical variables and $Z=(Z_1,Z_2,\cdots,Z_p)$ be a set of covariates. Let π_{ij} be the cell probability in i^{th} row and j^{th} column in two way contingency table of (X_1, X_2) and θ be log-odds ratio, $\log(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}})$.

We restrict that the missingness is confined to two categorical variables. A fully efficient fractional imputation(FEFI) method with logistic regression is proposed to impute missing values on X_1, X_2. The log-odds ratio and cell probability, π_{11} are estimated and their variance estimates are studied.

2. FEFI with Logistic Regression

There are three kinds of missing units with missing on X_1,missing on X_2 , and missing on both variables. Let $\underline{z}=(z_1,z_2,\cdots,z_p)$represent observed values for the covariates. Given the observed $X_2=v$ and $\underline{z}$, let $\varpi_{u|\underline{z}}$ be $P(X_1=u|X_2=v,Z_1=z_1,Z_2=z_2,\cdots,Z_p=z_p)$ for the units missing on X_1, $\phi_{u|\underline{z}}$ be $P(X_2=v|X_1=u,Z_1=z_1,Z_2=z_2,\cdots,Z_p=z_p)$ for the units missing on X_2, and $\phi_{u|\underline{z}}$ be$P(X_1=u,X_2=v|Z_1=z_1,Z_2=z_2,\cdots,Z_p=z_p)$ for both missing units, where $u,v=0,1$.

We can fit the following logistic regression model through complete cases to estimate $\varpi_{uv|\underline{z}}$,

$$\text{logit}(\varpi_{11|\underline{z}})=\beta_0+\alpha+\beta_{1z_1}+\beta_{2z_2}+\cdots+\beta_{pz_p}\ ,$$

then $\widehat{\varpi_{11}|\underline{z}}=\frac{\exp(\hat{\beta}_0+\hat{\alpha}+\widehat{\beta_{1z_1}}+\widehat{\beta_{2z_2}}+\ldots+\widehat{\beta_{pz_p}})}{1+\exp(\hat{\beta}_0+\hat{\alpha}+\widehat{\beta_{1z_1}}+\widehat{\beta_{2z_2}}+\ldots+\widehat{\beta_{pz_p}})}$ and $\widehat{\varpi_{01}|\underline{z}}=\frac{1}{1+\exp(\hat{\beta}_0+\hat{\alpha}+\widehat{\beta_{1z_1}}+\widehat{\beta_{2z_2}}+\ldots+\widehat{\beta_{pz_p}})}$.

The logistic regression model is set up in a same manner to estimate $\phi_{vu|\underline{z}}$ for the units missing on X_2.

Let's define three binary response variables,

$$Y_1 = \begin{cases} 1 & \text{if } X_1 = 1 \text{ and } X_2 = 1 \\ 0 & \text{otherwise,} \end{cases}$$
$$Y_2 = \begin{cases} 1 & \text{if } X_1 = 1 \text{ and } X_2 = 1 \\ 0 & \text{otherwise,} \end{cases}$$
$$Y_3 = \begin{cases} 1 & \text{if } X_1 = 1 \text{ and } X_2 = 1 \\ 0 & \text{otherwise,} \end{cases}$$

to set up logistic regression models for estimating of $\varphi_{uv|\underline{z}}$, Let κ_i be

$$\kappa_i = \log\ddot{y}\left(\Pr\left(Y_i = 1|\underline{z}\right)\right)$$

in the logistic regression model, where $i = 1,2,3$. After fitting three logistic regression models, we will get

$$\widehat{\varphi_{11|\underline{z}}} = \frac{\exp(\widehat{k_1})}{1+\exp(\widehat{k_1})+\exp(\widehat{k_2})+\exp(\widehat{k_3})},$$
$$\widehat{\varphi_{12|\underline{z}}} = \frac{\exp(\widehat{k_2})}{1+\exp(\widehat{k_1})+\exp(\widehat{k_2})+\exp(\widehat{k_3})},$$
$$\widehat{\varphi_{21|\underline{z}}} = \frac{\exp(\widehat{k_3})}{1+\exp(\widehat{k_1})+\exp(\widehat{k_2})+\exp(\widehat{k_3})},$$
$$\widehat{\varphi_{22|\underline{z}}} = \frac{1}{1+\exp(\widehat{k_1})+\exp(\widehat{k_2})+\exp(\widehat{k_3})}.$$

The estimates of $\varpi_{uv|\underline{z}}$, $\phi_{vu|\underline{z}}$, and $\varphi_{uv|\underline{z}}$ are used as weights in the FEFI procedure. Table 1 shows an example given the estimated values of weights and the 2×2 contingency table is made from Table 1 as shown in Table 2.

<Table 1> Example of FEFI

obs.	Observed X_1	Observed X_2	FEFI X_1	FEFI X_2	Weights
1	1	1	1	1	1
2	1	0	1	0	1
3	0	1	0	1	1
4	0	0	0	0	1
5	?	1	1	1	$\varpi_{11\mid z_5}$
			0	1	$\varpi_{01\mid z_5}$
6	0	?	0	1	$\phi_{01\mid z_6}$
			0	0	$\phi_{00\mid z_6}$
7	?	?	1	1	$\varphi_{11\mid z_7}$
			1	0	$\varphi_{10\mid z_7}$
			0	1	$\varphi_{01\mid z_7}$
			0	0	$\varphi_{00\mid z_7}$

<Table 2> FEFI cell counts for (X_1, X_2) from Table1

(1,1)	(1,0)	(0,1)	(0,0)
$1+\hat{\varpi}_{11\mid z_5}+\hat{\varphi}_{11\mid z_7}$	$1+\hat{\varphi}_{10\mid z_7}$	$1+\hat{\varpi}_{01\mid z_5}+\hat{\phi}_{01\mid z_6}+\hat{\varphi}_{01\mid z_7}$	$1+\hat{\phi}_{00\mid z_6}+\hat{\varphi}_{00\mid z_7}$

3. Variance Estimation

Most discussions of imputation methods and the EM algorithm concern point estimation of population quantities with missing values. A second concern is how to get standard errors of the point estimates obtained from the filled-in data by imputation methods and EM algorithm.

The resampling method is one of general approaches to account for the additional uncertainty due to nonresponse. Apply the imputation and analysis procedure repeatedly to resampled versions of the incomplete data. Two major resampling methods are the bootstrap and the jackknife. These methods are often easy to implement and have broad applicability, but they rely on large samples and are computationally intensive. Jackknife method is examined in this section, which is widely used in survey sampling applications.

Another approach is multiple imputation. We can create multiply imputed data sets that allow the additional uncertainty from imputation to be assessed. Multiple imputation(MI) was first proposed by Rubin(1978). Replacing each missing value by a vector of $D \geq 2$ imputed values. We impute several values for each missing

value instead of just one for the MI. D completed data sets can be created from the vectors of imputations: For example, the first set of imputed values are used to form the first completed data set. D sets of imputations are repeated random draws from the predictive distribution of the missing values. Standard complete-data methods are used to analyze each data set and D complete-data inferences can be combined to form one inference that properly reflects uncertainty due to nonresponse.

3.1 Jackknife Variance Estimation

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

$$\hat{\theta}_{jack} = \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}), \tag{1}$$

where $\bar{\theta} = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}^{(\setminus j)}$ Now $\hat{V}_{jack}$, the variance of $\hat{\theta}$ or $\hat{\theta}_{jack}$ is

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

Suppose some observations of the original sample S are incomplete on categorical variables. We can impute the missing data by the FEFI procedure with logistic regression in section 2. It is explained how to apply jackknife procedure to get the variance of the point estimation from the imputed data by FEFI methods. We repeat the following 1-3 steps n times to get the set of estimates $(\theta^{\widehat{(\setminus 1)}},\cdots,\theta^{\widehat{(\setminus n)}})$:

- Step1: Delete the j^{th} observation from S with incomplete some observations, yielding the sample . $S^{(\setminus j)}$.
- Step2: Fill in the missing data in $S^{(\setminus j)}$ by applying FEFI procedures introduced in section 2, yielding $\hat{S}^{(\setminus j)}$
- Step3: Compute $\theta^{\widehat{(\setminus j)}}$ on ^ $S^{\widehat{(\setminus j)}}$, which is the imputed jackknife sample.

Now we can use equations (2) for a consistent estimates of the variances of $\hat{\theta}$ For this Jackknife variance estimator, we need n times FEFI procedures.

3.2 MI Variance Estimation

It is reviewed how to combine D complete data inferences to get an estimate of θ, $\widehat{\theta_{MI}}$ and an estimate of the variance of $\hat{\theta}_{MI}$, $\hat{V}(\widehat{\theta_{MI}})$ Each data set completed by imputation is analyzed using the same complete-data method. Let $\hat{\theta}_d$ be the complete-data estimate of θ based on the d^{th} imputed data, where $d=1,\cdots,D$
The multiple imputation estimator of θ, $\hat{\theta}_{MI}$ is the average of D estimates of Θ from D imputed data sets.

$$\hat{\theta}_{MI}=\frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d. \tag{3}$$

$\widehat{\theta_{MI}}$ provides a valid estimate of μ and increases the efficiency of estimate over a single imputation estimator based on the stochastic regression imputation method.
Let $W_{d,}\hat{V}(\widehat{\theta_d})$ be the estimate of the variance of $\hat{\theta}_d$ based on the d^{th} imputed data. Now $\hat{V}(\hat{\theta}_{MI})$ has two components as follows:

1. The average within-imputation variance: $\overline{W_D}=\frac{1}{D}\sum_{d=1}^{D}W_d$ and $\overline{W_D}$ is the estimated total variance when there is no missing value.

2. The between-imputation componen: $B_D=\frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta_d}-\widehat{\theta_{MI}})^2$.

The total variability associated with $\hat{\theta}_{MI}$ is

$$\hat{V}(\widehat{\theta_{MI}})=T_D=\overline{W_D}+\frac{D+1}{D}B_D, \tag{4}$$

where $\frac{D+1}{D}$ is an adjustment for finite D.

4. Simulation Results

4.1 Simulation Design

There are four random variables, X_1,X_2,Z_1,Z_2, where X_1,X_2 denote two categorical response variables that have 0 or 1 binary values and Z_1,Z_2 have bivariate normal distribution with mean vector $\mu_d,d=1,2,3,4$ 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 random numbers 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, $\log(\frac{\pi_{11}\pi_{12}}{\pi_{12}\pi_{21}})$, is 0.8109.

Three types of data sets are generated to differ the association, R^2, between (X_1, X_2) and (Z_1, Z_2)varying μ_d in Table 3. 1000 data sets are generated per each type. R^2 was empirically computed by the following formula (5) from the generated data sets;

$$R^2 = 1 - \frac{|S_{XX} - S_{XZ}S_{ZZ}^{-1}S_{ZX}|}{S_{XX}}, \tag{5}$$

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

<Table 3> Four types of data sets

Type	μ_1'	μ_2'	μ_3'	μ_4'	R^2
1	(10,11)	(10,11)	(10,11)	(10,11)	1.00
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 and the missing probability of $X1$ and $X2$ are 0.2 under missing completely at random(MCAR) missing mechanism.

4.2 Simulation Results

The new method, which is fully effcient fractional imputation(FEFI) with iogistic regression, is compared with multiple impoutation(MI), complete caseanalysis(CC) and standard analysis for fully observed data set(Full). The values are averages of 1000 iog-odds point estimates and the values in parenthesis are standard deviations of 1000 log-odds estimates in Table 4. Table 4 shows that the new method and CC have similar performsnces which are close to the results of fully observed data set and the newmethod tends to have less standard deviations when R^2 is increased than CC.

Multiple imputed data sets for the categorical variables are generated by data augmentation with dirichlet flattening prior, c=1.05 using 'emCgm' function in S_Plus(2001). MI seems to give bias estimates of log-odds ratio. This bias may be reduced little bit as choosing another prior.

<Table 4> Point estimation of log-odds ratio

Type(R^2)	1(0.00)	2(0.31)	3(0.53)
New	0.833	0.824	0.825
	(0.3740)	(0.3592)	(0.3324)
MI	0.752	0.766	0.777
	(0.3443)	(0.3381)	(0.3186)
CC	0.833	0.826	0.823
	(0.3733)	(0.3774)	(0.3644)
Full	0.835	0.818	0.821
	(0.2921)	(0.2961)	(0.2901)

The values are averages of 1000 standard errors of log-odds ratio and the values in parenthesis are standard deviations of 1000 standard errors of log-odds in Table 5. 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 efficent to estimate standard errors than CC when R^2 is getting increased. MI has the biggest standard deviations of 1000 standard errors of log-odds ratio.

<Table 5> Estimation of S.E(log-odds ratio)

Type(R^2)	1(0.00)	2(0.31)	3(0.53)
New with	0.3742	0.3568	0.3431
Jackknife	(0.0147)	(0.0139)	(0.0120)
MI	0.3639	0.3499	0.3357
	(0.0326)	(0.0279)	(0.0223)
CC	0.3661	0.3662	0.3662
	(0.0133)	(0.0138)	(0.0125)
Full	0.2913	0.2911	0.2912
	(0.0050)	(0.0050)	(0.0048)

5. Conclusion

The values in Table 6 and Table 7 show that All three methods provide essentially unbiased estimates for the cell probabilities and their 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, FEFI with jackknife tends to provide smaller standard errors of cell proportion than MI in most cases.

<Table 6> Point estimation of π_{11}

Type(R^2)	1(0.00)	2(0.31)	3(0.53)
New	0.3019	0.2996	0.3004
	(0.03680)	(0.03748)	(0.03515)
MI	0.2970	0.2960	0.2974
	(0.03611)	(0.03684)	(0.03428)
CC	0.3022	0.2993	0.3003
	(0.04016)	(0.04163)	(0.04141)
Full	0.3026	0.2993	0.2998
	(0.03162)	(0.03368)	(0.03206)

We can conclude that the new method has always the best performances to estimate the cell probabilities. Although the imputation methods improve the

<Table 7> Estimation of S.E($\hat{\pi}_{11}$)

Type(R^2)	1(0.00)	2(0.31)	3(0.53)
New with	0.0374	0.0363	0.0355
Jackknife	(0.00138)	(0.00142)	(0.00132)
MI	0.0368	0.0359	0.0352
	(0.00239)	(0.00216)	(0.00182)
CC	0.0405	0.0403	0.0404
	(0.00189)	(0.00200)	(0.00196)
Full	0.0324	0.0323	0.0323
	(0.00097)	(0.00105)	(0.00100)

estimation of individual cell probabilities relative to complete-case analysis, it is not always true when we consider a measure of association between the two variables like log-odds ratio. 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 log-odds ratio. If the correlation between categorical variables and covariates is higher, the new method provides the better estimates of log-odds ratio.

References

1. Fitzmaurice, G. M., Laird, N. M., Hall, and Lipsitz, S. R. (1994). Analyzing Incomplete Longitudinal Binary Responses: A Likelihood-Based Approach. *Biometrics*, 50, 601-612.
2. Rubin, D. B. (1978). Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse . in *Proceedings of the Survey Research Methods Section, American Statistical Association* 1978, 20-34.
3. S-Plus 6.1 Manual: *Analyzing Data with Missing Values in S-Plus (2001)*. Insightful Corporation. Seattle, Washington.
4. Vach, W., and Schumacher, M. (1993). Logistic Regression with Incompletely Observed Categorical Covariates: A Comparison of Three Approaches. *Biometrika*, 80, 353-362.

[received date : Nov. 2006, accepted date : Dec. 2006]