A Power Transformation Method for Incomplete Two way Contingency Tables with Nonignorable Nonrespose

Seongyong Kim
Economics and Statistics Institution, Korea University Sejong Campus
2511 Sejong-Ro, Jochiwon-Eup,Yeongi-Gun
Chungnam, 339-700, Repubplic of Korea
E-mail: yaba96@korea.ac.kr

Yousung Park
Department of Statistics, Korea University
5-1 Anam-dong, Seongbuk-Gu
Seoul, 136-701, Republic of Korea
E-mail: your e-mail Address

Introduction

Categorical data usually summarized by a contingency table are analyzed to investigate the relationship between categorical response variables, or a categorical response variabe and explanatory variable. We often encounter non-responses in such analyses, yielding incomplete contingency table with partially classified counts or unclassified counts. Non-responses can be ignorable when they are missing completely at random, or missing at random (Little and Rubin 2002). However, non-response is nonignorable when they occur depending on their unobserved values in the sense that discarding them or mis-specifying their missing mechanism leads to large variances and biases in estimating parameters of interest (Chen 1972, Park and Brown 1994, Choi *et al.* 2009, Park and Choi 2010).

Fay (1986) used log-linear models to impute nonignorable nonresponse, and Baker and Laird (1988) indicated that the maximum likelihood estimates (ML) often give rise to estimates on the boundary solution of the parameter space and to imperfect fits for saturated models. They also provided the conditions that ML falls on the boundary solution and, in particular, proposed a simple condition of the boundary solution arising from a saturated log-linear model for a 2×2 contingency table (i.e, one response variable and one explanatory variable both with two categories). Baker *et al.* (1992) suggested the conditions for boundary solutions in two-way contingency tables where one or both response variables can be missing. However, it is not easy to use those conditions because they are not often in closed forms except special cases such as a two-way contingency table with the same number of categories for both response variables.

To overcome the boundary solution problem, Bayesian approaches have been suggested. Park and Brown (1994) and Park (1998) considered Bayesian models with empirical priors depending only on responses. Clogg *et al.* (1991) used a constant prior for an incomplete one-way contingency table. For two-way contigency tables, Choi *et al* (2009) and Park and Choi (2010) expanded Park and Brown (1994) by introducing the priors depending both on responses and non-responses.

To see when a boundary solution problem occurs in the ML estimation for two-way tables, we present an explicit condition expressed by ordered ratios of observed cell counts. This condition is very simple, reduced to Baker and Laird (1988) when a 2×2 contingency table is considered, and easily expanded to one-way and more-than-two way contingency tables. Using those conditions, we propose a frequentist approach to escape such a boundary solution by employing a power transformation as a link function instead of the usual log-linear or logit link function. However, this method may not be an answer for the boundary solution problem in a contingency table with more than two categories because the new condition is only a necessary condition for such a contingency table not to have a boundary

solution. Thus, we priorly distribute a value between 0 and 0.5 to each cell, where, in particular, 0.5 is called the Jeffrey prior, and then use the power link function to guarantee the necessary condition the ML is not on the boundary solution. The ML estimates under such a power link function are compared with those under the logit link function for two-way tables with nonignorable nonresponses using an empirical example and several scenarios of simulated data sets.

Conditions for boundary solutions

Let Y_1 and Y_2 be categorical variables indexed by I and J categories, respectively. We also let $R_1 = 1$ when Y_1 is observed and $R_1 = 2$ when Y_1 is not observed and, similarly $R_2 = 1$ for observed Y_2 and $R_2 = 2$ for unobserved Y_2. Then the full array of Y_1, Y_2, R_1, and R_2 constructs an $I \times J \times 2 \times 2$ contingency table which has completely classified counts, partially classified counts, and unclassified counts. To distinguish these three types of counts, let $y_{ijk\ell}$ be the count belonging to the ith category of Y_1, the jth category of Y_2, the kth value of R_1, and the ℓth value of R_2. Thus, y_{ij11} is used for the completely classified counts, y_{i+12} and y_{+j21} for respective column and row supplementary margins, and y_{++22} for unclassified counts.

We consider the following model to describe a general form of nonignorable nonresponse models which can be suffered from boundary solutions.

$$(1) \quad log\,(m_{ijk\ell}) = \lambda_0 + \lambda^i_{Y_1} + \lambda^j_{Y_2} + \lambda^k_{R_1} + \lambda^\ell_{R_2} + \lambda^{ij}_{Y_1Y_2} + \lambda^{ik}_{Y_1R_1} + \lambda^{i\ell}_{Y_1R_2} + \lambda^{j\ell}_{Y_2R_2} + \lambda^{jk}_{Y_2R_1} + \lambda_{R_1R_2}$$

where the sum of each λ-term across its respective superscript is zero and $m_{ijk\ell} = N \cdot \pi_{ijk\ell}$ is the excepted cell size. By assigning a symbol that lists the highest-order terms, we denote this model by $[Y_1Y_2, Y_1R_1, Y_1R_2, Y_2R_2, Y_2R_1, R_1R_2]$. It is well known that the interaction terms of Y_1R_2 and Y_2R_1 do not introduce any boundary solution in likelihood estimation (Baker *et al.* 1992, Park and Choi 2010). Thus, we focus following five nonignorable nonresponse models : $[Y_1Y_2, Y_1R_1, Y_2R_2, R_1R_2]$, $[Y_1Y_2Y_1R_1Y_1R_2, R_1R_2]$, $[Y_1Y_2Y_1R_2, Y_2R_2, R_1R_2]$, $[Y_1Y_2, Y_1R_1, R_1R_2]$, $[Y_1Y_2, Y_2R_2, R_1R_2]$.

We assume that $\{y_{ijk\ell}\}$ follow a multinomial distribution with observations, y_{ij11}, y_{i+12}, y_{+j21}, and y_{++22}, as given by

$$(2) \quad L = \sum_i \sum_j y_{ij11} log(\pi_{ij11}) + \sum_i y_{i+12} log(\pi_{i+12}) + \sum_j y_{+j21} log(\pi_{+j21}) + y_{++22} log(\pi_{++22})$$

where $\pi_{ijk\ell} = Pr[Y_1 = i, Y_2 = j, R_1 = k, R_2 = \ell]$ and $N = \sum_{i,j,k,\ell} y_{ijk\ell}$ is fixed.

Let $\beta_{ij} = m_{ij21}/m_{ij11}$, $\gamma_{ij} = m_{ij12}/m_{ij11}$. Under above five models, β_{ij} depends only on subscript i and γ_{ij} only on j. To stress these dependency, we denote them by $\beta_{i\cdot}$ and $\gamma_{\cdot j}$, respectively. Baker *et al.* (1992) showed that the ML estimates lie on the boundary of the parameter space if a $\hat{\beta}_{i\cdot}$ from $\sum_i \hat{m}_{ij11}\hat{\beta}_{i\cdot} = y_{+j21}$ is negative or a $\hat{\gamma}_{\cdot j}$ from $\sum_j \hat{m}_{ij11}\hat{\gamma}_{\cdot j} = y_{i+12}$ is negative. They also showed the closed form of ML estimates for m_{ij11}.

We now reduce a $I \times J \times 2 \times 2$ to a $2 \times J \times 2 \times 2$ or a $I \times 2 \times 2 \times 2$ contingency table by aggregating all the counts from the second category to the last category over the variables Y_1 and Y_2, respectively. Accordingly, we have observations, in a $2 \times J \times 2 \times 2$ contingency table, $y^*_{1j11} = y_{1j11}$, $y^*_{2j11} = \sum_{i=2}^I y_{ij11}$, $y^*_{1+12} = y_{1+12}$, $y^*_{2+12} = \sum_{i=2}^I y_{i+12}$, and the corresponding ML estimates are denoted by $m^*_{1jk\ell} = m_{1jk\ell}$ and $m^*_{2jk\ell} = \sum_{i=2}^I m_{ijk\ell}$. Similarly, for the $I \times 2 \times 2 \times 2$, observations are $y^*_{i111} = y_{i111}$, $y^*_{i211} = \sum_{j=2}^J y_{ij11}$, $y^*_{+121} = y_{+121}$, $y^*_{+221} = \sum_{j=2}^J y_{+j21}$, and the ML estimates are $m^*_{i1k\ell} = m_{i1k\ell}$ and $m^*_{i2k\ell} = \sum_{j=2}^J m_{ijk\ell}$.

Using these new ML estimates $m^*_{ijk\ell}$, define $\hat{\beta}^*_{ij} = \hat{m}^*_{ij21}/\hat{m}^*_{ij11}$ for $j = 1, 2$ for the aggregated $I \times 2 \times 2 \times 2$ and $\hat{\gamma}^*_{ij} = \hat{m}^*_{ij12}/\hat{m}^*_{ij11}$ for $i = 1, 2$ for $2 \times J \times 2 \times 2$. Then we have the following results.

Lemma 0.1. *For the above five nonignorable nonresponse models, (i) when $\hat{\beta}_{ij} = \hat{\beta}_{i\cdot}$, then $\hat{\beta}^*_{ij} = \hat{\beta}^*_{i\cdot}$ and $\hat{\beta}^*_{i\cdot} = \hat{\beta}_{i\cdot}$ and (ii) when $\hat{\gamma}_{ij} = \hat{\gamma}_{\cdot j}$, then $\hat{\gamma}^*_{ij} = \hat{\gamma}^*_{\cdot j}$ and $\hat{\gamma}^*_{\cdot j} = \hat{\gamma}_{\cdot j}$.*

This Lemma states that when a nonignirable nonresponse model includes the Y_1R_1 and/or Y_2R_2 terms in its log-linear model, then the respective positiveness of $\hat{\beta}_{i\cdot}$ and $\hat{\gamma}_{\cdot j}$ (i.e., whether or not the ML estimates fall on a boundary solution) can be checked from the aggregated $I \times 2 \times 2 \times 2$ and $2 \times J \times 2 \times 2$ tables, respectively instead of the full $I \times J \times 2 \times 2$ contingency table.

Moreover, this Lemma leads to the below simple conditions for a boundary solution by the aggregated contingency table but not by the original $I \times J \times 2 \times 2$ contingency table. To do this, for the aggregated $I \times 2 \times 2 \times 2$ contingency table, let $v_i = \hat{m}^*_{i111}/\hat{m}^*_{i211}$ for $i = 1, 2, \cdots, I$, and $v = y^*_{+121}/y^*_{+221}$. We also let $v_{max} = max\{v_i\}$ and $v_{min} = min\{v_i\}$. For the aggregate $2 \times J \times 2 \times 2$ contingency table, let $\omega_j = \hat{m}^*_{1j11}/\hat{m}^*_{2j11}$ for $j = 1, 2, \cdots, J$, $\omega = y^*_{1+12}/y^*_{2+12}$, $\omega_{max} = max\{\omega_j\}$ and $\omega_{min} = min\{\omega_j\}$.

Theorem 0.2. *For nonignorable nonresponse models in a $I \times J \times 2 \times 2$ contingency table, If ML esitmates under a log linear link function are not on the boundary of the parameter space, v lies between v_{min} and v_{max} and ω lies between ω_{min} and ω_{max}.*

This theorem states that the ML estimates under the nonignorable nonresponse models including the Y_1R_1 and/or Y_2R_2 terms are on the boundary of the parameter space whenever at least one of $v < v_{min}$, $v > v_{max}$, $\omega < \omega_{min}$, and $\omega > \omega_{max}$ is occurred.

A power transformation method

To overcome boundary solution problems, we propose a power transformation method using our the conditions for boundary solutions. For simple discussion, we assume that $v > v_{max}$ and $\omega > \omega_{max}$ and consider $[Y_1Y_2, Y_1R_1, Y_2R_2, R_1R_2]$ because other nonignorable nonresponse models are nested models of it. Define the following logit link functions: for all $i = 1, 2, \cdots, I$, $j = 1, 2, \cdots, J$, and $k, \ell = 1, 2$ except $i = M$, $j = 1$, and $k = \ell = 1$, and $i = 1$, $j = M$, and $k = \ell = 1$ (the subscript M corresponding to v_{max} and ω_{max} as before),

$$(3) \quad log\left(\frac{\pi_{ijk\ell}}{\pi_{IJ11}}\right) = \eta^i_{Y_1} + \eta^j_{Y_2} + \eta^k_{R_1} + \eta^\ell_{R_2} + \eta^{ik}_{Y_1R_1} + \eta^{j\ell}_{Y_2R_2} + \eta^{k\ell}_{R_1R_2}$$

where π_{IJ11} stands for a reference cell probability. For $i = M$ and $j = 1$, and $i = 1$ and $j = M$,

$$(4) \quad \begin{aligned} \frac{1}{\alpha_{M111}}\left(\left(\frac{\pi_{M111}}{\pi_{IJ11}}\right)^{\alpha_{M111}} - 1\right) &= \eta^M_{Y_1} + \eta^1_{Y_2} + \eta^1_{R_1} + \eta^1_{R_2} + \eta^{M1}_{Y_1R_1} + \eta^{11}_{Y_2R_2} + \eta^{11}_{R_1R_2} \\ \frac{1}{\alpha_{1M11}}\left(\left(\frac{\pi_{1M11}}{\pi_{IJ11}}\right)^{\alpha_{1M11}} - 1\right) &= \eta^1_{Y_1} + \eta^M_{Y_2} + \eta^1_{R_1} + \eta^1_{R_2} + \eta^{11}_{Y_1R_1} + \eta^{M1}_{Y_2R_2} + \eta^{11}_{R_1R_2} \end{aligned}$$

implying that we only use power link functions for m_{M111} and m_{1M11} that correspond to $v_{max} = m_{M111}/m_{M211}$ and $\omega_{max} = m_{1M11}/m_{2M11}$. Further, define

$$k^v = \frac{\exp\left[\frac{1}{\alpha_{M111}}\left(\left(\frac{\pi_{M111}}{\pi_{IJ11}}\right)^{\alpha_{M111}} - 1\right)\right]}{\frac{\pi_{M111}}{\pi_{IJ11}}} \text{ and } k^\omega = \frac{\exp\left[\frac{1}{\alpha_{1M11}}\left(\left(\frac{\pi_{1M11}}{\pi_{IJ11}}\right)^{\alpha_{1M11}} - 1\right)\right]}{\frac{\pi_{1M11}}{\pi_{IJ11}}}.$$

Using these notations, then we have

Theorem 0.3. *Suppose that $v > v_{max}$ and $\omega > \omega_{max}$ in a nonignorable nonresponse model. If α_{M111} and α_{1M11} are chosen to satisfy $k^v > v/v_{max}$ and $k^\omega > \omega/\omega_{max}$, then the ML estimates under the link functions of (3) and (4) satisfy the necessary conditions given in Theorem 0.2.*

This result implies that, when $v > v_{max}$, the first power link function given in (4) produces a new $v^*_{max} = k^v v_{max} > v$ so that $v_{min} < v < v^*_{max}$, and in a symmetric way, when $\omega > \omega_{max}$, the second

power function gives a new $\omega^*_{max} = k^\omega \omega_{max} > \omega$ satisfying $\omega_{min} < \omega < \omega^*_{max}$. Thus, the power link functions defined by (4) move v_{max} and ω_{max} into some values greater than v and ω by mutiplying k^v and k^ω, respectively.

Remark 0.1. *When $v < v_{min}$ and/or $\omega < \omega_{min}$, replace (4) by*

$$\frac{1}{\alpha_{N111}}\left(\left(\frac{\pi_{N111}}{\pi_{IJ11}}\right)^{\alpha_{N111}} - 1\right) = \eta^N_{Y_1} + \eta^1_{Y_2} + \eta^1_{R_1} + \eta^1_{R_2} + \eta^{N1}_{Y_1R_1} + \eta^{11}_{Y_2R_2} + \eta^{11}_{R_1R_2}$$

and

$$\frac{1}{\alpha_{1N11}}\left(\left(\frac{\pi_{1N11}}{\pi_{IJ11}}\right)^{\alpha_{1N11}} - 1\right) = \eta^1_{Y_1} + \eta^N_{Y_2} + \eta^1_{R_1} + \eta^1_{R_2} + \eta^{11}_{Y_1R_1} + \eta^{N1}_{Y_2R_2} + \eta^{11}_{R_1R_2},$$

where the subscript N is the category of Y_1 corresponding to v_{min} and of Y_2 corresponding to ω_{min} and, accordingly, re-define

$$k^v = \frac{\exp\left[\frac{1}{\alpha_{N111}}\left(\left(\frac{\pi_{N111}}{\pi_{IJ11}}\right)^{\alpha_{N111}} - 1\right)\right]}{\frac{\pi_{N111}}{\pi_{IJ11}}} \text{ and } k^\omega = \frac{\exp\left[\frac{1}{\alpha_{1N11}}\left(\left(\frac{\pi_{1N11}}{\pi_{IJ11}}\right)^{\alpha_{1N11}} - 1\right)\right]}{\frac{\pi_{1N11}}{\pi_{IJ11}}}.$$

Then take α_{N111} and α_{1N11} to satisfy $k^v v_{min} > v$ and $k^\omega \omega_{min} > \omega$.

Since the condition given in Theorem 0.2 is necessary for ML not to be on the boundary of the parameter space, the power link functions may not guarantee no boundary solution in likelihood inference. However, in $2 \times 2 \times 2 \times 2$ contingency tables, the condition given in Theorem 0.2 is a necessary and sufficient condition for the ML estimates to be free from a boundary solution as shown below

Corollary 0.4. *For nonignorable nonresponse $2 \times 2 \times 2 \times 2$ contingency tables, the ML estimates are not on the boundary of parameter space if and only if $v_{min} < v < v_{max}$ and $\omega_{min} < \omega < \omega_{max}$.*

Application and simulation

We compare power link functions with logit link functions using a real data and simulated data sets. Bayesian approaches have successfully proved their usefulness to overcome the boundary solution problem. Among others, the Jeffrey prior can be interpreted as a prior allocation of 0.5 observation to each cell and is appeared to have good performance in avoiding the boundary solution problem as we will show below. Thus, we include Bayesian methods with the Jeffrey prior in comparing power and logit link functions.

First, we consider the data for a prospective study of pregnant women to assess the relationship between perinatal factors and subsequent development of abnormalities in the offspring used by Baker *et al* (1992), and Park and Choi (2010). As a result, the ML estimates of expected cells under the logit link function are appeared to be biased because of the boundary solution problem although the logit model is saturated. However, the ML estimates under the power model are perfect fit with observations. The Bayesian estimates under the logit link function reveal imperfect fit with observation; the Bayesian estimates under the power link function are much closer to their corresponding observations than those under the logit link function.

For the simulation study, we generate $2 \times 2 \times 2$ contingency tables for model $[Y_1Y_2, Y_2R_2]$. We compare power link functions with logit link function as the response patterns between respondents and nonrespondents. In Table 1, the smaller M, the more the response patterns are different between respondents and nonrespondents.

The mean squared errors (MSE) are calculated for each missing cell under the logit and power link function. The results are summarized in Table 1. Nonignorable nonresponse models with power link

Table 1: MSEs for imputed cell expectations for model $[Y_1Y_2, Y_2R_2]$

	$M = 0.3$				$M = 0.5$			
	ML		Bayesian		ML		Bayesian	
cell expectation	logit	power	logit	power	logit	power	logit	power
m_{1112}	4134	1550	1570	840	2728	726	743	283
m_{1212}	3452	922	1416	405	2125	302	605	62
m_{2112}	3984	467	1529	67	2684	128	737	39
m_{2212}	4745	930	1710	283	3258	353	846	37
sum	16315	3870	6225	1595	10796	1509	2932	421
	$M = 0.7$				$M = 0.9$			
	ML		Bayesian		ML		Bayesian	
cell expectation	logit	power	logit	power	logit	power	logit	power
m_{1112}	1955	373	416	111	1465	195	241	64
m_{1212}	1439	92	301	23	1038	21	154	75
m_{2112}	1850	34	376	191	1389	63	193	384
m_{2212}	2374	114	462	32	1806	28	243	122
sum	7617	613	1555	356	5699	307	831	646

functions produces smaller MSEs than those with logit link functions both in the ML and Bayesian estimates. In particular, the ML estimates with the power link function are better than the Bayesian estimates with the logit link function. This is generally true for all other nonignorable nonresponse $2 \times 2 \times 2 \times 2$ contingency tables because an appropriately chosen power link function always can avoid the boundary solution problem as shown in Corollary 0.4. The Bayesian estimates under the power function have the smallest MSEs for all cases except $M = 0.9$. Note that all four estimates have smaller MSEs as the response patterns between respondents and nonrespondents becomes alike (i.e., M gets larger).

Data analysis and simulation study shows the power link functions performs betters the logit link functions in both ML estimation and Bayesian estimation.

References

[1] Agresti, A. (2002). *Categorical Data Analysis*. 2^{nd} Edition. New York: John Wiley & Sons, Inc.

[2] Baker, S. G., and Laird, N. M. (1988). Regression analysis for categorical vaiables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

[3] Baker, S. G., Rosenberger, W. F. and Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11, 643-657.

[4] Chen, T. (1972). Mixed-up frequencies and missing data in contingency tables. Unpublished Ph.D. dissertation, University of Chicago, Dept. of Statistics.

[5] Choi, B. S., Choi, J. W. and Park, Y. (2009). Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye State) Polls. *Survey Methodology*, 35, 37-51.

[6] Clarke, P. S. (2002). On boundary solutions and identifiability in categorical regression with non-ignorable non-response. *Biometrical Journal*, 44, 701-717.

[7] Clogg, C. C., Rubin, D. B., Schenker, N. and Schultz, B. (1991). Multiple imputation of industry and occupation codes in Census Public use-samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

[8] Dempster, A. P., Laird, N. M. and Rubin, D. M. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Sereis B*, 39, 1-38.

[9] Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.

[10] Gelman, A., Carlin, J. P., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*. 2^{nd} Edition. New York: Chapman and Hall/CRC.

[11] Little, J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data*. 2^{nd} Edition. New York: John Willey & Sons, Inc.

[12] Park, T., and Brown, M. B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, 89, 44-52.

[13] Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, 54, 1579-1690.

[14] Park, Y., and Choi, B. S. (2010). Bayesian analysis for incomplete multi-way contingency tables with nonignorable nonresponse. *Journal of Applied Statistics*, 37, 1439-1453.

[15] Smith, P. W. F., Skinner, C. J. and Clarke, P.S. (1999). Allowing for non-ignorable nonresponse in the analysis of voting intention data. *Applied Statistics*, 48, 563-577.