Note on "Obtaining the Maximum Likelihood Estimates in Incomplete $R \times C$ Contingency Tables..."

Michael Friendly, York University

Abstract

This note extends the construction of the design matrix used for estimating cell probabilities with ignorable missing data described by Lipsitz et al. (1998). A reformulation for the general case of an n-way table is described, and implemented in a SAS macro program. The macro constructs this design matrix and offset variable, estimates the cell probabilities, and returns a table with the estimates, their standard errors, and fitted cell frequencies.
Key words: categorical data; contingency table; missing data; generalized linear model; SAS macro.

1 Introduction

Lipsitz, Parzen and Molenberghs (1998, henceforth, LPM) describe a simple method for obtaining estimates of cell frequencies in a two-way contingency table with ignorable missing data (missing completely at random, or missing at random) on the row and column variables. They show that the cell probabilities may be estimated as a Poisson generalized linear model (GLM), with a structured design matrix and an offset containing various marginal totals. This approach may be carried out using standard statistical software for GLMs, such as `PROC GENMOD` in SAS, or `glm` in S-Plus, rather than requiring special purpose software (e.g., Duffy (1994), Espeland and Odoroff (1985), Vermunt (1997)).

The use of standard software relies on an equivalence between the multinomial, missing data likelihood and the Poisson likelihood of a GLM using a specially constructed design matrix, an offset variable, and identity link. Although the construction of these model matrices is conceptually simple, carrying out the construction is tedious and error-prone, particularly for larger tables than the 2×2 and 3×3 examples presented by LPM[1]. It would be useful, therefore, to have software to perform this construction automatically. Moreover, while LPM illustrate how their method can be extended to multiway tables with a $2 \times 2 \times 2$ example, they do not present the extension in general terms.

Here, we present a general representation for the structure of the required design matrix for maximum likelihood estimation of cell probabilities in multiway tables with ignorable missing data. This representation allows the construction for the general case to be programmed in SAS/IML (or any other matrix language). We describe a general-purpose SAS macro program, `MISSRC`, which takes only the contingency table as input, constructs the design matrix and offset variable, and estimates the cell probabilities using `PROC GENMOD`; it returns an output table with the estimates, their standard errors, and fitted cell frequencies. Similar programming could be used with other statistical packages capable of matrix operations.

[1]For example, LPM's equation (5.1) for their $2 \times 2 \times 2$ example contains an extraneous row of zeros in the design matrix.

	Y_2			
Y_1	0	1	2	3
0		z_{+1}	z_{+2}	z_{+3}
1	w_{1+}	u_{11}	u_{12}	u_{13}
2	w_{2+}	u_{21}	u_{22}	u_{23}
3	w_{3+}	u_{31}	u_{32}	u_{33}

Table 1: A 3×3 table with missing data

2 Design matrix for the Poisson linear model

2.1 Notation

For brevity, we introduce only the minimum context and notation (telegraphically, and slightly modified) from LPM to proceed to a general reformulation for the n-way case. We use $1 : t$ as shorthand for the sequence $1, 2, \ldots, t$, and $z_1 : z_t$ for $z_1, z_2, \ldots, z_t$. Let $Y_{i1} : Y_{in}$ refer to n discrete variables observed on subject $i, i = 1 : N$, where the j-th variable, Y_{ij}, can take on values $1 : J_j$, or be missing, which we represent by the value $J_j = 0$. Indicator variables, $R_{i1} : R_{in}$, are defined as $R_{ij} \equiv Y_{ij} > 0$, so that $R_{ij} = 1$ if Y_{ij} is observed and is 0 if Y_{ij} is missing. In the absence of missing data, the contingency table would be an array of dimension $J_1 \times J_2 \times \cdots \times J_n$. Allowing missing data on all variables data, the contingency table is of size $(J_1 + 1) \times (J_2 + 1) \times \cdots \times (J_n + 1)$, except that the cell where *all* variables are missing cannot occur, because such subjects would not be observed at all in the study.

In application, the data is summarized by its contingency table, shown using LPM's notation for the 3×3 case in Table 1, where the counts of the complete cases are denoted u_{jk}, the counts of cases observed only on Y_1 are denoted w_{j+}, and the counts of cases observed only on Y_2 are denoted z_{+k}.

For a two-way table, let $\mathbf{p} = \{p_{jk}\}$ be the $(J_1 J_2 - 1) \times 1$ vector of non-redundant multinomial cell probabilities (excluding $p_{J_1 J_2}$), and let $\mathbf{u} = \{u_{jk}\}$ be a $(J_1 J_2 \times 1)$ vector of non-missing counts. Similarly, let $\mathbf{w} = \{w_{j+}\}$ and $\mathbf{z} = \{z_{+k}\}$ be the $(J_1 \times 1)$ and $(J_2 \times 1)$ vectors of counts observed on only Y_1 and Y_2, respectively. Finally, stack the vectors of counts to form $\mathbf{f} = [\mathbf{u} \ / \ \mathbf{w} \ / \ \mathbf{z}]$, where the / operator means vertical concatenation.

LPM demonstrate that the MLEs of the cell probabilities may be estimated from a Poisson linear model for $\mathbf{f}$ of the form

$$\mathcal{E}(\mathbf{f}) = \mathbf{X}\mathbf{p} + \gamma \tag{1}$$

where $\mathbf{X}$ is the specially constructed design matrix and γ is an offset containing the totals u_{++}, w_{++}, z_{++}, required to impose the multinomial restrictions on the estimated cell probabilities. In their Appendix, LPM give, for two-way tables, the general form of the terms in (1) in partitioned form, with explicit formulas for the submatrices shown below,

$$\mathcal{E} \begin{pmatrix} \mathbf{f}_u \\ \mathbf{f}_w \\ \mathbf{f}_z \end{pmatrix} = \begin{bmatrix} \mathbf{X}_u \\ \mathbf{X}_w \\ \mathbf{X}_z \end{bmatrix} \mathbf{p} + \begin{pmatrix} \gamma_u \\ \gamma_w \\ \gamma_z \end{pmatrix} \tag{2}$$

Those expressions, however, do not shed any light on the structure of the problem, nor do they permit easy generalization to three-way and larger tables. We show below that the structure of $\mathbf{X}$ and γ may be derived by analogy with design matrices for factorial linear models.

Type	Y_1	Y_1	**count**
3	1	1	u_{11}
	1	2	u_{12}
	1	3	u_{13}
	2	1	u_{21}
	2	2	u_{22}
	2	3	u_{23}
	3	1	u_{31}
	3	2	u_{32}
	3	3	u_{33}
2	1	0	w_{1+}
	2	0	w_{2+}
	3	0	w_{3+}
1	0	1	z_{+1}
	0	2	z_{+2}
	0	3	z_{+3}

Table 2: A 3×3 table in profile form

2.2 Reformulation

To facilitate exposition, restructure the contingency table in profile form with the variables $Y_1, \ldots, Y_n$ given explicitly, as shown in Table 2. For an n-way table there are $2^n - 1$ distinct *types* of counts, corresponding to the the possible patterns of the missingness indicators $R_1 R_2 \cdots R_n$. If all n table variables are observed for a subject, then all $R_j = 1$. We write the counts for the subjects observed on all n variables as the $(J_1 J_2 \cdots J_n \times 1)$ vector $\mathbf{f}_{11\ldots1}$, with elements corresponding to the usual contingency table (with no missing data) for $(Y_1, Y_2, \ldots, Y_n)$. For the 3×3 example shown in Table 2,

$$\mathbf{f}_{11} = (u_{11}, u_{12}, u_{13}, u_{21}, u_{22}, u_{23}, u_{31}, u_{32}, u_{33})' \quad .$$

In general, we can write the counts associated with pattern $r_1 r_2 \cdots r_n$ as $\mathbf{f}_{r_1 r_2 \cdots r_n}$, where, if $R_j = 0$, we replace the subscript for Y_j in the counts with a '+'. For example, if $J_1 = J_2 = 3$ and $R_2 = 0$, then

$$\mathbf{f}_{10} = (w_{1+}, w_{2+}, w_{3+})' \quad .$$

For computational purposes, the $2^n - 1$ types of counts may be indexed by the decimal equivalent of $r_1 r_2 \cdots r_n$; for example, $\text{type}(\mathbf{f}_{11}) = 3$.

These types have an obvious analogous relation to the terms in a complete factorial design with no intercept whose model formula (in `glm` notation) is:

$$\begin{aligned} Y_1 * Y_2 * \cdots * Y_n - 1 \quad &= \quad Y_1 + Y_2 + \cdots Y_n + Y_{12} + Y_{13} + \cdots + Y_{(n-1)n} \\ &\quad + Y_{123} + Y_{124} + \cdots Y_{(n-2)(n-1)n} + \cdots \ \cdots + Y_{12\ldots n} \end{aligned}$$

where terms with one subscript represent variables observed only once, and correspond to main efects in a factorial design, terms with two subscripts represent non-missing on two variables, and correspond

to two-way terms in a factorial design, and so forth, up to the last term which represents the totally complete cases, and is analogous to the n-way interaction in the factorial design.

As shown in the Appendix, in the general case, we can write the (less-than-full-rank) model for $\mathcal{E}(\mathbf{f}_{r_1 r_2 \cdots r_n})$ as

$$\mathcal{E}(\mathbf{f}_{r_1 r_2 \cdots r_n}) = \mathbf{X}_{r_1 r_2 \cdots r_n} \mathbf{p} \quad ,$$

where

$$\mathbf{X}_{r_1 r_2 \cdots r_n} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_n \ , \tag{3}$$

and

$$\mathbf{A}_j = \begin{cases} \mathbf{I}_{J_j} & \text{if } r_j = 1 \\ \mathbf{1}'_{J_j} & \text{otherwise} \end{cases} \quad .$$

An equivalent full-rank model is obtained by substituting for the identity matrix any choice of contrast matrix $\mathbf{C}$, such that $\mathbf{C}'\,\mathbf{1} = \mathbf{0}$; but the parameter estimates depend on that choice. In the incomplete contingency table problem, LPM construct the design matrix so that the model (1) will have a Poisson likelihood equivalent to the multinomial likelihood. To achieve this, they force the total of the expected counts under the model in each block to equal the corresponding total of the observed counts, $g_{r_1:r_n} \equiv \sum_{r_j=1} f_{r_1 r_2 \cdots r_n}$, in that block, for which a necessary condition is $p_{++\cdots+} = 1$.

For each block, let $\mathbf{X}^-_{r_1:r_n}$ denote $\mathbf{X}_{r_1 r_2 \cdots r_n}$ dropping the last row and column (removing $p_{J_1:J_n}$ from the parameter vector $\mathbf{p}$), and add an equation to equate the expected total count in that block to the observed total. The final model, equivalent to (1), may be expressed in terms of a block for each type of the form

$$\mathcal{E}(\mathbf{f}_{r_1 r_2 \cdots r_n}) = g_{r_1:r_n} \begin{bmatrix} \mathbf{X}^-_{r_1:r_n} \\ -h(\mathbf{X}^-_{r_1:r_n}) \end{bmatrix} \mathbf{p} + \begin{pmatrix} \mathbf{0} \\ g_{r_1:r_n} \end{pmatrix} \ , \tag{4}$$

where $h(\mathbf{X}^-_{r_1:r_n})$ evaluates to a row vector with a 1 in each column where $\mathbf{X}^-$ has a 1 in any row and 0 otherwise, and the last term is the offset. For example, in a $2 \times 2 \times 2$ table, consider frequencies of type $3 = 011_2$, corresponding to non-missing observations on Y_2 and Y_3. Application of (3) and (4) give

$$\mathcal{E}(\mathbf{f}_{011}) = g_{011} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ -1 & -1 & -1 & -1 & -1 & -1 & 0 \end{bmatrix} \mathbf{p} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ g_{011} \end{pmatrix} .$$

3 The `MISSRC` Macro

The above scheme for generating the design matrix and offset vector is easily programmed in any matrix-oriented language. The SAS/IML product is a fully programmable matrix language, which, combined with the SAS macro facility, allows flexible, general procedures to be written.

The `MISSRC` macro takes as input a multiway contingency table in the profile form of Table 2 (only the Y and count variables are required). From this, it constructs the design matrix and offset variable, estimates the cell probabilities using `PROC GENMOD`, and returns a table with the estimates, their standard errors, and fitted cell frequencies. The program documentation and source code is available at the web address `http://www.math.yorku.ca/SCS/sasmac/missrc.html`. Two examples illustrate its use.

3.1 Example 1

Consider the 2×2 example from Little and Rubin (1987, p. 183) presented in LPM's Table 3. The lines below create a SAS data set with R and C as the table variables and frequency variable is named COUNT Missing values for R and C appear as '.'. The data set is in exactly the same format as shown in LPM's Table 6, or our Table 2 (with 0s for Y_1 and Y_2 replaced by '.').

```
data little;
input R C count @@;
cards;
1 1 100    1 2 50    1 . 30
2 1  75    2 2 75    2 . 60
. 1  28    . 2 60
;
```

The MISSRC macro is called with keyword arguments, with default values for all except VAR, which specifies the table variables.

```
%missrc(data=little, var=R C);
```

The macro call creates the DESIGN data set containing the design matrix, offset and frequency variables, as follows. Parameter names are constructed automatically from the levels of R and C.

OBS	COUNT	P11	P12	P21	OFFSET
1	100	300	0	0	0
2	50	0	300	0	0
3	75	0	0	300	0
4	75	-300	-300	-300	300
5	30	90	90	0	0
6	60	-90	-90	0	90
7	28	88	0	88	0
8	60	-88	0	-88	88

The following output data set is produced after the model (1) is fit using PROC GENMOD. Variable P is the observed cell probability for the complete-case data, ESTIMATE is the MLE, $\hat{p}_{jk}$, and FITTED is the estimated cell frequency.

R	C	COUNT	P	PARM	ESTIMATE	STDERR	FITTED
.	1	28	.		.	.	.
.	2	60	.		.	.	.
1	.	30	.		.	.	.
1	1	100	0.33333	P11	0.27947	0.022310	133.589
1	2	50	0.16667	P12	0.17402	0.020978	83.184
2	.	60	.		.	.	.
2	1	75	0.25000	P21	0.23872	0.022660	114.108
2	2	75	0.25000	P22*	0.30778	0.025298	147.120

The *-ed parameter is found as $\hat{p}_{JK} = 1 - \sum_{jk \neq JK} \hat{p}_{jk}$. Its standard error is found as $\{\mathbf{1}' \, \widehat{\mathbf{V}} \, \mathbf{1}\}^{1/2}$, where $\widehat{\mathbf{V}}$ is the estimated variance-covariance matrix of the non-redundant parameters.

3.2 Example 2

The three-way data used by LPM (Table 6) concern a longitudinal study of coronary risk factors in 4,858 school children, in which the binary response, obesity, was assessed in three years. As LPM note, the cases with complete data comprise only 1,770 (36.3%) of the sample, so the gain in estimation of the joint probabilities using the partial cases may be considerable.

```
data obese;
   input o77 o79 o81 count @@;
   label o77='Obese in 77'  o79='Obese in 79'  o81='Obese in 81';
cards;
0 0 0 1209     0 0 1  91     0 1 0  66     0 1 1   78
1 0 0   64     1 0 1  31     1 1 0  62     1 1 1  169
0 0 .  426     0 1 .  54     1 0 .  33     1 1 .  118
0 . 0  125     0 . 1  27     1 . 0   5     1 . 1   27
. 0 0  463     . 0 1  63     . 1 0  37     . 1 1   62
0 . .  583     1 . . 173
. 0 .  293     . 1 .  77
. . 0  381     . . 1 119
;
%missrc(data=obese, var=o77 o79 o81);
```

The output from `MISSRC` includes the `PROC GENMOD` results and other information not shown here. The final table of parameter estimates is shown below. We see that there were 169 children in the complete-data sample who were assessed as obese on all three occasions, likely the group of highest risk. The estimate for the total sample is over 3 times as great.

O77	O79	O81	COUNT	P	PARM	ESTIMATE	STDERR	FITTED
.	.	0	381	.		.	.	.
.	.	1	119	.		.	.	.
.	0	.	293	.		.	.	.
.	0	0	463	.		.	.	.
.	0	1	63	.		.	.	.
.	1	.	77	.		.	.	.
.	1	0	37	.		.	.	.
.	1	1	82	.		.	.	.
0	.	.	583	.		.	.	.
0	.	0	125	.		.	.	.
0	.	1	27	.		.	.	.
0	0	.	426	.		.	.	.
0	0	0	1209	0.68305	P000	0.66332	.0078223	3221.07
0	0	1	91	0.05141	P001	0.05778	.0048275	280.58
0	1	.	54	.		.	.	.
0	1	0	66	0.03729	P010	0.03480	.0037399	168.99
0	1	1	78	0.04407	P011	0.04394	.0041767	213.37
1	.	.	173	.		.	.	.
1	.	0	5	.		.	.	.

```
 1    .    1      27    .                    .          .               .
 1    0    .      33    .                    .          .               .
 1    0    0      64    0.03616    P100      0.03555    .0038920      172.65
 1    0    1      31    0.01751    P101      0.02068    .0032591      100.40
 1    1    .     118    .                    .          .               .
 1    1    0      62    0.03503    P110      0.03571    .0039286      173.43
 1    1    1     169    0.09548    P111*     0.10822    .0055613      525.51
```

A final advantage of this general formulation is that it does not require that *all* factorial combinations of missing and non-missing data actually occur. In the limiting case, when all data is complete, the estimates are just the observed sample proportions, as they should be. For example, we can restrict analysis to the complete-case data as follows:

```
data complete;
   set obese;
   where (o77^=. & o79^=. & o81^=.);

%missrc(data=complete, var=o77 o79 o81);
```

This gives the output below. Comparing the standard errors of p_{111}, we see that the relative efficiency of the missing data approach is 126% compared to the restricted complete-case sample. This is equivalent to the gain of an additional 460 subjects for the estimation of this probability.

```
O77  O79  O81   COUNT     P       PARM    ESTIMATE    STDERR    FITTED

 0    0    0    1209   0.68305   P000     0.68305   0.011059    1209
 0    0    1      91   0.05141   P001     0.05141   0.005249      91
 0    1    0      66   0.03729   P010     0.03729   0.004503      66
 0    1    1      78   0.04407   P011     0.04407   0.004879      78
 1    0    0      64   0.03616   P100     0.03616   0.004437      64
 1    0    1      31   0.01751   P101     0.01751   0.003118      31
 1    1    0      62   0.03503   P110     0.03503   0.004370      62
 1    1    1     169   0.09548   P111*    0.09548   0.006985     169
```

Appendix

It is well-known (e.g., Bock (1975, Section 5.3.2),Kurkjian and Zelen (1962)) that the model matrix for any fully-crossed factorial design may be generated as the n-fold Kronecker product of the one-way design matrices. Thus, for a linear model $\mathcal{E}(\mathbf{y}) = \mathbf{X}\beta$, with n crossed factors, $F_1, F_2, \ldots, F_n$, the less-than-full-rank model matrix may be constructed column-wise as

$$\mathbf{X}_{F_1 F_2 \cdots F_n} = [\mathbf{1}_{J_1} \,|\, \mathbf{I}_{J_1}] \otimes \; [\mathbf{1}_{J_2} \,|\, \mathbf{I}_{J_2}] \; \otimes \cdots \otimes \; [\mathbf{1}_{J_n} \,|\, \mathbf{I}_{J_n}] \,, \tag{5}$$

and the model is made estimable by placing restricitions on the unknown parameters, either by adding additional equations (rows of $\mathbf{X}$ and $\mathbf{y}$) or by reparameterization so that estimability is ensured (replacing $\mathbf{I}_J$ by a $J \times (J-1)$ conrast matrix $\mathbf{C}$ such that $\mathbf{C}'\,\mathbf{1} = \mathbf{0}$). The expansion of (5) gives, for a three-way design,

$$\begin{aligned} \mathbf{X}_{F_1 F_2 F_3} &= [\mathbf{1}_{J_1} \otimes \mathbf{1}_{J_2} \otimes \mathbf{1}_{J_3} \,|\, \mathbf{I}_{J_1} \otimes \mathbf{1}_{J_2} \otimes \mathbf{1}_{J_3} \,|\, \mathbf{1}_{J_1} \otimes \mathbf{I}_{J_2} \otimes \mathbf{1}_{J_3} \,|\, \mathbf{1}_{J_1} \otimes \mathbf{1}_{J_2} \otimes \mathbf{I}_{J_3} \,| \\ &\quad \mathbf{I}_{J_1} \otimes \mathbf{I}_{J_2} \otimes \mathbf{1}_{J_3} \,|\, \mathbf{I}_{J_1} \otimes \mathbf{1}_{J_2} \otimes \mathbf{I}_{J_3} \,|\, \mathbf{1}_{J_1} \otimes \mathbf{I}_{J_2} \otimes \mathbf{I}_{J_3} \,|\, \mathbf{I}_{J_1} \otimes \mathbf{I}_{J_2} \otimes \mathbf{I}_{J_3} \,] \\ &= [\mathbf{X}_{000} \,|\, \mathbf{X}_{100} \,|\, \mathbf{X}_{010} \,|\, \mathbf{X}_{001} \,|\, \mathbf{X}_{110} \,|\, \mathbf{X}_{101} \,|\, \mathbf{X}_{011} \,|\, \mathbf{X}_{111}] \,, \end{aligned}$$

where, in the last line the subscripts are symbolic basis indicators, with 1 and 0 representing the presence and absence, respectively, of the corresponding factor in each effect.

The present problem is analogous to the transpose of this situation, where the *rows* of the design matrix reflect the factorial structure of missing and non-missing observations on $Y_1, \ldots, Y_n$, which play the role of "factors". For example, in a 3-way table the design matrix, ignoring estimability constraints for the moment, arises from vertically stacking the design blocks corresponding to the 7 combinations of missing/non-missing on Y_1, Y_2, Y_3

$$\mathbf{X} = [\mathbf{X}_{111} \ / \ \mathbf{X}_{110} \ / \ \mathbf{X}_{101} \ / \ \mathbf{X}_{100} \ / \ \mathbf{X}_{011} \ / \ \mathbf{X}_{010} \ / \ \mathbf{X}_{001}] \ . \tag{6}$$

In the general case, this gives a model $\mathcal{E}(\mathbf{f}) = \mathbf{X}\mathbf{p}$, where each block in $\mathbf{X}$ is a matrix of $\prod_{j=1}^{n} J_j$ columns of the form

$$\mathbf{X}_{r_1 r_2 \cdots r_n} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_n \ ,$$

composed as

$$\mathbf{A}_j = \begin{cases} \mathbf{I}_{J_j} & \text{if } r_j = 1 \\ \mathbf{1}'_{J_j} & \text{otherwise} \end{cases} \ .$$

Acknowledgments

This work is supported by Grant 8150 from the National Sciences and Engineering Research Council of Canada. I am grateful to the Associate Editor and reviewer for helpful suggestions.

References

Bock, R. D. *Multivariate Statistical Methods in Behavioral Research.* McGraw Hill, New York, 1975.

Duffy, D. L. Loglin: A program for loglinear analysis of complete and incomplete count data. WWW document, 1994. `http://www.qimr.edu.au/davidD/loglin.html`.

Espeland, M. A. and Odoroff, C. L. Log-linear models for doubly sampled categorical data fitted by the EM algorithm. *JASA*, pages 663–670, 1985.

Kurkjian, B. and Zelen, M. A calculus for factorial arrangements. *Annals of Mathematical Statistics*, 33:600–619, 1962.

Lipsitz, S. R., Parzen, M., and Molenberghs, G. Obtaining the maximum likelihood estimates in incomplete contingency tables using a Poisson generalized linear model. *Journal of Computational and Statistical Graphics*, 7:356–376, 1998.

Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 1987.

Vermunt, J. K. *Log-linear Models for Event Histories.* Sage Publications, Thousand Oaks, CA, 1997.