WP. 41
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

ON THE IMPUTATION OF CATEGORICAL DATA SUBJECT TO EDIT RESTRICTION USING LOGLINEAR MODELS

Supporting Paper

Submitted by Statistics Netherlands[1]

I. INTRODUCTION

1. This paper describes a model-based approach for imputation of categorical variables under edit constraints. The models considered belong to the flexible and widely used (for categorical data) class of loglinear models. The edit constraints are equivalent to the constraint that some value combinations must be zero (e.g. married="yes" and age class= "0-10 year"). Such constraints are also known as structural zeros in the contingency table formed by all possible combinations of categories of all variables involved. The general approach is to estimate a loglinear model for the contingency table with structural zeros and then use the model based estimated cell probabilities (for the non-structural zeros) to impute for missing values.

2. Loglinear models usually do not consider more than 5 to 10 variables at a time, depending on the number of categories per variable. When an imputation model is build for the simultaneous imputation of all categorical variables in a social survey, the number of variables can greatly exceed these numbers. Consequently, the dimensionality of the contingency tables involved and the number of cells in these tables can become very large. The problem of imputation under edit constraints using loglinear models has already been described by Winkler (2003) who gives an example of a labour force survey where the number of cells is approaching 10^{46}. The currently known algorithms are infeasible for such high dimensions due to memory restrictions and time considerations.

3. Another problem originates in the relative small number of respondents for a survey compared to the enormous space of the complete contingency table. Individual cell probabilities cannot (reliably) be estimated because the number of observations in most cells will be small and often zero. To deal with this problem we will consider only relatively simple loglinear models, i.e. models that do not include interactions between, say, three or more variables.

4. In Section II the constrained loglinear model is introduced. Section III shows how to estimate the parameters of this model for contingency tables of moderate size, based on the fully observed records (the complete cases). In Section IV an estimation method for high dimensional tables is proposed. Section V

---

[1] Prepared by Frank Van den Eijkhof (fekf@cbs.nl), Ton de Waal and Jeroen Pannekoek. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

describes how to use the estimated model for imputation, and Section VI concludes the paper with some directions for extensions and future research.

II. CONSTRAINED LOGLINEAR MODELS

5. For notational simplicity we assume in Sections 2 and 3 that we are dealing with a two-dimensional contingency table. The spanning variables of the contingency table are X and Y, with I respectively J categories. Our approach can straightforwardly be extended to higher-dimensional tables. We assume that the data are generated by a multinomial probability process. This means that we have a sample of fixed size n and the expected number of observations in the cell corresponding to category i of variable X and category j of variable Y is

$$m_{ij} = np_{ij}, \tag{1}$$

with p_{ij} the probability of an observation in cell (i,j). Since $\sum_{ij} p_{ij} = 1$, we have for the total of the expected counts $\sum_{ij} m_{ij} = m_{++} = n$.

6. The standard loglinear model for the expected counts can be written as

$$\log(m_{ij}) = \mathrm{M}, \tag{2}$$

where M denotes a specific linear model (Agresti, 1990). A simple example is the so-called independence model. In this model the expected cell counts are given by

$$\log(m_{ij}) = m + l_i^X + l_j^Y. \tag{3}$$

The model can be written in multiplicative form as

$$m_{ij} = ab_i^X b_j^Y, \tag{4}$$

with $a = \exp(m)$, $b_i^X = \exp(l_i^X)$ and $b_j^Y = \exp(l_j^Y)$.

7. Without constraints the parameters in models (3) and (4) are not identifiable. The parameters in model (3) are identifiable up to an additive constant. Adding a constant g to each l_i^X and subtracting the same constant from each l_j^Y or from m yields an equivalent model (with the same values for m_{ij}) but with different parameter values. Similarly, the parameters in model (4) are identifiable up to a multiplicative constant. Multiplying each b_i^X by a constant c and dividing each b_j^Y or a by the same constant yields an equivalent model. To render the model identifiable, different constraints can be used. Most statistical software packages use $l_I^X = l_J^Y = 0$ implying that $m = \log(m_{IJ})$, $b_I^X = b_J^Y = 1$ and $a = m_{IJ}$.

8. The fact that the parameters in (2) and (4) are not uniquely determined allows redefining the models as

$$\log(m_{ij}) = l_i^X + l_j^Y \tag{5}$$

$$m_{ij} = b_i^X b_j^Y \tag{6}$$

where μ is set to zero and a constant is added to each l_i^X and another constant is added to each l_j^Y such that these two constants add up to μ. To make the parameters of (5) and (6) identifiable, we still need one more constraint, for instance setting one of the parameters of (5) equal to zero and one of the parameters of (6) equal to 1.

9. In a constrained loglinear model, some of the cell counts equal zero. We subdivide the cells in the contingency table into two classes: the cells for which the count has to be equal to zero (the structural zeros) and the other cells. The former set of cells is denoted by **Z**, the latter set of cells by **U**.

10. For a constrained version of the loglinear model (5)-(6), the expected cell counts are given by

$$m_{ij} = np_{ij} = 0 \qquad \text{for } (i, j) \in \mathbf{Z},$$

$$m_{ij} = np_{ij} = \exp(l_i^X + l_j^Y) = b_i^X b_j^Y \qquad \text{for } (i, j) \in \mathbf{U},$$

or

$$m_{ij} = d_{ij} b_i^X b_j^Y, \tag{7}$$

with $d_{ij} = 0$ for $(i, j) \in \mathbf{Z}$ and $d_{ij} = 1$ for $(i, j) \in \mathbf{U}$ and $\sum_{ij} d_{ij} p_{ij} = 1$ or equivalently $\sum_{ij} d_{ij} m_{ij} = m_{++} = n$. This model is called a quasi-independence model (see e.g. Bishop et al., 1975, Ch. 5, Goodman, 1968).

III. PARAMETER ESTIMATION IN MODERATE SIZE TABLES USING THE COMPLETE CASES

11. The parameters in a constrained loglinear model can be estimated by maximizing the multinomial likelihood or, equivalently, loglikelihood. The loglikelihood for the quasi-independence model is given by

$$\mathrm{L}(l_i^X, l_j^Y) = \sum_{ij} d_{ij} n_{ij} \log m_{ij} = \sum_{ij} d_{ij} n_{ij} (l_i^X + l_j^Y), \tag{8}$$

where we have omitted terms not involving the parameters since these are irrelevant for the estimation problem. Without these terms, (8) is called the *kernel* of the loglikelihood. The parameter estimates can be obtained by maximizing (8) under the constraint $\sum_{ij} d_{ij} m_{ij} = n$. To impose this constraint we introduce the Lagrangean

$$\mathrm{L}_0(l_i^X, l_j^Y) = \sum_{ij} d_{ij} n_{ij} (l_i^X + l_j^Y) - a(\sum_{ij} d_{ij} \exp(l_i^X + l_j^Y) - n). \tag{9}$$

12. By setting the derivatives of (9) with respect to the Lagrange multiplier a and the parameters equal to zero we obtain the likelihood equations

$$\partial L_0 / \partial a = \sum_{ij} d_{ij} m_{ij} - n = 0 \tag{10}$$

$$\partial L_0 / \partial l_i^X = n_{i+} - a \sum_j d_{ij} m_{ij} = 0 \tag{11}$$

$$\partial L_0 / \partial l_j^Y = n_{+j} - a \sum_i d_{ij} m_{ij} = 0 \tag{12}$$

By summing (11) over i (or summing (12) over j) and using (10) we see that a equals 1.

13. The likelihood equations (11) and (12) are similar to those for loglinear models for tables without structural zeros in that they equate certain margins of the estimated expected frequencies to the corresponding observed margins. The difference is that for the constrained table, the summations are over the cells in U only.

14. For the estimation of the parameters in quasi-loglinear models, several algorithms have been proposed in the literature. Two of these procedures are described below. The target of the first algorithm is the solution of the equations (11) and (12) in terms of estimated expected frequencies $\hat{m}_{ij}$. The maximum likelihood estimates of the parameters can be obtained from these expected frequencies afterwards, if so desired. The second algorithm directly produces the maximum likelihood estimates of the parameters (from which the estimated expected frequencies can be obtained).

15. The first algorithm is the iterative proportional fitting (ipf) algorithm. The algorithm begins with starting values $\hat{m}_{ij}^{(0)}$ for the estimated expected frequencies. Convenient starting values in the presence of structural zeros are: $\hat{m}_{ij}^{(0)} = d_{ij}$. These starting values are then updated according to

$$\hat{m}_{ij}^{(t-1)} = \hat{m}_{ij}^{(t-2)} \frac{n_{i+}}{\sum_j d_{ij} \hat{m}_{ij}^{(t-2)}} \qquad \forall\, i \tag{13}$$

$$\hat{m}_{ij}^{(t)} = \hat{m}_{ij}^{(t-1)} \frac{n_{+j}}{\sum_i d_{ij} \hat{m}_{ij}^{(t-1)}} \qquad \forall\, j \tag{14}$$

Thus, we multiply the current estimated expected frequencies alternating by factors depending only on *i* and factors depending only on *j*. This sequence can be expressed as

$$\hat{m}_{ij}^{(t)} = d_{ij}\, f_i^{(1)} f_j^{(2)} f_i^{(3)} f_j^{(4)} \cdots f_i^{(t-1)} f_j^{(t)} . \tag{15}$$

Note that by choosing the d_{ij} as starting values, the estimated frequencies for the structural zero cells remain zero throughout the iterative process. If we accumulate the factors f_i and f_j we obtain, after convergence, the estimated multiplicative model

$$\hat{m}_{ij} = d_{ij} \hat{b}_i^X \hat{b}_j^Y .$$

16. Expression (15) also suggests a more direct way of obtaining the parameter estimates, without explicitly recalculating the expected frequencies at each step. At some iteration *t*, we can rewrite (15) as

$$\hat{m}_{ij}^{(t)} = b_i^{X(t-1)} b_j^{Y(t)} = d_{ij}\, f_i^1 f_i^3 \cdots f_i^{(t-1)} f_j^2 f_j^4 \cdots f_j^{(t)} = b_i^{X(t-3)} b_j^{Y(t-2)} f_i^{(t-1)} f_j^{(t)}$$

and so,

$$b_i^{X(t-1)} = b_i^{X(t-3)} f_i^{(t-1)} = b_i^{X(t-3)} \frac{n_{i+}}{\sum_j d_{ij} b_i^{X(t-3)} b_j^{Y(t-2)}} = \frac{n_{i+}}{\sum_j d_{ij} b_j^{Y(t-2)}} , \tag{16}$$

and similarly,

$$b_j^{Y(t)} = \frac{n_{+j}}{\sum_i d_{ij} b_i^{X(t-1)}} \tag{17}$$

The iteration defined by (16)-(17) appears in Goodman (1968).

17. As mentioned before, the parameter estimates $\hat{b}_i^X$ and $\hat{b}_j^Y$ are determined only up to a multiplicative constant. To uniquely identify these parameters we can multiply the parameters $\hat{b}_j^Y$ by $1/\hat{b}_J^Y$ and the parameters $\hat{b}_i^X$ by $\hat{b}_J^Y$. In effect, we then have imposed the additional constraint $\hat{b}_J^Y = 1$.

IV. PARAMETER ESTIMATION FOR LARGE TABLES

18. The model proposed in Section 4 contains only two variables. For a larger set of variables, for instance variables v_1, v_2, v_3, v_4 with indices *i, j, k, l,* formula (16) extends to

$$b_i^{v_1(t)} = \frac{n_{i+++}}{\sum_{jkl} d_{ijkl} b_j^{v_2(t-1)} b_k^{v_3(t-1)} b_l^{v_4(t-1)}} \tag{18}$$

which updates the parameter for category *i* of variable v_1 in some iteration *t*, using the current estimates of the other parameters $b_j^{v_2(t-1)}$, $b_k^{v_3(t-1)}$ and $b_l^{v_4(t-1)}$. The denominator of (18) shows the problem of

dimensionality. The summation is over all combinations of categories of all variables except for variable v_1. For large numbers of variables, this summation becomes infeasible.

19. In general, formula (18) can be extended as follows. Suppose a set of J variables $V = \{v_1,...,v_j,...,v_J\}$, exists for which a loglinear model is to be estimated. Let the ith category of variable j be denoted by $i(v_j)$. Then we can define the set of category combinations (cells) of the J variables in V as $c(V) = \{i(v_1),...,i(v_J)\}$. Furthermore, we define V_k as the set of variables V with variable v_k excluded. The category combinations of the variables in V_k are $c(V_k) = \{i(v_1),...,i(v_{k-1}), i(v_{k+1}),...,i(v_J)\}$. With this notation, the generalized form of (18) can be expressed as

$$b_{i(v_k)}^{v_k(t)} = \frac{n_{i(v_k)+}}{f^{(c)}(i(v_k))} \tag{19}$$

with $n_{i(v_k)+} = \sum_{c(V_k)} n_{c(V)}$, and $f^{(c)}(i(v_k)) = \sum_{c(V_k)} d_{c(V)} \Pi_{v_j \in V_k} b_{i(v_j)}^{v_j(c)}$.

20. Again, the problem is in the summation over the combinations $c(V_k)$ in the expression for $f^{(c)}(i(v_k))$. In order to suggest a possible solution to this problem, we first re-express $f^{(c)}(i(v_k))$ as

$$f^{(c)}(i(v_k)) = C(V_k)\,\bar{f}^{(c)}(i(v_k))$$

where $C(V_k)$ is the number of category combinations in $c(V_k)$ and $\bar{f}^{(c)}(i(v_k)) = f^{(c)}(i(v_k))/C(V_k)$. We now can approximate the average value $\bar{f}^{(c)}(i(v_k))$ and thus the total $f^{(c)}(i(v_k))$ by randomly drawing, say $S(V)$, cells from the set of cells $c(V)$ using the uniform distribution. Denote this sampled set of cells by $s(V)$ and the corresponding set of $S(V_k)$ sampled cells excluding variable v_k by $s(V_k)$. An approximation for $\bar{f}^{(c)}(i(v_k))$ can then be obtained as

$$\hat{\bar{f}}^{(c)}(i(v_k)) = \frac{1}{S(V_k)} \sum_{s(V_k)} d_{s(V)} \Pi_{v_j \in V_k} b_{i(v_j)}^{v_j(c)} \tag{20}$$

and this approximation can be used to calculate (19).

V. IMPUTATION SUBJECT TO EDIT RESTRICTIONS

21. The model estimated using formulae (19), or (16) and (17), can be used as follows. Consider a dataset XY of respondents, containing complete respondent cases in set X and incomplete respondent cases in set Y. Assume some loglinear model M estimated using the data from X. The set Y can now be imputed per respondent by first considering a (relatively small) contingency table C for all possible combinations of values for the missing variables for the respondent. A cell in C is either forbidden due to edits (structural zeros), or has a conditional probability (given the observed values) that can be obtained from the model M.

22. Imputation can be performed using by randomly drawing from the combinations in C according to the conditional probabilities of these combinations.

VI. CONCLUSION

23. In this paper a method is described for estimation of loglinear model parameters under edit constraints when the number of cells is very large. While the model in this paper assumes independence between variables, addition of interactions between the variables only causes a linear increase of the required memory usage and computation time in terms of the number of parameters to be estimated. However, due to the sparseness of the data, the models are limited in the sense that interactions between many variables cannot be estimated reliably.

24. Future research will first focus on experimental results for the proposed parameter estimation on complete data, after which the remaining incomplete data is imputed using this model. This approach will allow a better understanding of the proposed sampling technique and its effect on imputation.

25. Secondly, the described complete-case parameter estimation will be used in an EM algorithm, which iterates between parameter estimation for the complete data (the M step) and imputation using the estimated model (the E step). In this way, the EM algorithm allows to use both complete and incomplete records for parameter estimation (Dempster, Laird and Rubin, 1977; Little and Rubin, 2002).

26. The purpose of these experiments is to gain more insight into the problems of dimensionality and sparseness with regard to parameter estimation for loglinear models under edit constraints and the usefulness of these models for imputation in surveys with large sets of categorical variables.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis.* Wiley.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge Massachusetts.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). *Maximum-Likelihood from Incomplete data via the EM Algorithm.* Journal of the Royal Statistical Society Series B. 39, pp. 1-38.

Goodman, L.A. (1968). *The Analysis of Cross-Classified Data: Independence, Quasi-Independence and Interaction in Contingency Tables with or without Missing Cells.* Journal of the American Statistical Association 63, pp. 1091-1113.

Little, R.J.A. and D.B. Rubin (2002). *Statistical Analysis with Missing Data.* Wiley, New York.

Winkler, W.E. (2003). *A Contingency-Table Model for Imputing Data Satisfying Analytic Constraints*. UN/ECE Work Session on Statistical Data Editing, Madrid.
