

**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

**SMOOTHING IMPUTATIONS FOR CATEGORICAL DATA IN THE LINEAR REGRESSION
PARADIGM**

Invited Paper

Submitted by the U.S. Census Bureau, United States¹

I. INTRODUCTION

1. Several authors (Little and Rubin 2002, Thibaudeau 2003, Winkler 2003) propose implementing the EM algorithm in the context of log-linear models to estimate conditional probabilities and predict missing data in contingency tables. However, log-linear modeling remains unappealing to many. One problem is the absence of an algebraic distinction between the dependent and independent variables in the log-linear paradigm.

2. Another problem is the difficulties encountered when attempting to analyse observations generated by a complex survey in the framework of an advanced log-linear model. Complex surveys involve elaborate variance/covariance structures. A log-linear analysis involving hierarchical structures may require complex linearization methods (or the design of complicated bootstrapping/jackknifing schemes to account for the design of the sample when estimating the variance of the estimators. The presence of missing data just compounds the problem. Applying the EM algorithm for estimating multinomial conditional probability in the context of a complex survey is far from simple. Mainstream textbooks typically only discuss completely flat designs.

3. Practitioners of all stripes may find the set-up of a simple linear regression more natural to work with. In this set-up, the response variables and the predictor variables are clearly identified and the causality system is explicit. In addition, complex covariance structures can easily be integrated.

4. In the paper we propose a systematic approach to estimate log-linear conditional probabilities in the linear regression paradigm, and in presence of missing data. Thibaudeau (2002) shows how to use such estimates to smoothen discrete imputations over small geographical areas. We propose an algorithm to produce estimates of the conditional probabilities by estimating the conditional log-odds. The log-odds parameterization leads to framing the estimation of the conditional probabilities in the linear regression paradigm. It is then more natural to account for complex sample design. Our approach applies to any set of hierarchical log-linear constraints.

¹ Prepared by Yves Thibaudeau and William E. Winkler, U.S. Census Bureau (yves.thibaudeau@census.gov, william.e.winkler@census.gov).

II. ESTIMATING CONDITIONAL PROBABILITIES FOR A 2X2X2X2 TABLE

5. We illustrate our approach in a simple situation. Yet this special case is complex enough so that extending our approach to more general situations is natural. We consider a $2 \times 2 \times 2 \times 2$ contingency table defined by four categorical variables denoted i, j, k, l . Therefore, $i, j, k, l = 1, 2$. We posit an all-2nd order interaction log-linear model to describe the process generating the table. In this framework, we let $u_{a,b,c,d}$ be the conditional probabilities of i and j being equal to a and b respectively, given k and l are equal to c and d :

$$u_{a,b,c,d} = \Pr(i=a, j=b | k=c, l=d) \quad (1)$$

6. We use the set of conditional probabilities $\{u_{a,b,c,d}\}_{a,b,c,d=1}^2$ to construct a set of log-odds $\{\mathbf{m}_m\}_{m=1}^{12}$. The log-odds in this set are defined as follows:

$$\mathbf{m}_1 = \log\left(\frac{u_{1,1,1,1}}{u_{2,2,1,1}}\right); \quad \mathbf{m}_2 = \log\left(\frac{u_{2,1,1,1}}{u_{2,2,1,1}}\right); \quad \mathbf{m}_3 = \log\left(\frac{u_{1,2,1,1}}{u_{2,2,1,1}}\right); \quad (2)$$

$$\mathbf{m}_4 = \log\left(\frac{u_{1,1,2,1}}{u_{2,2,2,1}}\right); \quad \dots \quad \mathbf{m}_{12} = \log\left(\frac{u_{1,2,2,2}}{u_{2,2,2,2}}\right)$$

7. From first principles, there is a one-to-one mapping between legitimate values for the elements of the set of conditional probabilities $\{u_{a,b,c,d}\}_{a,b,c,d=1}^2$ and elements of the set of log-odds $\{\mathbf{m}_m\}_{m=1}^{12}$.

8. To further explain our approach we express the log-odds defined in (2) in terms of a non-singular set of conditional probabilities. The cardinality of this set is exactly the number of degrees of freedom underlying the set $\{u_{a,b,c,d}\}_{a,b,c,d=1}^2$ when specifying an all-2nd order interactions log-linear model.

Appendix A sketches an algorithm to elicit these conditional probabilities (see also Thibaudeau 2003). Operating our algorithm in this special situation leads to define the set

$\{q_{1,1,1}, q_{2,1,1}, q_{1,2,1}, q_{2,2,1}, r_{1,1}, r_{2,1}, r_{1,2}\}$ where

$$\begin{aligned} q_{a,b,c} &= \mathbf{p}(i=1 | j=a, k=b, l=c) \\ r_{a,b} &= \mathbf{p}(j=1 | i=a, k=b, l=b) \end{aligned} \quad (3)$$

9. Again, given the all-2nd interactions log-linear model, we can write the log-odds defined in (2) as a deterministic function of these conditional probabilities. To show this we first facilitate the notation by taking the logit transformation of each conditional probability:

$$\mathbf{I}_1 = \text{logit}(q_{1,1,1}); \quad \mathbf{I}_2 = \text{logit}(q_{2,1,1}); \quad \mathbf{I}_3 = \text{logit}(q_{1,2,1}); \quad \mathbf{I}_4 = \text{logit}(q_{2,2,1}) \quad (4)$$

$$\mathbf{I}_5 = \text{logit}(r_{1,1}); \quad \mathbf{I}_6 = \text{logit}(r_{2,1}); \quad \mathbf{I}_7 = \text{logit}(r_{1,2})$$

10. Let $\underline{\mathbf{l}} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3, \mathbf{l}_4, \mathbf{l}_5, \mathbf{l}_6, \mathbf{l}_7]^t$. The range of $\underline{\mathbf{l}}$ is $(-\infty, \infty)^7$, or $(-\infty, \infty)$ as a short hand. The vector of log-odds is fully characterized by $\underline{\mathbf{l}}$ and consequently we denote this vector by $\underline{\mathbf{m}}(\underline{\mathbf{l}})$. The following linear equation formally defines $\underline{\mathbf{m}}(\underline{\mathbf{l}})$ and its characterization by $\underline{\mathbf{l}}$.

$$\underline{\mathbf{m}}(\underline{\mathbf{l}}) = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \mathbf{m}_3 \\ \mathbf{m}_4 \\ \mathbf{m}_5 \\ \mathbf{m}_6 \\ \mathbf{m}_7 \\ \mathbf{m}_8 \\ \mathbf{m}_9 \\ \mathbf{m}_{10} \\ \mathbf{m}_{11} \\ \mathbf{m}_{12} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 & 0 & 0 & 0 \\ -2 & 1 & 1 & 1 & -1 & 1 & 1 \\ -1 & 1 & 0 & 0 & -1 & 1 & 1 \\ -2 & 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \underline{\mathbf{l}} \quad (5)$$

11. Proposition (5) is algebraically derived from the structure of the all-2nd order log-linear model. It gives a representation of the parametric log-odds as linear combinations of the logits of the conditional probabilities elicited by our algorithm. In effect, (5) suggests expressing the observed log-odds in the linear regression paradigm. I.e., to predict the conditional probabilities $\{u_{a,b,c,d}\}_{a,b,c,d=1}^2$, (5) suggests regressing the observed log odds on the parametric log-odds $\underline{\mathbf{m}}(\underline{\mathbf{l}})$. To formulate this in a compact format, we introduce the following shorthand for representing the observed log-odds.

$$Y_1 = \log\left(\frac{X_{1,1,1,1}}{X_{2,2,1,1}}\right); Y_2 = \log\left(\frac{X_{2,1,1,1}}{X_{2,2,1,1}}\right); Y_3 = \log\left(\frac{X_{1,2,1,1}}{X_{2,2,1,1}}\right); \quad (6)$$

$$Y_4 = \log\left(\frac{X_{1,1,2,1}}{X_{2,2,2,1}}\right); \dots Y_{12} = \log\left(\frac{X_{1,2,2,2}}{X_{2,2,2,2}}\right)$$

12. Let $\underline{\mathbf{Y}} = [Y_1, Y_2, \dots, Y_{12}]^t$, we propose deriving estimates of the conditional probabilities $\{u_{a,b,c,d}\}_{a,b,c,d=1}^2$ through the following regression model.

$$\underline{\mathbf{Y}} = \underline{\mathbf{m}}(\underline{\mathbf{l}}) + \underline{\mathbf{y}} \quad (7)$$

13. $\underline{\mathbf{y}}$ is a multivariate error approximately normally distributed with mean zero and covariance matrix \mathbf{S} . The sampling plan determines the structure of \mathbf{S} .

III. WEIGHTED LEAST SQUARES FOR CONDITIONAL PROBABILITIES

14. So, (5) and (7) hint at estimating $\left\{u_{a,b,c,d}\right\}_{a,b,c,d=1}^2$: we first perform a weighted least-squares regression of \underline{Y} on $\underline{m}(\underline{I})$, according to the model in (7). After the regression we compute an estimate of $\left\{u_{a,b,c,d}\right\}_{a,b,c,d=1}^2$ by substituting in (5) $\hat{\underline{I}}$, the weighed least-squares estimator of \underline{I} , to obtain an estimate of $\underline{m}(\underline{I})$. Given $\underline{m}(\underline{I})$, an estimate for the conditional probabilities $\left\{u_{a,b,c,d}\right\}_{a,b,c,d=1}^2$ is constructed.

15. Therefore $\hat{\underline{I}}$, the least square estimator of \underline{I} , is pivotal to implement our suggested approach. We formally define $\hat{\underline{I}}$ to be any estimator of \underline{I} such that, for any vector \underline{t} in the open multivariate interval $(-\infty, \infty)^7$, the following holds:

$$\begin{aligned} & \left[\underline{Y} - \underline{m}(\hat{\underline{I}}) \right]^t \mathbf{S}^{-1} \left[\underline{Y} - \underline{m}(\hat{\underline{I}}) \right] \\ & \leq \left[\underline{Y} - \underline{m}(\underline{t}) \right]^t \mathbf{S}^{-1} \left[\underline{Y} - \underline{m}(\underline{t}) \right] \end{aligned} \tag{8}$$

IV. MISSING DATA

16. We generalize the weighed least-square regression described in the previous section to situations involving missing data. Again, for simplicity, we minimally expand the special case presented in the preceding section. The general case follows easily.

17. Suppose some units known to be in the collapsed cells defined by $k = I, l = I$ cannot be fully cross classified because one the value of i or j is missing. We assume i and j are never missing together. We give additional notation to formulate the linear regression in this set-up.

18. We use the customary dot notation to denote the marginal counts generated by missing i or j . For example, $X_{\square 1,1,1}$ is the marginal count over i , given $j, k, l = I$. The following definitions generalize the expression for the observed odds to situation where missing values of i or j are present when $k, l = I$.

$$\begin{aligned}
Z_1(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}) &= \log\left(\frac{X_{1,1,1,1} + \mathbf{a} X_{\square,1,1,1} + \mathbf{b} X_{1,\square,1,1}}{X_{2,2,1,1} + (1-\mathbf{d})Y_{\square,2,1,1} + (1-\mathbf{e})Y_{2,\square,1,1}}\right) \\
Z_2(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}) &= \log\left(\frac{X_{2,1,1,1} + (1-\mathbf{a})X_{\square,1,1,1} + \mathbf{e} X_{2,\square,1,1}}{X_{2,2,1,1} + (1-\mathbf{d})X_{\square,2,1,1} + (1-\mathbf{e})X_{2,\square,1,1}}\right) \\
Z_3(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}) &= \log\left(\frac{X_{2,1,1,1} + (1-\mathbf{a})X_{\square,1,1,1} + \mathbf{e} X_{2,\square,1,1}}{X_{2,2,1,1} + (1-\mathbf{d})X_{\square,2,1,1} + (1-\mathbf{e})X_{2,\square,1,1}}\right)
\end{aligned} \tag{9}$$

$$Z_m(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}) = Y_m; \quad m = 4, \dots, 12$$

19. The parameters $\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}$ represent weight serving to reapportion each marginal count between the cell counts. Since the overall marginal counts must remain unchanged, the condition $1 \leq \mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e} \leq 1$ applies in (9).

20. To further compact the notation let:

$$\underline{Z}(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}) = [Z_1(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}), Z_2(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}), \dots, Z_{12}(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e})]^t \tag{10}$$

21. The general regression model becomes

$$\underline{Z}(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}) = \underline{m}(\underline{l}) + \underline{y} \tag{11}$$

22. (11) hints at regressing $\underline{Z}(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e})$ on $\underline{m}(\underline{l})$ to produce an estimate of \underline{l} . From there, estimates of the conditional probabilities $\{u_{a,b,c,d}\}_{a,b,c,d=1}^2$ are derived as before. But in this case, the regression procedure must implicitly produce optimal value for the reapportionment parameters $\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}$.

23. In the context of regressing $\underline{Z}(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e})$ on $\underline{m}(\underline{l})$ with respect to all the parameters involved, we define an optimal class of estimators for \underline{l} . We call it the class of “general weighted least-squares estimators” (the class of GWLSE).

24. We use \underline{l}^* to denote any GWLSE for \underline{l} . A specific \underline{l}^* takes the form of a weighted least-squares estimator, as when there are no missing data. But it also minimizes the sum of squares with respect to the reapportionment variables $\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}$. A formal definition for \underline{l}^* follows.

25. \underline{l}^* is a GWLSE for \underline{l} if there exists $\mathbf{a}^*, \mathbf{b}^*, \mathbf{d}^*, \mathbf{e}^*$ such that for any $-\infty < \underline{t} < \infty$ and any $0 < \mathbf{g}, \mathbf{h}, \mathbf{i}, \mathbf{k} < 1$, the following holds:

$$\begin{aligned}
& \left[\underline{z}(\mathbf{a}^*, \mathbf{b}^*, \mathbf{d}^*, \mathbf{e}^*) - \underline{m}(\underline{l}^*) \right]^t \mathbf{S}^{-1} \left[\underline{z}(\mathbf{a}^*, \mathbf{b}^*, \mathbf{d}^*, \mathbf{e}^*) - \underline{m}(\underline{l}^*) \right] \\
& \leq \left[\underline{z}(\mathbf{g}, \mathbf{h}, \mathbf{i}, \mathbf{k}) - \underline{m}(\underline{t}) \right]^t \mathbf{S}^{-1} \left[\underline{z}(\mathbf{g}, \mathbf{e}, \mathbf{i}, \mathbf{k}) - \underline{m}(\underline{t}) \right]
\end{aligned} \tag{12}$$

26. Recall the EM algorithm implicitly computes estimates of parameters equivalent to $\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}$. These estimates are obtained through formal repeated E-steps –i.e. conditional expectation steps. While it does not integrate a formal E-step, our approach is in effect a special case of the EM algorithm.

References

- Di Zio, M., Scanu, M., Coppola, L., Luzi, O., and Ponti, A. (2004), “Bayesian Networks for Imputation,” *Journal of the Royal Statistical Society, A*, 167 (2), 309-322.
- Fellegi, I. P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17-35.
- Gelman, A., Carlin, J. B., Stern, H.S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, Chapman & Hall: London.
- Little, R. A., and Rubin, D. B., (1987), *Statistical Analysis with Missing Data*, John Wiley: New York.
- Rubin, D. B. (2003), “Iteratively Reweighted Least Squares,” in (Kotz, S., and Johnson, N. L.) *Encyclopedia of Statistics*, J. Wiley: New York.
- Schaefer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, CRC Press: London.
- Thibaudeau, Y. (2003), “An Algorithm for Computing Full Rank Minimal Sufficient Statistics with Applications to Confidentiality Protection,” *Monographs of Official Statistics*, Work Session on Data Confidentiality (EUROSTAT), Luxembourg.
- Thibaudeau, Y. (2002). “Model Explicit Item Imputation for Demographic Categories,” *Survey Methodology*, **28**, 135-143.
- Winkler, W. E. (2003), “A Contingency Table Model for Imputing Data Satisfying Analytic Constraints,” *American Statistical Association, Proc. Survey Research Methods Section*, CD-ROM, also research Report SRS 2003/07 at <http://www.census.gov/srd/www/byyear.html>.
