Uncertainty in statistical matching for discrete categorical variables

Pier Luigi Conti and Daniela Marella and Mauro Scanu

Abstract Statistical matching has the objective to estimate a joint distribution of two r.v. (Y,Z) when two sample surveys on (X,Y) and (X,Z) are available, X being a set of common variables in the two surveys. The aim of this paper is to analyze the uncertainty (due to the lack of joint sample information on (Y,Z)) in statistical matching for ordered categorical variables. The notion of uncertainty is first introduced, and a measure of uncertainty is then proposed. Moreover, the reduction of uncertainty in the statistical model due to the introduction of logical constraints is investigated and evaluated via simulation.

Key words: Statistical Matching, contingency tables, structural zeroes, nonidentifiability, uncertainty

1 Introduction

Let (X, Y, Z) be a three-dimensional random variable (r.v.), and let A and B be two independent samples of n_A and n_B i.i.d. records from (X, Y, Z), respectively. Assume that the marginal (bivariate) (X, Y) is observed in A, and that the marginal (bivariate) (X, Z) is independently observed in B. The main goal of statistical matching, at a macro level, consists in estimating the joint distribution of (X, Y, Z). Such a distribution is not identifiable due to the absence of joint information on Z and Y given X, see [3]. The main consequence of the lack of identifiability is that some

Pier Luigi Conti
Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma "La Sapienza", e-mail: pierluigi.conti@uniroma1.it

Daniela Marella
Dipartimento di Scienze dell'Educazione, Università "Roma Tre" e-mail: dmarella@uniroma3.it

Mauro Scanu
ISTAT, Roma e-mail: scanu@istat.it

parameters of the model cannot be estimated on the basis of the available sample information. For instance, in a parametric setting, instead of point estimates, one can only reasonably construct sets of "possible estimates", compatible with what can be actually estimated. These sets provide a representation of uncertainty about the model parameters. In this setting, the main task consists in constructing a coherent measure that can reasonably quantify the uncertainty about the (estimated) model. In this paper, we provide a precise definition of uncertainty on the (estimated) model, and construct a coherent measure that can reasonably quantify such an uncertainty. We confine ourselves to the case of ordered categorical variables. The case of discrete variables with nominal values is dealt with in [3].

2 Uncertainty in statistical matching for ordered categorical variable

Assume that, given a discrete r.v. X with I categories, Y and Z are discrete r.v.s too, with J and K categories, not necessarily ordered. With no loss of generality, the symbols $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, denote the categories taken by X, Y and Z, respectively. Let $\gamma_{jk|i}$ be the conditional probability $Pr(Y = j, Z = k|X = i)$, and denote by $\phi_{j|i} = Pr(Y = j|X = i)$ and $\psi_{k|i} = Pr(Z = k|X = i)$ the corresponding marginals, respectively. For real numbers a, b, define further the two quantities: $U(a, b) = \min(a, b)$, $L(a, b) = \max(0, a+b-1)$, then

$$L(\phi_{j|i}, \psi_{k|i}) \leq \gamma_{j,k|i} \leq U(\phi_{j|i}, \psi_{k|i}). \tag{1}$$

The interval (1) summarizes the pointwise uncertainty about the statistical model for every triple (i, j, k). It is intuitive to take the length of such an interval as a pointwise measure of uncertainty. Formally

$$\Delta^{jk|i} = U(\phi_{j|i}, \psi_{k|i}) - L(\phi_{j|i}, \psi_{k|i}) \tag{2}$$

The larger $\Delta^{jk|i}$ the more uncertain the statistical model generating the data w.r.t. (i, j, k). Conditionally on i, the pointwise uncertainty measures (2) can be summarized as follows

$$\Delta^{x=i} = \sum_{j=1}^{J} \sum_{k=1}^{K} \left\{ U(\phi_{j|i}, \psi_{k|i}) - L(\phi_{j|i}, \psi_{k|i}) \right\} \phi_{j|i} \psi_{k|i}, \tag{3}$$

representing the conditional uncertainty measure. Analogously, the overall uncertainty measure is given by

$$\Delta = \sum_{i=1}^{I} \Delta^{x=i} \xi_i \tag{4}$$

where ξ_i is the probability of the event $(X = i)$. Sharper results are obtained when the categories taken by (X,Y,Z) are ordered. For the sake of simplicity, we use the customary order for natural numbers. In this case, define with $H_{jk|i} = \sum_{y=1}^{j} \sum_{z=1}^{k} \gamma_{yz|i}$, $j = 1,\dots,J, k = 1,\dots,K, i = 1,\dots,I$, $F_{j|i} = \sum_{y=1}^{j} \phi_{y|i}$, $j = 1,\dots,J$, $i = 1,\dots,I$, $G_{k|i} = \sum_{z=1}^{k} \psi_{z|i}$, $k = 1,\dots,K$, $i = 1,\dots,I$, the cumulative d.f.'s are respectively. Then, the inequalities

$$L(F_{j|i}, G_{k|i}) \leq H_{jk|i} \leq U(F_{j|i}, G_{k|i}). \tag{5}$$

hold true. It is easy to prove that inequalities (5) imply sharper inequalities for the probabilities $\gamma_{jk|i}$. In this case, the pointwise uncertainty measure (2) is

$$\Delta^{jk|i} = U(F_{j|i}, G_{k|i}) - L(F_{j|i}, G_{k|i}). \tag{6}$$

and the conditional uncertainty measure (3) is

$$\Delta^{x=i} = \sum_{j=1}^{J} \sum_{k=1}^{K} \left\{ U(F_{j|i}, G_{k|i}) - L(F_{j|i}, G_{k|i}) \right\} \phi_{j|i} \psi_{k|i}. \tag{7}$$

The overall uncertainty measure Δ is the average of (7) with respect to the probabilities ξ_is. In several real cases, uncertainty about the joint distribution of Y and Z can be considerably reduced by introducing appropriate logical constraints among the values taken by Y and Z. Precisely, we consider constraints acting as *structural zeroes*, *i.e.* constraints that make equal to 0 some of the joint probabilities $\gamma_{jk|i} = Pr(Y = j, Z = k|X = i)$. Of course, this is equivalent to assume that logical constraints "reduce" the support of the joint distribution of Y and Z (given X), which is strictly smaller than the Cartesian product of the supports of Y and Z. To introduce the kind of constraints we will deal with, consider the support of (Y,Z) given X, which is a subset (either proper or improper) of $\{(j,k);\ j = 1,\dots,J; k = 1,\dots,K\}$. For each $j \in \{1,\dots,J\}$, define the two integers: k_j^+ = largest integer k such that $\gamma_{jk|i} > 0$; k_j^- = smallest integer k such that $\gamma_{jk|i} > 0$. Of course, there exist integers j_1, j_2 such that $k_{j_1}^+ = K$ and $k_{j_2}^- = 1$. The support of (Y,Z) (given X) is *Y-regular* if, for all $j = 1,\dots,J$: $\gamma_{jk|i} = 0$ if and only if $k > k_j^+$, or $k < k_j^-$. In this setting new extremal distributions $H_{jk|i}^+$, $H_{jk|i}^-$ obtained by suitable algorithms can be found. Furthermore, the unconstrained bounds in (5) will be reduced whenever the support of (Y,Z) given X is Y-regular. Z-regularity can be defined in a similar way, leading to similar results. $\Delta_c^{x=i}$ and Δ_c will be the conditional and unconditional measures of uncertainty under these constraints.

3 Estimation of the measure(s) of uncertainty

An important feature of the measure of estimation introduced so far is that they can be estimated on the basis of sample data. Let $n_{A,i}^x$ $(n_{B,i}^x)$ be the number of sample observations in sample A (B) such that $X = i$, and let $n^{xy}_{A,ij}$ ($n^{xz}_{B,ik}$) be the number of observations in sample A (B) such that $X = i$ and $Y = j$ ($X = i$ and $Z = k$), $i = 1,\dots,I$, $j = 1,\dots,J$, $k = 1,\dots,K$. Let $\widehat{F}_{j|i}$ and $\widehat{G}_{k|i}$ be the empirical cumulative distribution functions (e.c.d.f.s) of $F_{j|i}$, $G_{k|i}$.

The conditional and unconditional measures of uncertainty can be estimated by

$$\widehat{\Delta}_c^{x=i} = \sum_{j=1}^{J}\sum_{k=1}^{K}\left(\widehat{H}^{+}_{jk|i} - \widehat{H}^{-}_{jk|i}\right)\widehat{\phi}_{j|i}\,\widehat{\psi}_{k|i},, \qquad \widehat{\Delta}_c = \sum_{i=1}^{I}\widehat{\Delta}^{x=i}\,\widehat{\xi}_i$$

respectively. It can be proved that:

- The estimators of uncertainty measures are consistent:

$$\widehat{\Delta}_c^{x=i} \overset{a.s.}{\to} \Delta_c^{x=i} \;\; as\; n_A \to \infty, n_B \to \infty, \;\; i = 1,\dots,I;$$

$$\widehat{\Delta}_c \overset{a.s.}{\to} \Delta_c \;\; as\; n_A \to \infty,\; n_B \to \infty.$$

- The estimators of uncertainty measures are asymtpotically normal. Assume that $n_A/(n_A + n_B) \to \alpha$ as n_A, n_B go to infinity, with $0 < \alpha < 1$, and that $F_{j|i}$s, $G_{k|i}$s satisfy some differentiability conditions. Then, both

$$\sqrt{\frac{n^x_{A,i}n^x_{B,i}}{n^x_{A,i} + n^x_{B,i}}}(\widehat{\Delta}_c^{x=i} - \Delta_c^{x=i}) \text{ and } \sqrt{\frac{n_A n_B}{n_A + n_B}}(\widehat{\Delta}_c - \Delta_c)$$

  do have normal asymptotic distribution with mean zero and positive variance σ_i^2 and σ^2 respectively, as n_A, n_B tend to infinity.

The asymptotic variances σ_i^2s, σ^2 do have a complicate form, depending on the "true" $F_{j|i}$s, $G_{k|i}$s . However, they can be consistently estimated by bootstrap. The above results are useful to construct point and interval estimates of the uncertainty measures $\Delta_c^{x=i}$, Δ_c. They are also useful to test the hypothesis that the class of bivariate d.f.s with upper bounds $H^{+}_{jk|i}$s and lower bounds $H^{-}_{jk|i}$ is "narrow", when structural zeroes are considered.

References

1. Conti, P.L., Marella, D., Scanu, M.: How far from identifiability? A nonparametric approach to uncertainty in statistical matching under logical constraints. Technical Report n.22, DSPSA, Università di Roma "La Sapienza" (2009).
2. Conti, P.L., Marella, D., Scanu, M.: Uncertainty analysis in statistical matching. To be published on the Journal of Official Statistics (2012).
3. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching: Theory and Practice. Wiley, Chichester (2006).