REGRESSION ANALYSIS IN ESTIMATION OF CONDITIONAL PROBABILITIES FROM GROUPED QUALITATIVE DATA

A.I. MEDANI

in a wide class of problems, the events of greatest interest are those whose occurrence is conditional on the occurrence of other events. The probability that a variable Y assumes a certain value Y (= 1,2, ..., I) for an individual unit depends upon the value of the variable X (= 1,2, ..., J) assumed by the characteristics of the individual unit. Thus, the probability p is conditional and may be defined in terms of relative frequency :

$$P(Y/X) = \lim_{n_{.x} \longrightarrow \infty} \frac{n_{yx}}{n_{.x}} = a_{yx} \qquad (1)$$

where n_{yx} represents the number of individuals in the sample with a particular value of Y and a particular value of X. $n_{.x} = \sum_{y} n_{yx}$. a_{yx} indicates the conditional probability in question. If both Y and X are known for all individuals, the I J probabilities p (Y / X) could be estimated as follows :

$$\hat{a}_{y/x} = \frac{n_{yx}}{n_{.x}} \qquad (2)$$

But, since both the denominator and the numerator in the right-hand side of (2) are, in general, stochastic, the $\hat{a}_{y/x}$, as ratio-estimates, are biased. However, (2) ensures that the estimates are statistically consistent in the sense that

$$\sum_{y} \hat{a}_{yx} = 1$$

In many situations, values of Y and X are not known for individual Units, but their frequency distributions may be derived for a number of groups into which the units have been classified. In this case, the conditional probability may be defined as :

$$P(Y/X,g) = \lim_{n_{.xg} \to \infty} \frac{n_{yxg}}{n_{.xg}} \qquad (3)$$

where $n_{y \times g}$ represent the number of individuals belonging to group g with a particular value of X and a particular value of y, and

$$n_{.xg} = \sum_{y} n_{yxg}$$

If the expected value of y given x does not depend upon the group g to which the individual unit belongs, then

$$P(Y/X,g) = P(Y/X)$$

$$= \lim_{n_{.xg} \to \infty} \frac{n_{yxg}}{n_{.xg}} \qquad (4)$$

Hence, for any fixed sample, we may write :

$$\frac{n_{yxg}}{n_{.xg}} = a_{y/x} + e_{yxg} \qquad (5)$$

where $e_{y \times g}$ represents sampling error or the difference between actual and expected value of $\frac{n_{y \times g}}{n_{. \times g}}$

The equation in (5) may be rewritten as :

$$n_{xyg.} = a_{y/x}\ n_{.xg} + U_{yxg} \qquad (6)$$

where $U_{y \times g} = n_{. \times g}\ e_{y \times g}$ is a random term. Since $n_{y \times g}$ are unknown, summing (6) over x results in a relationship with known variables but unknown parameters,

$$n_{y.g} = \sum_x a_{y/x} \; n_{.xg} + U_{y.g} \qquad (7)$$

$$\text{with } U_{y.g} \sum_x U_{yxg}$$

In matrix notation, equation (7) may be written as

$$V = WA + U \qquad (8)$$

where

$$V = \begin{bmatrix} n_{1.1} & \cdots & n_{I.1} \\ \vdots & & \vdots \\ n_{1.G} & \cdots & n_{I.G} \end{bmatrix}, W = \begin{bmatrix} n_{.1\,1} & \cdots & n_{.I\,1} \\ \vdots & & \vdots \\ n_{.1\,G} & & n_{.IG} \end{bmatrix}$$

$$A = \begin{bmatrix} a_{1/1} & \cdots & a_{I/1} \\ \vdots & & \vdots \\ a_{j/1} & \cdots & a_{I/j} \end{bmatrix}, U = \begin{bmatrix} U_{1.1} & \cdots & U_{j.1} \\ \vdots & & \vdots \\ U_{1.G} & \cdots & U_{I.G} \end{bmatrix}$$

One would like to obtain an estimate of A which is most efficient under the following side conditions :

1. The estimates $\hat{a}_{yx}$ of a_{yx} when summed over y add up to 1.

2. The estimates $\hat{n}_{y.x}$ of the $n_{y.g}$ when summed over all groups g, add up to observed totals $n_{y1.}$

3. The elements of the matrix $\hat{A}$ should slao satisfy the condition $0 \leqslant \hat{a}_{yx} \leqslant_1$

The generalized least-squares (G.L.S.) estimate $\hat{A}$ of A is given by :

$$\hat{A} = (W' HW)^{-1} W' HV \qquad (9)$$

where $H = Z'Z$ and

$$Z = \begin{bmatrix} Z_{1,11} & \cdots & Z_{1,1G} \\ \vdots & & \vdots \\ Z_{1,1G} & \cdots & Z_{1,GG} \\ \vdots & & \vdots \\ Z_{I,11} & \cdots & Z_{K,1G} \\ \vdots & & \vdots \\ Z_{1,G1} & \cdots & Z_{1,GG} \end{bmatrix}$$

The elements of the Z matrix will be considered later.

The first side condition under which A is to be estimated is

$$\sum_{y} \hat{a}_{yx} = 1 = \sum_{y} a_{yx}$$

and this may be expressed in matrix notation as

$$\hat{A} I_i = I_j = A I_i \qquad (10)$$

where I_i and I_j are unit vectors of lengths i and j, respectively.

Estimate $\hat{A}$ of A according to (9) satisfies the condition in (10). This may be proved as follows :

Premultiply both sides of (9) by W'HW and post-multiply them by I_i results in

$$W' H W \hat{A} I_i = W' H V I_i \qquad (11)$$

But by definition

$$V I_i = W I_j = \begin{bmatrix} n \ldots 1 \\ n \ldots G \end{bmatrix} \qquad (12)$$

us, from (12)

$$W'H V I_i = W'HW I_j \qquad (13)$$

follows from (11) and (13) that

$$W'H W \hat{A} I_i = W H V I_i = W'H W I_j \qquad (14)$$

The equation in (14) implies that condition (10) holds.

The second side condition is implied by the following identity

$$\sum_G \hat{n}_{y.g} = \hat{n}_{y..} = n_{y..}$$

in matrix notation by

$$I_G' \hat{V} = I_G' V \qquad (15)$$

is a unit vector of length G.

In general, this condition is not satisfied by the estimate $\hat{V}$ of V derived from as given by (9) :

$$I_G' \hat{V} = I_G' W \hat{A} = I_G' W (W'HW)^{-1} W'H \neq I_G' V \qquad (16)$$

Equation (15) would only hold if

$$I_G' \hat{U} = 0 \qquad (17)$$

where $\hat{U}$ is the G.L.S. estimate of u for the set of observations.

In order to ensure that condition (15) is satisfied, one could either try choose the matrix Z (and consequently H) in such a manner that the last inequality in (16) changes into an equality, or to obtain least squares estimates with condition 2 as an additional constraint.

Restrictions on Z. The matrix Z is applied to the set of equations to obtain an estimate $\hat{A}$ of A with minimum variance. This may be done by the application of single least squares to the equations in (8) pre-multiplied by Z such that the resulting variance-covariance matrix Z U U′Z′ is a scalor (σ^2) times a unit matrix. This method gives simultaneous estimates of all the IJ parameters a_{yix} constituting A. However, such an ambitious approach does not seem to be necessary. One may assume that the variances of the disturbance terms relating to different values of Y for units belonging to different groups are zero :

$$E(U_{y,g} \cdot U_{y'g'}) = O \text{ for } Y \neq Y' \text{ and } g = g' \text{ simultaneously} \quad (18)$$

A more stronger assumption is that the expected values of these covariances are also zero for disturbances relating to the same Y in different groups and for those relating to different Y_{is} in the same group, namely :

$$E\,(U_{y.g} \cdot Y_{y.g}') = \sigma^2_{y.g} \text{ for } Y = Y' \text{ and } g = g' \text{ simultaneously} \quad (19)$$

Accepting the loss in efficiency in the parameter estimates due to this assumption, (19) may be an extension of (18) that is worth the savings due to the resulting simplification. The equation in (19) means that for estimating A, each Y may be dealt with separately, according to

$$V_y = W_{ay} - U_y \; (Y = 1,2,\ldots..,\, I) \quad (20)$$

where

$$V_y = \begin{bmatrix} n_{y-1} \\ \vdots \\ n_{y.G} \end{bmatrix}, a_y = \begin{bmatrix} a_{y1} \\ \vdots \\ a_{yn} \end{bmatrix}, U_y = \begin{bmatrix} n_{y.1} \\ \vdots \\ n_{y.G} \end{bmatrix}$$

being columns of V, A and U respectively, and W is the same as before.

Restrictions on U. In order to obtain efficient estimates of a_{ylx} satisfying condition 1, one should minimize

$$S = U_y' H_y U_y - 2 \lambda I_G' U_y$$

where λ is a lagrange multiplier.

In order to economize on the length of the derivation, one may start with = I (unit matrix of order G) and then it will be easy to generalize for any H_y.

The necessary conditions for minimizing S with $\Omega = 1$ are :

$$\frac{\partial S}{\partial a_y} = -2 V_y' W + 2 a_y' W' W + 2 \lambda I_G' W = 0 \qquad (22)$$

after transposition and indicating a second estimate of a_y by $\hat{a}_y$:

$$W' V_y = W' W \hat{a}_y + \lambda W' I_G \qquad (23)$$

in addition :

$$I_G' V_y = I_G' W \hat{a}_y \qquad (24)$$

quations !23) and (24) may be combined into :

$$\begin{bmatrix} W' \\ I_G' \end{bmatrix} V_y = \begin{bmatrix} W'W & W' I_G \\ I_G' W & 0 \end{bmatrix} \begin{bmatrix} \hat{a}_y \\ \lambda \end{bmatrix} \qquad (25)$$

Provided that the first matrix at the right-hand side of (25) is nonsingular, the solution of the coefficient vector, given by λ would be :

$$\begin{bmatrix} \hat{a}_y \\ \lambda \end{bmatrix} \begin{bmatrix} W'W & W' I_G \\ I_G' W & 0 \end{bmatrix}^{-1} \begin{bmatrix} W' \\ I_G' \end{bmatrix} V_y \qquad (26)$$

From (26) $\hat{a}_y$ is given explicitly by :

$$\hat{\hat{a}}_y = \left[(W'W)^{-1}\left\{I - \frac{W'I_G\, I_G'\, W(W'W)^{-1}}{I_G'\, W(W'W)^{-1}W'I_G}\right\}W' \cdot \frac{(W'W)^{-1}W'I_G\, I_G'}{I_G'\, W(W'W)^{-1}\, W'I_G}\right]V_y \qquad (27)$$

By virtue of (27) $I'_G\, W\, \hat{a}_y = I_G'\, V_y$, that is (16) is satisfied.

In the general case where $H_y \neq I$, W' is replaced by $W'\, H_y$ resulting in :

$$\hat{\hat{a}}_y = \left[(W' H_y W)^{-1}\left\{I - \frac{W' H_y\, I_G\, I_G' W(W' H_y W)^{-1}}{I_G' W(W' H_y W)^{-1} W'}\right\}W' H_y \cdot \frac{(W' H_y W)^{-1} W' H_y\, I_G\, I_G'}{I_G' W\,(W' H_y W)^{-1} W' H_y\, I_G}\right] V_y \qquad (28)$$

or in matrix notation :

$$\hat{\hat{A}} = F\, V \qquad (29)$$

where F is the expression between brackets in the right-hand side of (28). One may easily verify that

$$I_G'\, W\, \hat{\hat{A}} = \left[I_G'\, W\, F\right]\, V = I_G'\, V \qquad (30)$$

so that condition 1 is satisfied.

In order to obtain most efficient estimates $\hat{\hat{a}}_y$ under the said side-conditions, one should choose,

$$H_y = U_y\, U_y' \qquad (31)$$

However, the U_y are unknown, so that we have to content ourselves with estimates of U_y. The estimates of U_y corresponding to the estimates of a_y according to (27) is

$$\hat{\hat{U}}_y = V_y - W\, \hat{\hat{a}}_y \qquad (32)$$

Restrictions on A. In addition to the two sets of constraints, the elements of the matrix should also satisfy the inequality

$$O \leqslant a_{ylx} \leqslant 1 \tag{33}$$

in matrix notation :

$$O \leqslant A \leqslant E \tag{34}$$

Which O and E are matrices of order IxS.

Estimation of $\hat{\hat{A}}$ according to (31) does not preclude possible violation of Negative estimates or estimates exceeding one might appear frequently.

The problem of estimating A, taking account of (34) is really one of quadratic programming.

The most obvious way of handling negative values of $\hat{\hat{a}}_{ylx}$ or values of $\hat{\hat{a}}_{ylx}$ greater than one is to replace one or more negative[1] values of $\hat{\hat{a}}_{yix}$ by 0 and values of $\hat{\hat{a}}_{ylx}$ greater than one by 1. The latter cannot be applied to more than one X for any y, implying zero values of $\hat{\hat{a}}_{ylx}$ for all other x for the same y. The former one reduces the W matrix of the explanatory variables, since the X which appear to be redundant for any particular Y drops out. Hence, all a_{yix} which are not put equal to zero *a priori* could estimated by using the reduced W matrix, and the estimates will still satisfy the additivity condition in (15). However, in such a case, one can no longer ensure that the first additivity condition in (10) would hold simultaneously.

Because of the unsystematic nature of the occurrence of $\hat{\hat{a}}_{ylx}$ not satisfying (B), the resulting violation of (10) can not be readily amended. Hence, where negative and/or larger than unity values of $\hat{\hat{a}}_{ylx}$ may occur, it would be preferable to aggregate X and/or Y in such a manner and to such an extent that these anamolies would vansh . If this does not remove these anamolies, one could draw up a preliminary table for $\hat{\hat{n}}_{ylx}$ for all groups g, estimated according to :

$$\hat{\hat{n}}_{y/x} = \hat{\hat{a}}_{y/x}\ n_{.x} \tag{35}$$

---

(1) Replacing all originally negative values of $\hat{\hat{a}}_{ylx}$ by zero might not be necessary, since it might be sufficient to restrict this operation to some of the naegative $\hat{\hat{a}}_{ylx}$ and at the same time «cure» the negativity of the others.

which may involve negative values of $\hat{\hat{n}}_{ylx}$ corresponding to negative values of $\hat{\hat{a}}_{ylx}$. This may give rise to excess values of horizontal and vertical marginal totals. These surpluses could be eliminated by distributing the negative $\hat{\hat{n}}_{ylx}$ per row (column) over the non-zero cells in proportion to the marginal vertical (horizontal) totals. In general, the original horizontal (vertical) marginal totals will differ from the resulting vertical (horizontal) sums. The resulting differences between the two could be distributed horizontally (vertically) over the non-zero cells in proportion to the marginal vertical (horizontal) totals, till the differences between the horizontal (vertical) sums of the cell frequencies and the marginal horizontal (vertical) totals would become generally smaller than the unit in which the $n_{y|x}$ are expressed(2) The final estimates of a_{ylx} may be obtained as

$$\tilde{a}_{y/x} = \frac{\tilde{n}_{yx.}}{n_{.x}} \qquad (36)$$

where $\tilde{n}_{yx.}$ are the call frequencies obtained in the last stage of the interative procedure outlined above, some of the $\tilde{n}_{yx.}$ and consequently the corresponding $\tilde{a}_{ylx}$ may have values of zero.

The model outlined above is especially applicable to situations where Y and X are attributes. It could also be applied to quantifiable variables where the effect of X on Y is not linear, and in particular not monotonic, porvided that both X and Y may be classified into groups. Such situations arise frequently in Socio-cultural relationships, featuring variables which are largely attributes or at least should be so considered.

---

(2) This method has been proposed by W, E, Deming, *Adjustment of data* (1954).