MISCLASSIFICATION OF CATEGORICAL DATA

Charles D. Cowan, U.S. Bureau of the Census

1. INTRODUCTION

There are several situations where misclassification of categorical data can be observed. Dual system estimation studies (used for evaluation of the coverage of a census) match records from a survey to a census, where both studies have collected basic demographic variables like age, race, sex, and ethnicity. Studies are also conducted matching survey results to administrative records, such as health studies which match survey data to patient records. Reinterview studies are conducted to evaluate the accuracy of censuses and surveys, or to monitor the work of interviewers. Finally, panel studies designed to measure change over time of certain variables will often collect the same demographic information if there is a sufficiently long time between interviews.

2. STUDYING MISCLASSIFICATION

In this paper, a method will be investigated for analyzing crossclassification data with misclassification present. In this investigation, the emphasis will be on estimation of various parameters of interest, such as the true proportions of respondents or cases falling into the categories under study. It will be shown that using repeated observations on a qualitative variable, the true proportions falling into the categories of this variable usually cannot be estimated from the crossclassified table, except in very special circumstances.

3. A MODEL FOR MISCLASSIFICATION

There is relatively little research available on misclassification, most likely because of the few opportunities available to study data collected in the same way multiple times on the same variable. In the research that has been done, different authors have used different models to study the phenomenon of observing misclassified data. The one underlying characteristic of these studies is that some assumption must be employed or device used to be able to estimate the true proportions in the categories being studied. Press (1968) assumes that the rates at which errors are made in observing (or reporting) the data are known. Nordheim (1984) assumes not that the rates are known, but rather that the relative rates (the odds) of misclassification are known. This latter approach works well because it reduces the number of parameters to be estimated, making Nordheim's model estimable, but still assumes that the researcher will have certain knowledge about the relation between the rates at which errors are being made. Chen (1974) studies misclassification and uses double sampling to obtain a resolution on a subset of records so as to be able to estimate the proportions in the categories of the variable being studied.

This paper will take a slightly different approach by trying to analyze the crossclassification of the dual or multiple observations of a variable by using the EM algorithm. The EM algorithm is an iterative procedure that consists of an expection step and a maximization step; it can be used to calculate parameters in a likelihood. A probability model that describes how the crossclassification was generated will be presented, and then the EM algorithm will be used to estimate the value of the parameters in the model.

The data observed in a study of misclassification can be summarized in a two-way table (or higher dimensional if there are multiple observations on a variable). There are r rows and r columns, and the cells of the tables are comprised of counts Z_{ij}; for i=j the cell totals represent counts of cases that agree in classification between the first observation and the second observation. For cells where $i \neq j$, Z_{ij} is the count of cases classified in the i^{th} category the first enumeration and the j^{th} category the second enumeration or observation. The Z_{ij} represent counts of the way individuals are tabulated, not their true status; i.e. the Z_{ij} are not counts giving the correct classification of individuals unless there is no misclassification. Assume there is an underlying variable, X_{ijk}, which is a count of the true number of cases in category k but observed or denoted as i at the first observation and observed or denoted as being in category j at the second observation. We never observe the number of cases truly in category k unless we have a situation with no misclassification. What we do observe are the Z_{ij}, which are the sums of the true counts of cases in categories k = 1, ...,r, observed in the (i,j) cell, expressed as:

$$Z_{ij} = \sum_{k=1}^{r} X_{ijk} \quad , \quad k = 1,\ldots,r \quad (1)$$

The X_{ijk} will be referred to as the complete data (which we do not observe), and Z_{ij} will be referred to as the observed data.

We need a mechanism for estimating the proportion of cases in a category, k, and so we postulate a probability model such that the probability that an individual we observe is from category k is p_k, and that there is further a probability that an individual will be categorized as being in category i the first observation and j the second observation that can be expressed as f_{ijk}, which is the probability that a case which is truly from category k is observed as being in category i the first observation and category j the second observation. We assume that misclassification will be differential between groups being observed, and that the event of being misclassified in the second observation is not necessarily independent of how the case was classified in the first observation. With these definitions, we consider sets of classification probabilities for each category k, and have the standard restriction:

$$\sum_{k}^{r} p_k = 1.0$$

$$\sum_{i}^{r}\sum_{j}^{r} f_{ijk} = 1.0 \text{ for } k = 1,\ldots,r \tag{2}$$

With these assumptions and the restriction (2), one could express the likelihood for the X_{ijk} as a product multinomial.

$$L \propto \prod_{k}^{r}\prod_{i}^{r}\prod_{j}^{r} (p_k f_{ijk})^{X_{ijk}} \tag{3}$$

The equation (3) is a joint probability of observing individuals who come from certain categories k and are observed in categories (i,j,k).

The problem, of course, is that we do not observe the X_{ijk}; we only observe the sums of the X_{ijk}, the Z_{ij}, and the probabilities of the Z_{ij} are convolutions developed from (3). They cannot be expressed in closed form, and it will be shown below that estimation of the parameters in the likelihood (3) do not involve the Z_{ij}, but instead require at least some of the X_{ijk} values.

4. THE EM ALGORITHM

To obtain estimates of the proportion of individuals in each of the categories for the variable under study, and the parameters in the model, we turn to the EM algorithm. A complete description of the EM algorithm can be found in Dempster, Laird, and Rubin (1977).

For the problem under consideration in this paper, the solutions to both steps can be found in the works of other authors. The M-step, estimation of the parameters by maximization of the likelihood (3), given the restriction set forth in (2), is a well known result that can be found in almost any text on qualitative methods, such as Bishop, Fienberg, and Holland (1975). Maximum likelihood estimates (MLE's) for the parameters are given by:

$$p_k = X_{++k} \Big/ \sum_{k}^{r} X_{++k} \quad \text{for } k = 1,\ldots,r$$

$$f_{ijk} = X_{ijk} / X_{++k} \quad \text{for} \quad \begin{array}{l} i = 1,\ldots,r \\ j = 1,\ldots,r \\ k = 1,\ldots,r \end{array} \tag{4}$$

$$\text{where} \quad X_{++k} = \sum_{i}^{r}\sum_{j}^{r} X_{ijk}$$

There are r-1 estimates of p_k, since restriction (2) would give us

$$p_r = 1 - \sum_{k}^{r-1} p_k \tag{5}$$

and $r\,(r^2-1)$ estimates of f_{ijk}, with f_{rrk} being estimated using restriction (2) again for k=1, ..., r. There are $(r-1)[1+r(r+1)] = r^3-1)]$ parameters total to be estimated. In a simple dichotomy, there would be seven parameters to be estimated. Restrictions on the model would reduce the number of parameters to be estimated.

The development of the E-step is less obvious. One must condition on the observed data to calulate the conditional expected values of the X_{ijk}. Work by Birch (1963) and Bishop, Fienberg, and Holland (1975), Cowan (1984), and Cowan and Malec (1984) show that for the product multinomial, the conditional expectations for the unobserved variables can be expressed as:

$$E(X_{ijk} \mid Z_{ij}, f_{ijk}, p_k) = Z_{ij} p_k f_{ijk} \Big/ \sum_{k}^{r} p_k f_{ijk} \tag{6}$$

There are $r^2(r-1)$ separate conditional expected values estimated. The remaining conditional expected values do not need to be estimated separately since they can be obtained by subtraction using (1). In the dichotomous case, there are four conditional expected values to be estimated. The parameters and conditional expected values to be estimated are presented for the dichotomous case in Table 1 below.

6. ESTIMATION

When one studies the data actually on hand, one discovers that there are only r^2-1 degrees of freedom available for estimation in the twoway crossclassification of the data. There are r^2 cells in the table, but the cell values must sum to a fixed value, N, the total sample size or population size. With only r^2-1 degrees of freedom, but r^3-1 parameters to be estimated, there is no unique solution for the parameters. The likelihood is a hill or pair of hills with a flat ridge at the top. The EM algorithm is guaranteed to converge, and does, but to a point on this ridge. Different arbitrary starting points used for the algorithm lead to different but equally likely solutions. Since the solutions are equally likely, there is no way to choose between the solutions, and so the problem is indeterminate. Other examples of this type of indeterminacy in solution can be found in the EM literature and in other areas like "errors-in-variables" models in regression.

There are two approaches that can be taken to obtain a solution to the problem of estimating the model parameters: double sampling and restriction of the model. The double sampling method is to draw a sample for adjudication only from the off diagonal cells. This will be described more fully in an example that follows for the dichotomous case.

A second method to obtain a solution is to restrict the model in some way. The most obvious way is to consider the classification events independent of one another; that is, classification to category i in the first observation is independent of classification to category j in the second observation. This reduces the parameter space in each level k of the complete data table from r^2-1 to 2(r-1). In general, let g_{ik} be the probability of a correct classification for class k in the first data set, and h_{jk} be the probability of a correct classification for class k in the second data set. Then the joint classification probabilites can be reexpressed as:

$$f_{ijk} = g_{ik}h_{jk} \qquad (7)$$

with the restriction that

$$\sum_i g_{ik} = 1.0 \text{ and } \sum_j h_{jk} = 1.0$$

For the dichotomous case we have

$$f_{11k} = g_{1k}h_{1k} \qquad (7')$$

$$f_{12k} = g_{1k}h_{2k} = g_{1k}(1-h_{1k})$$

$$f_{21k} = g_{2k}h_{1k} = (1-g_{1k})h_{1k}$$

$$f_{22k} = g_{2k}h_{2k} = (1-g_{1k})(1-h_{1k})$$

and we see there are only two parameters for each level of k rather than three to be estimated. With this restriction we now have (r-1)(2r+1) parameters to be estimated, but still have only (r^2-1) degrees of freedom, so the solution is still indeterminate. From this point there are two approaches that can be taken to make the problem tractible.

The first approach is double sampling. To make the problem estimable, one could take all the cases tallied in the off-diagonal cells, recheck or revisit these cases, adjudicate between responses and allocate the cases to the cells in the complete data table. This approach has the advantage that in almost all situations there will be relatively few cases off the main diagonal of this table, so that rechecking these cases will be inexpensive since few cases have to be resolved. The other advantage relative to double sampling is that the off-diagonal cells have much more information about misclassification than do the diagonal cells, since these cases are by definition misclassified at least once whereas in the diagonal cells most cases will not likely be misclassified.

Taking this approach adds $r(r-1)^2$ degrees of freedom to the r^2-1 degrees of freedom we had originally in the two way classification table. We now have $(r-1)(r^2+1)$ degrees of freedom, but need to estimate (r-1)(2r+1) parameters in the independence model, (and $(r-1)(r^2+r+1)$ parameters in the unconstrained model). For $r > 1$, and r an integer.

$$r^2+r+1 > r^2+1 > 2r+1$$

(with equality holding only for r=2), so the unconstrained model still cannot be resolved (in terms of estimation of parameters), but the model involving independence of classification can be.

In the EM algorithm, by forcing the conditional expectations of the off-diagonal cells to be equal to the values determined in the adjudication, (i.e. the X_{ijk}, $i \neq j$, are known) one gets two solutions in the iterative process. One solution is the correct solution with the maximum likelihood estimators (MLE's) of the parameters being exactly what they should be. The other solution is the mirror solution to the correct solution set.

If one cannot double sample those cases which fall in the off diagonal cells, one is left with only the original r^2-1 degrees of freedom. The only estimable models in this situation are those models with fewer parameters than degrees of freedom. Two types of restrictions suggest themselves in addition to the independence model. One is to set all parameters for a class equal across the two sources, as

$$g_{ik} = h_{ik} \text{ for } i = 1,\ldots,r \qquad (8)$$

The other method is to set the probability for a correct classification to a constant regardless of the category, but allow it to differ by source. Of necessity, to make the problem tractible, we would also make the probability of a misclassification a constant regardless of the category incorrectly specified. This can be expressed as:

$$\begin{aligned} g_{ik} &= s && \text{for } i = k \\ g_{ik} &= (1-s)/(r-1) && \text{for } i \neq k \\ h_{jk} &= t && \text{for } j = k \\ h_{jk} &= (1-t)/(r-1) && \text{for } j \neq k \end{aligned} \qquad (9)$$

We can attempt to use the EM algorithm to solve for the MLE's of the parameters in the restricted proplems. For the first restricted independence model, that described by (8), table 2 presents the conditional expected values and estimators for parameters in the model where parameters are restricted across sources within classes.

The alternative restriction, using the restrictions in (9), is presented in table 3. This is the case where classification attempts are independent, and classification parameters are the same between classes but differ between sources. Examples of where either type of restrictions may be appropriate are given in the next section.

6. EXAMPLES

To test the methods described in the last section, a computer program was written that would perform the iterations required for the EM algorithm. Different versions of the programs were written to reflect the different restrictions and concommitant estimators presented in tables 1, 2, and 3. The programs accept a two-by-two table with cross-classified data and attempt a series of runs of the EM algorithm. Each attempt begins with a different starting point and iterates until it converges. In no case did any example take more than 39 iterations to converge; the criterion for convergence was that no parameter estimate changed from the estimate of the same parameter in the prior iteration by more than 0.00001. In most cases it took less than 10 iterations to achieve convergence.

The first model considered was the model with the fewest restrictions. An artificial data set was constructed using parameters $p_1 = .90$, $g_{11} = .88$, $g_{22} = .92$, $h_{11} = .94$, $h_{22} = .96$, and $N = 10,000$. The artificial data set was used so that the correct convergence point would be known, and furthermore so that the off-diagonal values would be known for use in determining the convergence properties of the use of the double sampling information. In other words, all values of the complete data were generated and summed to form the Z_{ij} values which were used as input to the computer program; at the same time the values of the X_{ijk} were retained so that the off-diagonal values (all X_{ijk} where $i \neq j$) could be used as input to study the double sampling solutions. The Z_{ig} values generated are presented in Table 4 below. Starting values used were vectors of the form (p_1, g_{11}, g_{22}, h_{11},h_{22}); starting vectors used were (.1, .1, .1, .1, .1), (.3, .3, .3, .3, .3), (.5, .5, .5, .5, .5), (.7, .7, .7, .7, .7), and (.9, .9, .9, .9, .9).

Table 5 presents the results of application of the EM algorithm to the data in table 6. It's easily seen that different starting values used in the starting vectors lead to widely divergent results for the parameter estimates. All of the vectors which result are equiprobable since each represents a point on the ridge at the top of the likelihood. In this case the use of a vector consisting of all probabilities set equal to 0.5 equally splits the Z_{ij} into two sets with $X_{ij1} = X_{ij2}$, which leads immediately to a convergence point.

The last line in table 5 presents the results of the EM algorithm when the true parameter values are used as starting values. Again, the algorithm converges immediately, in this case to the true values. One would expect this result since the parameter values input would generate the correct X_{ijk} values which would then be used to estimate the parameter values, and these would be unchanged from the initial values.

Table 6 gives the results for the EM algorithm for the Z_{ij} values presented in table 4 and the same starting values presented in table 5. In this case the algorithm was changed so that the off-diagonal elements in the complete data (the X_{ijk}, $i \neq j$) were used in the X_{ijk} table for estimation of the parameters, rather than the conditional expected values. As can be seen from table 6, by supplying the off-diagonal elements (the values used were the correct X_{ijk} values used to tabulate to the Z_{ij} values), the algorithm converges to one of two points. Starting values where all values are less than or equal to 0.5 lead to convergence, and convergence to the same result but to a result that makes no sense. The result for starting values of .1, .3, or .5 lead to estimation of correct classification values of less than 0.5. But starting values above 0.5 lead to convergence at a single point, the correct point, with all correct classification rates above 0.5. As noted in the last section, there are two points of convergence, even though the off-diagonal values are fixed, because the likelihood function has a mirror image. Attempts to climb the likelihood (the method used by the EM algorithm) will lead to one peak or the other in the multidimensional bimodel distribution. A choice can be made between the two solutions as long as one would expect all correct classification rates to be greater than 0.5.

Examples of the convergence properties of the EM algorithm for the two restricted models described above are not given here because of space limitations. A more extensive version of this paper including the examples are available from the author.

7. CONCLUSIONS

It has been demonstrated that using a general model of misclassification, the parameters in the model are not usually estimable without some assumptions on the process by which misclassification occurs. As a minimum, one must either conduct a double sampling study to make estimates of appropriate allocations for each cell in the misclassification table, or one must assume independence between classification attempts and either double sample the off-diagonal cells or further restrict the model. It has also been shown (by counter-example) that not all restrictions of the parameter space will serve to make the problem estimable.

The really outstanding feature of the examples in this paper is that without the use of the EM algorithm, one could be severely misled about the proportion of cases that fall in a particular category. Modeling of the misclassification process can give a much better appreciation of the quality of the data being analyzed and should be attempted when the data are available.

8. REFERENCES

Birch, M.W. (1963) "Maximum Likelihood in Three-Way Contingency Tables," Journal of the Royal Statistical Society, Series B, 25:202-233.

Bishop, Yvonne M.M., Fienberg, Stephen E., and Holland, Paul W. (1975) Discrete Multivariate Analysis, The MIT Press, Cambridge, Massachusetts.

Chen, T. Timothy (1979) "Log-Linear Models for Categorical Data with Misclassification and Double Sampling," Journal of the American Statistical Association, 74:481-488.

Cowan, Charles D. (1984) The Effects of Misclassification on Estimates from Capture-Recapture, Unpublished doctoral dissertation, The George Washington University, Washington, D.C.

Cowan, Charles D. and Malec, Donald J. Capture-Recapture Models When Both Sources Have Clustered Observations, Proceedings of the Section on Survey Methods, American Statistical Association Meetings, 1984.

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

Nordheim, Eric V. "Inference from Non Randomly Missing Categorical Data: An Example From a Genetic Study on Turner's Syndrom," Journal of the American Statistical Association 79:772-780.

Press, S. James (1968) "Estimating from Misclassified Data," Journal of the American Statistical Association, 63:123-133.

Table 1: Parameters and Conditional Expected Values to be Estimated by the EM Algorithm

Parameter	Conditional Expected Value
1) $\hat{p}_1 = X_{++1}/N$	1) $\hat{X}_{111} = Z_{11}p_1f_{111}/(p_1f_{111} + p_2f_{112})$
2) $\hat{f}_{111} = X_{111}/X_{++1}$	2) $\hat{X}_{121} = Z_{12}p_1f_{121}/(p_1f_{121} + p_2f_{122})$
3) $\hat{f}_{121} = X_{121}/X_{++1}$	3) $\hat{X}_{211} = Z_{21}p_1f_{211}/(p_1f_{211} + p_2f_{212})$
4) $\hat{f}_{211} = X_{211}/X_{++1}$	4) $\hat{X}_{221} = Z_{22}p_1f_{221}/(p_1f_{221} + p_2f_{222})$
5) $\hat{f}_{112} = X_{112}/X_{++2}$	where $X_{++1} = \sum_i^2 \sum_j^2 X_{ij1}$
6) $\hat{f}_{122} = X_{122}/X_{++2}$	
7) $\hat{f}_{212} = X_{212}/X_{++2}$	and $N = \sum_k^2 X_{++k}$

Table 2: Parameters and Conditional Expected Values to be Estimated by the EM Algorithm/with Classification Attempts Independent Between Sources and Classification Parameters/Equal Across Sources But Differing Between Classes ($g_{1k}=h_{1k}$) for Dichotomous Case

Parameters	Conditional Expected Values
1) $\hat{p}_1 = X_{++1}/N$	1) $\hat{X}_{111} = Z_{11}p_1g_{11}^2/(p_1g_{11}^2 + p_2g_{12}^2)$
2) $\hat{g}_{11} = (X_{1+1} + X_{+11})/2X_{++1}$	2) $\hat{X}_{121} = Z_{12}p_1g_{11}g_{21}/(p_1g_{11}g_{21} + p_2g_{12}g_{22})$
3) $\hat{g}_{12} = (X_{1+2} + X_{+12})/2X_{++2}$	3) $\hat{X}_{211} = Z_{21}p_1g_{11}g_{21}/(p_1g_{11}g_{21} + p_2g_{12}g_{22})$
	4) $\hat{X}_{221} = Z_{22}p_1g_{21}^2/(p_1g_{21}^2 + p_2g_{22}^2)$

where the plus subscript denotes summation over categories for that indicator, and $N = \sum_k^r X_{++k}$

Table 3. Parameters and Conditional Expected Values to be Estimated by the EM Algorithm with Classification Attempts Independent Between Sources and Classification Parameters Equal When Classification is to Correct Category But Differing Between Sources (g_{kk} = s and h_{kk} = t) for Dichotomous Case

Parameters	Conditional Expected Values
1) $\hat{p}_1 = X_{++1}/N$	1) $\hat{X}_{111} = Z_{11}p_1st/(p_1st + p_2(1-s)(1-t))$
2) $s = (X_{1+1} + X_{2+2})/N$	2) $\hat{X}_{121} = Z_{12}p_1s(1-t)/(p_1s(1-t) + p_2(1-s)t)$
3) $t = (X_{+11} + X_{+22})/N$	3) $\hat{X}_{211} = Z_{21}p_1(1-s)t/(p_1(1-s)t + p_2s(1-t))$
	4) $\hat{X}_{221} = Z_{22}p_1(1-s)(1-t)/(p_1(1-s)(1-t) + p_2st)$

Table 4: Z_{ij} Values Used in First Example of EM Algorithm With Restriction: Independence Between Classification Attempts

		Second Data Source		
		Category		
		1	2	Total
First Data Source	Category 1	7289	691	7980
	Category 2	1031	989	2020
	Total	8320	1680	10000

Table 5: Results of EM Algorithm for Data in Table 4 for Different Starting Values

	Parameters					
Start	p_1	g_{11}	g_{22}	h_{11}	h_{22}	Number of Iterations
.10	.1089	.0705	.1017	.1131	.0787	7
.30	.1869	.2516	.3152	.0764	.0492	13
.50	.5000	.7980	.8320	.2020	.1680	1
.70	.8131	.9236	.9508	.7484	.6849	11
.90	.8911	.8869	.9213	.9295	.8983	6
True	.9	.88	.92	.94	.96	1

Table 6: Results of EM Algorithm for data in Table 4 for Different Starting Values with Double Sampling Values (X_{ijk}, $i \neq j$) Substituted for Conditional Expected Values

	Parameters					
Start	p_1	g_{11}	g_{22}	h_{11}	h_{22}	Number of Iterations
.1	.3253	.3906	.5013	.0056	.0086	26
.3	.3253	.3906	.5013	.0056	.0086	16
.5	.3253	.3906	.5013	.0056	.0086	41
.7	.9000	.8800	.9200	.9400	.9600	10
.9	.9000	.8800	.9200	.9400	.9600	6
True	.9000	.8800	.9200	.9400	.9600	1