AN EVALUATION OF CATEGORICAL DATA ANALYSIS METHODOLOGY WHEN APPLIED TO COUNTY ESTIMATES

Nancy J. Carter, California State University, Chico
Douglas C. Bond, U.S. Department of Agriculture

Introduction

Agricultural estimates at the county-level have been of interest for many years. They have generally been derived from population censuses, special surveys, or by using some nonprobability-based technique.

The need for improved methodology for setting county-level estimates stems from the fact that censuses and special surveys are usually very expensive. As a result, in many states, data for county-level estimates are collected from nonprobability surveys and the estimates are constructed by hand computation. Frequently, there is no sound statistical basis for the estimation techniques employed. For example, instead of using a probability-based approach, a bookkeeping type of method may be used, with the primary aim of this procedure being to avoid wide deviations from previous year estimates which were themselves the product of a similar procedure. As a result, there is usually no way to measure the precision of the estimates. Even in those states that have a large probability-based survey and computerized summary system, the process may be tedious and subjective. It is possible, given the methods and small sample sizes currently used, that the precision and accuracy of a number of county estimates are not good.

In recent years, the problem of deriving small area (such as county-level) estimates from survey data has been receiving increased attention. A number of new methods for estimation have been developed and evaluated by research statisticians in demography and health statistics. Noel Purcell in his 1979 Ph.D. dissertation (Purcell, 1979) used a categorical data analysis (CDA) approach to try to develop efficient estimators for small domains. The evaluation of this CDA approach to SRS county-level agricultural estimates was the subject of the research reported in this paper. First, the CDA method will be explained and the estimators introducted. Next an evaluation of the methodology will be given. Finally, the results will be summarized and recommendations given for future work.

Description of Categorical Data Analysis for Small Area Estimation

The most extensive study of CDA for small area estimation is presented in Purcell's thesis (1979). A summary of his work can be found in a paper by Purcell and Kish (1980). Purcell's notation will be used in the following discussion and report.

The CDA approach to county-level agricultural estimation was evaluated on data gatherd on harvested acreage in North Carolina for certain crops and land uses. Data have been collected in North Carolina for several years in a multiple frame, stratified, probability A & P survey designed to gather information from every county. A paper has been published by Ford (1981) on using these data to derive direct, synthetic, and composite estimates. Also, Ford, Bond, and Carter (1983) published a paper on further research using these data in a model that includes historical trends in acreage and production since 1972. Hence, a substantial amount of information has been gathered and evaluated (for other purposes) for North Carolina. In addition, a relatively recent census of agriculture (1978) was done in North Carolina. This, along with the other information just mentioned, made North Carolina a good state for evaluation of CDA estimation.

The CDA estimation approach requires two data structures: an association structure and an allocation structure. The association structure consists of data that are broken into categories of the variable of interest, crosstabulated by associated variables and small areas. These data are normally obtained at some previous time, usually from a census. The allocation structure consists of data, again broken into categories of the variable fo interest and crosstabulated by associated variables; but accumulated over small areas. These data are usually obtained from a current large scale survey. The allocation structure may include additional current information, such as data at the small area level accumulated over the categories of the variable of interest and of the associated variables.

For this research, the goal was to estimate the number of harvested acres for certain crops and land uses for each of the 100 counties in North Carolina. The association structure consisted of the 12000 cells of the crosstabulation of the categories $i(i=1,...,20)$ of the variable of interest, certain crops and land uses, by the categories $g(g=1,...,6)$ of the associated variable, farm size, by the counties, subscripted by h. The number of acres in each cell is denoted by N_{hig}. The allocation structure for a given year consists of a crosstabulation of crops and land uses by farm size, at the state level, obtained from the A&P survey for the particular year. Each cell of the allocation structure has a count $m_{.ig}$, where the dot denotes summation over a subscript. The allocation structure may include additional information on current accurate county-level data on total farmland.

To estimate X_{hi}, the number of harvested acres for the twenty categories of crops and land uses, the association structure is adjusted in such a way that all interactions of variables are preserved, except those that are respecified by the allocation structure (the crops and land use by farm size margin). Then the adjusted association structure counts, denoted by X_{hig}, are summed over the associated variable (farm size) to obtain the county by crop and land use margin, whose cells, X_{hi}, are the desired estimates.

There are a number of ways to adjust the association structure, depending on the amount of information available in the association and allocation structures. Three cases were investigated by this research project.

Case 1 A full association structure was used which consisted of the 1978 North Carolina Census of Agriculture data. The allocation structure for a given year consisted of estimates for the same categories at the state level. These estimates, as mentioned previously, came from the A&P survey for the particular year. By any of the three methods - minimizing a weighted sum of squares, maximum likelihood, or minimizing a discriminant information criterion - the following estimate is obtained for the adjusted association structure counts:

$$x_{hig} = \frac{N_{hig}}{N_{.ig}} m_{.ig}$$

Recall that N_{hig} = number of harvested acres from the 1978 census for a particular county, crop, and farm size; $N_{.ig}$ = total # of harvested acres in North Carolina, based on the 1978 Census, for a particular crop and farm size; and $m_{.ig}$ = # of harvested acres in North Carolina for particular crop and farmsize, based on the appropriate year's A&P survey.

Then the estimator of X_{hi}, the number of harvested acres for a particular crop or land use, at the county level is

$$X_{hi} = \sum_{g} X_{hig} = \sum_{g} \frac{N_{hig}}{N_{.ig}} m_{.ig} \ .$$

Case 2 Only an incomplete association structure is availalbe for this case. The association structure is a dummy structure, where total farmland data, crosstabulated by county and total farmland stratum (the hg margin of the association structure, are substituted at each level i (crops and land uses) of the association structure. These data came from the 1978 Census of Agriculture. Thus, for our problem, the 12000 cells of the association matrix were assumed to have counts $N_{h.g}$ where $N_{h.g}$ = total harvested acres for county h and farm size g. The allocation structure is the same as in Case 1 - statelevel A&P estimates for all the ig categories of the association structure.

The estimator for the adjusted association structure is formed for this case as:

$$x_{hig} = \frac{N_{h.g}}{N_{..g}} m_{.ig} \ .$$

Hence, the county-level estimator is

$$X_{hi} = \sum_{g} \frac{N_{h.g}}{N_{..g}} m_{.ig} \ .$$

Case 3 All of the information of Case 1 is available for this case. The association structure is the same as in Case 1, where Census of Agriculture data was used, and the allocation structure includes the information in the allocation structure of Cases 1 and 2. In addition, the allocation structure contains current accurate county-level data on total farmland. This came from the A&P survey for the current year. Estimators for the adjusted association structure are constructed using the method of iterative proportional fitting (Deming and Stephen, 1940). See Carter and Bond (1985) for a description of the exact procedure used. Then as in Cases 1 and 2, the resulting estimator X_{hig} is summed over the associated variable to obtain the county-level estimator:

$$X_{hi} = \sum_{g} X_{hig} \ .$$

This last case was of great interest since it utilized the most information of the three cases, and because it was the most accurate method in Purcell's application.

Evaluation of County Estimates

The three estimators described in the last section were determined and evaluated using the 1978 Census of Agriculture and the 1981, 1982, and 1983 A&P data.

An evaluation technique was needed in order to compare the three types of estimators. Methods of estimation of the variance-covariance matrix of the CDA estimates are given by Purcell. However, all of these methods are rather complex, and Purcell did not actually compute any variance-covariance matrices in his evaluation of CDA estimates. He points out that the variances of these estimates depend mainly on the variances of the allocation structure estimates, which can be controlled with sample design. The bias is therefore a more important source of error, and Purcell gives methods for estimating its size. The bias measure used for evaluation purposes in this report is the percentage absolute relative difference (%ARD). For this type of evaluation, the CDA county estimates are compared with values from a current census, or from independent sources, by computing the %ARD:

$$\%ARD = \frac{|x_{hi} - X_{hi}|}{X_{hi}} \times 100 \ ,$$

where x_{hi} = CDA county estimate, and X_{hi} = the "true" county-level value.

The %ARD formed the major part of the evaluation of the three estimators. The three estimators were compared with respect to the mean, median, and standard deviation of this measure across the 100 counties for seven different crops.

In addition, the Pearson product-moment correlation coefficients were computed to examine the degree of the relationship between the three different estimates and their respective "true" values.

As just stated, both the %ARD and Pearson correlation coefficients require the "true" county-level value before they can be computed. The official SRS county estimates were used as the "true" county-level values. The problem with using these estimates in the evaluation is that virtually none have check data for them. Only cotton, tobacco and peanuts can be verified by ASCS figures. Peanuts were planted in less than half of the 100 counties in all three study years. Therefore, despite the fact that the SRS estimates may be subjective and possibly inaccurate, with no measure of reliability, the CDA estimates were evaluated in comparison with these figures.

Results and Conclusions

On examining the results of both the %ARD and correlation analyses, it was clear the Case 1 and Case 3 estimators consistently and significantly outperformed the Case 2 estimator. This result agrees with the intuitive belief that a full association structure should be superior to a partial association structure. There was, however, no clear pattern in the results which demonstrated superiority between the Case 1 and Case 3 estimators. Since the Case 1 estimator is easier to compute and requires less information the logical conclusion is to recommend using Case 1 estimates when the necessary association and allocation structures are available. However, caution is warranted before such a recommendation is made. In Purcell's dissertation the Case 3 estimator was generally superior to the Case 1 estimator (especially in the later years of his analysis period (10 years)). The time span for this analysis was only five years and only included predictions for three years (81, 82, 83), using one census, and one state.

Generally speaking the CDA approach to agricultural county estimation seems promising. Further research needs to be done on the problem of nondisclosure, which results in missing cells in the association structure, and on the possibility of a composite estimator using the CDA estimates in combination with other estimators.

Bibliography

Carter, N.J. and D.C. Bond. (1985). "An Evaluation of Categorical Data Analysis Methodology for County Estimates in North Carolina, unpublished report, California State University, Chico, California.

Deming, W.E. and F.F. Stephan. (1970). "On a Least Squares Adjustment of A Sampled Frequency Table When the Expected Marginal Totals Are Known," Annals of Mathematical Statistics, 11, pp. 427-444.

Ford, B.L. (1981). "The Development of County Estimates in North Carolina," Statistical Reporting Service, USDA, Washington, D.C.

Ford, B.L., D. Bond, and N.J. Carter. (1983). "Combining Historical and Current Data to Make District and County Estimates for North Carolina," Statistical Reporting Service, USDA, Washington, D.C.

Purcell, N.J. (1979). "Efficient Estimation for Small Domains: A Categorical Data Analysis Approach," unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Purcell, N.J. and L. Kish (1980). "Postcensal Estimates for Local Areas (or Domains)," International Statistical Review, 48, pp. 3-18.