THE EFFECT OF A DISPROPORTIONATE, STRATIFIED DESIGN ON PRINCIPAL COMPONENT ANALYSIS USED FOR VARIABLE ELIMINATION

Robert D. Tortora, U.S. Department of Agriculture

Abstract

The effects of using unweighted or reweighted data from a disproportionate stratified survey design on a method based on principal component analysis to eliminate variables are studied. A model incorporating dummy variables to account for stratum membership is taken as the standard. The unweighted data added more unnecessary variables. For reweighted data, more pertinent variables were missed. In addition, the similarity between the complete and reduced data sets was low for the reweighted data. Thus, the analyst is cautioned to study the effect of preparing data before performing a multivariate analysis.

1. Introduction

Data from a sample survey that is primarily designed for descriptive statistics is often used for multivariate analysis. Typically, a large number of population parameters are estimated by these descriptive statistics. The survey design may be complex, i.e., not self-weighting and the observations must be appropriately weighted in order to obtain unbiased estimates of the parameters. For ease of computation, methods of adjusting the survey data to make the design self-weighting have been discussed by various authors, including Kish (1965) and Murthy (1967). These discussions have been limited to the problem of parameter estimation. However, the issue is also of concern in multivariate analysis.

Beddington and Smith (1977) have analyzed the problem of estimating the correlation matrix for sample designs. Their results indicate that with proportional allocation there will be little or no impact on the multivariate analysis. However, it is often the case that the analyst has data from a disproportionate design and must develop a model or uncover relationships for the entire population either by choice or by force (insufficient sample sizes per stratum for the number of explanatory variables). A model over all strata must be developed. Thus, the data analyst must select a procedure on which to base the analysis which should not bear on the final results. Various procedures are available to develop the model. They include (P1) ignore the disproportionate design and analyze unweighted data, (P2) reweight the data (Jones, Sheatsley and Stinchcombe, 1979) and proceed as if the sample is a simple random sample, (P3) introduce dummy variables (d.v.'s) to account for stratum membership (Draper and Smith, 1966), (P4) random elimination, (P5) random duplication, and (P6) random elimination and duplication of the data to obtain a self-weighting design (Kish, 1965).

Because the use of P4, P5, and P6 are dependent on the particular sample eliminated and/or duplicated we will only consider the first three procedures. Thus, only procedures that avoid replication of the results are considered.

The impact of the first two procedures, using P3 as the standard is measured in the sequel. It is important for the reader to understand that the author does not consider the use of the d.v. approach as optimal for disproportionate survey designs. The best approach, assuming sufficient sample size, would be to conduct the multivariate analysis in each stratum. However, if the sample size is not sufficient the d.v. approach is a reasonable, well understood alternative for the analysis. Comparisons will be made on actual survey data to study the effects of P1, P2, and P3 on discarding variables using a method based on principal component analysis, since one is often concerned with developing a model where a reduced number of variables account for most of the variation in the data. The desire is to obtain a model with only the pertinent variables. It would be unfortunate if the variables returned were in the model because of improper weighting (or absence of weighting) and not because they account for a large part of the variation.

We make the following assumptions:
1) the data is the result of a single stage, disproportionate, stratified, survey design,
2) there are insufficient observations within each stratum to conduct a separate analysis by stratum, and
3) the d.v. approach is the standard since it produces an "average" multiple regression over the strata (Kendall, 1975).

A description of the reweighting procedure, the d.v. approach, the method of variable elimination and the data used are discussed below. A comparison of the results obtained for the removal of redundant variables is made.

2. The Data

The data used in this paper was obtained from a survey of farmers and ranchers in North and South Dakota conducted by the National Opinion Research Center in cooperation with USDA (Jones, et.al., 1979). The primary purpose of this survey was to find out the farmers' and ranchers' understanding and attitudes towards crop and livestock surveys and to improve USDA's understanding of the data needs, concerns and motivation for farmer and rancher participation in surveys.

A disproportionate stratified sample of farmers and ranchers was drawn from the list frames in North and South Dakota. Table I gives more details about the sample design of the survey.

Table I

Sample Design Characteristics in North Dakota (ND) and South Dakota (SD)

Stratum	Population Size	Sample Size
1: SD large-scale livestock producers: operators with either 1,000 or more head of cattle and 400 or more head of hogs.	597	134
2: SD small-scale cattle producers: operators with no hogs and with fewer than 1000 head of cattle.	17,904	279
3: SD small-scale hog producers: operators with no cattle and with fewer than 400 head of hogs.	1,087	73
4: SD small-scale producers of both cattle and hogs: operators with 1-999 head of cattle and 1-399 hogs.	11,075	160
5: SD small-scale crop producers: operators with no livestock and fewer than 500 acres of crops	6,584	78
6: SD large scale crop producers: operators with no livestock and 500 or more acres of crops.	1,255	27
7: ND large-scale cattle producers: operators with 500 or more head of cattle.	518	138
8: ND small-scale cattle producers: operators with less than 500 head of cattle.	23,856	372
9: ND small-scale crop producers: operators with no (or unknown numbers of) cattle and fewer than 500 acres planted to crops.	9,463	125
10: ND large-scale crop producers: operators with no (or unknown numbers of) cattle and more than 500 or more acres planted to crops.	7,623	194

Notice that the design is heavily weighted toward the large operators. The sampling fractions in these strata (1,6,7,10) ranging from two to ten times larger than the sampling fractions for the small-scale operators. The sample size was adequate for parameter estimation within each stratum but not large enough to permit multivariate analysis in each stratum without subjectively eliminating variables.

In order to conduct various methodological studies two versions of the questionnaire were developed. There were several identical items on the two questionnaires, but each questionnaire explored some different areas and also allowed for the measurement of the effects of question wording and ordering. One version allowed the respondent to describe his past numerical participation rate (number of times responded to surveys/number of times asked to respond) during the previous year. The data from this version of the questionnaire is used for further analysis in this paper. Only those respondents who indicated that they had been asked to participate in at least one survey during the year prior to interview were included in the data set. A total of 630 producers were interviewed on this version of the questionnaire.

Nineteen variables (Table II) were considered for procedures P1 and P2. However, for P3, nine additional d.v.'s were added to account for the 10 strata in the sample design. The variables can be classified into the following categories: (1) six background information variables such as total number of cattle, total cropland acres etc., (2) thirteen Crop and Livestock Evaluation (CLE) variables such as source of agriculture information, usefulness of agricultural statistics, attitudes about confidentially of survey data, and (3) for procedure P3, the nine d.v.'s.

Table II

Variable Description

	Variable Number	Description
Background Information Variables	1	Age of farm operator
	2	Education of farm operator
	3	Total acres of cropland
	4	Total number of cattle
	5	Total number of pigs
	6	Total number of crops
Crop & Livestock Evaluation Variables	7	USDA divulge data to private company
	8	USDA divulge data to another government agency
	9	Number of sources of farm information
	10	Influence of farm organization on participation
	11	Impact of Crop and Livestock reports
	12	Use of Crop and Livestock reports by others aiding farmers
	13	Capability of Crop and Livestock reports to harm farmers
	14	Number of groups that use Crop and Livestock reports to harm farmers
	15	Why farmers and ranchers participate in surveys
	16	Usefulness of Crop and Livestock reports for farm management
	17	Who benefits most from Crop and Livestock reports
	18	Accuracy of Crop and Livestock reports
	19	Geographic use of Crop and Livestock reports

We separate the variables into these categories because the first relates farm and farm operator characteristics and are in a sense given. They cannot be impacted by any programs, say, to improve survey participation rates. On the other hand, changes in the CLE variables have the possibility of impact on survey participation. For example, if the confidentiality variable accounts for a large part of the variation it may be possible to improve the interview introduction and also initiate a public relations program to assure confidentiality with the hope of improving survey response rates. This second category represents the variables the analyst is often concerned with detecting since their importance can cause changes in management and fiscal policy towards improving survey participation.

3. Unweighted and Reweighted Data

Unweighted data is usually used when conducting a multivariate analysis when the data comes from a proportional allocation. However, the design may be disproportionate and the analysis conducted on unweighted data. If the variables associated with the model are dependent on stratum membership the under-or over- representation of certain subpopulations may effect the outcome of the analysis.

On the other hand, it is natural for the analyst to consider reweighting the data in attempting to avoid this under- or over- representation problem. For the purpose of this paper we will use the method of reweighting presented in Jones, et.al. (1979). Procedure P2 uses a method developed by Kish (p. 429, 1965) to measure the increase in variance caused by disproportionate allocation when proportionate allocation is optimum. Under the constraint that the reweighted sample size is equal to the raw sample size n, the relative efficiency of the sample is computed by using,

$$nV^2 = \Sigma\, W_h\, k_h'\, (1 - f/k_h')$$

where $W_h = N_h/N$, the stratum weight, k_h' equals the initial element weight, and f equals n/N, the overall sampling fraction. For the data described in Section 2 the relative efficiency

is .7655 or just over 75 percent of that of a proportionate sample of equal size. Final weights, i.e., those values attached to the data to reweight it, are the product of the initial weights and the relative efficiency of the sample. These values are summarized in Table III.

Table III

Initial and Final Weights for Data

Stratum	Initial Weights	Final Weights
1	.094	.072
2	1.349	1.033
3	.313	.240
4	1.455	1.114
5	1.774	1.358
6	.200	.158
7	.079	.060
8	1.348	1.032
9	1.591	1.218
10	.826	.632

A more detailed description of this procedure can be found in Jones, et.al, (1979).

Notice that the use of weighted data that produces unbiased estimates over the entire population is purposely omitted since these initial weights are close to the weights used in P2.

P2 allows for slightly easier computation of estimates of population parameters since it avoids computing estimates for each stratum and then combining these into an estimate for the population. Thus, P2 allows for the use of statistical software packages in which the data is assumed to come from a simple random sample. Unbiased estimates of variance are not obtained using these weights. So, even if the slightly easier procedure to estimate population parameters is used the estimation of variances can be troublesome. The design effect must be calculated in order to compute these estimates of variability.

If a multivariate analysis is being conducted the reweighted data may be used for the elimination of redundant variables. Does the reweighting have an impact on the final variables retained for further analysis? That is, are variables eliminated (or retained) because a reweighting procedure was used? The data analyzed and discussed in Section 6 indicate a positive answer to the question.

4. The Dummy Variable Approach

The dummy variable or pseudo-variable approach is useful for modeling when some of the independent variables are classifactory rather than continuous (Here the dependent variable might be the probability of a farm operartor participating in a survey.). Draper and Smith (1966) use this technique in regression analysis to account for data that occurs at two or more distinct levels. These variables then take account of the fact that separate deterministic effects are produced on these different levels. For K levels K-1 dummy variables are required. For example, suppose we have three strata from which responses have been obtained. Then two dummy variables, Z_1 and Z_2 say, are required to account for the strata. They are:

$$(Z_1, Z_2) = (1, 0) \text{ for stratum 1}$$
$$= (0, 1) \text{ for stratum 2}$$
$$= (0, 0) \text{ for stratum 3.}$$

Kendall (1975) has shown that these dummy variables produce a regression line with slope the weighted average of the lines, had regressions been calculated for each stratum. Thus, as Beddington and Smith recommend, it would be appropriate to conduct the analysis by stratum. Unfortunately, there is often an insufficient parameter to sample size relationship. Kendall (1978) suggests 10 observations per independent variable, to conduct such an analysis. Therefore, the use of dummy variables presents a viable alternative in this situation.

5. Variable Elimination and Principal Component Analysis (PCA)

Variable elimination is important to the data analyst because redundant or colinear variables are removed. Variables are often present that complicate the analysis yet do not provide additional knowledge. Thus, by eliminating these extraneous variables efficiencies are realized with a consolidated measurement instrument and with fewer variables to be analyzed, particularly as future investigations are conducted. In this paper a method described by Jolliffe (1972, 1973) is used for eliminating variables via a PCA. Jolliffe studied this method using 587 artificial data sets (1972) and four real data sets (1973). It was found to perform as well as or better than various other methods of variable elimination.

A PCA is performed on all p variables, and the eigenvalues inspected. If p' eigenvalues are less than 0.7 (a value determined empirically) the corresponding eigenvectors are considered in turn, starting with the eigenvector associated with the smallest eigenvalue, then the eigenvector corresponding to the second smallest eigenvalue and so on until all eigenvectors with corresponding eigenvalues less than 0.7 have been considered. One variable is then associated with each of the p' eigenvectors, namely that variable which has the largest coefficient in the eigenvector under consideration and which has not already been associated with a previously considered component. The p' variables associated with the p' eigenvectors are then eliminated. The remaining p-p' variables are retained for further analysis.

In order to compare principal components for the full and reduced sets of data the product moment correlations between the full and reduced set of data are computed (Jolliffe, 1973).

Suppose the entire set of data contains n observations measured on k variables $x_1, x_2, \ldots, x_k$. All analysis is done on the correlation matrix and the sample correlation r_{ij} between each pair of variables (s_i, x_j) are computed.

Any principal component is a linear combination of the variables in the set. For the entire set of p variables it can be written as

$$y = a_1 x_1 + \ldots + a_p x_p$$

where the a_i's are constant. For the reduced set of p-p' variables it can be written as

$$z = b_1 x_1 + \ldots + b_p x_p$$

where the b_i's are constant, but here all p' constants corresponding to eliminated variables are zero.

Using the n observations for y and z the correlation coefficient between them can be calculated, call it r. If the first k components are of interest for the full data set, then the similarity between components for the entire and the reduced data set is defined by

$$Q = \left(\sum_{i=1}^{k} q_i r(i) \right) / \left(\sum_{i-1}^{k} q_i 0 \right)$$

where r(i) is the maximum value of r between the i^{th} component for the full set of data and any component for the reduced set and q_i is the proportion of the total variation accounted for by the i^{th} component in the entire data set. So the similarity between components is the weighted sum of correlations between components and the weights are proportional to the amount of variation explained by the first few components of the entire data set.

6. Comparison of the Three Procedures on Variable Elimination

A PCA was conducted for each procedure. For the unweighted data 13 variables were retained, the PCA on the reweighted data retained nine variables, and the PCA when the 28 variables were included resulted in the retention of 19 variables. Table IV gives the variables retained by category of variable.

Table IV

Variables Retained by Category

Procedure	Background Information	Crop and Livestock Evaluation Variables	D.V.
P1	3 4 5 6	8 9 10 11 12 16 17 18 19	
P2	4 5 7	9 10 11 15 16 18	
P3	2 3 6	9 10 11 14 15 16 18 19	1 2 3 4 5 6 7 9

Comparing P1 with P3 we see that P1 retains 2 of the backgroud information variables that P3 retained, but adds two unnecessary background information variables. Six of the nine CLE variables retained by P1 match with P3. There are three variables retained in P1 that are not in P3 and also two variables (14 and 15) were missed by P1. Procedure P2 also has six variables matching with P3, it adds 2 (7 and 15) unnecessary CLE variables and missed two variables (14 and 19). Notice that d.v. eight was eliminated by the PCA. This combines strata 8 and 10, the small-scale cattle operations in ND.

Table V measures the similarity between all variables and the reduced set of variables by procedure. Nine components were used for comparison since P2 retained the fewest (9) variables.

Procedures P1 and P3 are nearly equivalent with a weighted average of correlations of .893 and .868, respectively. P2 falls dramatically

Table V

Measures of Similarity, r, Q, Between Components for all Variables and Reduced Set of Variables by Procedure

	P1	P2	P3
r_1	.827	.430	.891
r_2	.998	.352	.960
r_3	.998	.525	.984
r_4	.879	.531	.837
r_5	.872	.390	.903
r_6	.912	.735	.586
r_7	.986	.164	.859
r_8	.837	.973	.996
r_9	.788	.230	.731
Q	.893	.503	.868

below P1 and P3 with a weighted average of .503. Examination of the individual correlations for P2 indicates that the correlations for P1 and P3 are about twice as large as the correlations for P2 in six of the nine components.

Summarizing the results of the PCA we see that P1 matches 8, adds 5 unnecessary, and misses three variables and P2 matches six, adds three unnecessary, and misses five variables when compared to P3. The components retained by P2 are not as similar to the full data set as the components retained by P1 and P3.

7. Summary

The effect of the three procedures to prepare survey data for analysis were examined for a method of variable elimination based on principal component analysis. The data was obtained from a single-stage, disproportionate, stratified design and the analysis was conducted on (P1) unweighted data, (P2) reweighted data, and (P3) additional dummy variables to account for stratum membership (the standard).

It was found that P1 came closest to matching the variables retained in P3, but is also added the most extraneous variables. There was a high similarity between the complete and reduced data sets for P1 and P3 while P2 retained little similarity. Thus, the potential for effecting the results because of the procedure selected by the analyst prior to analysis is demonstrated. The procedure to prepare the data causes variables to be retained or eliminated without sufficient statistical justification. The author recommends for stratified, disproportionate designs where there are insufficient observations to conduct a multivariate analysis by stratum that the d.v. approach be used to obtain models for the entire population.

8. References

1. Beddington, A. and Smith, T.M.F. (1977). The Effect of Survey Design on Multivariate Analysis. Model Fitting. Edited by O'Muircheartaigh and Payne, Jong Wiley & Sons, New York.

2. Kish, L. (1965). Survey Sampling, John Wiley & Sons, New York.

3. Kendall, M. (1975). Multivariate Analysis, Hafner Press, New York.

4. Kendall, M. (1978). Personal communication.

5. Jolliffe, I.T. (1972). Discarding variables in a principal component analysis I: Artificial data. Applied Statistics, 21 160-173.

6. Jolliffe, I.T. (1973). Discarding variables in a principal component analysis II: Real data. Applied Statistics, 22, 21-31.

7. Jones, D., Sheatsley, P., and Stinchcombe, A. (1979). Dakota Farmers and Ranchers Evaluate Crop and Livestock Surveys. National Opinion Research Center Report No. 128, Chicago, Illinois.

8. Murthy, M.N. (1967). Sampling Theory and Methods, Statistical Publishing Society, Calcutta.