A MULTIVARIATE ANALYSIS OF FARM COSTS AND RETURNS SURVEY DATA

W. W. Donaldson, U.S.D.A., and R.E. Bargmann, University of Georgia
W.W. Donaldson, Rm. 4818 South Building, U.S.D.A.,Washington, D.C. 20250

KEYWORDS : Prediction, Transformation, Cross-validation, Variable selection

ABSTRACT

The United States Department of Agriculture annually conducts the Farm Costs and Returns Survey. This survey collects economic data at the individual farm level. There are approximately 500 possible responses per questionnaire. In this project, techniques of data transformation and multivariate analysis are used to impute responses for a subset of questions from the remaining responses on that questionnaire. The data base contains the responses for 4,500 1987 version 1 questionnaires. Relationships are developed for a homogeneous subgroup of data.

Summary

There exists a subset of 1987 Farm Cost and Returns Survey questionnaire items for which their mean totals can be imputed with sufficient accuracy. Techniques have been developed which permit analysis of very large data at a much reduced cost, with little increase in the time necessary to perform an analysis. Adaptations of the programs have been used to aid in the analysis of other survey data sets.

Introduction

Much of the information contained in the introductory paragraphs below was quoted or paraphrased from Staff Report Number AGES 89-1, 1987 Farm Costs and Returns Survey Data:Selected State and Region Highlights (1989).

The United States Department of Agriculture(USDA) annually conducts the Farm Costs and Returns Survey (FCRS). This survey collects economic data at the individual farm level. The survey is conducted in the 48 contiguous states. Data collection begins in mid-February and ends near the end of March. During this time, over 24,000 farms and ranches are contacted. The farms that are to be contacted are chosen from two sample frames: list and area. The list sample frame contains a list of most large farms and a less complete list of smaller farms. The area sample frame is used to augment the list sample frame. All farms in the population not included in the list sample frame are in the area sample frame.

Farms that are chosen are assigned a version of the FCRS questionnaire. In 1987, there were seven versions of the questionnaire. Version 1 contains questions about farm expenditures. Versions 2 through 7 contain questions about expenditures in addition to questions concerning specific crops. Approximately one-half of the sample receives the Version 1 questionnaire; the other versions of the questionnaire are distributed to the remaining farms.

There are nearly 500 items on a version 1 questionnaire. Some of the topics that are covered concern land uses, livestock and crop production, and farm production expenses.

Not every farm contacted will be included in the survey. A farm is included in the survey if "$1000 or more agricultural products were sold or would normally be sold during the year." Each farm that is included represents farms of similar size and type.

Statement of Problem

The purpose of this study was to determine if there exists a subset of version 1 questionnaire item responses that can be imputed from the remaining questionnaire item responses with minimal information loss.

Data

Complete 1987 FCRS version 1 list frame data were analyzed. This data set consisted of the responses for 4492 farms. Each questionnaire has approximately 500 items. For the average questionnaire, approximately 80% of the 500 responses are zero. The distribution of responses to most items was skewed to the right; for many questions the modal response was zero and the range of responses was very large. Responses of the 4492 farms were recorded as four-byte integers, and represented 10 megabytes of data.

Not all the questionnaire responses were analyzed. For an item to be included in the analysis three criteria had to be met:

1) five percent or more of the responses had to be non-zero;
2) the number of possible different responses had to be greater than two; and
3) responses had to be on at least an ordinal scale.

Responses to different items were combined if corresponding questions pertained to similar material, and one or more of the questions had very few non-zero responses at the national level.

Questionnaire responses were used to construct relatively homogeneous

groups. A homogeneous group is characterized by a set of farms that have similar response patterns. One advantage of such grouping is the possibility of including items which, in the complete data set, may have too few non-zero responses. It can also be expected that homogeneous groups of data would give rise to highly concentrated clusters of items. A typical subset consisted of 600 or more farms, and approximately 220 to 250 items were retained in such subsets for analysis.

Data sets were broken into two parts. One part was use to build relationships; the other part was used to test the performances of the models.

Hardware and Software Requirements

All computer work was done on 386 class personal computers. Originally, a Compaq Deskpro 386/16 was used for program development and applications. Later, a Compaq Deskpro 386/25 with 10 megabytes of RAM was used to implement the programs. A 386 class machine with two megabytes of RAM is the minimum needed for analyzing a data set of this size. Without the use of a virtual disk, one of the programs can take up to four hours to run to completion.

Software was developed using Borland's Turbo Pascal versions 4 and 5. Standard statistical packages were not used because of the size of data set. Programs may be run interactively or in batch mode.

Methodology

1. Scaling

The main purpose of this project was the detection of relationships among items. Some approximate normalization is obviously required for the construction of meaningful correlation and regression matrices. One of the oldest techniques [Fechner(1860)] was used for this purpose. It is somewhat similar to the Probit transformation. In effect, it assigns responses to each scalable item into a number of classes (20 or fewer were used in the present case) and divides a standard normal curve into slices with relative class frequencies equal to the observed ones. Raw scores falling into a given interval are replaced by the expected value under the corresponding slice of the standard normal distribution. The resulting scaled scores have mean zero, but variance less than one ("Sheppard correction").

2. Correlations and Item Clustering

Correlation matrices are obtained from the scaled scores, typically of order 220 by 220. Item clusters (clusters of variables) and predictor sets are identified on the basis of these correlation matrices. To obtain predominantly positive correlations, some of the items are reflected in sign, following a well known reflection technique [Thurstone(1948)]. Items that were reflected are clearly marked in the displays.

The cluster search program is interactive. At every stage the user has a choice whether a cluster should begin with a specified item, or whether the program should select a pair of items that has the largest correlation among those items not yet assigned to clusters. At each stage, a new cluster is started with one pair of items. The program searches the unassigned items to find the one which has the highest correlation with the center of the first pair. Detailed information on the three items (correlation matrix, description, number of non-zeros, etc.) is displayed to the user who is then asked whether the third item is to be accepted or rejected.

This human intervention is deemed essential in the analysis of datasets of this magnitude and heterogeneity. If, conceivably after several rejections, a third member is accepted by a user, the program will obtain the cluster center (variable that has equal correlation with the members of the cluster) and then, again, searches the remainder of the correlation matrix for that item which has the largest correlation with the cluster center. For a given cluster the number of items is limited to twelve.

At any stage, the user may terminate the cluster and then look at a detailed log of the session up to that point. The program permits interruption of the session and continuation at a later time, at the end of each terminated cluster. A special program prepares more detailed information on the items in a cluster (see Fig. 1 - 3).

3. Prediction

The program which identifies potential predictors for given predictands has a structure similar to the variable-cluster program. In the beginning it searches all items to find the one with the highest correlation with a specified predictand. The user may accept or reject this predictor. The program finds, among the remaining set, that variable which, together with the first accepted predictor, produces the highest multiple correlation with the predictand. Again, the user is prompted to decide on acceptance or rejection. As is well known in these stepwise procedures, some very unsuitable items are often chosen on the basis of some of the formalistic criteria for item selection; interactive

choice of acceptance or rejection is essential. Some items are rejected automatically, if their correlation with the predictand is very small, even though they produce maximum multiple correlation within the given predictor set. Details on each predictor set are available, very similar to the reports generated by the cluster program.

Independently, the user may now decide to use the foregoing information to obtain regression equations between selected predictors and a predictand. A separate program uses the regression equation thus obtained and applies it to the data in a different subset. (Odd- and even-numbered farm records were chosen as parallel sets). The predicted scaled scores, for each farm record, are then reconverted into raw scores, and distributions of observed and imputed scores are compared after some smoothing (at this stage, a Gamma distribution is fitted to the two data sets). Some selected comparisons are shown in the following section.

Results

The results for one homogeneous subgroup are presented. This group consists of farms that are similar in the number of farms they represent. The complete data set contains 1172 farm records.

For this data set, there are 238 items that are imputable. The information obtained from the predictor search program was used to identify 102 predictands and the corresponding predictor sets. These items were chosen as being imputable because the multiple correlation was reasonably high.

Of the 102 items chosen 19 had R^2 values greater than 0.90, 30 had R^2 values between 0.80 and 0.90, and 53 had R^2 values of less than 0.80. A high R^2 value did not guarantee good performance.

For the 102 items, comparison between predicted and actual responses showed that 25 had predicted means within 10 % of the actual mean. Of these 25 , seven had R^2 values greater than 0.90. Eighteen of the items had predicted means within 10 to 20 % of the actual mean. Of these, five had R^2 values greater than 0.9. The remaining 59 items had predicted means that differed from the actual means by more than 20% . Of these 59 predictands, 7 had R^2 values greater than 0.9. The following table summarizes the results.

Percent Predicted Mean differs from Actual Mean

	< 10%	10%< <20%	20%<
$R^2 > 0.9$	7	5	7
$0.8<R^2<0.9$	12	9	9
$R^2 < 0.8$	6	4	43

The number of items that had predicted means within 10% of the actual mean is distributed across all three R^2 categories. There seem to be two major reasons for this occurrence:

1) The variance of the variable to be predicted is small and

2) The item has few non-zero responses.

The predictor-search program constructed contingency tables which can be used to illustrate these phenomena. Ideally, the predictor-predictand pairs should form a tight cluster around the diagonal of the contingency table. Figure 1 depicts two variables that form a cluster close to ideal.

In figure 1 there is, as in most items, a high concentration of data in the upper lefthand corner, the remaining predictand-predictor pairs tend to cluster around the diagonal of the contingency table.

Figure 2 is representative of an item that has extremely skewed distributions in both the predictor and the predictand. In this set, five predictors were used for prediction. The R^2 value was equal to 0.7525 and the percent difference between the actual and predicted means was 7.17%.

Another example concerns a predictand-predictor set with an R^2 value that is very high, but the predicted mean is very different from the actual mean. Figure 3 shows the reason for this. The predictor-predictand pairs cluster into two groups. There is a very large number of zero responses in both variables. The non-zero responses also form a cluster; hence, the correlation for all responses is quite high. Within the non-zero cluster the relationship between the variables is not very strong.

The value of R^2 for this pair is 0.9308, for all responses. If only those response pairs that have non-zeros in both are included, the correlation drops to 0.587. For this pair, the difference between the predicted mean and the actual mean is 34.34%.

Detailed reports are prepared by one of the programs (GAMFIT), to compare observed and predicted raw scores for each item. Table 1 shows a part of this report comparing the observed and predicted values for a 1987 FCRS item

The "Standardized Median Difference" is a nonparametric analog of a pooled t statistic. One-half the distance between the 87'th and 13'th percentile is used as an equivalent of a standard deviation. If this is denoted by **s** , and numbers of non-zeros are denoted by n_o and n_p for observed and predicted values, then:

$$z = (Md_o - Md_p)/[1.57\{s_o^2/n_o + s_p^2/n_p\}]^{1/2}$$

(the factor 1.57 being included since the variance of a median of a sample from a standard normal distribution is approximately 1.57/n). An absolute value of z greater than 2 indicates that the two medians are significantly different.

The following table summarizes the breakdown of Standardized Median Differences(SMD).

	SMD => 2	SMD < 2
$R^2 > 0.90$	0	17
$0.80 < R^2 <= 0.90$	1	26
$R^2 <= 0.80$	8	29

For the majority of the imputed items the SMD is less than two. For some of the items a small SMD value is the result of a strong relationship between the predictand and predictor variables. In other instances the SMD appears to be favorably small because the variation of the predictand is large.

Conclusions

This report describes an on-going study of very large data sets. It shows the feasibility of using 386 class personal computers for analyses which would be extremely costly on a main frame. The following procedures have been performed on large data sets at minimal cost and execution time:

1) Construction of large covariance matrices.
2) Efficient identification of clusters and predictor sets among a large number of candidates.
3) Generation of detailed reports, and comparison of predicted and actual data.

These are procedures that cannot possibly be done manually, because of time restraints. By designing the programs to be used on a personal computer, the programs become more accessible; it is not necessary for the user to have access to a mainframe computer. It takes approximately one day to analyze 250 variables, with 1170 experimental units. Most of that time is spent searching for predictor variables. Computer run-time costs are minimal.

The developed programs may also be adapted to work with other survey data sets. The software is designed to work with a maximum of 495 variables. If the number of variables for an alternate data set does not exceed this limit, adaptation is quite simple.

For the chosen data set, some strong statistical relationships were identified. When using the Standardized Median Difference test statistic, for many items the predicted and actual medians were not statistically different. But it should be noted that one measure is not enough to determine the quality of prediction. Comparison of additional predicted and observed percentiles would be more informative.

REFERENCES

(1) Fechner, Elemente der Psychophysik, Berlin und Leipzig: Breitkopf und Haertel, 1860.

(2) Hogg, R.V. and A.T. Craig. Introduction to Mathematical Statistics. New York, New York: Macmillan Publishing Co., 1978

(3) Johnson, R.A. and D.W. Wichern. Applied Multivariate Statistical Analysis. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1988.

(4) Neter J., W. Wasserman, and M.H. Kutner. Applied Linear Regression Models. Homewood, Illinois: Richard D. Irwin, Inc., 1983.

(5) Thurstone, L.L. Multiple Factor Analysis. University of Chicago Press, 1947

(6) U.S. Department of Agriculture, National Agricultural Statistics Service, Economic Research Service. 1987 Farm Costs and Returns Survey Data: Selected State and Region Highlights. 1989

Figure 1

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	325	8	4	4	5		4	2	3	1		1	1		1	1	1	1		2	
1	2	5	3	1															1		
2			6	1	3	2															
3				7	3	1		1													
4					2	4	3	2			1										
5						4	2	3	2				1								
6						1	1	3	4	1		1			1						
7						1	3	2		3	2				1						
8	1								1	3	4	1		1				1			
9									1	2	2	4	2				1				
10									3	1	1	2	3	2							
11										1	2			2	3	1	2	1			
12										1	1	1	2	2		3	1				1
13												2	1	2	3	3	2	1	1		
14													3	1	3	3	1	1			
15														3		1	1	4	2	1	
16												1			1		3	2	2	2	1
17																1	1	2	5	3	
18																			2	4	6
19																				1	2

Correlation between non-zeros ≈ 0.9098

Figure 2

	0	1	2	3	4	5	6	7	8	9
0	434									
1	7	1								
2	6	1	1							
3	8	1								
4	6	1			1					
5	4	1	1	1			1	1		
6	6			1			1			
7	5		2				1			
8	6	1				2		1		
9	5		2		1					
10	4				2			1	1	
11	5					1		1	1	
12	5						2		1	
13	5				1			1		1
14	5			1		1	1			
15	5			1		1		1		
16	4								1	3
17	4			2					1	1
18	5					1			1	1
19	3				1					

Correlation between non-zeros ≈ 0.5780

Figure 3

	0	1	2	3	4	5	6	7	8	9
0	528									
1		2	2	1	1					
2	2		1	1	1				1	
3			1	3	3		1			
4		1	1		1	1	1	1		
5	1	1		1			1	1	1	
6						2	1	3		
7	1		1			3				1
8		1					1	1	3	
9		1					1	1	1	2
10										2

Correlation between non-zeros ≈ 0.5870

Table 1

Sample Size = 586

Number of farms with non-zeros in the observed set = 60
in the predicted set = 52

RAW SCORES

(Numbers in parentheses are ranks corresponding to 5'th, 10'th...95th percentiles of non-zero responses.
Note that observed zero responses rank from 1 to 526, predicted from 1 to 534)

OBS:	1000(528)	2200(531)	8000(534)	14000(537)	15000(540)
PRD:	6497(536)	6497(538)	22007(541)	22007(544)	22007(546)
OBS:	18000(543)	20000(546)	25800(549)	30000(552)	37435(556)
PRD:	22007(549)	36365(552)	36365(554)	36365(557)	36365(560)
OBS:	41000(559)	52000(562)	60000(565)	75000(568)	90000(571)
PRD:	51288(562)	51288(565)	51288(567)	51288(570)	91321(573)
OBS:	99000(574)	120000(577)	158872(580)	240000(583)	
PRD:	91321(575)	123560(578)	123560(581)	123560(583)	

Comparison between observed and predicted values
SMOOTHED SCORES (GAMMA)

OBS:	9887(.05)	11297(.10)	12914(.15)	14773(.20)	16916(.25)
PRD:	12662(.05)	14028(.10)	15556(.15)	17270(.20)	19198(.25)
OBS:	19393(.30)	22266(.35)	25616(.40)	29542(.45)	34177(.50)
PRD:	21375(.30)	23844(.35)	26659(.40)	29889(.45)	33621(.50)
OBS:	39692(.55)	46329(.60)	54425(.65)	64485(.70)	77301(.75)
PRD:	37976(.55)	43113(.60)	49263(.65)	56763(.70)	66149(.75)
OBS:	94226(.80)	117839(.85)	154066(.90)	222408(.95)	
PRD:	78324(.80)	95028(.85)	120202(.90)	210892(.95)	

WEIGHTED MEANS

Observed	Predicted
6725.57	5412.50

Median,obs = 37435.000 Median,pred. = 36365.00000

1/2 (P87-P13),obs. = 59724.000 1/2 (P87-P13),pred. = 58531.500

Standardized Median Difference = 0.076