Paper 261-26

Variable Reduction for Modeling using PROC VARCLUS

Bryan D. Nelson, Fingerhut Companies Incorporated, Minnetonka, MN

ABSTRACT

Most direct mail and e-commerce companies have hundreds if not thousands of variables for each customer on their database. When statisticians try to build segmentation or other types of models with a large number of variables, it becomes difficult to figure out the correct relationships between the dependent and independent variables. In fact, when redundant variables are included in some of the model building procedures, the model can degrade the segmentation model by: destabilizing the parameter estimates, increasing computation time, confounding the interpretation, and increasing the amount of time spent building a segmentation model. PROC VARCLUS can help a statistician quickly reduce the number of variables used to build a segmentation model. PROC VARCLUS will cluster variables - it will find groups of variables that are as correlated as possible among themselves and as uncorrelated as possible with variables in other clusters. The algorithm used by PROC VARCLUS is binary and divisive - all variables start in one cluster. If the second eigenvalue is above the current threshold (i.e. there is more than one dominant dimension) then the cluster is split. By default, PROC VARCLUS does a non-hierarchical version where variables can be reassigned to other clusters.

INTRODUCTION

When there are hundreds or even thousands of variables that can be used to create segmentation models, it becomes difficult to determine the correct relationships between variables. Some of the variables are highly correlated with one another. Including these highly correlated variables in the modeling process often increases the amount of time spent by the statistician finding a segmentation model that meets Marketing and business needs. In order to speed up the modeling process, the predictor variables should be grouped into similar clusters. A few variables can then be selected from each cluster - this way the analyst can quickly reduce the number of variables and speed up the modeling process.

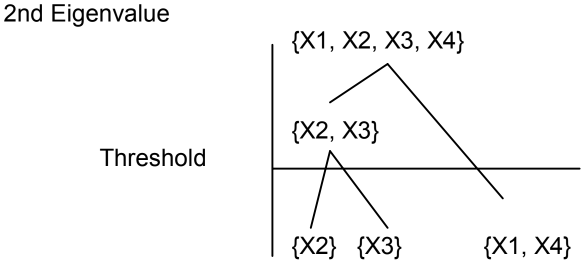
DIMENSION REDUCTION

In high dimensional data sets, identifying irrelevant inputs is more difficult than identifying redundant inputs. A good strategy is to first reduce redundancy and then tackle irrelevancy in a lower dimension space.

PROC VARCLUS is closely related to principal component analysis and can be used as an alternative method for eliminating redundant dimensions (SAS/STAT® User's Guide, page 1642). This type of variable clustering will find groups of variables that are as correlated as possible among themselves and as uncorrelated as possible with variables in other clusters. If the second eigenvalue for the cluster is greater than a specified threshold, the cluster is split into two different dimensions.

The reassignment of variables to clusters occurs in two phases. The first is a nearest component sorting phase, similar in principle to the nearest centroid sorting algorithms described by Anderberg (1973). In each iteration, the cluster components are computed and each variable is assigned to the component with which it has the highest squared correlation (SAS/STAT® User's Guide, pages 1642-1643). The second phase involves a search algorithm in which each variable in turn is tested to see if assigning it to a different cluster increases the amount of variance explained. If a variable is reassigned during the search phase, the components of the two clusters involved are recomputed before the next variable is tested (SAS/STAT® User's Guide, pages 1642-1643).

Divisive Clustering



Note: If the second eigenvalue is larger than the specified threshold, more than one dimension exists in the cluster. In the above example, the initial cluster is broken up into three different clusters.

Larger eigenvalue thresholds result in fewer clusters, and smaller thresholds (such as one or less) yield more clusters. To account for sampling variability, smaller values such as .7 have been suggested (Jackson 1991).

PROC VARCLUS SPECIFICATIONS

As with many other SAS procedures, PROC VARCLUS has a countless number of different specifications. The basic specification is:

```
PROC VARCLUS MAXEIGEN = .7
OUTTREE=FORTREE  SHORT;
VAR PREDICTORVARIABLES; RUN;
```

The maxeigen value is the threshold for identifying additional dimensions within a cluster. The outtree=fortree option creates a data set which can be used in PROC TREE to print a tree diagram. Short suppresses printing of the cluster structure, scoring coefficients, and intercluster correlations.

OUTPUT

Once completed, the output will include the total number of clusters created, the number of variables used in the analysis, the number of observations, and the maxeigen threshold used to create the clusters. Below is an example of the cluster output from PROC VARCLUS.

```
Oblique Principal Component Cluster Analysis
9997 Observations      PROPORTION  =       0             28 Variables
MAXEIGEN    =  1.2

                           Cluster summary for 1 cluster(s)

                                 Cluster    Variation    Proportion    Second
        Cluster     Members    Variation   Explained    Explained    Eigenvalue
        -------------------------------------------------------------------------
          1            28          28.0000    3.03272       0.1083        2.8960


          Total variation explained = 3.032724 Proportion = 0.1083
          Cluster 1 will be split.

                           Cluster summary for 2 cluster(s)

                                 Cluster    Variation    Proportion    Second
        Cluster     Members    Variation   Explained    Explained    Eigenvalue
        -------------------------------------------------------------------------
          1            17          17.0000    2.90641       0.1710        2.0299
          2            11          11.0000    2.57240       0.2339        1.4742

             Total variation explained = 5.478813 Proportion = 0.1957
```

The end of the output will show the number of final clusters PROC VARCLUS has created. PROC VARCLUS will also show which variables have been assigned to the various clusters. Below is an example of the output:

```
                          R-Squared    R-Squared
                            Own          Next         1-R**2
Variable                  Cluster      Closest        Ratio
Cluster 1                 ------------- -------------  ----------
Variable 1                  0.3654       0.1324       0.7314
Variable 2                  0.6832       0.6618       0.9368
Variable 3                  0.6477       0.4387       0.6277
Variable 4                  0.7268       0.4294       0.4787
Cluster 2                 ------------- -------------  ---------
Variable 5                  0.8923       0.4287       0.1886
Variable 6                  0.5860       0.5562       0.9329
Variable 7                  0.7142       0.5832       0.6856
```

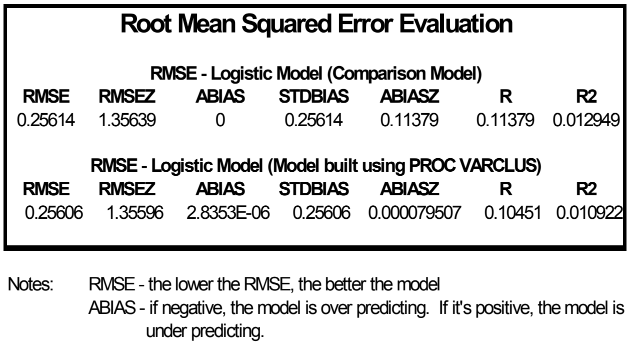
The analyst can then begin selecting variables from each cluster - if the cluster contains variables which do not make any sense in the final model, the cluster can be ignored. A variable selected from each cluster should have a high correlation with its own cluster and a low correlation with the other clusters (Logistic Regression Modeling, pages 56-57). The 1-R**2 ratio can be used to select these types of variables. The formula for this ratio is:

$$1\text{-}R^{**}2 \text{ ratio} = \frac{1-R^2 \text{ own cluster}}{1-R^2 \text{ next closest}} = \frac{1-\uparrow}{1-\downarrow} \Rightarrow \frac{\downarrow}{\uparrow} \Rightarrow \downarrow$$

If a cluster has several variables, two or more variables can be selected from the cluster.

RESULTS

The PROC VARCLUS method was used to build a response model for one of Fingerhut's affiliates. Customer characteristics from a historical mailing were used to predict the result of that mailing (purchase or no purchase). After using the PROC VARCLUS method for paring down the number of variables, various variable selection techniques were used to build the final segmentation model. This model was compared against the traditional method used by Fingerhut to build a segmentation model. Fingerhut's traditional method currently does not take into account a method similar to PROC VARCLUS. The model built using the PROC VARCLUS method was built four times faster than the traditional method. Even though less time was used to build the model, the model built using PROC VARCLUS segmented customers just as well as the traditional segmentation model. PROC VARCLUS was much faster since it eliminated many highly correlated variables from the modeling process. Below are some of the results:



Root Mean Squared Error Evaluation

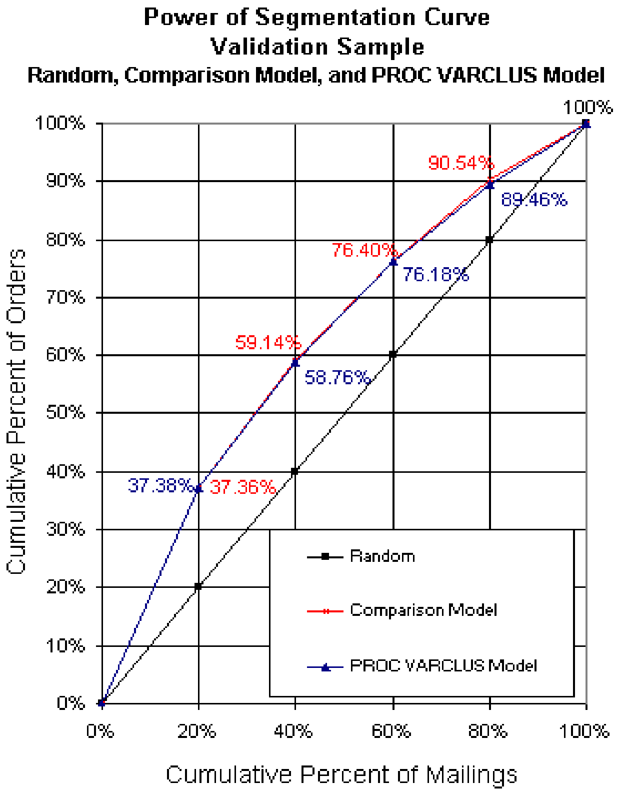
RMSE - Logistic Model (Comparison Model)

RMSE	RMSEZ	ABIAS	STDBIAS	ABIASZ	R	R2
0.25614	1.35639	0	0.25614	0.11379	0.11379	0.012949

RMSE - Logistic Model (Model built using PROC VARCLUS)

RMSE	RMSEZ	ABIAS	STDBIAS	ABIASZ	R	R2
0.25606	1.35596	2.8353E-06	0.25606	0.000079507	0.10451	0.010922

Notes: RMSE - the lower the RMSE, the better the model
ABIAS - if negative, the model is over predicting. If it's positive, the model is under predicting.



Power of Segmentation Curve
Validation Sample
Random, Comparison Model, and PROC VARCLUS Model

Note: If 40% of the customer list was mailed and no response model was used, only 40% of the buyers would be selected. However, if the comparison or PROC VARCLUS model were used to select the best 40% of the customer list, this would capture approximately 60% of the total number of buyers.

CONCLUSION

By using PROC VARCLUS, statisticians are able to build segmentation models quickly without any deterioration in the quality of the model. PROC VARCLUS helps to reduce the redundancy of variables which are used to build new segmentation models - this helps to reduce the amount of time the statistician needs to finalize the model. If several models are built on the same universe, each model could be built from the variables selected from PROC VARCLUS. PROC VARCLUS can also help determine whether or not new variables that are created by the modeling area are unique variables.

BACKGROUND

By default, PROC VARCLUS begins with all variables in a single cluster (SAS/STAT® User's Guide, page 1651). It then repeats the following steps:

1. A cluster is chosen for splitting.
2. The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation, and assigning each variable to the rotated component with which it has the higher squared correlation.
3. Variables are iteratively reassigned to clusters to maximize the variance accounted for by the cluster components.

Execution Time

n = number of observations
v = number of variables
c = number of clusters

The time required by PROC VARCLUS to analyze a data set varies greatly - it depends on whether centroid or principal components are used as well as other specifications along with the number of variables in the data set. The amount of memory needed (in bytes) is approximately equal to v2+2vc+20v+15c.

REFERENCES

SAS Institute Inc. (1999), *Logistic Regression Modeling Course Notes*, Cary, NC: SAS Institute Inc.

Jackson, J.E. (1991), *A Users Guide to Principal Components*, New York, John Wiley & Sons

SAS Institute Inc. (1999), *SAS/STAT® User's Guide*, Cary, NC: SAS Institute Inc.

Anderberg, M.R. (1973), *Cluster Analysis for Applications*, New York: Academic Press, Inc.

CONTACT INFORMATION

Bryan D. Nelson
Business Intelligence
Fingerhut Companies, Inc.
4400 Baker Road
Minnetonka, MN 55343
tel: (952) 936-5286
email: bryan.nelson@fingerhut.com