Abstract

This discussion covers the processing and aggregation of consumer panel diary data and briefly introduces how such data can be used to describe a market using the SAS system. The discussion begins with an introduction to panel data and then covers some data processing techniques. Next, the discussion turns to the kinds of inferences one can draw about households in the market, and the kinds of inferences one can draw about each of the brands. A final section focuses on more advanced techniques, which are illustrated by building a product space.

Introduction

The level of aggregation of a set of data in large part determines the type of inferences which can be drawn from that data. In general, high levels of data aggregation, such as weekly sales totals, permit only the most general types of model building. As an example, one can build a market response model which predicts share as a function of unit price using weekly scanner data.

There are a set of important questions, however, which cannot be answered at such a high level of aggregation. For example; what type of people tend to buy our product? How deal-prone are the buyers of the larger packages we sell? Does our 30% market share stem from the fact that 30% of the market are loyal purchasers, or from the fact that everybody in the market buys our brand 30% of the time? What is our penetration rate? The repeat purchase rate? What types of switching patterns are exhibited by consumers and what does this imply about the way the brands are perceived by the market? In the following paper I would like to demonstrate how SAS can be used to answer such questions.

Typical consumer panel data come in rectangular files with one observation for each purchase event. A typical panel file might include the following variables:

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIABLE NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household ID</td>
<td>ID</td>
</tr>
<tr>
<td>Week and Day</td>
<td>TIME</td>
</tr>
<tr>
<td>UPC for Purchased Item</td>
<td>UPC</td>
</tr>
<tr>
<td>Units Purchased</td>
<td>UNITS</td>
</tr>
<tr>
<td>Volume Purchased</td>
<td>VOLUME</td>
</tr>
<tr>
<td>Price Paid</td>
<td>PRICE</td>
</tr>
<tr>
<td>Coupon Origin (if any)</td>
<td>CORIGIN</td>
</tr>
<tr>
<td>Coupon Value (if any)</td>
<td>CVALUE</td>
</tr>
<tr>
<td>Item on Feature? (1=Y, 0=N)</td>
<td>FEATURE</td>
</tr>
<tr>
<td>Item on Display? (1=Y, 0=N)</td>
<td>DISPLAY</td>
</tr>
<tr>
<td>PriceCut (if any)</td>
<td>PRICEC</td>
</tr>
</tbody>
</table>

Frequently the first step in analyzing panel data is to decide which UPCs are of primary interest. Once this decision has been made, one can use the SAS data step to extract a subset of the panel records which involve those UPCs. Such an extraction is easily achieved, as we see below:

```sas
PROC SORT DATA=ORIGINAL.PANBL OUT=SORTED;
    BY UPC;
    DATA UPXS;
    INPUT UPC @@ ;
    CARDS;
    1000010000 1000010001 1000010002
    1000010003 9000010001 9000010002
    DATA EXTRACT;
    MERGE SORTED UPCSIN=WANTED ;
    BY UPC ;
    IF WANTED ;
    DOLLARS = UNITS*PRICE ;
```

We begin by sorting the panel data. Then, a DATA step reads in the variable UPC. These UPCs are the ones which we are interested in grabbing from the sorted panel data. The second DATA step subsets the panel data, using only the relevant UPCs. Now, the SAS data set EXTRACT contains the panel records for purchases which correspond to the five relevant UPCs.

Simple Inferences about Households

Many of the questions listed at the beginning of this article require that the data be reorganized from one record per purchase occasion, to one record per household or family. This can be easily done with either a DATA step or with PROC SUMMARY. In this section we illustrate this process by highlighting PROC SUMMARY and the use of the variables we can create using it.

The following code illustrates the latter technique. Let us assume that we are running this job for the brand manager of a brand which corresponds to UPC 1000010001. In that case, we begin by further subsetting the EXTRACT data set.

```sas
DATA MYBRAND ; SET EXTRACT ;
    IF UPC=1000010001 ;
```

The next step is to run PROC SUMMARY against data set MYBRAND.
PROC SUMMARY DATA=MYBRAND;
CLASS ID;
VAR UNITS VOLUME PRICE CVALUE FEATURE DISPLAY;
OUTPUT OUT=FAMILIES
MEAN=MUNITS MVOL MPRICE MCVALUE MFEAT MDISP
SUM=SUNITS SVOL SPRICE SCVALUE SPEAT SDISP;

PROC SUMMARY will create a new dataset called FAMILIES which will contain highly useful information. To begin with, the number of cases in the data set FAMILIES tells us the penetration rate for the brand in question. If the original panel consisted of a sample of 1,000 families, and there are 300 cases in the FAMILIES data set, the penetration rate for the time period of the panel is 30 percent. The variable MUNITs provides the average number of units purchased by each family when this brand was bought. One could easily run PROC MEANS against the FAMILIES data set to discover an overall average for the number of units bought per family when the brand in question is purchased. Note that a brand with a high MUNITs score but a low penetration would ordinarily be called a niche or specialty brand, while a brand with a high penetration rate but a low MUNITs score is probably a change-of-pace brand.

The variable MPRICE tells you the average purchase price paid per family. Since the original FEATURE variable is coded as 1 if the brand was featured, and 0 if it was not, the variable MFEAT tells you the proportion of purchase occasions each family bought when the brand was on FEATURE. If you sorted the FAMILIES data set by MFEAT, you would have at the beginning of the data set the families who were unlikely to buy on feature, and at the end of the data set, the families who were most 'feature-prone.' Of course the same technique could be used for the price-cut (MPRICE-C) variable, the coupon (MCVALUE) and the display (MNDISP) variable. For example, sorting on the MCVALUE variable lets you identify which families pick up on your coupon drops. Such information can be used as input to other statistical procedures, such as regression and so forth.

So far we have used the MEAN= variables which we created by PROC SUMMARY. Next, we consider sum variables such as SUNITS. SUNITS provides the total number of units bought per family during the time period of the panel. A simple bar chart of SUNITS would be informative with respect to the purchase rates occurring in the market. For example, if the panel data cover one year, than SUNITS=12 implies a purchase rate of once per month.

Sorting on SUNITS allows you to determine usage segments and then use this information in other models. If demographics (or psychographics, etc) are provided with the panel data, you can merge this information with the FAMILIES data set as follows:

DATA ID;
MERGE FAMILIES (IN=WANTED)
MASTER.DEMOS
IF WANTED;
IF _TYPE_ NE 0;
PROC CORR;
VAR SUNITS;
WITH INCOME AGE etc;

Here we can use PROC CORR to find out which demographic variables correlate with high usage of our brand. As an alternative, we could divide the sample into two groups using the median on SUNITS and compare the two groups on the several demographic variables.

Simple Inference about Brands

The basic unit of analysis in the preceding section was the household, or family. Now let us consider some analyses with respect to the brands themselves. To do this, we need to return to our original EXTRACT data set and again process it using PROC SUMMARY. Only now our class variable will be the UPC rather than the family ID.

PROC SUMMARY DATA=EXTRACT;
CLASS UPC;
VAR UNITS VOLUME PRICE CVALUE FEATURE DISPLAY DOLLARS;
OUTPUT OUT=BRANDS
MEAN=MUNITs MVOL MPRICE MCVALUE MFEAT MDISP MDOLLARS;

If one were to simply use PROC PRINT to dump the BRANDS data set, the variables MUNITs, MPRICE, MCVALUE, and MDOLLARS would show you the average units per purchase occasion, average price paid, average average coupon value per shopping trip. and the average amount spent for each brand per trip. If FEATURE and DISPLAY were coded 0 and 1, MFEAT and MDISP would provide the proportion of purchases made on feature and display for each brand.

While this information is useful, sometimes we are not interested in what happens during an average trip, but instead in the overall performance for the brand. Then we should use units as a FREQ variable as below:

382
PROC SUMMARY DATA=EXTRACT;
CLASS UPC;
FREQ UNITS;
VAR PRICE CVALUE FEATURE
DISPLAY DOLLARS;
OUTPUT OUT=BRANDS2
MEAN= MSPRICE MCVALUE MFEAT
MDISP MDOLLARS;
SUM = SPRICE SCVALUE SF'EAT
SDOLLARS;

The variable MPRICE contains the average unit price for each brand, and MCVALUE contains average unit coupon redemption. MFEAT and MDISP provide the percentage of units sold on feature and display. SDOLLARS provides the gross revenue for each brand. Of course, you could divide the SDOLLARS figure for each brand by the SDOLLARS figure for the _TYPE_ 0 record to arrive at the dollar share for each brand.

Some Advanced Inferences

This section describes some more advanced techniques for using information. We start off with a technique to infer the product space for the set of brands which we extracted. This example requires the use of PROC ALSCAL, which is part of the SUGI library.

The dimensions or axes of a product space indicate the product attributes to which consumers attend. As such, the space can be used for positioning. A more thorough discussion of this technique can be found in Green, Carmone and Smith (1989), while an alternative technique based on clustering has been presented by Fraser and Bradford (1983).

The first step required to build the product space is to build a two way table which cross tabulates the brand bought on purchase occasion t with the brand bought on purchase occasion t+1. Here we require a matrix of row conditional probabilities, where the rows represent any diary entry except the last for each family, and the columns represent the brand of their next entry. In other words, aggregating across all diary entries for all families, if brand A was purchased on a particular occasion, what is the probability that brand B was purchased on the next purchase occasion? Such a matrix is called a transition matrix. A sample is given below:

<table>
<thead>
<tr>
<th>Occasion t</th>
<th>Occasion t + 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>pCA</td>
</tr>
<tr>
<td>B</td>
<td>pCA</td>
</tr>
<tr>
<td>C</td>
<td>pCA</td>
</tr>
</tbody>
</table>

A typical entry in the transition matrix, say PCB|A), is the probability of a purchase of B on occasion t+1 given that a purchase of A had been made on occasion t. A combination of SORT and DATA steps, and PROC FREQ can be used to create such a table. To start, sort EXTRACT:

PROC SORT DATA=EXTRACT OUT=PREFREQ;
BY ID TIME;

At this point we are ready to process PREFREQ, which now contains the ordered purchase histories of each family.

DATA PREFREQ;
KEEP OCC1 OCC1P1;
SET PREFREQ;
BY ID;
LAGBRAND = LAG(UPC);
IF NOT FIRST.ID THEN DO;
OCC1 = LAGBRAND;
OCC1P1 = UPC;
OUTPUT;
END;

It should be noted that this code will not work if there are any zeroes in the transition matrix. The data set PREFREQ has two important variables; OCC1 (occasion t) and OCC1P1 (occasion t plus one). PROC FREQ will produce conditional probabilities for us once PREFREQ has been appropriately sorted.

PROC SORT DATA=PREFREQ OUT=PREFREQ2;
BY OCC1;
PROC FREQ DATA=PREFREQ2;
BY OCC1;
TABLES OCC1P1 / OUT = MAT1 NOPRINT;
PROC TRANSPOSE DATA=MAT1 OUT=MAT2;
VAR PERCENT;
BY OCC1;

The transition matrix has now been captured in data set MAT2, which is available for mapping. In the data set, the columns of the table will take on the variable names COL1, COL2, etc. Each refers to a different brand as well. In markets where there is only a small number of brands, the table can be visually inspected for patterns. As an example, if the larger conditional probabilities can be found within the same product format, it would then seem that the market is 'format primary', which simply means that consumers first decide which format to buy, and then
switch around within a format either for variety or because of deals. Conversely, if the larger conditional probabilities can be found within the same manufacturers brands, we can conclude that the market is "manufacturer primary".

When there are large number of brands, or when the pattern of probabilities is not so obvious, it becomes necessary to build a picture of the market in the form of a product space. A product space can be created using multidimensional scaling. There is a SUGI procedure called ALSCAL which is ideally suited to this end. To utilize ALSCAL on the MATRIX data set, we would utilize the following code:

```
PROC ALSCAL DATA=MATRIX LEVEL=ORDINAL
DIMENS=2 SIMILAR SHAPE=ASYMMETR
MODEL = ASYMCAL ;
VAR COLL = COLn ;
```

You would replace the n with the number of brands in the EXTRACT data set. Here we have specified two dimensions, but in some markets more than two dimensions are required.

The transition matrix (MAT2) can also be used for stochastic modeling. A thorough, but mathematically advanced, introduction to stochastic modeling is given in a book by Massey, Montgomery & Morrison (1970). Such models provide quantitative estimation for brands of variety seeking, loyalty, learning, competitive structure, and the degree of heterogeneity of taste in the market. While we have used PROC SUMMARY to aggregate across families and therefore provide brand summary information, there are also advanced techniques which rely on family data aggregated over purchase occasions.

As an example, consider a rectangular matrix of 0's and 1's where the rows of the matrix are panel families, the columns are the brands of interest, and an entry in the array is assigned a zero if the family never purchases the brand and the entry is given a one if the family purchased the brand at least once during the period under investigation. One technique appropriate for this type of data is pick-any analysis, discussed by Holbrook, Moore & Winer (1986). Note that one could easily create such a data set by transforming N to 0 or 1, and by using PROC TRANSPOSE to reshape the FAMILIES data set.

Now suppose we change the matrix somewhat so that each element represents the number of purchases for each brand for each family. Correspondence analysis is one type of analysis appropriate to such data (see Hoffman & Franke 1985 for an introduction). Here we note that PROC CORRESP is available for this technique. Note here that we could use the N variable from the FAMILIES data set 'as is', followed by reshaping through PROC TRANSPOSE.

The variable N in the dataset FAMILIES can also be used for more advanced model fitting techniques. In order to utilize it as such one can use PROC DISCRETE (Geaghan, Gates & Williams 1983). It has been found that the purchase incidences of particular brands frequently follow the negative binomial distribution. One can then compare the actual distribution of the variable N with the theoretical distribution under the negative binomial distribution. This comparison allows you to detect when there is too little repeat buying, which is of course especially relevant for new brands. Using the N variable created above is described in a fairly simple book by Ehrenberg (1972).

In order to analyze N using PROC DISCRETE, you must use PROC FREQ to produce a table on the N variable from the FAMILIES dataset. The tables data set is then read into PROC DISCRETE.

References