Conjoint and Beyond: Clustering and Logistic Regression of Attribute Importance

Leslie A. Christensen, The Goodyear Tire & Rubber Company, Akron, Ohio

Abstract

This paper will cover a case study that examines conjoint utilities by using logistic regression to analyze clusters based on attribute importance. The method encompasses PROC PLAN, PROC OPTEX, PROC TRANSREG, PROC FASTCLUS and PROC PROBIT. It will demonstrate the code used to perform the analysis and how to interpret the SAS output. It is intended for market researchers with limited experience of the topic.

Definition of Conjoint

Conjoint is the study of trade-offs. Conjoint is based on two concepts: attributes, e.g. foods and price; and levels, e.g. hamburgers, sandwiches, tacos, chicken and 59¢, 69¢. A combination of attribute levels, e.g. 59¢ tacos, is called a product concept. In a conjoint study, consumers rate product concepts as opposed to rating each individual attribute of a product. For example, a conjoint study may ask “How likely are you to buy a low fat hamburger?” Whereas another approach may ask “Which type of location do you prefer?: fast food with playground or fast food without playground.” and “Which food do you prefer?: hamburger, chicken, tacos”, etc. The results of conjoint are sets of utilities that measure how a consumer’s preference ranges among an attribute’s levels and a set of importance ratings that measure how a consumer’s preference ranges across the attributes. If desired, conjoint can measure intent or some other variable instead of preference.

Purpose

Analyzing overall and a-priori segments may lead to misleading results. For instance, an attribute may have a low overall utility but be very important to a particular segment of the market that was not predefined by the researcher.

Rather than be an end result, the conjoint attribute importance ratings themselves can be used to further analyze market segments. The conjoint findings can be merged with the standard questionnaire portion of a survey and further analyzed. This paper will work through an example step-by-step from design through analysis. The example will use SAS® software code to demonstrate how to:

1. design a main effects conjoint study,
2. calculate attribute importance ratings per respondent,
3. cluster respondents based on their attribute importance ratings,
4. describe the clusters using logistic regression based on additional survey data.

Where possible, the paper will suggest alternative SAS PROCs available should the reader’s data differ slightly from the author’s example.

Design a Conjoint Study

Real world constraints such as time, money, and respondent fatigue often restrict the researcher to use a fractional, partial, conjoint design. While a fractional design could be chosen randomly, it is best to choose a design using a method that will assure a main effects model. A main effects model is one in which each attribute will be evaluated independently but no interactions will be assessed.

As in any modeling effort, it is the researcher’s responsibility to assure that choosing a main effects model will not result in confounding results. That is the researcher should feel comfortable that the wrong decisions will not be made due to not having measured interactions. The SAS system can handle a conjoint with interactions, e.g. brand x price; but an interaction model is beyond the scope of this paper. Refer to SAS/QC® Software: Reference manual for more details.

A main effects model can be designed three ways using SAS. The simplest approach is to use the SAS ADX® menu system. ADX is part of the SAS/QC product. However, the ADX menu system will only allow the design to have attributes with two or three levels.

Thus the next approach is to use what the ADX menus use, PROC FACTEX. The caveat with PROC FACTEX is that it is only easy to use when you have balanced designs, e.g. all attributes having the same number of levels. PROC FACTEX can be used with unbalanced designs by employing a bridging technique, but this becomes complicated if the design has many different levels.

The most generic approach which our example will use is PROC OPTEX. PROC OPTEX in
Statistics, Data Analysis, and Modeling

conjunction with PROC PLAN will create a best attempt at an orthogonal, main effects, design. When running PROC OPTEX for conjoint studies, use the G-efficiency number as your base for evaluating the design. The G-efficiency statistic relates to the variance of the total utility. If the G-efficiency statistic equals 1.0 then the design is orthogonal. Be aware to check the resulting design of PROC OPTEX output. OPTEX can sometimes contain duplicate product concepts. If this happens, select a different random seed and try again.

If your design is not complicated, use of ADX Menus or PROC FACTEX is encouraged. The following code will create an approximation to an orthogonal 4x2x2x2 design.

```sas
PROC PLAN ORDERED;
  FACTORS MENU=4 PLAY=2 FAT=2 CAL=2 / NOPRINT;
  OUTPUT OUT=MYDB,SUGI01;
  MENU CVALS=('TACOS' 'HAMBURGERS'
              'SANDWICHES' 'CHICKEN')
  PLAY CVALS=('YES' 'NO')
  FAT CVALS=('YES' 'NO')
  CAL CVALS=('YES' 'NO');
RUN;
PROC OPTEX DATA=MYDB_SUGIO1 SEED=123;
  CLASS MENU PLAY FAT CAL;
  MODEL MENU PLAY FAT CAL;
  GENERATE N=16 ITER=300
      METHOD=M_FEDEROV;
  OUTPUT OUT=MYDB.SUGI02;
RUN;
QUIT;
```

In PROC OPTEX, N= is the number of product concepts the respondent will be asked to rate. When creating a design for conjoint, the number of product concepts the respondent will rate must be greater than or equal to the design's degrees of freedom. PROC OPTEX will permit interactions in the model statement. Further discussion of PROC OPTEX is beyond the scope of this paper. Refer to the SAS/QC® Software: Reference manual or Marketing Research: Practical Applications Using the SAS® System Course Notes.

Calculate Attribute Importance
When analyzing conjoint product concept results two results are most commonly sought. These are: 1. attribute importance across attributes and 2. utilities within attributes. Code to calculate both is included in the steps need to prepare the data for clustering.

First, our data needs to be formatted so that PROC TRANSREG will run correctly. In the example, data was entered with one observation per respondent. PROC TRANSREG requires the data to be one observation per product concept. The SAS code below is an example of how to read in and reformat the raw data. It is important to mention that all raw data should be examined for data entry errors and missing values before running PROC TRANSREG.

```sas
DATA RAWCNJ (KEEP=MENU PLAY FAT CAL
             SUBJ RESPNO)
  RAWID (KEEP=HH KIDS FFFREQ
          MARITAL AGE SEX
          RESPNO);
  INFILE 'H:\L123\SUGI.TXT';
  RESPNO= _N_; DO 1= 1 TO 16;
    INPUT MENU $1. PLAY $1. FAT $1.
    CAL $1. SUBJ 2. @;
    OUTPUT RAWCNJ;
  END;
  INPUT HH KIDS FFFREQ MARITAL $
      AGE SEX $;
  OUTPUT RAWID;
RUN;
PROC SORT DATA=RAWCNJ;
  BY DESCENDING MENU PLAY FAT CAL
      RESPNO;
RUN;
PROC TRANSPOSE DATA=RAWCNJ
  OUT=RAWTRANS;
  VAR SUBJ;
  BY DESCENDING MENU PLAY FAT CAL;
RUN;
```

PROC TRANSPOSE uses a weight variable. In our example, all data collected will be used in the analysis. Thus the following DATA step can be used to append the variable.

```sas
DATA RAWTRANS;
  SET RAWTRANS;
  WEIGHTS = 1;
RUN;
```

Now the data are ready to run through PROC TRANSREG.
PROC TRANSREG DATA=RAWTRANS SHORT
UTILITIES METHOD=MORALS
OUTTEST=UTILS;
MODEL LINEAR(COL:)=CLASS(MENU PLAY
FAT CAL / ZERO = SUM);
WEIGHT WEIGHTS;
OUTPUT DAPPROXIMATIONS IREPLACE
OUT=RESULTS;
RUN;

The printed output will look something like this.
(This example has had some columns removed.)
Utilities Table Based on the Usual Degrees of Freedom

<table>
<thead>
<tr>
<th>Label</th>
<th>Utility</th>
<th>Importance (%) Utility Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.56250000</td>
<td></td>
</tr>
<tr>
<td>MENUc</td>
<td>0.43750000</td>
<td>35.714 chicken</td>
</tr>
<tr>
<td>MENUh</td>
<td>-.56250000</td>
<td>7.143 hamburgers</td>
</tr>
<tr>
<td>MENUs</td>
<td>0.18750000</td>
<td>46.429 sandwiches</td>
</tr>
<tr>
<td>MENUt</td>
<td>-.06250000</td>
<td>10.714 tacos</td>
</tr>
<tr>
<td>PLAYn</td>
<td>-.10000000</td>
<td>7.143 no playground</td>
</tr>
<tr>
<td>PLAYy</td>
<td>0.10000000</td>
<td>7.143 playground</td>
</tr>
<tr>
<td>FATn</td>
<td>-.65000000</td>
<td>46.429 no low fat items</td>
</tr>
<tr>
<td>FATy</td>
<td>0.65000000</td>
<td>46.429 low fat items</td>
</tr>
<tr>
<td>CALn</td>
<td>-.15000000</td>
<td>10.714 no low cal items</td>
</tr>
<tr>
<td>CALy</td>
<td>0.15000000</td>
<td>10.714 low cal items</td>
</tr>
</tbody>
</table>

You can see in this example, that respondent numbers 2 and 3 are only concerned with MENU. Yet an average, via PROC MEANS, of all attribute importance ratings yields the overall importance of MENU is 50%. Cluster analysis will let us see if there are groups of respondents with similar importance ratings.

### Definition of Clusters
Cluster analysis is a technique that groups respondents into homogenous groups. Cluster analysis works on the principle of distances. Cluster analysis is an iterative technique which sometimes converges to an optimal solution. To find more information on clustering techniques in general, refer to a multivariate techniques text.

#### Cluster Respondents
The next step of the analysis is to use PROC FASTCLUS to cluster the attribute importance ratings found in the IMPORT data set. PROC FASTCLUS is a non-hierarchial clustering technique. SAS recommends 100 or more observations for PROC FASTCLUS. Since the attribute importance values from PROC TRANSREG sum to zero, the input to PROC FASTCLUS is already standardized.

To cluster importance values, submit the following code repeatedly changing the MAXCLUSTERS= option. Examine each iteration's cluster summary output and stop when a positive peak of the cubic clustering criterion statistic is reached. (You may need to use your best judgement to determine a manageable number of clusters that is a good approximation of the variation in importance values.)

```sas
PROC FASTCLUS DATA=IMPORT
MAXCLUSTERS=4 RANDOM=44445
REPLACE=FULL OUT=CLUS;
VAR MENU PLAY FAT CAL;
LABEL _DEPVAR_ ="SUBJECT";
RUN;
```

The resulting data set will look like this.

<table>
<thead>
<tr>
<th>OBS</th>
<th>RESPNO</th>
<th>MENU</th>
<th>PLAY</th>
<th>FAT</th>
<th>CAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>COL1</td>
<td>35.714</td>
<td>7.143</td>
<td>46.426</td>
<td>10.714</td>
</tr>
<tr>
<td>2</td>
<td>COL2</td>
<td>100.000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>COL3</td>
<td>100.000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

To cluster importance values, submit the following code repeatedly changing the MAXCLUSTERS= option. Examine each iteration's cluster summary output and stop when a positive peak of the cubic clustering criterion statistic is reached. (You may need to use your best judgement to determine a manageable number of clusters that is a good approximation of the variation in importance values.)

```sas
PROC FASTCLUS DATA=IMPORT
MAXCLUSTERS=4 RANDOM=44445
REPLACE=FULL OUT=CLUS;
VAR MENU PLAY FAT CAL;
LABEL _DEPVAR_ ="SUBJECT";
RUN;
```

In the output of FASTCLUS, the cubic clustering criterion statistic is found in the cluster summary statistics.
Statistics, Data Analysis, and Modeling

Pseudo F Statistic = 39.19
Approximate Expected Over-All R-Squared = 0.41425
Cubic Clustering Criterion = 2.673

WARNING: The two above values are invalid for correlated variables.

Cluster Means

<table>
<thead>
<tr>
<th>Cluster</th>
<th>MENU</th>
<th>PLAY</th>
<th>FAT</th>
<th>CAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.46</td>
<td>2.14</td>
<td>38.20</td>
<td>12.70</td>
</tr>
<tr>
<td>2</td>
<td>68.75</td>
<td>12.70</td>
<td>10.16</td>
<td>8.39</td>
</tr>
</tbody>
</table>

Once the number of clusters is determined, the cluster means can be examined. The cluster means table translates into market terms. A high attribute mean indicates that the attribute is very important to the cluster. Likewise, a very low mean indicates that the attribute is not important to the cluster.

For example, cluster 1 is more interested in the health content of fast food than the other cluster. Likewise, cluster 2 is more concerned about the type of menu and least concerned about the healthiness of the items. At this point, the researcher may choose to name the clusters. For example, cluster 1 could be named “healthy eaters.” Depending on the study’s sampling frame, the size of the clusters may relate to the size of the market segments.

Sometimes the attribute importance ratings do not fall into a manageable amount of clusters. If this is the case, try PROC CLUSTER using different linkage options. Should clusters still not converge to a manageable amount, analysis will stop at this point.

Description of Logistic Regression

As clusters are categorical variables, normal linear regression is not appropriate to model them. Instead, probit can be used to model the probability of cluster membership. Probit is a variation of logistic regression. Logistic regression models the probability of an event. In the case of conjoint, the event is group membership. It is most commonly used with dichotomous variables, e.g., ‘member’ and ‘non-member’. Probit can be used with both continuous and categorical dependent variables.

Describe the Clusters

Before describing the clusters, the survey questionnaire data needs merged to the clusters. After merging, dichotomous variables for indicating cluster membership need defined.

To accomplish this, we will first set up a variable called RESPNO in the CLUS data set that can be used for merging.

```
DATA CLUSID;
SET CLUS;
LENGTH RESPNO 4.;
N = SUBSTR(DEPVAR,4,2);
RESPNO = PUT(N,2.);
RUN;
```

PROC SORT DATA=RAWID;
BY RESPNO;
RUN;
PROC SORT DATA=CLUSID;
BY RESPNO;
RUN;
DATA CLUSRAW;
MERGE CLUSID RAWID;
BY RESPNO;
RUN;

Now data set CLUSRAW can be used for the logistic regression analysis.

```
DATA CLUSRAW;
SET CLUSRAW;
IF CLUSTER=1 THEN CLUS1=0;
ELSE CLUS1=1;
IF CLUSTER=2 THEN CLUS2=0;
ELSE CLUS2=1;
*continue with this type of logic if data has more clusters;
RUN;
```

Now PROC PROBIT can be used to describe each cluster. PROC PROBIT supports the CLASS statement. Thus character variables can be used.

The following code will be used to examine the "healthy eaters". PROC PROBIT defaults such that the lowest value of the dependent variable is the event. Therefore in our example, set the membership variable to 0=member and 1=non-member. Also, it is easier to read the output if the data has SAS formats.

```
PROC PROBIT DATA=CLUSRAW;
CLASS CLUS1;
MODEL CLUS1 = KIDS / D=LOGISTIC;
RUN;
```
Interpreting PROBIT output is a two-step process. First, examine the overall p-value statistic for each variable, Pr>Chi. If the p-value is less than or equal to 0.05 then there is a 95% confidence that the variable has an effect on cluster membership. You may wish to select a different level of confidence. In our example, KIDS meets a 90% criterion.

For continuous variables that affect cluster membership, look at the parameter estimate coefficient. If the coefficient is positive then the probability of membership increases as the value of the variable increases. If the estimate is negative, the probability of membership decreases as the variable increases.

For categorical variables that affect cluster membership, examine the estimate for each level of the variable and interpret in the same manner as continuous variables.

As in other modeling efforts, start your model with several variables and then eliminate non-significant ones one at a time. In our example, the continuous variable KIDS was the only variable that contributed to the likelihood of cluster membership.

Conclusion
The result of analyzing the "healthy eaters" cluster is that KIDS affected cluster membership. KIDS was a continuous variable indicating the number of children age 10 and under living in the household. The negative parameter estimate of -0.6 for KIDS indicated that the probability of a household eating at a fast food restaurant based on healthy food choices decreased for every kid age 10 and under.

The result of finding clusters and profiling them is more meaningful than overall average attribute importance ratings. The table below compares the overall ratings results to the cluster results of our example.