Sawtooth Software

RESEARCH PAPER SERIES

Analysis of Traditional Conjoint Using Microsoft Excel™: An Introductory Example

Bryan K. Orme,
Sawtooth Software, Inc.,
2002

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

Analysis of Traditional Conjoint Using Excel™: An Introductory Example

In this example, we assume the reader has a basic understanding of multiple regression analysis.

A traditional conjoint analysis is really just a multiple regression problem. The respondent's ratings for the product concepts form the *dependent* variable. The characteristics of the product (the attribute levels) are the *independent* (predictor) variables. The estimated *betas* associated with the independent variables are the *utilities* (preference scores) for the levels. The *R-Square* for the regression characterizes the internal consistency of the respondent.

Consider a conjoint analysis problem with three attributes, each with levels as follows:

Brand
A
B
C

Color
Red
Blue

Price
$50
$100
$150

For simplicity, let's consider a *full-factorial* experimental design. A full-factorial design includes all possible combinations of the attributes. There are (3)·(2)·(3) = 18 possible product concepts (commonly called *cards*) that can be created from these three attributes. Further assume that respondents rate each of the 18 product concepts on a score from 0 to 10, where 10 represents the highest degree of preference.

Assume the data for one respondent are as follows (as if in an Excel spreadsheet):

	A	B	C	D	E
1	Card#	Brand	Color	Price	Preference
2	1	1	1	1	5
3	2	1	1	2	5
4	3	1	1	3	0
5	4	1	2	1	8
6	5	1	2	2	5
7	6	1	2	3	2
8	7	2	1	1	7
9	8	2	1	2	5
10	9	2	1	3	3
11	10	2	2	1	9
12	11	2	2	2	6
13	12	2	2	3	5
14	13	3	1	1	10
15	14	3	1	2	7
16	15	3	1	3	5
17	16	3	2	1	9
18	17	3	2	2	7
19	18	3	2	3	6

The first card is made up of level 1 of each of the attributes, or (Brand A, Red, $50). The respondent rated that card a “5” on the preference scale.

After collecting the respondent data, the next step is to code the data in an appropriate manner for estimating utilities using multiple regression. We use a procedure called *dummy coding* for the independent variables (the product characteristics). In its simplest form, dummy coding uses a “1” to reflect the presence of a feature, and a “0” to represent its absence. For example, we can code the Brand attribute as three separate columns.

	Brand A	Brand B	Brand C
If Brand is “A”, then dummy codes =	1	0	0
If Brand is “B”, then dummy codes =	0	1	0
If Brand is “C”, then dummy codes =	0	0	1

Applying dummy-coding for all attributes results in an array of columns as follows:

	A	B	C	D	E	F	G	H	I	J
1	Card #	A	B	C	Red	Blue	$50	$100	$150	Preference
2	1	1	0	0	1	0	1	0	0	5
3	2	1	0	0	1	0	0	1	0	5
4	3	1	0	0	1	0	0	0	1	0
5	4	1	0	0	0	1	1	0	0	8
6	5	1	0	0	0	1	0	1	0	5
7	6	1	0	0	0	1	0	0	1	2
8	7	0	1	0	1	0	1	0	0	7
9	8	0	1	0	1	0	0	1	0	5
10	9	0	1	0	1	0	0	0	1	3
11	10	0	1	0	0	1	1	0	0	9
12	11	0	1	0	0	1	0	1	0	6
13	12	0	1	0	0	1	0	0	1	5
14	13	0	0	1	1	0	1	0	0	10
15	14	0	0	1	1	0	0	1	0	7
16	15	0	0	1	1	0	0	0	1	5
17	16	0	0	1	0	1	1	0	0	9
18	17	0	0	1	0	1	0	1	0	7
19	18	0	0	1	0	1	0	0	1	6

Again, we see that card 1 is defined as (Brand A, Red, $50), but we have expanded the layout to reflect dummy coding.

To this point, the coding has been very straightforward. But, there is one complication that must be resolved. In multiple regression analysis, no independent variable may be perfectly predictable based on the state of any other independent variable or combination of independent variables. If so, the regression procedure could not separate the effects of the confounded variables. We have that problem with the data above, since, for example, we can perfectly predict the state of brand A based on the states for brands B and C. This situation is termed *linear dependency*.

To resolve this linear dependency, we omit one column from each attribute. It really doesn't matter which column (level) we drop, and for this example we have excluded the first level for each attribute, to produce the modified data table below:

	A	B	C	D	E	F	G
1	Card #	B	C	Blue	$100	$150	Preference
2	1	0	0	0	0	0	5
3	2	0	0	0	1	0	5
4	3	0	0	0	0	1	0
5	4	0	0	1	0	0	8
6	5	0	0	1	1	0	5
7	6	0	0	1	0	1	2
8	7	1	0	0	0	0	7
9	8	1	0	0	1	0	5
10	9	1	0	0	0	1	3
11	10	1	0	1	0	0	9
12	11	1	0	1	1	0	6
13	12	1	0	1	0	1	5
14	13	0	1	0	0	0	10
15	14	0	1	0	1	0	7
16	15	0	1	0	0	1	5
17	16	0	1	1	0	0	9
18	17	0	1	1	1	0	7
19	18	0	1	1	0	1	6

Even though it appears that one level from each attribute is missing from the data, they are really implicitly included as *reference levels* for each attribute. The explicitly coded levels are estimated as contrasts with respect to the omitted levels, which are defined as "0."

Microsoft Excel™ (we have used Excel from Office 2000 in this example) offers a simple multiple regression tool, under ***Tools + Data Analysis + Regression*** (you must have installed the Analysis Toolpak add-in). Using the tool, specify the preference score (column G) as the dependent variable (the *Input Y Range*) and the five dummy-coded attribute columns (columns B through F) as independent variables (the *Input X range*). You should also make sure a constant is estimated (this usually happens by default).

The mathematical expression of the model is as follows:

$$Y = b_1(\text{Brand B}) + b_2(\text{Brand C}) + b_3(\text{Blue}) + b_4(\$100) + b_5(\$150) + \text{constant} + e$$

where:

Y = respondent's preference for the product concept,
b_1 through b_5 are beta weights (utilities) for the features,
e is an error term, and
the reference levels are equal to "0."

The solution minimizes the sum of squares of the errors over all observations. A portion of the output from Excel is as follows:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.948902
R Square	0.900415
Adjusted R Square	0.858921
Standard Error	0.942809
Observations	18

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.833333	0.544331	10.71652	1.69E-07	4.647338	7.019329	4.647338	7.019329
X Variable 1	1.666667	0.544331	3.061862	0.009865	0.480671	2.852662	0.480671	2.852662
X Variable 2	3.166667	0.544331	5.817538	8.24E-05	1.980671	4.352662	1.980671	4.352662
X Variable 3	1.111111	0.444444	2.5	0.027915	0.14275	2.079472	0.14275	2.079472
X Variable 4	-2.16667	0.544331	-3.98042	0.001825	-3.35266	-0.98067	-3.35266	-0.98067
X Variable 5	-4.5	0.544331	-8.26703	2.68E-06	-5.686	-3.314	-5.686	-3.314

Using that output (after rounding to two decimals places of precision), the utilities (Coefficients) are:

Brand	
A	0.00
B	1.67
C	3.17

Color	
Red	0.00
Blue	1.11

Price	
$50	0.00
$100	-2.17
$150	-4.50

The constant is 5.83, and the fit for this respondent (R-Square) is 0.90. The fit ranges from a low of 0 to a high of 1.0. The standard errors of the coefficients (betas) reflect how precisely we are able to estimate the betas with this design. Lower standard errors are better. The remaining statistics presented in Excel's output are beyond the scope of this paper, and are generally not of much use when considering individual-level conjoint analysis problems.

Notes:

One can easily generalize how many parameters (independent variables plus the constant) are involved in any conjoint analysis problem as #Levels - #Attributes + 1. Most traditional conjoint analysis problems solve a separate regression equation for each respondent. Therefore, to estimate utilities, the respondent must have evaluated at least as many cards as parameters to be estimated. When the respondent answers the minimum number of conjoint cards to enable estimation, this is called a *saturated design*. While such a design is easiest on the respondent, it leaves no room for respondent error. It also always yields an R-square of 100, and therefore no ability to assess respondent consistency.

Most good conjoint designs in practice include more observations than parameters to be estimated (usually 1.5 to 3 times more). The design above has three times as many cards (observations) as parameters to be estimated. These designs usually lead to more stable estimates of respondent utilities than saturated designs.

Also note that in practice (except with the smallest problems), asking respondents to evaluate all possible combinations of the attribute levels is usually not practical. Design catalogs and computer programs are available to find efficient *fractional factorial* designs. Fractional factorial designs show just an efficient subset of the possible combinations, and still provide enough degrees of freedom to estimate utilities.

The standard errors for the Color attribute are lower than for Brand and Price (recall that lower standard errors imply greater precision of the beta estimate). Because Color only has two levels (as compared to three each for Brand and Price), each color level has more representation within the design. Therefore, more information is provided for each color level than is provided for any level of the three-level attributes.

Acknowledgements:

Excel and Office are trademarks of Microsoft Corporation.