Chapter 8

Traditional Conjoint Analysis with Excel

A traditional conjoint analysis may be thought of as a multiple regression problem. The respondent's ratings for the product concepts are observations on the dependent variable. The characteristics of the product or attribute levels are observations on the independent or predictor variables. The estimated regression coefficients associated with the independent variables are the part-worth utilities or preference scores for the levels. The R^2 for the regression characterizes the internal consistency of the respondent.

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

Brand	*Color*	*Price*
A	Red	$50
B	Blue	$100
C		$150

For simplicity, let us consider a full-factorial experimental design. A full-factorial design includes all possible combinations of the attributes. There are 18 possible product concepts or cards that can be created from these three attributes:

$$3\,\text{brands} \times 2\,\text{colors} \times 3\,\text{prices} = 18\,\text{cards}$$

Further assume that respondents rate each of the 18 product concepts on a scale from 0 to 10, where 10 represents the highest degree of preference. Exhibit 8.1 shows the experimental design.

We can use Microsoft Excel to analyze data from traditional conjoint questionnaires. This chapter shows how to code, organize, and analyze data from one hypothetical respondent, working with spreadsheets and spreadsheet functions. Multiple regression functions come from the Excel Analysis ToolPak add-in.

Reprinted from Orme, B. (2006) *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research.* Madison, Wis.: Research Publishers LLC.

Card	Brand	Color	Price ($)
1	A	Red	50
2	A	Red	100
3	A	Red	150
4	A	Blue	50
5	A	Blue	100
6	A	Blue	150
7	B	Red	50
8	B	Red	100
9	B	Red	150
10	B	Blue	50
11	B	Blue	100
12	B	Blue	150
13	C	Red	50
14	C	Red	100
15	C	Red	150
16	C	Blue	50
17	C	Blue	100
18	C	Blue	150

Exhibit 8.1. Full-factorial experimental design

8.1 Data Organization and Coding

Assume the data for one respondent have been entered into an Excel spreadsheet, illustrated in exhibit 8.2. The first card is made up of the first level on each of the attributes: (Brand A, Red, $50). The respondent rated that card a 5 on the preference scale. The second card has the first level on brand and color and the second level on price: (Brand A, Red, $100). This card gets a 5 on the preference scale. And so on.

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 or 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. The brand attribute would be coded as three separate columns, color as two columns, and price as three columns. Applying dummy coding results in an array of columns as illustrated in exhibit 8.3. 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 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 of Brand B and Brand C. This situation is called 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 a modified data table, as illustrated by exhibit 8.4.

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.

	A	B	C	D	E
1	**Card**	**Brand**	**Color**	**Price**	**Preference**
2	1	1	1	50	5
3	2	1	1	100	5
4	3	1	1	150	0
5	4	1	2	50	8
6	5	1	2	100	5
7	6	1	2	150	2
8	7	2	1	50	7
9	8	2	1	100	5
10	9	2	1	150	3
11	10	2	2	50	9
12	11	2	2	100	6
13	12	2	2	150	5
14	13	3	1	50	10
15	14	3	1	100	7
16	15	3	1	150	5
17	16	3	2	50	9
18	17	3	2	100	7
19	18	3	2	150	6

Exhibit 8.2. Excel spreadsheet with conjoint data

	H	I	J	K	L	M	N	O	P	Q
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

Exhibit 8.3. Excel spreadsheet with coded data

	S	T	U	V	W	X	Y
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

Exhibit 8.4. Modified data table for analysis with Excel

8.2 Multiple Regression Analysis

Microsoft Excel offers a simple multiple regression tool (under Tools + Data Analysis + Regression with the Analysis Toolpak add-in installed). Using the tool, we can specify the preference score (column Y) as the dependent variable (Input Y Range) and the five dummy-coded attribute columns (columns T through X) as independent variables (Input X range). You should also make sure a constant is estimated; this usually happens by default (by not checking the box labeled "Constant is zero").

The mathematical expression of the model is as follows:

$$Y = b_0 + b_1(\text{Brand B}) + b_2(\text{Brand C}) + b_3(\text{Blue}) + b_4(\$100) + b_5(\$150) + e$$

where Y is the respondent's preference for the product concept, b_0 is the constant or intercept term, b_1 through b_5 are beta weights (part-worth utilities) for the features, and e is an error term. In this formulation of the model, coefficients for 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 illustrated in exhibit 8.5. Using that output (after rounding to two decimal places of precision), the utilities (coefficients) are the following:

Brand	*Color*	*Price*
A = 0.00	Red = 0.00	$ 50 = 0.00
B = 1.67	Blue = 1.11	$100 = -2.17
C = 3.17		$150 = -4.50

The constant or intercept term is 5.83, and the fit for this respondent $R^2 = 0.90$. The fit values range from a low of 0 to a high of 1.0. The standard errors of the regression coefficients (betas) reflect how precisely we are able to estimate those coefficients with this design. Lower standard errors are better. The remaining statistics presented in Excel's output are beyond the scope of this chapter and are generally not of much use when considering individual-level conjoint analysis problems.

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^2 of 1, and therefore no ability to assess respondent consistency.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.94890196
R Square	0.90041494
Adjusted R Sq	0.85892116
Standard Erro	0.94280904
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	96.4444444	19.2888889	21.7	1.2511E-05
Residual	12	10.6666667	0.88888889		
Total	17	107.111111			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	5.83333333	0.54433105	10.7165176	1.6872E-07
X Variable 1	1.66666667	0.54433105	3.06186218	0.00986485
X Variable 2	3.16666667	0.54433105	5.81753814	8.2445E-05
X Variable 3	1.11111111	0.44444444	2.5	0.0279154
X Variable 4	-2.16666667	0.54433105	-3.98042083	0.0018249
X Variable 5	-4.5	0.54433105	-8.26702788	2.6823E-06

Exhibit 8.5. Conjoint analysis with multiple regression in Excel

One can easily determine the number of parameters to be estimated in a traditional conjoint analysis:

$$\#\,\mathbf{parameters\ to\ be\ estimated} = (\#\,\mathbf{levels}) - (\#\,\mathbf{attributes}) + 1$$

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.

Only in the smallest of problems (such as our 18-card example) would we ask people to respond to all possible combinations of attribute levels. Large full-factorial designs are not practical. Fortunately, design catalogs and computer programs are available to find efficient fractional-factorial designs. Fractional-factorial designs show an efficient subset of the possible combinations and provide enough information to estimate utilities.

In our worked example, 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 the three-level attributes.