How to Get More Value from Your Survey Data

Discover four advanced analysis techniques that make survey research more effective

Table of contents

Introduction

Like many survey researchers, your day-to-day data analysis tools and techniques are likely to include cross-tabulations, bar charts, and finding mean differences between groups. However, these methods, while valuable, may be too simplistic to enable you to derive the most value from your survey data.

This white paper introduces you to four types of advanced analysis—cluster, factor, reliability, and regression—that can help you gain important insights that you might miss using more basic methods. By expanding your survey analysis toolkit, you can delve deeper into your data to increase your understanding of survey responses and respondents, create better measures of important concepts, and make more accurate predictions about behaviors and attitudes.

Here is a brief overview of the four techniques:

- **Cluster analysis** is used to discover similar groups, or segments, of respondents. Segmentation enables you to focus sales and marketing efforts on defined groups. You can also use subgroups in analyses, to be more sensitive to differences between respondents.
- **Factor** and **reliability analysis** enable you to combine several questions into a more valid and reliable measure of an important concept. They also help you isolate survey questions that may be redundant or unnecessary.
- **Regression analysis** is used to create predictive behavior models that include many predictor variables simultaneously. Regression analysis enables you to identify the best predictors, so you can focus on them in future actions.

If you are new to these techniques, it may help to focus on the benefits of each method, rather than on the technical details, as you read through this paper. When you're ready to try these advanced techniques in your own analysis, begin with the method that is most suitable for your data, or with which you are the most comfortable.

SPSS software and technologies are used throughout the paper to illustrate how to apply advanced analysis methods to typical survey data. Each example includes advice on using the technique and interpreting the results and output.

Descriptive survey research

All survey researchers use descriptive statistical methods to summarize data and get a description of the responses to questions. These methods include frequency tables, cross-tabulations (stub and banner tables), and finding mean differences between groups or correlations between questions. For example, Figure 1 is a typical frequency table showing overall customer satisfaction for the hypothetical BSI hotel chain.

I am a satisfied BSI customer

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Disagree	20	2.1	2.1	2.1
Disagree	130	13.6	13.6	15.7
Neither Agree Nor Disagree	360	37.7	37.7	53.4
Agree	300	31.4	31.4	84.8
Strongly Agree	145	15.2	15.2	100.0
Total	955	100.0	100.0	

Figure 1: This standard frequency table shows attitudes for overall customer satisfaction.

Though this type of analysis is necessary in any survey project, a series of such tables, or even cross-tabulations, only provides limited information about the attitudes and behaviors of the respondents. That is because the real world is actually multivariate, meaning that many factors play a part in an individual response. Bivariate approaches, such as analyzing data one question at a time, or using a cross-tabulation table to determine whether two questions relate, create an oversimplified view of the customer.

The table in Figure 1 doesn't provide information about which customers give similar answers to several questions. Analyzing only one or two questions makes it impossible to see which set of questions measures similar concepts. In addition, using a table to see how responses to one question help predict responses to a second question ignores the factors that influence the second question. Thus, using a rating of overall service to predict satisfaction at the hotel chain ignores other factors, such as frequency of stay, rating of restaurants, and rating of room quality.

The advanced survey methods covered in this paper enable you to analyze many survey questions simultaneously, in order to cluster respondents, group questions, and make predictions with greater accuracy.

Segmenting respondents with cluster analysis

Cluster analysis enables you to group respondents with similar behaviors, preferences, or characteristics into clusters, or segments. Through segmentation, you gain a greater understanding of important similarities and differences between your respondents. You can use this information to develop targeted marketing strategies, or to provide subgroups for analysis. In the case of survey data, clustering enables researchers to group respondents who provide similar responses on several questions.

Clustering, or segmentation, is a multivariate technique that analyzes responses to several questions in order to find similar respondents. Clustering is based on the concept of creating groups based on their proximity to, or distance from, each other. Respondents within a cluster, therefore, are relatively homogenous.

There are two types of cluster analysis:

- Hierarchical: Observations are joined in a cluster and remain so throughout the clustering
- Non-hierarchical: Cases can switch clusters as the clustering proceeds. The most common non-hierarchical method is K-means.

Cluster analysis requires you to:

- Check the number of respondents in each cluster, as clusters of only a few respondents are not very useful
- Assess whether the clusters make sense, and whether their characteristics are easy to understand and describe
- Validate the clusters by analyzing how they relate to other variables

The following cluster analysis example uses the Two-Step Cluster procedure in SPSS, which incorporates statistical criteria to determine the optimal number of clusters.

In this example, we apply the Two-Step Cluster method to customer survey data from the hypothetical BSI hotel chain. Though we measured the survey questions on various scales, this does not present a problem, as the Two-Step Cluster method is able to use data on nominal, ordinal, or interval scales. We intend to create clusters of customers based on the following criteria:

- Frequency of stay at BSI hotels
- Length of customer relationship
- Usefulness of Internet access
- Customer company type
- Degree of involvement in company's decision to use BSI hotels
- Importance of travel to job

The first output from the Two-Step method tells us how many clusters were found, and how many respondents are in each cluster. From the table in Figure 2, we see that the Two-Step method found three natural groups or clusters in the data, based on the responses to the six questions above.

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	180	26.9%	18.8%
	2	240	35.8%	25.1%
	3	250	37.3%	26.2%
	Combined	670	100.0%	70.2%
Excluded Cases		285		29.8%
Total		955		100.0%

Figure 2: The Two-Step Cluster method identified three clusters of customers, with approximately the same number of customers in each cluster.

Additional information from the Two-Step method enables us to understand what type of customer each cluster represents. In Figure 3 on the next page we see, for example, that customers in Cluster 1 travel more for work than customers in the other two clusters, and that they frequently stay at BSI hotels. Measured by length of customer relationship, however, customers in Cluster 1 are not the most loyal.

Cluster	How long have you been staying at BSI hotels?	Travel an integral prt of wrk	How frequently do stay at BSI hotels?	Usefulness of internet services
	Mean	Mean	Mean	Mean
1	3.44	4.36	5.33	3.33
2	3.00	3.69	4.23	3.52
3	3.82	4.08	4.50	3.66
Combined	3.43	4.01	4.63	3.52

Figure 3: This table displays the mean of various questions by cluster, and demonstrates that customers in Cluster 1 stay more frequently at BSI hotels and travel more for work than customers in the other two clusters.

Once you are satisfied that you understand the clusters and their characteristics, you can create descriptive labels. In this case, we label Cluster 1 customers "Frequent BSI Road Warriors." Cluster 2 customers are "Less Frequent Travelers," and Cluster 3 customers are "Loyal BSI Road Warriors."

Using the clusters in analysis

Now that we have identified the clusters of respondents, we can use them in analyses and reports. For example, to see how overall customer satisfaction relates to cluster membership, we use a clustered bar chart (Figure 4). The chart shows that that customers in Cluster 3 are more satisfied with BSI hotels than are customers in Cluster 1. Since Cluster 1 customers stay at BSI hotels more often than customers in the other clusters, these results may be cause for action with that segment. Knowing the characteristics of the most and least satisfied customers can help you make important business decisions.

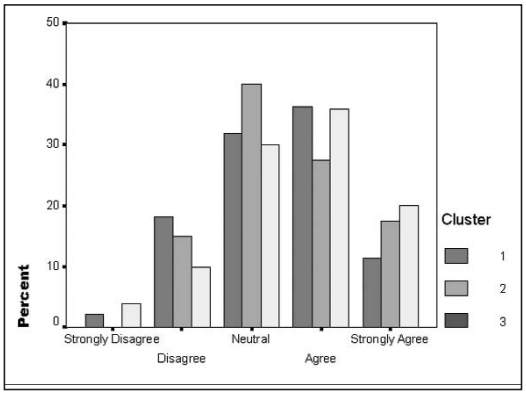

Figure 4: This bar chart shows differences in overall satisfaction between the three clusters of customers. Customers in Cluster 3 are the most satisfied, while customers in Cluster 1 are the least satisfied.

Presenting your results

The Custom Tables procedure in SPSS enables you to create attractive and complex stub- and banner-type tables. The table in Figure 5 shows the mean of customer satisfaction by cluster membership. For each cluster, we nest the customers' willingness to stay again at a BSI hotel. Each row includes questions about use of hotel spas and billing problem resolution. The Custom Tables procedure enables you to condense many types of information into an attractive and compact table.

			TwoStep Cluster					
			Frequent BSI Road Warriors		Less Frequent Travelers		Loyal BSI Road Warriors	
			Would you stay with BSI again?		Would you stay with BSI again?		Would you stay with BSI again?	
			Yes	No	Yes	No	Yes	No
I am a satisfied BSI customer	Use spa facilities?	No	3.9	2.0	3.8	2.0	4.0	2.5
		Yes	3.6	.	3.6	2.0	3.6	1.0
	Ever called BSI to resolve a problem?	Yes	3.8	2.0	3.9	2.0	3.6	2.0
		No	3.6	2.0	3.6	2.0	4.2	1.7

Figure 5: The Custom Tables procedure enables you to create complex tables, such as the one above. This table illustrates the relationship between the customers in each cluster and overall satisfaction, and includes several nested questions.

Grouping questions with factor analysis

Most questionnaires include several questions about each key topic. A questionnaire used to measure patient satisfaction with an HMO, for example, might include several questions about physician care, as well as several questions about the performance of support staff, such as x-ray technicians. While it is helpful to look at each question individually, it is often possible to create more reliable and valid measures by using the responses to several questions simultaneously. Compound measures of critical concepts can make your analyses more powerful.

Factor analysis enables you to discover clusters or groups of questions about similar concepts, based on correlations or covariances between questions. You can use the factors to:

- Create scales or compound measures composed of several questions
- Reduce the number of questions on a questionnaire by, for example, identifying questions that measure the same concept
- Understand the relationships between several questions simultaneously

Factor analysis is a general linear model (GLM) technique, which means that it assumes data are measured on an interval scale. As is common in survey research, we can easily use variables measured on five-, six-, or seven-point scales.

How SPSS helps you find the appropriate analysis technique

The SPSS Statistics Coach makes it easy to select the most appropriate analysis method for your data. In this case, we want to group questions, but we're not sure which method to use. From the "Help" menu, we select "Statistics Coach" to display the first screen shown in Figure 6. Then, we choose the "Identify groups of similar variables" option, click "Next," and choose the data type. SPSS immediately opens the Factor Analysis dialog box, shown in the second screen.

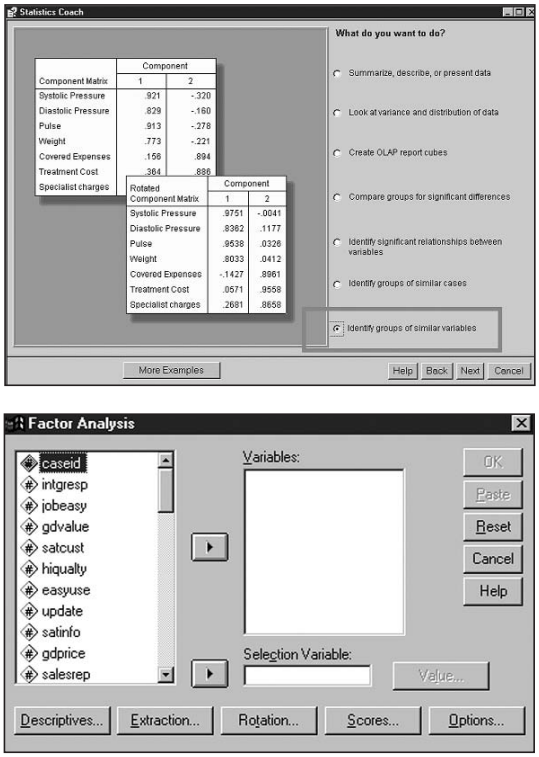

Figure 6: The screenshot at top depicts the Statistics Coach dialog box with the "Identify groups of similar variables" option selected. After we entered our data type, SPSS automatically opened the Factor Analysis dialog box at bottom.

Factor analysis includes two distinct steps. The first step involves extracting a small number of factors from the data. Think of the factors as underlying attitudes reflected in answers to specific questions. There are several extraction methods; principal components extraction and principal axis factoring are used most frequently. In the second step, the factors are rotated to ease interpretation. Varimax is the rotation method used most frequently with survey data.

In this example, we apply the principal components extraction and varimax rotation methods to the following statements taken from the survey of BSI hotel customers. These statements had a five-point response scale, ranging from "Strongly Disagree" to "Strongly Agree:"

- BSI services are a good value
- BSI offers high-quality services
- BSI makes it easy to make reservations
- BSI makes my job easier
- I am satisfied with BSI dining facilities
- BSI facilities are up to date
- BSI rooms are appropriately priced

The resulting table (Figure 7) contains the two factors extracted from the survey statements. The procedure automatically identifies factors that explain more variance than individual statements. The two factors in the table account for approximately 59 percent of the total variance among the statements, which is quite satisfactory.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.047	43.529	43.529	3.047	43.529	43.529	2.516	35.941	35.941
2	1.093	15.614	59.143	1.093	15.614	59.143	1.624	23.202	59.143
3	.859	12.268	71.411						
4	.670	9.568	80.978						
5	.544	7.772	88.751						
6	.463	6.608	95.359						
7	.325	4.641	100.000						

Extraction Method: Principal Component Analysis.

Figure 7: This total variance table shows the number of factors selected and the degree of variance for each, before and after rotation.

The most critical output of our factor analysis, however, is the rotated components table (Figure 8), which shows the loading, or correlation, between each question and the two extracted factors. Questions with high loadings on one factor and low loadings on other factors are associated with the high-loading factor. For example, the statement "BSI makes it easy to make reservations" is associated with Factor 1 because it has a high correlation (.716) with that factor and a low correlation (.260) with Factor 2.

The table in Figure 8 shows that Factor 1 is associated with the first five statements. Factor 2 is clearly associated with the two statements about pricing and value.

Rotated Component Matrix

	Component	
	1	2
BSI makes it easy to make reservations	.716	.260
BSI offers high-quality services	.711	.306
Business Suites makes job easier	.703	.198
BSI facilities are kept up to date	.626	-.413
Satisfied with dining facilities?	.596	.194
BSI rooms are priced right	.148	.842
BSI services are a good value	.485	.712

Figure 8: This rotated component matrix illustrates the relationship between the factors and the survey statements. By analyzing the statements associated with each factor, we can see that Factor 1 measures perceived quality, while Factor 2 measures perceived value.

At this point in your analysis, you can create descriptive factor labels to use in reports and subsequent analyses. In this case, we label Factor 1 "Quality," since it measures various aspects of the quality of hotel services and facilities. Factor 2 is labeled "Value," since it measures satisfaction with room price and the perceived value of services.

Though the individual responses to each statement may be useful, we now have two factors that are more valid measures of the quality and value of the hotel chain than any single statement. That is the essential benefit that factor analysis provides.

To use factors in future analyses, create a combined measure of the questions or statements associated with each factor. There are two ways to accomplish this:

- Automatically create factor scores with the factor procedure using standardized (z) scores
- Compute a new variable by adding the raw responses for the statements associated with each factor, and dividing by the number of questions, to create a mean score

Determining the reliability of factors

It's important to verify that survey responses are valid and reliable. The same is true for the factors that you discover. Fortunately, factors are by definition valid, because all of the associated questions or statements measure aspects of the same concept. It is still important, however, to establish the reliability of the factor.

The reliability analysis procedure in SPSS enables you to determine whether a set of survey questions, items, or statements forms a reliable scale. This means that the items measure a single concept with reasonably high intercorrelations. To perform reliability analysis, you don't need more assumptions about the data than you do for factor analysis. As an added benefit, the output is usually easy to interpret.

To demonstrate the reliability analysis capabilities of SPSS, we apply the technique to the "Quality" factor that we discovered using factor analysis. The screenshot in Figure 9 (on the next page) shows the main output. The key value in the output is Cronbach's alpha, which in this case is .6946. This statistic varies from zero to one, and though alpha has several interpretations, the cutoff value is more useful in determining whether a scale is reliable. The standard rule of thumb is that alpha must be greater than approximately .70 to conclude that the scale is reliable.

Since the output shows that alpha for the "Quality" factor is just below .70, we could decide that it's close enough for use in subsequent analyses. The output, however, also suggests which items can be removed from the reliability scale to increase alpha. For example, if we omit the "Update" item (which represents the statement, "BSI facilities are kept up to date"), alpha increases to more than .70.

Reliability

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

Statistics for	Mean	Variance	Std Dev	N of Variables
SCALE	18.1818	12.0721	3.4745	5

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
JOBEASY	14.1515	8.1653	.5260	.6142
HIQUALTY	14.2727	8.6400	.5615	.6120
EASYUSE	14.7576	7.8561	.5581	.5984
DINESAT	14.8485	7.6794	.4209	.6655
UPDATE	14.6970	9.2400	.2551	.7268

Reliability Coefficients

N of Cases = 495.0 N of Items = 5

Alpha = .6946

Figure 9: The reliability analysis output indicates that the alpha value of the "Quality" factor is less than the .70 needed to form a reliable scale. If we remove the "Update" item, the alpha value will increase and make the scale more reliable.

Using a factor in analysis

Once you determine that a factor is valid and reliable, what can you do with it? The answer is simple. Treat the factor as you would treat a new variable, and use it as you would use any question in which you have confidence. Relationships are typically more clear and distinct with factors than with individual questions or statements.

To illustrate, we examine the relationship between perceived quality and frequency of stay at BSI hotels. First, we use the single statement about quality (Figure 10). We then perform the same analysis using the "Quality" factor (Figure 11).

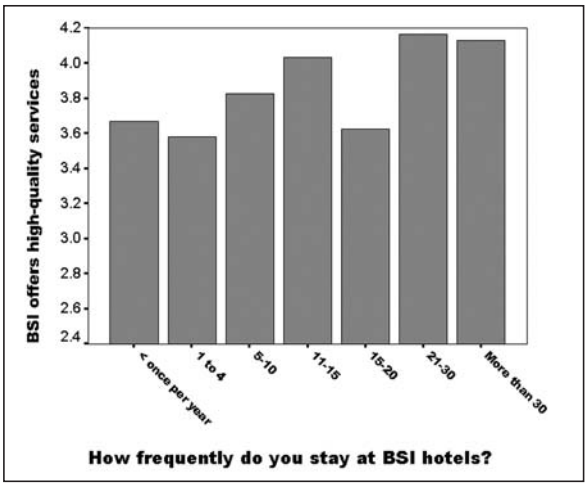

Figure 10: This chart shows the results of analysis using the single statement about quality. The results indicate that customers who stay at BSI hotels more frequently are somewhat more likely to agree that the chain offers high-quality services.

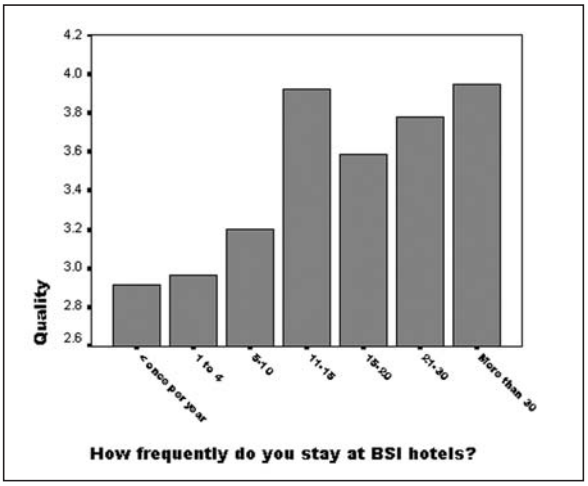

Figure 11: This chart shows the results of analysis using the "Quality" factor. The results give a stronger indication that customers who stay at BSI hotels more frequently are more likely to have higher perceptions of quality, and that quality is likely a key component of customer loyalty.

Both charts show a positive relationship between perceived quality and frequency of stay. Notice, however, that the relationship is much stronger for the "Quality" factor than for the single statement about quality. Hotel quality is comprised of several aspects of service, and single questions by definition don't do well at capturing all aspects of a general concept. This example is a direct illustration of the benefits of using factor and reliability analysis together.

Making predictions with regression

Multiple regression, a general linear model technique, is the most popular method for studying the relationship between an outcome variable and several predictor, or independent, variables. It is often used with survey data, because it enables you to combine many variables into one predictive equation. In addition, multiple regression helps to determine the unique role of each variable in predicting the outcome, provides a measure of the total explanatory power of the model (R2), and provides an estimate of whether a variable is a statistically significant predictor or not.

Multiple regression is often called linear regression, because a linear, or straight-line, relationship between predictors and outcome is assumed. Relationships between variables may not always be linear, but it is best to assume that they are in order to create a useful model. As with factor analysis, multiple regression works best with variables measured on an interval scale, but you can also use typical survey response scales.

In general, regression analysis should include only variables that may be good predictors, or variables that you want to include for reasons that are practical (the customer type variable is important for your business, for example) or theoretical (previous work suggests that customer gender is a key predictor). Though it's possible to include dozens of predictor variables in a regression equation, it's best to be more selective.

The first steps in regression analysis are to identify the predictors and the variable or question you want to predict. In this example, we apply regression techniques to the hypothetical BSI data to predict overall customer satisfaction.

We use nine predictors, including the "Quality" factor identified earlier, to demonstrate how to link regression and factor analysis.

The first results produced by SPSS are in the two tables shown in Figure 12. The R2 value is .583, so more than half of the variation in customer satisfaction responses is explained by these nine predictors. The second table shows that the set of predictors is statistically significant at predicting customer satisfaction at the .01 level of significance (because the "Sig." value is below this number). This means that there is a less than one in one hundred chance that our regression results occurred by chance.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.768	.590	.583	.654

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	335.453	9	37.273	87.177	.000
	Residual	233.016	545	.428		
	Total	568.468	554			

Figure 12: These model summary and ANOVA tables show the results of regression analysis to predict overall customer satisfaction.

The next step is to determine which of the questions are significant predictors in the multivariate model (Figure 13). For this, we check the "Sig." column for values below approximately .05. Seven of the nine questions meet this standard (one just barely). This is useful information, since it tells us which variables are significant, and which are not. We see that satisfaction with room price and facilities are not important when predicting customer satisfaction.

We then use the nonstandardized coefficients in the B column to assess the effect of each predictor. For example, in the B column for "BSI services are a good value," the value .468 means that for every one unit increase in the response to this question, satisfaction increases .468 units. To put it another way, if we compare customers who agree (score=4) that BSI services are a good value to customers who strongly

agree (score=5), we then predict that overall satisfaction will be about .47 units higher for the latter group. This relationship controls for the other variables in the model. In other words, we can make this absolute statement because multiple regression automatically controls for all variables in the model. By using the B coefficients for all significant predictors, we can create a prediction equation to use for overall customer satisfaction.

Note that the B coefficient for frequency of stay is negative. This means that customers who stay more often at BSI hotels are less satisfied than infrequent customers, after we take other factors into account. This could be a key finding for BSI analysts and management to explore.

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	.054	.198		.275	.783
BSI services are a good value	.468	.041	.521	11.444	.000
BSI facilities are kept up to date	.034	.033	.033	1.034	.301
Am satisfd with information received on services	.062	.032	.062	1.955	.051
BSI rooms are priced right	.021	.040	.022	.530	.597
Usefulness of internet services	-.051	.019	-.079	-2.642	.008
How frequently do stay at BSI hotels?	-.070	.018	-.112	-3.861	.000
GOVT	.156	.078	.059	2.010	.045
CORP	.146	.069	.066	2.116	.035
Quality	.520	.052	.351	9.934	.000

a. Dependent Variable: I am a satisfied BSI customer

Figure 13: This regression analysis table highlights the significant predictors of customer satisfaction.

The next step is to determine the relative importance of the variables using the Beta coefficient from the table in Figure 14. Beta values, which vary from –1 to 1, are important to use when questions have different response scales. The higher the absolute value of Beta, the more important the variable is in predicting customer satisfaction. In this case, we see that the most important predictors are the perceived value of services and the "Quality" factor from our factor analysis example.

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	.054	.198		.275	.783
BSI services are a good value	.468	.041	.521	11.444	.000
BSI facilities are kept up to date	.034	.033	.033	1.034	.301
Am satisfd with information received on services	.062	.032	.062	1.955	.051
BSI rooms are priced right	.021	.040	.022	.530	.597
Usefulness of internet services	-.051	.019	-.079	-2.642	.008
How frequently do stay at BSI hotels?	-.070	.018	-.112	-3.861	.000
GOVT	.156	.078	.059	2.010	.045
CORP	.146	.069	.066	2.116	.035
Quality	.520	.052	.351	9.934	.000

a. Dependent Variable: I am a satisfied BSI customer

Figure 14: This table highlights the beta values of the two most important predictors.

These results could enable the BSI hotel chain to develop an efficient and effective strategy for improving customer satisfaction, with a focus on only the most relevant features. The chain could also use this information to make predictions about dissatisfied customers, and target future marketing efforts appropriately.

In the BSI example, regression enables us to understand that customer satisfaction is related more to the perceived value of services than to the price of rooms. Although "BSI services are a good value" is the best predictor, "BSI rooms are priced right" is not even a significant predictor. This is surprising, since we might assume that price has an impact on satisfaction. The discovery suggests that price as compared to perceived value is more important to customers than price alone. Using regression analysis, we gain critical insight into the relationship between these two factors and the real-world attitudes of BSI customers.

Summary

As demonstrated in the examples given in this paper, advanced survey analysis methods enable you to extract detailed, valuable information about the attitudes and behavior of survey respondents.

SPSS offers powerful techniques—such as the cluster, factor, reliability, and regression methods discussed here—that enable you to discover relationships you might miss using more basic software and methods. When it's time to present your results, SPSS provides the professional graphics and reporting capabilities you need to effectively communicate complex ideas.

Any survey researcher can use the advanced methods covered in this paper, because SPSS makes analysis intuitive and straightforward. The more you use the methods presented here, the more comfortable you will become, and the more insight you will gain from your survey data. Further, these advanced methods can give your organization a competitive advantage, and enable you to make more informed decisions in a timely and effective manner.

About SPSS Inc.

SPSS Inc. [NASDAQ: SPSS] is the world's leading provider of predictive analytics software and solutions. The company's predictive analytics technology connects data to effective action by drawing reliable conclusions about current conditions and future events. More than 250,000 commercial, academic, and public sector organizations rely on SPSS technology to help increase revenue, reduce costs, improve processes, and detect and prevent fraud. Founded in 1968, SPSS is headquartered in Chicago, Illinois. To learn more, please visit **www.spss.com**. For SPSS office locations and telephone numbers, go to **www.spss.com/worldwide**.

SPSS

To learn more, please visit www.spss.com. For SPSS office locations and telephone numbers, go to www.spss.com/worldwide.