Basic Stats:

A Supplement to Multivariate Data Analysis

Multivariate Data Analysis

Pearson Prentice Hall Publishing

Table of Contents

Basic Stats:

A Supplement to *Multivariate Data Analysis*

LEARNING OBJECTIVES

In the course of completing this supplement, you will be introduced to the following:

- The basics of simple regression in terms of parameter estimation.
- Determination of statistical significance for both the overall regression model and individual estimated coefficients.
- Differences between the t-test and analysis of variance (ANOVA).
- Fundamentals elements of Bayesian Estimation.
- Estimation of part-worth values and interaction effects in conjoint analysis.
- Calculation of the measure of association or similarity for correspondence analysis.

PREVIEW

This supplement to the text *Multivariate Data Analysis* provides additional coverage of some basic concepts that are the foundations for techniques discussed in the text. The authors felt that readers may benefit from further review of certain topics not covered in the text, but critical to a complete understanding of some of the multivariate techniques. There will be some overlap with material in the chapters so as to fully integrate the concepts.

The supplement is not intended to be a comprehensive "primer" on all of the fundamental statistical concepts, but only those selected issues that add insight into the multivariate techniques discussed in the text. We encourage readers to complement this supplement with other discussions of the basic concepts as needed.

The supplement focuses on concepts related to five multivariate techniques: multiple regression, multivariate analysis of variance, conjoint analysis, correspondence analysis and structural equation modeling. For multiple regression, the discussion first focuses on the basic process of estimation, from setting the baseline standard to estimating a regression coefficient. The discussion then address the process of establishing statistical significance for both the overall regression model, based on the predicted values, and the individual estimated coefficients. The discussion then shifts to the fundamentals underlying the t-test and analysis of variance, which are extended in MANOVA. In each case the basic method of calculation, plus assessing statistical significance, are discussed. For conjoint analysis, two topics emerge. The first is the process of estimating part-worths and interaction effects. Simple examples are utilized to demonstrate the both the process as well as the practical implications. The second topic is a brief introduction to Bayesian estimation, which is emerging as an estimation technique quite common in conjoint analysis. Next, the calculation of the chi-square value and its transformation into a measure of similarity or association for correspondence analysis is discussed. Emphasis is placed on the rationale of moving from counts (nonmetric data) to a metric measure of similarity conducive to analysis. Finally, the path model of SEM is examined as a way to "decomposing" the correlations between constructs into direct and indirect effects. It should be noted that the discussion of additional topics associated with structural equation modeling (SEM) are contained in a separate document available on www.mvstats.com.

KEY TERMS

Before beginning the supplement, review the key terms to develop an understanding of the concepts and terminology used. Throughout the supplement the key terms appear in **boldface.** Other points of emphasis in the chapter and key term cross-references are *italicized*.

Alpha (α) Significance level associated with the statistical testing of the differences between two or more groups. Typically, small values, such as .05 or .01, are specified to minimize the possibility of making a *Type I error.*

Analysis of variance (ANOVA) Statistical technique used to determine whether samples from two or more groups come from populations with equal means (i.e., Do the group means differ significantly?). Analysis of variance examines one dependent measure, whereas multivariate analysis of variance compares group differences on two or more dependent variables.

Chi-square value Method of analyzing data in a contingency table by comparing the actual cell frequencies to an expected cell frequency. The expected cell frequency is based on the marginal probabilities of its row and column (probability of a row and column among all rows and columns).

Coefficient of determination (R^2) Measure of the proportion of the variance of the dependent variable about its mean that is explained by the independent, or predictor, variables. The coefficient can vary between 0 and 1. If the regression model is properly applied and estimated, the researcher can assume that the higher the value of R^2, the greater the explanatory power of the regression equation, and therefore the better the prediction of the dependent variable.

Correlation coefficient (*r*) Coefficient that indicates the strength of the association between any two metric variables. The sign (+ or –) indicates the direction of the relationship. The value can range from +1 to –1, with +1 indicating a perfect positive relationship, 0 indicating no relationship, and –1 indicating a perfect negative or reverse relationship (as one variable grows larger, the other variable grows smaller).

Critical value Value of a test statistic (*t* test, *F* test) that denotes a specified *significance level.* For example, 1.96 denotes a .05 significance level for the *t* test with large sample sizes.

Experimentwide error rate The combined or overall error rate that results from performing multiple *t* tests or *F* tests that are related (e.g., *t* tests among a series of correlated variable pairs or a series of *t*-tests among the pairs of categories in a multichotomous variable).

Factor Nonmetric independent variable, also referred to as a *treatment* or experimental variable.

Intercept (b_0) Value on the *Y* axis (dependent variable axis) where the line defined by the regression equation $Y = b_0 + b_1X_1$ crosses the axis. It is described by the constant term b_0 in the regression equation. In addition to its role in prediction, the intercept may have a managerial interpretation. If the complete absence of the independent variable has meaning, then the intercept represents that amount. For example, when estimating sales from past advertising expenditures, the intercept represents the level of sales expected if advertising is eliminated. But in many instances the constant has only predictive value because in no situation are all independent variables absent. An example is predicting product preference based on consumer attitudes. All individuals have some level of attitude, so the intercept has no managerial use, but it still aids in prediction.

Least squares Estimation procedure used in simple and multiple regression whereby the *regression coefficients* are estimated so as to minimize the total sum of the squared *residuals.*

Multiple regression Regression model with two or more independent variables.

Null hypothesis Hypothesis with samples that come from populations with equal means (i.e., the group means are equal) for either a dependent variable (univariate test) or a set of dependent variables (multivariate test). The null hypothesis can be accepted or rejected depending on the results of a test of statistical significance.

Part correlation Value that measures the strength of the relationship between a dependent and a single independent variable when the predictive effects of the other independent variables in the regression model are removed. The objective is to portray the unique predictive effect due to a single independent variable among a set of independent variables. Differs from the *partial correlation coefficient,* which is concerned with incremental predictive effect.

Part-worth Estimate from conjoint analysis of the overall preference or utility associated with each level of each factor used to define the product or service.

Partial correlation coefficient Value that measures the strength of the relationship between the criterion or dependent variable and a single independent variable when the effects of the other independent variables in the model are held constant. For example, $r_{Y,X2,\ X1}$ measures the variation in *Y* associated with X_2 when the effect of X_1 on both X_2 and *Y* is held constant. This value is used in sequential variable selection methods of regression model estimation to identify the independent variable with the greatest incremental predictive power beyond the independent variables already in the regression model.

Prediction error Difference between the actual and predicted values of the dependent variable for each observation in the sample (see *residual*).

Regression coefficient (*b*$_n$) Numerical value of the parameter estimate directly associated with an independent variable; for example, in the model $Y = b_0 + b_1X_1$, the value b_1 is the regression coefficient for the variable X_1. The regression coefficient represents the amount of change in the dependent variable for a one-unit change in the independent variable. In the multiple predictor model (e.g., $Y = b_0 + b_1X_1 + b_2X_2$), the regression coefficients are partial coefficients because each takes into account not only the relationships between *Y* and X_1 and between *Y* and X_2, but also between X_1 and X_2. The coefficient is not limited in range, as it is based on both the degree of association and the scale units of the independent variable. For instance, two variables with the same association to *Y* would have different coefficients if one independent variable was measured on a 7-point scale and another was based on a 100-point scale.

Residual (*e* or ε) Error in predicting the sample data. Seldom are predictions perfect. We assume that random error will occur, but that this error is an estimate of the true random error in the population (ε), not just the error in prediction for our sample (*e*). We assume that the error in the population we are estimating is distributed with a mean of 0 and a constant (homoscedastic) variance.

Semipartial correlation See *part correlation coefficient.*

Significance level (alpha) Commonly referred to as the level of statistical significance, the significance level represents the probability the researcher is willing to accept that the estimated coefficient is classified as different from zero when it actually is not. This is also known as *Type I error*. The most widely used level of significance is .05, although researchers use levels ranging from .01 (more demanding) to .10 (less conservative and easier to find significance).

Similarity Data used to determine which objects are the most similar to each other and which are the most dissimilar. Implicit in similarities measurement is the ability to compare all pairs of objects.

Simple regression Regression model with a single independent variable, also known as bivariate regression.

Standard error Expected distribution of an estimated regression coefficient. The standard error is similar to the standard deviation of any set of data values, but instead denotes the expected range of the coefficient across multiple samples of the data. It is useful in statistical tests of significance that test to see whether the coefficient is significantly different from zero (i.e., whether the expected range of the coefficient contains the value of zero at a given level of confidence). The *t* value of a *regression coefficient* is the coefficient divided by its standard error.

Standard error of the estimate (SE_E) Measure of the variation in the predicted values that can be used to develop confidence intervals around any predicted value. It is similar to the standard deviation of a variable around its mean, but instead is the expected distribution of predicted values that would occur if multiple samples of the data were taken.

Sum of squared errors (SS_E) Sum of the squared prediction errors (*residuals*) across all observations. It is used to denote the variance in the dependent variable not yet accounted for by the regression model. If no independent variables are used for prediction, it becomes the squared errors using the mean as the predicted value and thus equals the *total sum of squares.*

***t* statistic** Test statistic that assesses the statistical significance between two groups on a single dependent variable (see *t* test). It can be used to assess the difference between two groups or one group from a specified value. Also tests for the difference of an estimated parameter (e.g., *regression coefficient*) from a specified value (i.e., zero).

***t* test** Test to assess the statistical significance of the difference between two sample means for a single dependent variable. The *t* test is a special case of *ANOVA* for two groups or levels of a treatment variable.

Total sum of squares (SS_T) Total amount of variation that exists to be explained by the independent variables. This baseline value is calculated by summing the squared differences between the mean and actual values for the dependent variable across all observations.

Treatment Independent variable (*factor*) that a researcher manipulates to see the effect (if any) on the dependent variables. The treatment variable can have several levels. For example, different intensities of advertising appeals might be manipulated to see the effect on consumer believability.

Type I error Probability of rejecting the null hypothesis when it should be accepted, that is, concluding that two means are significantly different when in fact they are the same. Small values of *alpha* (e.g., .05 or .01), also denoted as α, lead to rejection of the null hypothesis and acceptance of the alternative hypothesis that population means are not equal.

Type II error Probability of failing to reject the null hypothesis when it should be rejected, that is, concluding that two means are not significantly different when in fact they are different. Also known as the beta (β) error.

FUNDAMENTALS OF MULTIPLE REGRESSION

Multiple regression is the most widely used of the multivariate techniques and its principles are the foundation of so many analytical techniques and applications we see today. As discussed in Chapter 4, multiple regression is an extension of simple regression by extending the variate of independent variables from one to multiple variables. So the principles of simple regression are essential in multiple regression as well. The following sections provide a basic introduction to the fundamental concepts of (1) understanding regression model development and interpretation based on the predictive fit of the independent variable(s), (2) establishing statistical significance of parameter estimates of the regression model and (3) partitioning correlation among variables to identify unique and shared effects essential for multiple regression.

THE BASICS OF SIMPLE REGRESSION

The objective of regression analysis is to predict a single dependent variable from the knowledge of one or more independent variables. When the problem involves a single independent variable, the statistical technique is called **simple regression.** When the problem involves two or more independent variables, it is termed **multiple regression.** The following discussion will describe briefly the basic procedure and concepts underlying simple regression. The discussion will first show how regression estimates the relationship between independent and dependent variables. The following topics are covered:

1. *Setting a baseline prediction* without an independent variable, using only the mean of the dependent variable
2. *Prediction using a single independent variable*—simple regression

Setting a Baseline: Prediction Without an Independent Variable

Before estimating the first regression equation, let us start by calculating the baseline against which we will compare the predictive ability of our regression models. The baseline should represent our best prediction without the use of any independent variables. We could use any number of options (e.g., perfect prediction, a prespecified value, or one of the measures of central tendency, such as mean, median, or mode). The baseline predictor used in regression is the simple mean of the dependent variable, which has several desirable properties we will discuss next.

The researcher must still answer one question: *How accurate is the prediction?* Because the mean will not perfectly predict each value of the dependent variable, we must create some way to assess predictive accuracy that can be used with both the baseline prediction and the regression models we create. The customary way to assess the accuracy of any prediction is to examine the errors in predicting the dependent variable.

Although we might expect to obtain a useful measure of prediction accuracy by simply adding the errors, this approach would not be useful because the errors from using the mean value always sum to zero. Therefore, the simple sum of errors would never change, no matter how well or poorly we predicted the dependent variable when using the mean. To overcome this problem, we square each error and then add the results together. This total, referred to as the **sum of squared errors (SS_E),** provides a measure of prediction accuracy that will vary according to the amount of **prediction errors.** *The objective is to obtain the smallest possible sum of squared errors as our measure of prediction accuracy.*

We choose the arithmetic average (mean) because it will always produce a smaller sum of squared errors than any other measure of central tendency, including the median, mode, any other single data value, or any other more sophisticated statistical measure. (Interested readers are encouraged to try to find a better predicting value than the mean.)

Prediction Using a Single Independent Variable: Simple Regression

As researchers, we are always interested in improving our predictions. In the preceding section, we learned that the mean of the dependent variable is the best predictor if we do not use any independent variables. As researchers, however, we are searching for one or more additional variables (independent variables) that might improve on this baseline prediction. When we are looking for just one independent variable, it is referred to as simple regression. This procedure for predicting data (just as the average predicts data) uses the same rule: minimize the sum of squared errors of prediction. *The researcher's objective for simple regression is to find an independent variable that will improve on the baseline prediction.*

The Role of the Correlation Coefficient Although we may have any number of independent variables, the question facing the researcher is: Which one to choose? We could try each variable and see which one gave us the best predictions, but this approach is quite infeasible even when the number of possible independent variables is quite small. Rather, we can rely on the concept of association, represented by the **correlation coefficient.** Two variables are said to be correlated if changes in one variable are associated with changes in the other variable. In this way, as one variable changes, we know how the other variable is changing. The concept of association, represented by the correlation coefficient (*r*), is fundamental to regression analysis by describing the relationship between two variables.

How will this correlation coefficient help improve our predictions? When we use the mean as the baseline prediction we should notice one fact: The mean value never changes. As such, the mean has a correlation of zero with the actual values. How do we improve on this method? We want a variable that, rather than having just a single value, has values that are high when the dependent variable is high and low values when the dependent variable is low. If we can find a variable that shows a similar pattern (a correlation) to the dependent variable, we should be able to improve on our prediction using just the mean. The more similar (higher correlation), the better the prediction will be.

Adding a Constant or Intercept Term In making the prediction of the dependent variable, we may find that we can improve our accuracy by using a constant in the regression model. Known as the **intercept,** it represents the value of the dependent variable when all of the independent variables have a value of zero. Graphically it represents the point at which the line depicting the regression model crosses the *Y* axis, hence the term *intercept*.

An illustration of using the intercept term is depicted in Table 1 for some hypothetical data with a single independent variable X_1. If we find that as X_1 increases by one unit, the dependent variable increases (on the average) by two, we could then make predictions for each value of the independent variable. For example, when X_1 had a value of 4, we would predict a value of 8 (see Table 1a). Thus, the predicted value is always two times the value of X_1 ($2X_1$). However, we often find that the prediction is improved by adding a constant value. In Table 1a we can see that the simple prediction of $2 \times X_1$ is wrong by 2 in every case. Therefore, changing our description to add a constant of 2 to each prediction gives us perfect predictions in all cases (see Table 1b). We will see that when estimating a regression equation, it is usually beneficial to include a constant, which is termed the *intercept.*

Estimating the Simple Regression Equation We can select the "best" independent variable based on the correlation coefficient because the higher the correlation coefficient, the stronger the relationship and hence the greater the predictive accuracy. In the regression equation, we represent the intercept as b_0. The amount of change in the dependent variable due to the independent variable is represented by the term b_1, also known as a **regression coefficient.** Using a mathematical procedure known as **least squares**, we can estimate the values of b_0 and b_1 such that the sum of the squared errors of prediction is minimized. The prediction error, the difference between the actual and predicted values of the dependent variable, is termed the **residual (*e* or ε).**

Interpreting the Simple Regression Model With the intercept and regression coefficient estimated by the least squares procedure, attention now turns to interpretation of these two values:

- *Regression coefficient.* Represents the estimated change in the dependent variable for a unit change of the independent variable. If the regression coefficient is found to be statistically significant (i.e., the coefficient is significantly different from zero), the value of the regression coefficient indicates the extent to which the independent variable is associated with the dependent variable.
- *Intercept.* Interpretation of the intercept is somewhat different. The intercept has explanatory value only within the range of values for the independent variable(s). Moreover, its interpretation is based on the characteristics of the independent variable:
  - In simple terms, the intercept has interpretive value only if zero is a conceptually valid value for the independent variable (i.e., the independent variable can have a value of zero and still maintain its practical relevance). For example, assume that the independent variable is advertising dollars. If it is realistic that, in some situations, no advertising is done, then the intercept will represent the value of the dependent variable when advertising is zero.
  - If the independent value represents a measure that never can have a true value of zero (e.g., attitudes or perceptions), the intercept aids in improving the prediction process, but has no explanatory value.

For some special situations where the specific relationship is known to pass through the origin, the intercept term may be suppressed (called *regression through the origin*). In these cases, the interpretation of the residuals and the regression coefficients changes slightly.

A Simple Example

To illustrate the basic principles of simple regression, we will use the example introduced in Chapter 4 concerning credit card usage. In our discussion we will repeat some of the information provided in the chapter so that the reader can see how simple regression forms the basis for multiple regression.

As described earlier, a small study of eight families investigated their credit card usage and three other factors that might impact credit card usage. The purpose of the study was to determine which factors affected the number of credit cards used. The three potential factors were family size, family income, and number of automobiles owned. A sample of eight families was used (see Table 2). In the terminology of regression analysis, the dependent variable (Y) is the number of credit cards used and the three independent variables (V_1, V_2, and V_3) are family size, family income, and number of automobiles owned, respectively.

The arithmetic average (the mean) of the dependent variable (number of credit cards used) is seven (see Table 3). Our baseline prediction can then be stated as "The predicted number of credit cards used by a family is seven." We can also write this prediction as a regression equation as follows:

Predicted number of credit cards = Average number of credit cards

or

$$\hat{Y} = \bar{Y}$$

With our baseline prediction of each family using seven credit cards, we overestimate the number of credit cards used by family 1 by three. Thus, the error is –3. If this procedure were followed for each family, some estimates would be too high, others would be too low, and still others might be exactly correct. For our survey of eight families, using the average as our baseline prediction gives us the best single predictor of the number of credit cards, with a sum of squared errors of 22 (see Table 3). In our discussion of simple and multiple regression, we use prediction by the mean as a baseline for comparison because it represents the best possible prediction *without using any independent variables.*

In our survey we also collected information on three measures that could act as independent variables. We know that without using any of these independent variables, we can best describe the number of credit cards used as the mean value, 7. But can we do better? Will one of the three independent variables provide information that enables us to make better predictions than achieved by using just the mean?

Using the three independent variables we can try to improve our predictions by reducing the prediction errors. To do so, the prediction errors in the number of credit cards used must be associated (correlated) with one of the potential independent variables (V_1, V_2, or V_3). If V_i is correlated with credit card usage, we can use this relationship to predict the number of credit cards as follows:

Predicted number of credit cards = Change in number of credit cards used × Value of V_i associated with unit change in V_i

or

$$\hat{Y} = b_1 V_1$$

Table 4 contains a correlation matrix depicting the association between the dependent (Y) variable and independent (V_1, V_2, or V_3) variables that can be used in selecting the best independent variable. Looking down the first column, we can see that V_1, family size, has the highest correlation with the dependent variable and is thus the best candidate for our first simple regression. The correlation matrix also contains the correlations among the independent variables, which we will see is important in multiple regression (two or more independent variables).

We can now estimate our first simple regression model for the sample of eight families and see how well the description fits our data. The regression model can be stated as follows:

Predicted number of credit cards used = Intercept +
Change in number of credit cards used
associated with a unit change in family
size ×
Family Size

or

$$\hat{Y} = b_0 + b_1 V_1$$

For this example, the appropriate values are a constant (b_0) of 2.87 and a regression coefficient (b_1) of .97 for family size.

Our regression model predicting credit card holdings indicates that for each additional family member, the credit card holdings are higher on average by .97. The constant 2.87 can be interpreted only within the range of values for the independent variable. In this case, a family size of zero is not possible, so the intercept alone does not have practical meaning. However, this impossibility does not invalidate its use, because it aids in the prediction of credit card usage for each possible family size (in our example from 1 to 5). The simple regression equation and the resulting predictions and residuals for each of the eight families are shown in Table 5.

Because we have used the same criterion (minimizing the sum of squared errors or *least squares*), we can determine whether our knowledge of family size helps us better predict credit card holdings by comparing the simple regression prediction with the baseline prediction. The sum of squared errors using the average (the baseline) was 22; using our new procedure with a single independent variable, the sum of squared errors decreases to 5.50 (see Table 5). By using the least squares procedure and a single independent variable, we see that our new approach, simple regression, is markedly better at predicting than using just the average.

This is a basic example of the process underlying simple regression. The discussion of multiple regression extends this process by adding additional independent variables to improve prediction.

ESTABLISHING STATISTICAL SIGNIFICANCE

Because we used only a sample of observations for estimating a regression equation, we can expect that the regression coefficients will vary if we select another sample of observations and estimate another regression equation. We don't want to take repeated samples, so we need an empirical test to see whether the regression coefficient we estimated has any real value (i.e., is it different from zero?) or could we possibly expect it to equal zero in another sample. To address this issue, regression analysis allows for the statistical testing of the intercept and regression coefficient(s) to determine whether they are significantly different from zero (i.e., they do have an impact that we can expect with a specified probability to be different from zero across any number of samples of observations).

Assessing Prediction Accuracy

If the sum of squared errors (SS_E) represents a measure of our prediction errors, we should also be able to determine a measure of our prediction success, which we can term the **sum of squares regression (SS_R).** Together, these two measures should equal the **total sum of squares (SS_T),** the same value as our baseline prediction. As the researcher adds independent variables, the total sum of squares can now be divided into (1) the sum of squares predicted by the independent variable(s), which is the sum of squares regression (SS_R), and (2) the sum of squared errors (SS_E)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \ddot{y}_i)^2 + \sum_{i=1}^{n}(\ddot{y}_i - \bar{y})^2$$

$$SS_T = SS_E + SS_R$$

where
$\bar{y} = average\ of\ all\ observations$
$y_i = value\ of\ individual\ observation\ i$
$\ddot{y}_i = predicted\ value\ of\ observation\ i$

We can use this division of the total sum of squares to approximate how much improvement the regression variate makes in describing the dependent variable. Recall that the mean of the dependent variable is our best baseline estimate. We know that it is not an extremely accurate estimate, but it is the best prediction available without using any other variables. The question now becomes: *Does the predictive accuracy increase when the regression equation is used rather than the baseline prediction?* We can quantify this improvement by the following:

Sum of squares total (baseline prediction)	SS_{Total}		SS_T
–Sum of squares error (simple regression)	$-SS_{Error}$	or	SS_E
Sum of squares explained (simple regression)	$SS_{Regression}$		SS_R

The sum of squares explained (SS_R) thus represents an improvement in prediction over the baseline prediction. Another way to express this level of prediction accuracy is the **coefficient of determination (*R^2*),** the ratio of the sum of squares regression to the total sum of squares, as shown in the following equation:

$$Coefficient\ of\ determination\ (R^2) = \frac{Sum\ of\ squares\ regression}{Total\ sum\ of\ squares}$$

If the regression model perfectly predicted the dependent variable, R^2 = 1.0. But if it gave no better predictions than using the average (baseline prediction), R^2 = 0. Thus, the R^2 value is a single measure of overall predictive accuracy representing the following:

- The combined effect of the entire variate in prediction, even when the regression equation contains more than one independent variable
- Simply the squared correlation of the actual and predicted values

When the coefficient of correlation (*r*) is used to assess the relationship between dependent and independent variables, the sign of the correlation coefficient (*–r, +r*) denotes the slope of the regression line. However, the strength of the relationship is represented by R^2, which is, of course, always positive. When discussions mention the variation of the dependent variable, they are referring to this total sum of squares that the regression analysis attempts to predict with one or more independent variables.

Establishing a Confidence Interval for the Regression Coefficients and the Predicted Value

With the expected variation in the intercept and regression coefficient across samples, we must also expect the predicted value to vary each time we select another sample of observations and estimate another regression equation. Thus, we would like to estimate the range of predicted values that we might expect, rather than relying just on the single (point) estimate. The point estimate is our best estimate of the dependent variable for this sample of observations and can be shown to be the average prediction for any given value of the independent variable.

From this point estimate, we can also calculate the range of predicted values across repeated samples based on a measure of the prediction errors we expect to make. Known as the **standard error of the estimate (SE_E),** this measure can be defined simply as the expected standard deviation of the prediction errors. For any set of values of a variable, we can construct a confidence interval for that variable about its mean value by adding (plus and minus) a certain number of standard deviations. For example, adding ±1.96 standard deviations to the mean defines a range for large samples that includes 95 percent of the values of a variable.

We can follow a similar method for the predictions from a regression model. Using a point estimate, we can add (plus and minus) a certain number of standard errors of the estimate (depending on the confidence level desired and sample size) to establish the upper and lower bounds for our predictions made with any independent variable(s). The standard error of the estimate (SE_E) is calculated by

$$SE_E = \sqrt{\frac{Sum\ of\ squared\ errors}{Sample\ size - 2}}$$

The number of SE_Es to use in deriving the confidence interval is determined by the level of significance (α) and the sample size (N), which give a *t value*. The confidence interval is then calculated with the lower limit being equal to the predicted value minus ($SE_E \times t$ value) and the upper limit calculated as the predicted value plus ($SE_E \times t$ value).

Significance Testing in the Simple Regression Example

Testing for the significance of a regression coefficient requires that we first understand the hypotheses underlying the regression coefficient and constant. Then we can examine the significance levels for the coefficient and constant.

Understanding the Hypotheses When Testing Regression Coefficients. A simple regression model implies hypotheses about two estimated parameters: the constant and regression coefficient.

> *Hypothesis 1. The intercept (constant term) value is due to sampling error and thus the estimate of the population value is zero.*
>
> *Hypothesis 2. The regression coefficient indicating an increase in the independent variable is associated with a change in the dependent variable also does not differ significantly from zero.*

Assessing the Significance Level. The appropriate test is the *t* test, which is commonly available on regression analysis programs. The *t* value of a coefficient is the coefficient divided by the **standard error**. Thus, the *t* value represents the number of standard errors that the coefficient is different from zero. For example, a regression coefficient of 2.5 with a standard error of .5 would have a *t* value of 5.0 (i.e., the regression coefficient is 5 standard errors from zero). To determine whether the coefficient is significantly different from zero, the computed *t* value is compared to the table value for the sample size and confidence level selected. If our value is greater than the table value, we can be confident (at our selected confidence level) that the coefficient does have a statistically significant effect in the regression variate.

Most programs calculate the **significance level** for each regression coefficient's *t* value, showing the significance level at which the confidence interval would include zero. The researcher can then assess whether this level meets the desired level of significance. For example, if the statistical significance of the coefficient is .02, then we would say it was significant at the .05 level (because it is less than .05), but not significant at the .01 level.

From a practical point of view, significance testing of the constant term is necessary only when used for explanatory value. If it is conceptually impossible for observations to exist with all the independent variables measured at zero, the constant term is outside the data and acts only to position the model.

The researcher should remember that the statistical test of the regression coefficient and the constant is to ensure—across all the possible samples that could be drawn—the estimated parameters would be different from zero within a specified level of acceptable error.

Significance Testing in the Credit Card Example

In our credit card example, the baseline prediction is the average number of credit cards held by our sampled families and is the best prediction available without using any other variables. The baseline prediction using the mean was measured by calculating the squared sum of errors for the baseline (sum of squares = 22). Now that we have fitted a regression model using family size, does it explain the variation better than the average? We know it is somewhat better because the sum of squared errors is now reduced to 5.50. We can look at how well our model predicts by examining this improvement.

Sum of squares total (baseline prediction)	SS_{Total}		SS_T	22.0
–Sum of squares error (simple regression)	$-SS_{Error}$	or	SS_E	–5.5
Sum of squares explained (simple regression)	$SS_{Regression}$		SS_R	16.5

Therefore, we explained 16.5 squared errors by changing from the average to the simple regression model using family size. It is an improvement of 75 percent (16.5 ÷ 22 = .75) over the baseline. We thus state that the coefficient of determination (R^2) for this regression equation is .75, meaning that it explained 75 percent of the possible variation in the dependent variable. Also remember that the R^2 value is simply the squared correlation of the actual and predicted values.

For our simple regression model, SE_E equals +.957 (the square root of the value of 5.50 divided by 6). The confidence interval for the predictions is constructed by selecting the number of standard errors to add (plus and minus) by looking in a table for the *t* distribution and selecting the value for a given confidence level and sample size. In our example, the *t* value for a 95% confidence level with 6 degrees of freedom (sample size minus the number of coefficients, or 8 – 2 = 6) is 2.447. The amount added (plus and minus) to the predicted value is then (.957 × 2.447), or 2.34. If we substitute the average family size (4.25) into the regression equation, the predicted value is 6.99 (it differs from the average of 7 only because of rounding). The expected range of credit cards becomes 4.65 (6.99 – 2.34) to 9.33 (6.99 + 2.34). The confidence interval can be applied to any predicted value of credit cards.

The regression equation for credit card usage seen earlier was:

$$Y = b_0 + b_1V_1$$

or

$$Y = 2.87 + .971(\text{family size})$$

This simple regression model requires testing two hypotheses for the each estimated parameter (the constant value of 2.87 and the regression coefficient of .971). These hypotheses (commonly called the null hypothesis) can be stated formally as:

> *Hypothesis 1. The intercept (constant term) value of 2.87 is due to sampling error, and the real constant term appropriate to the population is zero.*
>
> *Hypothesis 2. The regression coefficient of .971(indicating that an increase of one unit in family size is associated with an increase in the average number of credit cards held by .971) also does not differ significantly from zero.*

With these hypotheses, we are testing whether the constant term and the regression coefficient have an impact different from zero. If they are found not to differ significantly from zero, we would assume that they should not be used for purposes of prediction or explanation.

Using the *t* test for the simple regression example, we can assess whether either the constant or the regression coefficient is significantly different from zero. In this example, the intercept has no explanatory value because in no instances do all independent variables have values of zero (e.g., family size cannot be zero). Thus, statistical significance is not an issue in interpretation. If the regression coefficient is found to occur solely because of sampling error (i.e., zero fell within the calculated confidence interval), we would conclude that family size has no generalizable impact on the number of credit cards held beyond this sample. Note that this test is not one of any exact value of the coefficient but rather of whether it has any generalizable value beyond this sample.

In our example, the standard error of family size in the simple regression model is .229. The calculated *t* value is 4.24 (calculated as .971 ÷ .229), which has a probability of .005. If we are using a significance level of .05, then the coefficient is significantly different from zero. If we interpret the value of .005 directly, it means that we can be sure with a high degree of certainty (99.5%) that the coefficient is different from zero and thus should be included in the regression equation.

<u>Summary of Significance Testing</u>

Significance testing of regression coefficients provides the researcher with an empirical assessment of their "true" impact. Although it is not a test of validity, it does determine whether the impacts represented by the coefficients are generalizable to other samples from this population. One note concerning the variation in regression coefficients is that many times researchers forget that the estimated coefficients in their regression analysis are specific to the sample used in estimation. They are the best estimates for that sample of observations, but as the preceding results showed, the coefficients can vary quite markedly from sample to sample. This potential variation points to the need for concerted efforts to validate any regression analysis on a different sample(s). In doing so, the researcher must expect the coefficients to vary, but the attempt is to demonstrate that the relationship generally holds in other samples so that the results can be assumed to be generalizable to any sample drawn from the population.

CALCULATING UNIQUE AND SHARED VARIANCE AMONG INDEPENDENT VARIABLES

The basis for estimating all regression relationships is the correlation, which measures the association between two variables. In regression analysis, the correlations between the independent variables and the dependent variable provide the basis for forming the regression variate by estimating regression coefficients (weights) for each independent variable that maximize the prediction (explained variance) of the dependent variable. When the variate contains only a single independent variable, the calculation of regression coefficients is straightforward and based on the bivariate (or zero order) correlation between the independent and dependent variables. The percentage of explained variance of the dependent variable is simply the square of the bivariate correlation.

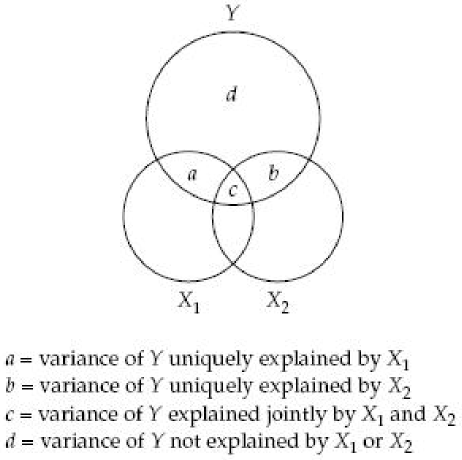
The Role of Partial and Part Correlations

As independent variables are added to the variate, however, the calculations must also consider the intercorrelations among independent variables. If the independent variables are correlated, then they share some of their predictive power. Because we use only the prediction of the overall variate, the shared variance must not be double counted by using just the bivariate correlations. Thus, we calculate two additional forms of the correlation to represent these shared effects:

1. The **partial correlation coefficient** is the correlation of an independent (X_i) and dependent (Y) variable when the effects of the other independent variable(s) have been removed from both X_i and Y.
2. The **part** or **semipartial correlation** reflects the correlation between an independent and dependent variable while controlling for the predictive effects of all other independent variables on X_i.

The two forms of correlation differ in that the partial correlation removes the effects of other independent variables from X_i and Y, whereas the part correlation removes the effects only from X_i. The partial correlation represents the incremental predictive effect of one independent variable from the collective effect of all others and is used to identify independent variables that have the greatest incremental predictive power when a set of independent variables is already in the regression variate. The part correlation represents the unique relationship predicted by an independent variable after the predictions shared with all other independent variables are taken out. Thus, the part correlation is used in apportioning variance among the independent variables. Squaring the part correlation gives the unique variance explained by the independent variable.

The figure below portrays the shared and unique variance among two correlated independent variables.



a = variance of *Y* uniquely explained by X_1
b = variance of *Y* uniquely explained by X_2
c = variance of *Y* explained jointly by X_1 and X_2
d = variance of *Y* not explained by X_1 or X_2

The variance associated with the partial correlation of X_2 controlling for X_1 can be represented as $b \div (d + b)$, where $d + b$ represents the unexplained variance after accounting for X_1. The part correlation of X_2 controlling for X_1 is $b \div (a + b + c + d)$, where $a + b + c + d$ represents the total variance of *Y,* and *b* is the amount uniquely explained by X_2.

The analyst can also determine the shared and unique variance for independent variables through simple calculations. The part correlation between the dependent variable (*Y*) and an independent variable (X_1) while controlling for a second independent variable (X_2) is calculated by the following equation:

$$Part\ correlation\ of\ Y, X_1, given\ X_2 = \frac{Corr\ of\ Y, X_1 - (Corr\ of\ Y, X_2\ \times Corr\ of\ X_1, X_2)}{\sqrt{1.0 -\ (Corr\ of\ X_1, X_2)}}$$

A simple example with two independent variables (X_1 and X_2) illustrates the calculation of both shared and unique variance of the dependent variable (Y). The direct correlations and the correlation between X_1 and X_2 are shown in the following correlation matrix:

	Y	X_1	X_2
Y	1.0		
X_1	.60	1.0	
X_2	.50	.70	1.0

The direct correlations of .60 and .50 represent fairly strong relationships with Y, but the correlation of .70 between X_1 and X_2 means that a substantial portion of this predictive power may be shared. The part correlation of X_1 and Y controlling for X_2 ($r_{Y,X_1(X_2)}$) and the unique variance predicted by X_1 can be calculated as:

$$r_{Y,X_1(X_2)} = \frac{.60 - (.50 \times .70)}{\sqrt{1.0 - .70^2}} = .35$$

Thus, the unique variance predicted by X_1 = $.35^2$ = .1225

Because the direct correlation of X_1 and Y is .60, we also know that the total variance predicted by X_1 is $.60^2$, or .36. If the unique variance is .1225, then the shared variance must be .2375 (.36 – .1225).

We can calculate the unique variance explained by X_2 and confirm the amount of shared variance by the following:

$$r_{Y,X_2(X_1)} = \frac{.50 - (.60 \times .70)}{\sqrt{1.0 - .70^2}} = .11$$

From this calculation, the unique variance predicted by X_2 = $.11^2$ = .0125

With the total variance explained by X_2 being $.50^2$, or .25, the shared variance is calculated as .2375 (.25 – .0125). This result confirms the amount found in the calculations for X_1. Thus, the total variance (R^2) explained by the two independent variables is

Unique variance explained by X_1	.1225
Unique variance explained by X_2	.0125
Shared variance explained by X_1 and X_2	.2375
Total variance explained by X_1 and X_2	.3725

These calculations can be extended to more than two variables, but as the number of independent variables increases, it is easier to allow the statistical programs to perform the calculations. The calculation of shared and unique variance illustrates the effects of multicollinearity on the ability of the independent variables to predict the dependent variable.

FUNDAMENTALS OF MANOVA

The following discussion addresses the two most common types of univariate procedures for assessing group differences, the *t*-test, which compares a dependent variable across two groups, and ANOVA, which is used whenever the number of groups is two or more. These two techniques are the basis on which we extend the testing of multiple dependent variables between groups in MANOVA.

THE *t*-TEST

The ***t*-test** assesses the statistical significance of the difference between two independent sample means for a single dependent variable. In describing the basic elements of the *t* test and other tests of group differences, we will address two topics: design of the analysis and statistical testing.

Analysis Design

The difference in group mean scores is the result of assigning observations (e.g., respondents) to one of the two groups based on their value of a nonmetric variable known as a **factor** (also known as a **treatment**). A factor is a nonmetric variable, many times employed in an **experimental design** where it is manipulated with prespecified categories or levels that are proposed to reflect differences in a dependent variable. A factor can also just be an observed nonmetric variable, such as gender. In either instance, the analysis is fundamentally the same.

An example of a simple experimental design will be used to illustrate such an analysis. A researcher is interested in how two different advertising messages—one informational and the other emotional—affect the appeal of the advertisement. To assess the possible differences, two advertisements reflecting the different appeals are prepared. Respondents are then randomly selected to receive either the informational or emotional advertisement. After examining the advertisement, each respondent is asked to rate the appeal of the message on a 10-point scale, with 1 being poor and 10 being excellent.

The two different advertising messages represent a single experimental factor with two levels (informational versus emotional). The appeal rating becomes the dependent variable. The objective is to determine whether the respondents viewing the informational advertisement have a significantly different appeal rating than those viewing the advertisement with the emotional message. In this instance the factor was experimentally manipulated (i.e., the two levels of message type were created by the researcher), but the same basic process could be used to examine difference on a dependent variable for any two groups of respondents (e.g., male versus female, customer versus noncustomer, etc.). With the respondents assigned to groups based on their value of the factor, the next step is to assess whether the differences between the groups in terms of the dependent variable are statistically significant.

Statistical Testing

To determine whether the treatment has an effect (i.e., Did the two advertisements have different levels of appeal?), a statistical test is performed on the differences between the mean scores (i.e., appeal rating) for each group (those viewing the emotional advertisements versus those viewing the informational advertisements).

Calculating the t Statistic The measure used is the ***t* statistic,** defined in this case as the ratio of the difference between the sample means ($\mu_1 - \mu_2$) to their standard error. The **standard error** is an estimate of the difference between means to be expected because of sampling error. If the actual difference between the group means is sufficiently larger than the standard error, then we can conclude that these differences are statistically significant. We will address what level of the *t* statistic is needed for statistical significance in the next section, but first we can express the calculation in the following equation:

$$t\ statistic = \frac{\mu_1 - \mu_2}{SE_{\mu_1 \mu_2}}$$

where

μ_1 = mean of group 1

μ_2 = mean of group 2

SE$\mu_1\mu_2$ = standard error of the difference in group means

In our example of testing advertising messages, we would first calculate the mean score of the appeal rating for each group of respondents (informational versus emotional) and then find the difference in their mean scores ($\mu_{\text{informational}} – \mu_{\text{emotional}}$). By forming the ratio of the actual difference between the means to the difference expected due to sampling error (the standard error), we quantify the amount of the actual impact of the treatment that is due to random sampling error. In other words, the *t* value, or *t* statistic, represents the group difference in terms of standard errors.

Interpreting the t Statistic How large must the *t* value be to consider the difference statistically significant (i.e., the difference was not due to sampling variability, but represents a true difference)? This determination is made by comparing the *t* statistic to the **critical value** of the *t* statistic (t_{crit}). We determine the critical value (t_{crit}) for our *t* statistic and test the statistical significance of the observed differences by the following procedure:

1. Compute the *t* statistic as the ratio of the difference between sample means to their standard error.
2. Specify a **Type I error** level (denoted as **alpha, ⍰**, or **significance level**), which indicates the probability level the researcher will accept in concluding that the group means are different when in fact they are not.
3. Determine the critical value (t_{crit}) by referring to the *t* distribution with $N_1 + N_2 – 2$ degrees of freedom and a specified ⍰ level, where N_1 and N_2 are sample sizes. Although the researcher can use the statistical tables to find the exact value, several typical values are used when the total sample size ($N_1 + N_2$) is at least greater than 50. The following are some widely used α levels and the corresponding t_{crit} values:

a (Significance) Level	t_{crit} Value
.10	1.64
.05	1.96
.01	2.58

4. If the absolute value of the computed *t* statistic exceeds t_{crit}, the researcher can conclude that the two groups do exhibit differences in group means on the dependent measure (i.e., $\mu_1 = \alpha_2$), with a Type I error probability of α. The researcher can then examine the actual mean values to determine which group is higher on the dependent value.

The computer software of today provides the calculated *t* value and the associated significance level, making interpretation even easier. The researcher merely needs to see whether the significance level meets or exceeds the specified

Type I error level set by the researcher.

The *t*-test is widely used because it works with small group sizes and is quite easy to apply and interpret. It does face a couple of limitations: (1) it only accommodates two groups; and (2) it can only assess one independent variable at a time. To remove either or both of these restrictions, the researcher can utilize analysis of variance, which can test independent variables with more than two groups as well as simultaneously assessing two or more independent variables.

ANALYSIS OF VARIANCE

In our example for the *t*-test, a researcher exposed two groups of respondents to different advertising messages and subsequently asked them to rate the appeal of the advertisements on a 10-point scale. Suppose we are interested in evaluating three advertising messages rather than two. Respondents would be randomly assigned to one of three groups, resulting in three sample means to compare. To analyze these data, we might be tempted to conduct separate *t*-tests for the difference between each pair of means (i.e., group 1 versus group 2; group 1 versus group 3; and group 2 versus group 3).

However, multiple *t*-tests inflate the overall Type I error rate (we discuss this issue in more detail in the next section). **Analysis of variance (ANOVA)** avoids this Type I error inflation due to making multiple comparisons of treatment groups by determining in a single test whether the entire set of sample means suggests that the samples were drawn from the same general population. That is, ANOVA is used to determine the probability that differences in means across several groups are due solely to sampling error.

Analysis Design

ANOVA provides considerably more flexibility in testing for group differences than found in the *t*-test. Even though a *t*-test can be performed with ANOVA, the researcher also has the ability to test for differences between more than two groups as well as test more than one independent variable. Factors are not limited to just two levels, but instead can have as many levels (groups) as desired. Moreover, the ability to analyze more than one independent variable enables the researcher more analytical insight into complex research questions that could not be addressed by analyzing only one independent variable at a time.

With this increased flexibility comes additional issues, however. Most important is the sample size requirements from increasing either the number of levels or the number of independent variables. For each group, a researcher may wish to have a sample size of 20 or so observations. As such, increasing the number of levels in any factor requires an increase in sample size. Moreover, analyzing multiple factors can create a situation of large sample size requirements rather quickly. Remember, when two or more factors are included in the analysis, the number of groups formed is the product of the number of levels, not their sum (i.e., Number of groups = Number of Levels$_{\text{Factor 1}}$ × Number of Levels$_{\text{Factor 2}}$). A simple example illustrates the issue.

The two levels of advertising message (informational and emotional) would require a total sample of 50 if the researcher desired 25 respondents per cell. Now, let's assume that a second factor is added for the color of the advertisement with three levels (1 = color, 2 = black-and-white, and 3 = combination of both). If both factors are now included in the analysis, the number of groups increases to six (Number of groups = 2 × 3) and the sample size increases to 150 respondents (Sample size = 6 groups × 25 respondents per group). So we can see that adding even just a three-level factor can increase the complexity and sample required.

Thus, researchers should be careful when determining both the number of levels for a factor as well as the number of factors included, especially when analyzing field surveys, where the ability to get the necessary sample size per cell is much more difficult than in controlled settings.

Statistical Testing

The logic of an ANOVA statistical test is fairly straightforward. As the name *analysis of variance* implies, two independent estimates of the variance for the dependent variable are compared. The first reflects the general variability of respondents within the groups (MS_W) and the second represents the differences between groups attributable to the treatment effects (MS_B):

- *Within-groups estimate of variance* (MS_W: mean square within groups): This estimate of the average respondent variability on the dependent variable within a treatment group is based on deviations of individual scores from their respective group means. MS_W is comparable to the standard error between two means calculated in the *t* test as it represents variability within groups. The value MS_W is sometimes referred to as the error variance.
- *Between-groups estimate of variance* (MS_B: mean square between groups): The second estimate of variance is the variability of the treatment group means on the dependent variable. It is based on deviations of group means from the overall grand

mean of all scores. Under the **null hypothesis** of no treatment effects (i.e., $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$), this variance estimate, unlike MS_W, reflects any treatment effects that exist; that is, differences in treatment means increase the expected value of MS_B. Note that any number of groups can be accommodated.

Calculating the F Statistic. The ratio of MS_B to MS_W is a measure of how much variance is attributable to the different treatments versus the variance expected from random sampling. The ratio of MS_B to MS_W is similar in concept to the *t* value, but in this case gives us a value for the *F* statistic.

$$F\ statistic = \frac{MS_B}{MS_W}$$

Because differences between the groups inflate MS_B, large values of the *F* statistic lead to rejection of the null hypothesis of no difference in means across groups. If the analysis has several different treatments (independent variables), then estimates of MS_B are calculated for each treatment and *F* statistics are calculated for each treatment. This approach allows for the separate assessment of each treatment.

Interpreting the F Statistic. To determine whether the *F* statistic is sufficiently large to support rejection of the null hypothesis (meaning that differences are present between the groups), follow a process similar to the *t* test:

1. Determine the critical value for the *F* statistic (F_{crit}) by referring to the *F* distribution with $(k - 1)$ and $(N - k)$ degrees of freedom for a specified level of ⍰ (where $N = N_1 + \ldots + N_k$ and k = number of groups). As with the *t* test, a researcher can use certain *F* values as general guidelines when sample sizes are relatively large.

   These values are just the t_{crit} value squared, resulting in the following values:

a (Significance) Level	F_{crit} Value
.10	2.68
.05	3.84
.01	6.63

2. Calculate the *F* statistic or find the calculated *F* value provided by the computer program.

3. If the value of the calculated *F* statistic exceeds F_{crit}, conclude that the means across all groups are not all equal. Again, the computer programs provide the *F* value and the associated significance level, so the researcher can directly assess whether it meets an acceptable level.

Examination of the group means then enables the researcher to assess the relative standing of each group on the dependent measure. Although the *F* statistic test assesses the null hypothesis of equal means, it does not address the question of which means are different. For example, in a three-group situation, all three groups may differ significantly, or two may be equal but differ from the third. To assess these differences, the researcher can employ either planned comparisons or *post hoc* tests.

FUNDAMENTALS OF CONJOINT ANALYSIS

Conjoint analysis represents a somewhat unique analytical technique among multivariate methods due to use of repeated responses from each respondent. As such, it presents unique estimation issues in deriving model estimates for each individual. The following sections describe not only the process of estimating part-worths, but also the approach use in identifying interaction effects in the results.

ESTIMATING PART-WORTHS

How do we estimate the **part-worths** for each level when we have only rankings or ratings of the stimuli? In the following discussion we examine how the evaluations of each stimulus can be used to estimate the part-worths for each level and ultimately define the importance of each attribute as well.

Assuming that a basic model (an additive model) applies, the impact of each level is calculated as the difference (deviation) from the overall mean ranking. (Readers may note that this approach is analogous to multiple regression with dummy variables or ANOVA.)

We can apply this basic approach to all factors and calculate the part-worths of each level in four steps:

Step 1: Square the deviations and find their sum across all levels.

Step 2: Calculate a standardizing value that is equal to the total number of levels divided by the sum of squared deviations.

Step 3: Standardize each squared deviation by multiplying it by the standardizing value.

Step 4: Estimate the part-worth by taking the square root of the standardized squared deviation.

We should note that these calculations are done for each respondent separately and the results from one respondent do not affect the results for any other respondent. This approach differs markedly from such techniques as regression or ANOVA where we deal with correlations across all respondents or group differences.

A Simple Example

To illustrate a simple conjoint analysis, we can examine the responses for respondent 1. If we first focus on the rankings for each attribute, we see that the rankings for the stimuli with the phosphate-free ingredients are the highest possible (1, 2, 3, and 4), whereas the phosphate-based ingredient has the four lowest ranks (5, 6, 7, and 8). Thus, a phosphate-free ingredient is clearly preferred over a phosphate-based cleanser. These results can be contrasted to rankings for each brand level, which show a mixture of high and low rankings for each brand.

For example, the average ranks for the two cleanser ingredients (phosphate-free versus phosphate-based) for respondent 1 are:

Phosphate-free:	(1 + 2 + 3 + 4)/4 = 2.5
Phosphate-based:	(5 + 6 + 7 + 8)/4 = 6.5

With the average rank of the eight stimuli of 4.5 [(1 + 2 + 3 + 4 + 5 + 6 + 7 + 8)/8 = 36/8 = 4.5], the phosphate-free level would then have a deviation of –2.0 (2.5 – 4.5) from the overall average, whereas the phosphate-based level would have a deviation of +2.0 (6.5 – 4.5). The average ranks and deviations for each factor from the overall average rank (4.5) for respondents 1 and 2 are given in Table A-6.

In this example, we use smaller numbers to indicate higher ranks and a more preferred stimulus (e.g., 1 = most preferred). When the preference measure is inversely related to preference, such as here, we reverse the signs of the deviations in the part-worth calculations so that positive deviations will be associated with part-worths indicating greater preference.

Let us examine how we would calculate the part-worth of the first level of ingredients (phosphate-free) for respondent 1 following the four step process described earlier. The calculations for each step are as follows:

Step 1: The deviations from 2.5 are squared. The squared deviations are summed (10.5).

Step 2: The number of levels is 6 (three factors with two levels apiece). Thus, the standardizing value is calculated as .571 (6/10.5 = .571).

Step 3: The squared deviation for phosphate-free (2^2 = 4; remember that we reverse signs) is then multiplied by .571 to get 2.284 (2^2 x .571 = 2.284).

Step 4: Finally, to calculate the part-worth for this level, we then take the square root of 2.284, for a value of 1.1511. This process yields part-worths for each level for respondents 1 and 2, as shown in Table A-7.

ESTIMATING INTERACTION EFFECTS ON PART-WORTH ESTIMATES

Interactions are first identified by unique patterns within the preference scores of a respondent. If they are not included in the additive model, they can markedly affect the estimated preference structure. We will return to our industrial cleanser example to illustrate how interactions are reflected in the preference scores of a respondent.

Obtaining Preference Evaluations

In our earlier example of an industrial cleanser, we can posit a situation where the respondent makes choices in which interactions appear to influence the choices. Assume a third respondent made the following preference ordering:

Rankings of Stimuli Formed by Three Factors (1 = most preferred, 8 = least preferred)

		Form	
		Liquid	Powder
Brand	Ingredients		
HBAT	Phosphate-Free	1	2
	Phosphate-Based	3	4
Generic	Phosphate-Free	7	8
	Phosphate-Based	5	6

> Assume that the “true” preference structure for this respondent should reflect a preference for the HBAT brand, liquid over powder, and phosphate-free over phosphate-based cleansers. However, a bad experience with a generic cleanser made the respondent select phosphate-based over phosphate-free only if it was a generic brand. This choice “goes against” the general preferences and is reflected in an interaction effect between the factors of brand and ingredients.

Estimating the Conjoint Model

If we utilize only an additive model, we obtain the following part-worth estimates:

Part-Worth Estimates for Respondent 3: Additive Model

Form		*Ingredients*		*Brand*	
Liquid	Powder	Phosphate-Free	Phosphate-Based	HBAT	Generic
.42	–.42	0.0	0.0	1.68	–1.68

When we examine the part-worth estimates, we see that the values have been confounded by the interaction. Most noticeably, both levels in the ingredients factor have part-worths of 0.0, yet we know that the respondent actually preferred phosphate-free. The estimates are misleading because the main effects of brand and ingredients are confounded by the interactions shown by the reversal of the preference order when the generic brand was involved.

The impact extends past just the part-worth estimates, as we can see how it also affects the overall utility scores and the predicted rankings for each stimulus. The following table compares the actual preference ranks with the calculated and predicted ranks using the additive model.

	HBAT				*Generic*			
	Phosphate-Free		*Phosphate-Based*		*Phosphate-Free*		*Phosphate-Based*	
	Liquid	Powder	Liquid	Powder	Liquid	Powder	Liquid	Powder
Actual Rank	1	2	3	4	7	8	5	6
Calculated Utility	2.10	1.26	2.10	1.26	–1.26	–2.10	–1.26	–2.10
Predicted Rank[a]	1.5	3.5	1.5	3.5	5.5	7.5	5.5	7.5

[a] Higher utility scores represent higher preference and thus higher rank. Also, tied ranks are shown as mean rank (e.g., two stimuli tied for 1 and 2 are given rank of 1.5).

The predictions are obviously less accurate, given that we know interactions exist. If we were to proceed with only an additive model, we would violate one of the principal assumptions and potentially make quite inaccurate predictions.

Identifying Interactions

Examining for first-order interactions is a reasonably simple task with two steps. Using the previous example with three factors, we will illustrate each step:

- Form 3 two-way matrices of preference order. In each matrix, sum the two preference orders for the third factor.

For example, the first matrix might be the combinations of form and ingredients with each cell containing the sum of the two preferences for brand. We illustrate the process with two of the possible matrices here:

	Matrix 1[a]		Matrix 2[b]	
	Form		Ingredients	
Brand	Liquid	Powder	Phosphate-Free	Phosphate-Based
HBAT	1 + 3	2 + 4	1 + 2	3 + 4
Generic	7 + 5	8 + 6	7 + 8	5 + 6

[a] Values are the two preference rankings for Ingredients.

[b] Values are the two preference rankings for Form.

- To check for interactions, the diagonal values are then added and the difference is calculated. If the total is zero, then no interaction exists.

For matrix 1, the two diagonals both equal 18 (1 + 3 + 8 + 6 = 7 + 5 + 2 + 4). But in matrix 2, the total of the diagonals is not equal (1 + 2 + 5 + 6 ≠ 7 + 8 + 3+ 4). This difference indicates an interaction between brand and ingredients, just as described earlier. The degree of difference indicates the strength of the interaction.

As the difference becomes greater, the impact of the interaction increases, and it is up to the researcher to decide when the interactions pose enough problems in prediction to warrant the increased complexity of estimating coefficients for the interaction terms.

Basic Concepts in Bayesian Estimation

The use of Bayesian estimation has found its way into many of the multivariate techniques we have discussed in the text: factor analysis, multiple regression, discriminant analysis, logistic regression, MANOVA, cluster analysis and structural equation modeling, it has found particular use in conjoint analysis. While it is beyond the scope of this section to detail the advantages of using Bayesian estimation or its statistical principles, we will discuss in general terms how the process works in estimating the probability of an event.

Let's examine a simple example to illustrate how this approach works in estimating the probability of an event happening. Assume that a firm is trying to understand the impact of their Loyalty program in getting individuals to purchase an upgraded warranty program. The question involves whether to continue to support the Loyalty program as a way to increased purchases of the upgrades. A survey of customers found the following results:

	Likelihood Probability of:	
Type of Customer	Purchasing an Upgrade	Not Purchasing an Upgrade
Member of Loyalty Program	.40	.60
Not in the Loyalty Program	.10	.90

If we just look at these results, we see that members of the Loyalty program are 4 times as likely as nonmembers to purchase an upgrade. This figure represents the likelihood probability (i.e., the probability of purchasing an upgrade based on the type of customer) that we can estimate directly from the data.

The likelihood probability is only part of the analysis, because we still need to know one more probability: How likely are customers to join the Loyalty program? This prior probability describes the probability of any single customer joining the Loyalty program.

If we estimate that 10 percent of our customers are members of the Loyalty program, we can now estimate the probabilities of any type of customer purchasing an upgrade. We do so by multiplying the prior probability times the likelihood probability to obtain the joint probability. For our example, this calculation results in the following:

JOINT PROBABILITIES OF PURCHASING AN UPGRADE

		Joint Probability .		
Type of Customer	Prior Probability	Purchase	No Purchase	Total
Member of Loyalty Program	10%	.04	.06	.10
Not in Loyalty Program	90%	.09	.81	.90
Total		.13	.87	1.00

Now we can see that even though members of the Loyalty program purchase upgrades at a much higher rate than nonmembers, the relatively small proportion of customers in the Loyalty program (10%) makes them a minority of upgrade purchasers. As a matter of fact, Loyalty program members only purchase about 30 percent (.04 ÷ .13 = .307) of the upgrades.

Bayesian estimation provides advantages in certain situations in that we can estimate the two probabilities using different assumptions and/or sets of respondents. For example, when we can assume that there is heterogeneity among respondents, we can “divide” the probability of an outcome into a probability general to the entire sample and then use individual-level data to try and estimate more unique effects for each individual. The robustness of Bayesian estimation plus its ability to handle heterogeneity among the respondents has made it a popular alternative to the conventional estimation procedure in many multivariate techniques.

FUNDAMENTALS OF CORRESPONDENCE ANALYSIS

Correspondence analysis is a unique multivariate technique in that it uses only non-metric data. While researchers may have a preference for use of metric data because of the greater amount of information that can be conveyed, the practical reality is that a large number of many research situations involve non-metric data. This creates a need for some form of multivariate analysis that provides the researcher a method of analyzing and portraying the association between these types of measures.

The following section details the process of creating a measure of association that is the basis for correspondence analysis. It is based on one of the most basic of statistical concepts – chi square. We describe briefly how a chi-square measure is calculated in a simple example and then its transformation into a measure of association for individual categories of two or more nonmetric variables.

CALCULATING A MEASURE OF ASSOCIATION OR SIMILARITY

Correspondence analysis uses one of the most basic statistical concepts, chi-square, to standardize the cell frequency values of the contingency table and form the basis for association or similarity. Chi-square is a standardized measure of actual cell frequencies compared to expected cell frequencies. In cross-tabulated data, each cell contains the values for a specific row/column combination. An example of cross-tabulated data reflecting product sales (Product A, B or C) by age category (Young Adults, Middle Age or Mature Individuals) is shown in Table 8. The chi-square procedure proceeds in four steps to calculate a chi-square value for each cell and then transform it into a measure of association.

Step 1: Calculate Expected Sales

The first step is to calculate the expected value for a cell as if no association existed. The expected sales are defined as the joint probability of the column and row combination. This joint probability is calculated as the marginal probability for the column (column total ÷ overall total) times the marginal probability for the row (row total ÷ overall total). This value is then multiplied by the overall total. For any cell, the expected value can be simplified to the following equation:

$$Expected\ cell\ count = \frac{Column\ total\ of\ cell\ x\ Row\ total\ of\ cell}{Overall\ total}$$

This calculation represents the expected cell frequency given the proportions for the row and column totals. In our simple example, the expected sales for Young Adults buying Product A is 21.82 units, as shown in the following calculation:

$$Expected\ sales_{Young\ Adults,Product\ A} = \frac{60\ x\ 80}{220} = 21.82$$

This calculation is performed for each cell, with the results shown in Table 9. The expected cell counts provide a basis for comparison with the actual cell counts and allow for the calculation of a standardized measure of association used in constructing the perceptual map.

Step 2: Difference in Expected and Actual Cell Counts

The next step is to calculate the difference between the expected and actual cell counts as follows:

Difference = Expected cell count – Actual cell count

The amount of difference denotes the strength of the association and the sign (positive for less association than expected and negative for greater association than expected) represented in this value. It is important to note that the sign is actually reversed from the type of association—a negative sign means positive association (actual cell counts exceeded expected), and vice versa.

Again, in our example of the cell for Young Adults purchasing Product A, the difference is 1.82 (21.82 – 20.00). The positive difference indicates that the actual sales are lower than expected for this age group–product combination, meaning fewer sales than would be expected (a negative association). Cells in which negative differences occur indicate positive associations (that cell actually bought more than expected). The differences for each cell are also shown in Table 9.

Step 3: Calculate the Chi-Square Value

The final step is to standardize the differences across cells so that comparisons can be easily made. Standardization is required because it would be much easier for differences to occur if the cell frequency was high compared to a cell with only a few sales. So we standardize the differences to form a chi-square value by dividing each squared difference by the expected sales value. Thus, the **chi-square value** for a cell is calculated as:

$$Chi - square\ (\mathrm{X}^2) for\ a\ cell = \frac{Difference^2}{Expected\ cell\ count}$$

For our example cell, the chi-square value would be:

$$Chi - square\ (\mathrm{X}^2)_{Young\ Adults, Product\ A} = \frac{(1.82)^2}{21.82} = .15$$

The calculated values for the other cells are also shown in Table 9.

Step 4: Create a Measure of Association

The final step is to convert the chi-square value into a **similarity** measure. The chi-square value denotes the degree or amount of similarity or association, but the process of calculating the chi-square (squaring the difference) removes the direction of the similarity. To restore the directionality, we use the sign of the original difference. In order to make the similarity measure more intuitive (i.e., positive values are greater association and negative values are less association) we also reverse the sign of the original difference. The result is a measure that acts just like the similarity measures used in earlier examples. Negative values indicate less association (similarity) and positive values indicate greater association.

In our example, the chi-square value for Young Adults purchasing Product A of .15 would be stated as a similarity value of –.15 because the difference (1.82) was positive. This added negative sign is necessary because the chi-square calculation squares the difference, which eliminates the negative signs. The chi-square values for each cell are also shown in Table 9.

The cells with large positive similarity values (indicating a positive association) are Young Adults/Product B (17.58), Middle Age/Product A (11.62), and Mature Individuals/Product C (12.10). Each of these pairs of categories should be close together on a perceptual map. Cells with large negative similarity values (meaning that expected sales exceeded actual sales, or a negative association) were Young Adults/Product C (–1.94), Middle Age/Product B (–2.47), and Mature Individuals/Product A (–1.17). Where possible, these categories should be far apart on the map.

SUMMARY

Correspondence analysis provides a means of portraying the relationships between objects based on cross-tabulation data. In doing so, the researcher can utilize nonmetric data in ways comparable to the metric measures of similarity underlying multidimensional scaling (MDS).

FUNDAMENTALS OF STRUCTURAL EQUATION MODELING

The use of SEM is predicated on a strong theoretical model by which latent constructs are defined (measurement model) and these constructs are related to each other through a series of dependence relationships (structural model). The emphasis on strong theoretical support for any proposed model underlies the confirmatory nature of most SEM applications.

But many times overlooked is exactly how the proposed structural model is translated into structural relationships and how their estimation is interrelated. Path analysis is the process wherein the structural relationships are expressed as direct and indirect effects in order to facilitate estimation. The importance of understanding this process is not so that the research can understand the estimation process, but instead to understand how model specification (and respecification) impacts the entire set of structural relationships. We will first illustrate the process of using path analysis for estimating relationships in SEM analyses. Then we will discuss the role that model specification has in defining direct and indirect effects and classification of effects as causal versus spurious. We will see how this designation impacts the estimation of structural model.

ESTIMATING RELATIONSHIPS USING PATH ANALYSIS

What was the purpose of developing the path diagram? Path diagrams are the basis for path analysis, the procedure for empirical estimation of the strength of each relationship (path) depicted in the path diagram. Path analysis calculates the strength of the relationships using only a correlation or covariance matrix as input. We will describe the basic process in the following section, using a simple example to illustrate how the estimates are actually computed.

Identifying Paths

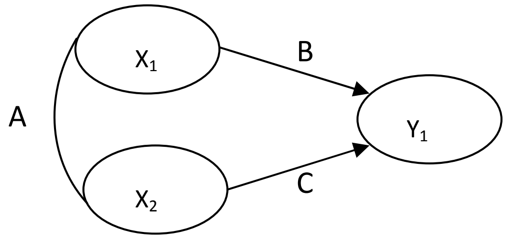
The first step is to identify all relationships that connect any two constructs. Path analysis enables us to decompose the simple (bivariate) correlation between any two variables into the sum of the compound paths connecting these points. The number and types of compound paths between any two variables are strictly a function of the model proposed by the researcher.

A compound path is a path along the arrows of a path diagram that follow three rules:

1. After going forward on an arrow, the path cannot go backward again; but the path can go backward as many times as necessary before going forward.
2. The path cannot go through the same variable more than once.
3. The path can include only one curved arrow (correlated variable pair).

When applying these rules, each path or arrow represents a path. If only one arrow links two constructs (path analysis can also be conducted with variables), then the relationship between those two is equal to the parameter estimate between those two constructs. For now, this relationship can be called a direct relationship. If there are multiple arrows linking one construct to another as in X → Y → Z, then the effect of X on Z seem quite complicated but an example makes it easy to follow:

The path model below portrays a simple model with two exogenous constructs (X_1 and X_2) causally related to the endogenous construct (Y_1). The correlational path A is X_1 correlated with X_2, path B is the effect of X_1 predicting Y_1, and path C shows the effect of X_2 predicting Y_1.



The value for Y_1 can be stated simply with a regression-like equation:

$$Y_1 = b_1X_1 + \ b_2X_1$$

We can now identify the direct and indirect paths in our model. For ease in referring to the paths, the causal paths are labeled A, B, and C.

Direct Paths	*Indirect Paths*
A = X_1 to X_2	
B = X_1 to Y_1	AC = X_1 to Y_1
C = X_2 to Y_1	AB = X_2 to Y_1

Estimating the Relationship

With the direct and indirect paths now defined, we can represent the correlation between each construct as the sum of the direct and indirect paths. The three unique correlations among the constructs can be shown to be composed of direct and indirect paths as follows:

$$Corr_{X1\,X2} = A$$

$$Corr_{X1\,Y1} = B + AC$$

$$Corr_{X2\,Y1} = C + AB$$

First, the correlation of X_1 and X_2 is simply equal to A. The correlation of X_1 and Y_1 (Corr$_{X1,Y1}$) can be represented as two paths: B and AC. The symbol B represents the direct path from X_1 to Y_1, and the other path (a compound path) follows the curved arrow from X_1 to X_2 and then to Y_1. Likewise, the correlation of X_2 and Y_1 can be shown to be composed of two causal paths: C and AB.

Once all the correlations are defined in terms of paths, the values of the observed correlations can be substituted and the equations solved for each separate path. The paths then represent either the causal relationships between constructs (similar to a regression coefficient) or correlational estimates.

Assuming that the correlations among the three constructs are as follows: Corr$_{X1\ X2}$ = .50, Corr$_{X1\ Y1}$ = .60 and Corr$_{X2\ Y1}$ = .70, we can solve the equations for each correlation (see below) and estimate the causal relationships represented by the coefficients b_1 and b_2. We know that A equals .50, so we can substitute this value into the other equations. By solving these two equations, we get values of B(b_1) = .33 and C(b_2) = .53. This approach enables path analysis to solve for any causal relationship based only on the correlations among the constructs and the specified causal model.

Solving for the Structural Coefficients

$.50 = A$

$.60 = B + AC$

$.70 = C + AB$

Substituting A = .50

$.60 = B + .50C$

$.70 = C + .50B$

Solving for B and C

$B = .33$

$C = .53$

As you can see from this simple example, if we change the path model in some way, the causal relationships will change as well. Such a change provides the basis for modifying the model to achieve better fit, if theoretically justified.

With these simple rules, the larger model can now be modeled simultaneously, using correlations or covariances as the input data. We should note that when used in a larger model, we can solve for any number of interrelated equations. Thus, dependent variables in one relationship can easily be independent variables in another relationship. No matter how large the path diagram gets or how many relationships are included, path analysis provides a way to analyze the set of relationships.

UNDERSTANDING DIRECT AND INDIRECT EFFECTS

While path analysis plays a key role in estimating the effects represented in a structural model, it also provides additional insight into not only the direct effects of one construct versus another, but all of the myriad set of indirect effects between any two constructs. While direct effects can always be considered causal if a dependence relationship is specified, indirect effects require further examination to determine if they are causal (directly attributable to a dependence relationship) or non-causal (meaning that they represent relationship between constructs, but it cannot be attributed to a specific causal process).

Identification of Causal versus Non-Causal effects

The prior section discussed the process of identifying all of the direct and indirect effects between any two constructs by a series of rules for compound paths. Here we will discuss how to categorize them into causal versus non-causal and then illustrate their use in understanding the implications of model specification.

An important question is: Why is the distinction important? The parameter estimates are made without any distinction as described above. But the estimated parameters in the structural model reflect only the direct effect of one construct on another. What about all of the indirect effects, which can be substantial? Just because a construct is not directly related to another construct does not mean that there is no impact. Thus, we need a method to distinguish between these myriad types of indirect effects and be able to understand if we can infer any dependence relationship (causal) attributable to them even though they are indirect.

Assume we are identifying the possible effects of A → B. Causal effects are of two types: a direct causal effect (A → B) or an indirect causal effect (A → C → B). In either the direct or indirect compound path, only dependence relationships are present and the direction of the relationships is never reversed. Non-causal (sometimes referred to as spurious effects) can arise from three conditions: common effects (C → A and C →B), correlated effects (C → A and C is correlated with B) and reciprocal effects (A → and B → A).

Decomposing Effects into Causal versus Noncausal

We will use a simple example of four constructs all related to each other (see Table 10). We can see that there are only direct dependence relationships in this example (i.e., no correlational relationships among exogenous constructs). So we might presuppose that all of the effects (direct and indirect) would be causal as well. But as we will see, that is not the case. The total set of effects for each relationship is shown in Table 10.

C_1 → C_2. Let's start with the simple relationship between C_1 and C_2. Using the rules for identifying compound paths earlier, we can see that there is only one possible effect – the direct effect of C_1 on C_2. There are no possible indirect or noncausal effects, so the path estimate for $P_{2,1}$ represents the total effects of C_1 on C_2.

C_1 → C_3. The next relationship is C_1 with C_3. Here we can see two effects: the direct effect ($P_{3,1}$) and the indirect effect ($P_{3,2}$ x $P_{2,1}$). Since the direction of the paths never reverses in the indirect effect, it can be categorized as causal. So the direct and indirect effects are both causal effects.

C_2 → C_3. This relationship introduces the first noncausal effects we have seen. There is the direct effect of $B_{3,2}$, but there is also the noncausal effect (due to common cause) seen in $B_{3,1}$ x $B_{2,1}$. Here we see the result of two causal effects creating a noncausal effect since they both originate from a common construct (C_1 → C_2 and C_1 → C_3).

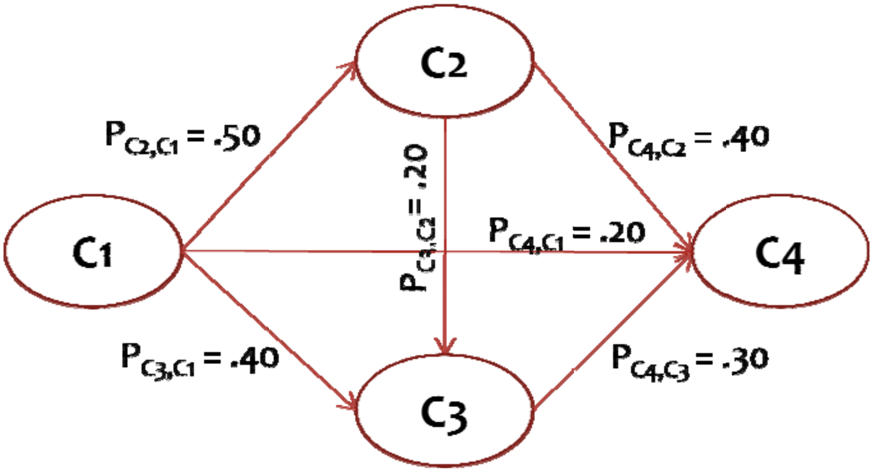
C_1 → C_4. In this relationship we will see the potential for numerous indirect causal effects in addition to direct effects. In addition to the direct effect (B 4,1), we see three other indirect effects that are also causal: $B_{4,2}$ x $B_{2,1}$; $B_{4,3}$ x $B_{3,1}$; and $B_{4,3}$ x $B_{3,2}$ x $B_{2,1}$.

C_3 → C_4. This final relationship we will examine has only one causal effects (B 4,2), but four different noncausal effects, all a result of C_1 or C_2 acting as common causes. The two noncausal effects associated with C_1 are $B_{4,1}$ x $B_{3,1}$ and $B_{4,1}$ x $B_{2,1}$ x $B_{3,2}$. The two other noncausal effects are associated with C_2 (B $_{4,2}$ x B $_{3,2}$ and $B_{4,2}$ x $B_{2,1}$ x $B_{3,1}$).

The remaining relationship is C_2 → C_4. See if you can identify the causal and noncausal effects. Hint: There are all three types of effects – direct and indirect causal effects and noncausal effects as well.

Calculating Indirect Effects

In the previous section we discussed the identification and categorization of both direct and indirect effects for any pair of constructs. The next step is to calculate the amount of the effect based on the path estimates of the model. For our purposes, assume this path model with estimates as follows:



For the relationship between C1 → C4, there are both direct and indirect effects. The direct effects are shown directly by the path estimate $P_{C4,C1}$ = .20. But what about the indirect effects? How are they calculated?

The size of an indirect effect is a function of the direct effects that make it up. SEM software typically produces a table showing the total of the indirect effects implied by a model. But to see the size of each effect you can compute them by multiplying the direct effects in the compound path.

For instance, one of the indirect effects for C1 → C4 is $P_{C2,C1} \times P_{C4,C2}$. We can then calculate the amount of this effect as .50 × .40 = .20. Thus, the complete set of indirect effects for C1 → C4 can be calculated as follows:

$P_{C2,C1} \times P_{C4,C2} = .50 \times .40 = .20$

$P_{C3,C1} \times P_{C4,C3} = .40 \times .30 = .12$

$P_{C2,C1} \times P_{C3,C2} \times P_{C4,C3} = .50 \times .20 \times .30 = .03$

Total Indirect Effects = .20 + .12 + .03 = .35

So in this example, the total effect of C1 → C4 equals the direct and indirect effects, or .20 + .35 = .55. It is interesting to note that in this example the indirect effects are greater than the direct effects.

Impact of Model Respecification

The impact of model respecification on both the parameter estimates and the causal/noncausal effects can be seen in our example as well. Look back at the C_3 → C_4 relationship. What happens if we eliminate the C1 → C4 relationship? Does it impact the C3 → C4 relationship in any way? If we look back at the indirect effects, we can see that two of the four noncausal effects would be eliminated ($B_{4,1}$ x $B_{3,1}$ and $B_{4,1}$ x $B_{2,1}$ x $B_{3,2}$). How would this impact the model? If these effects were substantial but eliminated when the C_1 → C_4 path was eliminated, then most likely the C_3 → C_4 relationship would be underestimated, resulting in a larger residual for this covariance and overall poorer model fit. Plus, a number of other effects that used this path would be eliminated as well. This illustrates how the removal or addition of a path in the structural model can impact not only that direct relationship (e.g., C_1 → C_4), but many other relationships as well.

SUMMARY

The path model not only represents the structural relationships between constructs, but also provides a means of depicting the direct and indirect effects implied in the structural relationships. A “working knowledge” of the direct and indirect effects of any path model gives the researcher not only the basis for understanding the foundations of model estimation, but also insight into the “total” effects of one construct upon another. Moreover, the indirect effects can be further subdivided into casual and non-causal/spurious to provide greater specificity into the types of effects involved. Finally, an understanding of the indirect effects allows for greater understanding of the implications of model respecification, either through addition or deletion of a direct relationship.

TABLE 1 Improving Prediction Accuracy by Adding an Intercept in a Regression Equation

(A) Prediction Without the Intercept

Prediction Equation: $Y = 2X_1$

Value of X_1	**Dependent Variable**	**Prediction**	**Prediction Error**
1	4	2	2
2	6	4	2
3	8	6	2
4	10	8	2
5	12	10	2

(B) Prediction with an Intercept of 2.0

Prediction Equation: $Y = 2.0 + 2X_1$

Value of X_1	**Variable**	**Dependent Prediction**	**Prediction Error**
1	4	4	0
2	6	6	0
3	8	8	0
4	10	10	0
5	12	12	0

TABLE 2 Credit Card Usage Survey Results

Family ID	Number of Credit Cards Used (Y)	Family Size (V_1)	Family Income (\$000) (V_2)	Number of Automobiles Owned (V_3)
1	4	2	14	1
2	6	2	16	2
3	6	4	14	2
4	7	4	17	1
5	8	5	18	3
6	7	5	21	2
7	8	6	17	1
8	10	6	25	2

TABLE 3 Baseline Prediction Using the Mean of the Dependent Variable

Regression Variate: $Y = y$

Prediction Equation: $Y = 7$

Family ID	Number of Credit Cards Used	Baseline Prediction[a]	Prediction Error[b]	Prediction Error Squared
1	4	7	–3	9
2	6	7	–1	1
3	6	7	–1	1
4	7	7	0	0
5	8	7	+1	1
6	7	7	0	0
7	8	7	+1	1
8	10	7	+3	9
Total	56		0	22

[a] Average number of credit cards used = 56 ÷ 8 = 7.

[b] Prediction error refers to the actual value of the dependent variable minus the predicted value.

TABLE 4 Correlation Matrix for the Credit Card Usage Study

	Variable	**Y**	**V_1**	**V_2**	**V_3**
Y	Number of Credit Cards Used	1.000			
V_1	Family Size	.866	1.000		
V_2	Family Income	.829	.673	1.000	
V_3	Number of Automobiles	.342	.192	.301	1.000

TABLE 5 Simple Regression Results Using Family Size as the Independent Variable

Regression Variate: $Y = b_0 + b_1V_1$

Prediction Equation: $Y = 2.87 + .97V_1$

Family ID	**Number of Credit Cards Used**	**Family Size (V_1)**	**Simple Regression Prediction**	**Prediction Error**	**Prediction Error Squared**
1	4	2	4.81	−.81	.66
2	6	2	4.81	1.19	1.42
3	6	4	6.75	−.75	.56
4	7	4	6.75	.25	.06
5	8	5	7.72	.28	.08
6	7	5	7.72	−.72	.52
7	8	6	8.69	−.69	.48
8	10	6	8.69	1.31	1.72
Total					5.50

TABLE 6 Average Ranks and Deviations for Respondent 1

Factor Level per Attribute	Ranks Across Stimuli	Average Rank of Level	Deviation from Overall Average Rank[a]
Form			
Liquid	1, 2, 5, 6	3.5	–1.0
Powder	3, 4, 7, 8	5.5	+1.0
Ingredients			
Phosphate-free	1, 2, 3, 4	2.5	–2.0
Phosphate-based	5, 6, 7, 8	6.5	+2.0
Brand			
HBAT	1, 3, 5, 7	4.0	–.5
Generic	2, 4, 6, 8	5.0	+.5

[a] Deviation calculated as Deviation = Average rank of level – Overall average rank (4.5). Note that negative deviations imply more preferred rankings.

TABLE 7 Estimated Part-Worths and Factor Importance for Respondent 1

	Estimating Part-Worths				***Calculating Factor Importance***	
Factor Level	**Reversed Deviation**[a]	**Squared Deviation**	**Standardized Deviation**[b]	**Estimated Part-Worth**[c]	**Range of Part- Worths**	**Factor Importance**[d]
Form						
Liquid	+1.0	1.0	+.571	+.756	1.512	28.6%
Powder	–1.0	1.0	–.571	–.756		
Ingredients						
Phosphate-free	+2.0	4.0	+2.284	+1.511	3.022	57.1%
Phosphate-based	–2.0	4.0	–2.284	–1.511		
Brand						
HBAT	+.5	.25	+.143	+.378	.756	14.3%
Generic	–.5	.25	–.143	–.378		
Sum of Squared Deviations		10.5				
Standardizing Value[e]		.571				
Sum of Part-Worth Ranges					5.290	

[a] Deviations are reversed to indicate higher preference for lower ranks. The sign of deviation is used to indicate sign of estimated part-worth.

[b] Standardized deviation is equal to the squared deviation times the standardizing value.

[c] Estimated part-worth is equal to the square root of the standardized deviation.

[d] Factor importance is equal to the range of a factor divided by the sum of ranges across all factors, multiplied by 100 to get a percentage.

[e] Standardizing value is equal to the number of levels (2 + 2 + 2 = 6) divided by the sum of the squared deviations.

TABLE 8 Cross-Tabulated Data Detailing Product Sales by Age Category

	Product Sales			
Age Category	**A**	**B**	**C**	**Total**
Young Adults	20	20	20	60
Middle Age	40	10	40	90
Mature Individuals	20	10	40	70
Total	80	40	100	220

TABLE 9 Calculating Chi-Square Similarity Values for Cross-Tabulated Data

	Product Sales			
Age Category	**A**	**B**	**C**	**Total**
Young Adults				
Sales	20	20	20	60
Column Percentage	25%	50%	20%	27%
Row Percentage	33%	33%	33%	100%
Expected Sales[a]	21.82	10.91	27.27	60
Difference[b]	1.82	–9.09	7.27	—
Chi-Square Value[c]	.15	7.58	1.94	9.67
Middle Age				
Sales	40	10	40	90
Column Percentage	50%	25%	40%	41%
Row Percentage	44%	11%	44%	100%
Expected Sales	32.73	16.36	40.91	90
Difference	–7.27	6.36	.91	—
Chi-Square Value	1.62	2.47	.02	4.11
Mature Individuals				
Sales	20	10	40	70
Column Percentage	25%	25%	40%	32%
Row Percentage	29%	14%	57%	100%
Expected Sales	25.45	12.73	31.82	70
Difference	5.45	2.73	–8.18	—
Chi-Square Value	1.17	.58	2.10	3.85
Total				
Sales	80	40	100	220
Column Percentage	100%	100%	100%	100%
Row Percentage	36%	18%	46%	100%
Expected Sales	80	40	100	220
Difference	—	—	—	—
Chi-Square Value	2.94	10.63	4.06	17.63

[a] Expected sales = (Row total × Column total) / Overall total

Example: $\text{Cell}_{\text{Young Adults, Product A}} = (60 \times 80) / 220 = 21.82$

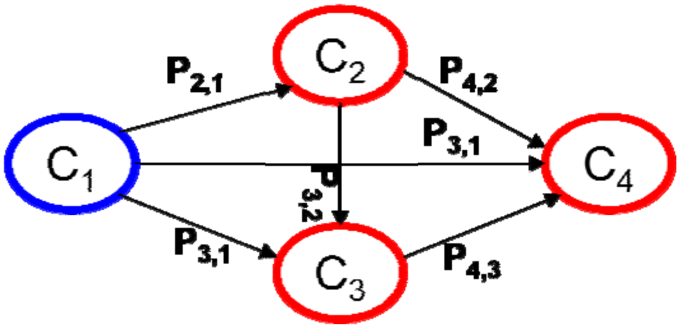
[b] Difference = Expected sales – Actual sales

Example: $\text{Cell}_{\text{Young Adults, Product A}} = 21.82 - 20.00 = 1.82$

[c] Chi-Square Value

Example: $\text{Cell}_{\text{Young Adults, Product A}} = 1.82^2 / 21.82 = .15$

Table 10 Identifying Direct and Indirect Effects



Relationship	Effects		
	Direct (Causal)	Indirect (Causal)	Indirect (Noncausal)
C_1 → C_2	$P_{2,1}$	None	None
C_1 → C_3	$P_{3,1}$	$P_{3,2}$ x $P_{2,1}$	None
C_2 → C_3	$P_{3,2}$	None	$P_{3,1}$ x $P_{2,1}$
C_1 → C_4	$P_{4,1}$	$P_{4,2}$ x $P_{2,1}$ $P_{4,3}$ x $P_{3,1}$ $P_{4,3}$ x $P_{3,2}$ x $P_{2,1}$	None
C_3 → C_4	$P_{4,3}$	None	$P_{4,1}$ x $P_{3,1}$ $P_{4,1}$ x $P_{2,1}$ x $P_{3,2}$ $P_{4,2}$ x $P_{3,2}$ $P_{4,2}$ x $P_{2,1}$ x $P_{3,1}$