Multivariate Data Analysis

Joseph F. Hair Jr. William C. Black
Barry J. Babin Rolph E. Anderson

Seventh Edition

Pearson New International Edition

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsoned.co.uk

ISBN 10: 1-292-02190-X
ISBN 13: 978-1-292-02190-4

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Printed in the United States of America

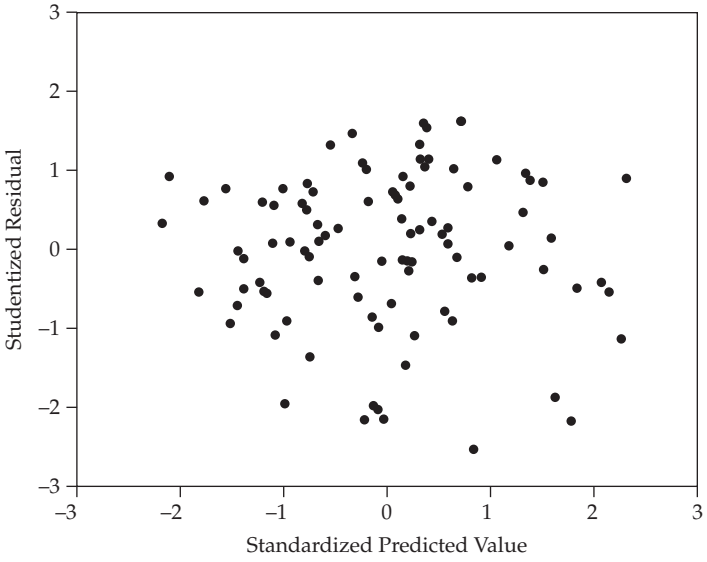

FIGURE 10 Analysis of Standardized Residuals

so, we use the partial regression plot for each independent variable in the equation. In Figure 11 we see that the relationships for X_6, X_9, and X_{12} are reasonably well defined; that is, they have strong and significant effects in the regression equation. Variables X_7 and X_{11} are less well defined, both in slope and scatter of the points, thus explaining their lesser effect in the equation (evidenced by the smaller coefficient, beta value, and significance level). For all five variables, no nonlinear pattern is shown, thus meeting the assumption of linearity for each independent variable.

Homoscedasticity. The next assumption deals with the constancy of the residuals across values of the independent variables. Our analysis is again through examination of the residuals (Figure 10), which shows no pattern of increasing or decreasing residuals. This finding indicates homoscedasticity in the multivariate (the set of independent variables) case.

Independence of the Residuals. The third assumption deals with the effect of carryover from one observation to another, thus making the residual not independent. When carryover is found in such instances as time series data, the researcher must identify the potential sequencing variables (such as time in a time series problem) and plot the residuals by this variable. For example, assume that the identification number represents the order in which we collect our responses. We could plot the residuals and see whether a pattern emerges.

In our example, several variables, including the identification number and each independent variable, were tried and no consistent pattern was found. We must use the residuals in this analysis, not the original dependent variable values, because the focus is on the prediction errors, not the relationship captured in the regression equation.

Normality. The final assumption we will check is normality of the error term of the variate with a visual examination of the normal probability plots of the residuals.

As shown in Figure 12, the values fall along the diagonal with no substantial or systematic departures; thus, the residuals are considered to represent a normal distribution. The regression variate is found to meet the assumption of normality.

Applying Remedies for Assumption Violations. After testing for violations of the four basic assumptions of multivariate regression for both individual variables and the regression variate, the researcher should assess the impact of any remedies on the results.

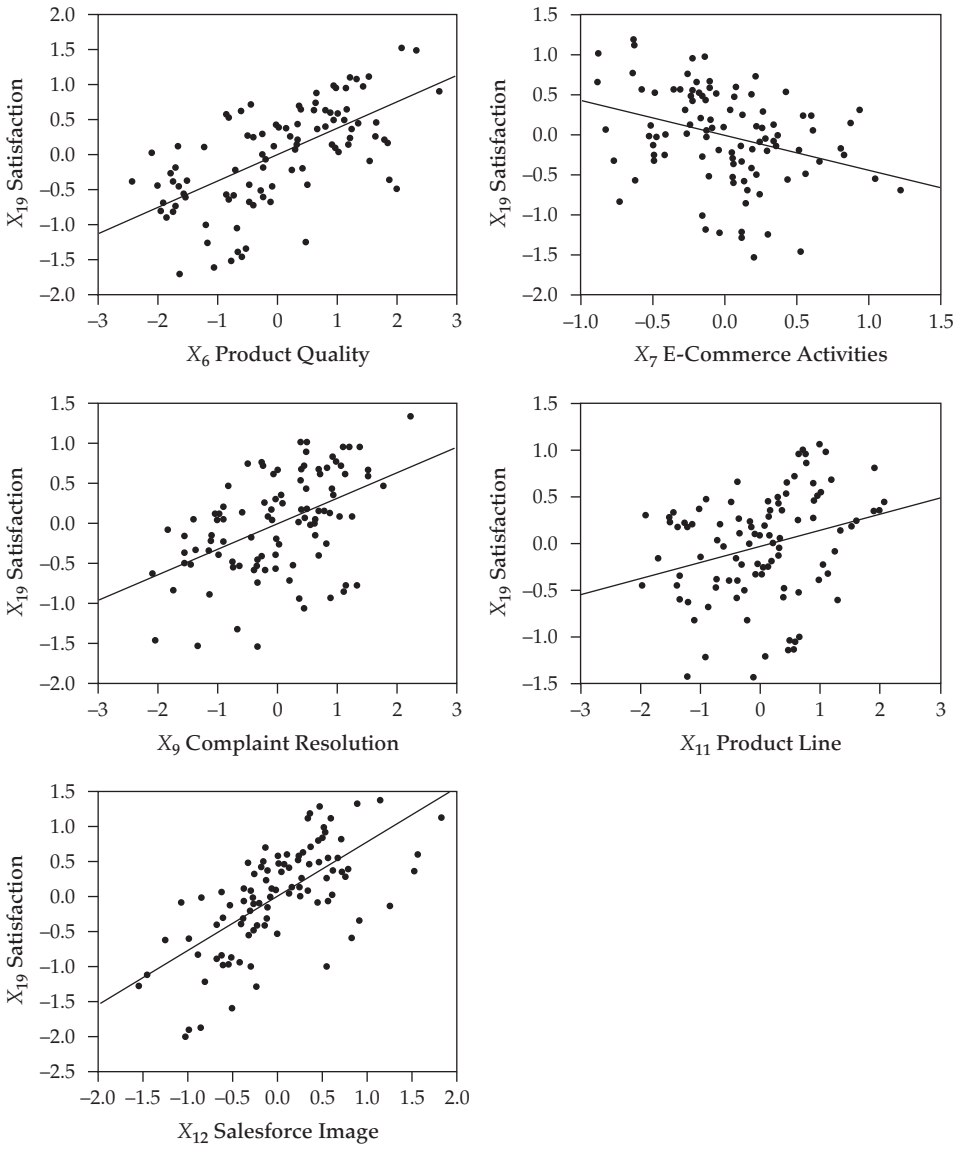

FIGURE 11 Standardized Partial Regression Plots

In an examination of individual variables, the only remedies needed are the transformations of X_6, X_7, X_{12}, X_{13}, X_{16}, and X_{17}. A set of differing transformations were used, including the squared term (X_6 and X_{16}), logarithm (X_7), cubed term (X_{13}), and inverse (X_{16}). Only in the case of X_{12} did the transformation not achieve normality. If we substitute these variables for their original values and reestimate the regression equation with a stepwise procedure, we achieve almost identical results. The same variables enter the equation with no substantive differences in either the estimated coefficients or overall model fit as assessed with R^2 and standard error of the estimate. The independent variables not in the equation still show nonsignificant levels for entry—even those that were transformed. Thus, in this case, the remedies for violating the assumptions improved the prediction slightly but did not alter the substantive findings.

IDENTIFYING OUTLIERS AS INFLUENTIAL OBSERVATIONS For our final analysis, we attempt to identify any observations that are influential (having a disproportionate impact on the regression

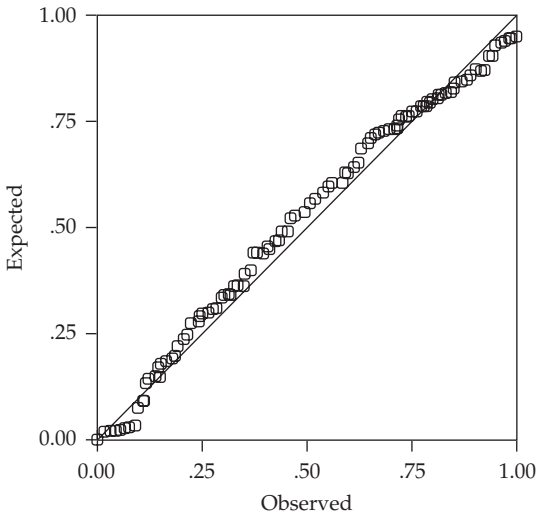

FIGURE 12 Normal Probability Plot: Standardized Residuals

results) and determine whether they should be excluded from the analysis. Although more detailed procedures are available for identifying outliers as influential observations, we address the use of residuals in identifying outliers in the following section.

The most basic diagnostic tool involves the residuals and identification of any outliers—that is, observations not predicted well by the regression equation that have large residuals. Figure 13 shows the studentized residuals for each observation. Because the values correspond to *t* values, upper and lower limits can be set once the desired confidence interval has been established. Perhaps the most widely used level is the 95% confidence interval ($\alpha = .05$). The corresponding *t* value is 1.96, thus identifying statistically significant residuals as those with residuals greater than this value (1.96). Seven observations can be seen in Figure 13 (2, 10, 20, 45, 52, 80, and 99) to have significant residuals and thus be classified as outliers. Outliers are important because they are observations not represented by the regression equation for one or more reasons, any one of which may be an influential effect on the equation that requires a remedy.

Examination of the residuals also can be done through the partial regression plots (see Figure 11). These plots help to identify influential observations for each independent–dependent variable relationship. Consistently across each graph in Figure 11, the points at the lower portion are those observations identified as having high negative residuals (observations 2, 10, 20, 45, 52, 80, and 99 in Figure 13). These points are not well represented by the relationship and thus affect the partial correlation as well.

More detailed analyses to ascertain whether any of the observations can be classified as influential observations, as well as what may be the possible remedies, are discussed in the supplement to this chapter available on the Web at www.pearsonhighered.com/hair or www.mvstats.com.

Stage 5: Interpreting the Regression Variate

With the model estimation completed, the regression variate specified, and the diagnostic tests that confirm the appropriateness of the results administered, we can now examine our predictive equation based on five independent variables (X_6, X_7, X_9, X_{11}, and X_{12}).

INTERPRETATION OF THE REGRESSION COEFFICIENTS The first task is to evaluate the regression coefficients for the estimated signs, focusing on those of unexpected direction.

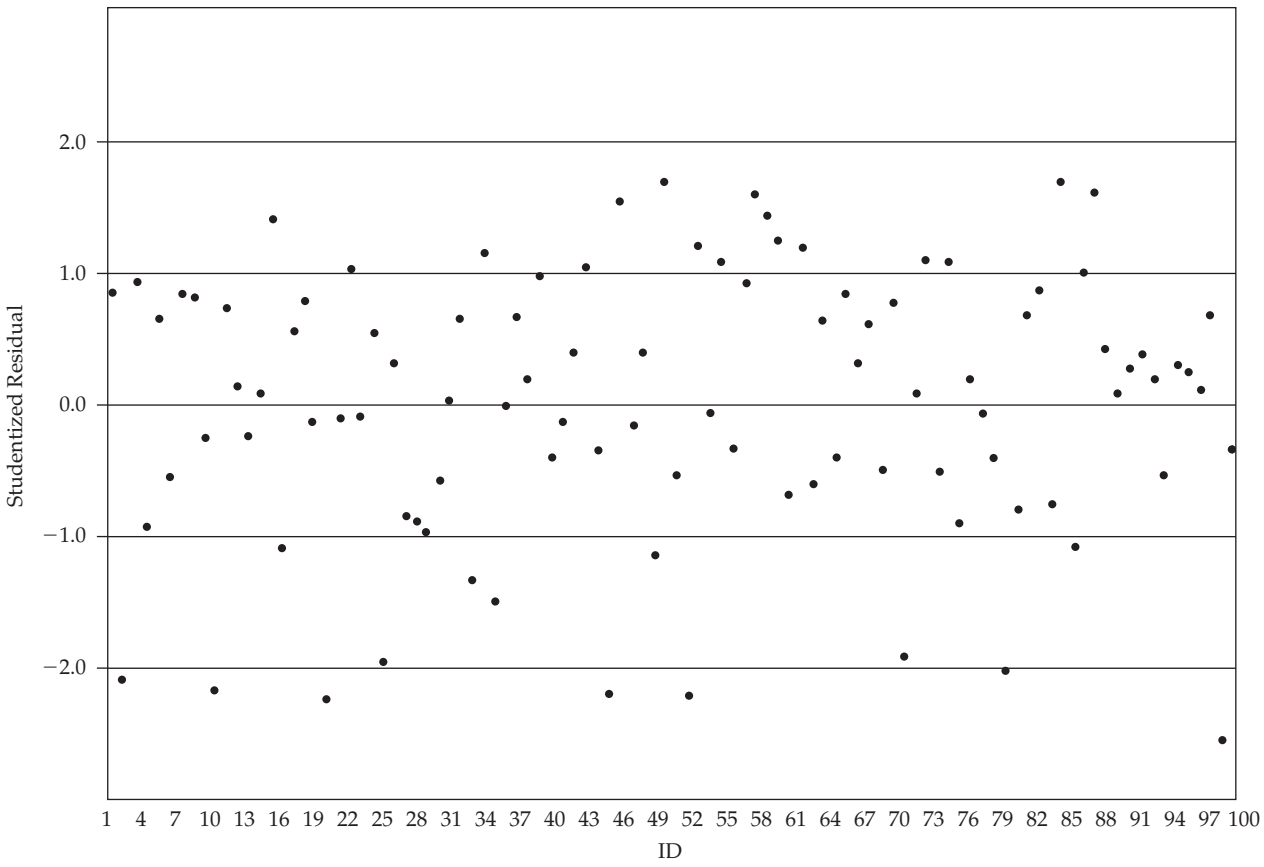

FIGURE 13 Plot of Studentized Residuals

The section of Table 11 headed "Variables Entered into the Regression Equation" yields the prediction equation from the column labeled "Regression Coefficient: B." From this column, we read the constant term (−1.151) and the coefficients (.319, .369, .775, −.417, and .174) for X_9, X_6, X_{12}, X_7, and X_{11}, respectively. The predictive equation would be written

$$Y = -1.151 + .319X_9 + .369X_6 + .775X_{12} + (-.417)X_7 + .174X_{11}$$

Note: The coefficient of X_7 is included in parentheses to avoid confusion due to the negative value of the coefficient.

With this equation, the expected customer satisfaction level for any customer could be calculated if that customer's evaluations of HBAT are known. For illustration, let us assume that a customer rated HBAT with a value of 6.0 for each of these five measures. The predicted customer satisfaction level for that customer would be

$$\begin{aligned}\text{Predicted Customer} &= -1.151 + .319 \times 6 + .369 \times 6 + .775 \times 6 + (-.417) \times 6 \\ &\quad + .174 \times 6 \\ \text{Satisfaction} &= -1.151 + 1.914 + 2.214 + 4.650 - 2.502 + 1.044 \\ &= 6.169\end{aligned}$$

We first start with an interpretation of the constant. It is statistically significant (significance = .023), thus making a substantive contribution to the prediction. However, because in our situation it is

highly unlikely that any respondent would have zero ratings on all the HBAT perceptions, the constant merely plays a part in the prediction process and provides no insight for interpretation.

In viewing the regression coefficients, the sign is an indication of the relationship (positive or negative) between the independent and dependent variables. All of the variables except one have positive coefficients. Of particular note is the reversed sign for X_7 (E-Commerce), suggesting that an increase in perceptions on this variable has a negative impact on predicted customer satisfaction. All the other variables have positive coefficients, meaning that more positive perceptions of HBAT (higher values) increase customer satisfaction.

Does X_7, then, somehow operate differently from the other variables? In this instance, the bivariate correlation between X_7 and customer satisfaction is positive, indicating that when considered separately, X_7 has a positive relationship with customer satisfaction, just as the other variables. We will discuss in the following section the impact of multicollinearity on the reversal of signs for estimated coefficients.

ASSESSING VARIABLE IMPORTANCE In addition to providing a basis for predicting customer satisfaction, the regression coefficients also provide a means of assessing the relative importance of the individual variables in the overall prediction of customer satisfaction. When all the variables are expressed in a standardized scale, then the regression coefficients represent relative importance. However, in other instances the beta weight is the preferred measure of relative importance.

In this situation, all the variables are expressed on the same scale, but we will use the beta coefficients for comparison between independent variables. In Table 11, the beta coefficients are listed in the column headed "Regression Coefficients: Beta." The researcher can make direct comparisons among the variables to determine their relative importance in the regression variate. For our example, X_{12} (Salesforce Image) was the most important, followed by X_6 (Product Quality), X_9 (Complaint Resolution), X_7 (E-Commerce), and finally X_{11} (Product Line). With a steady decline in size of the beta coefficients across the variables, it is difficult to categorize variables as high, low, or otherwise. However, viewing the relative magnitudes does indicate that, for example, X_{12} (Salesforce Image) shows a more marked effect (three times as much) than X_{11} (Product Line). Thus, to the extent that salesforce image can be increased uniquely from other perceptions, it represents the most direct way, *ceterus paribus,* of increasing customer satisfaction.

MEASURING THE DEGREE AND IMPACT OF MULTICOLLINEARITY In any interpretation of the regression variate, the researcher must be aware of the impact of multicollinearity. As discussed earlier, highly collinear variables can distort the results substantially or make them quite unstable and thus not generalizable. Two measures are available for testing the impact of collinearity: (1) calculating the tolerance and VIF values and (2) using the condition indices and decomposing the regression coefficient variance (see the supplement to this chapter available on the Web at www.pearsonhighered.com/hair or www.mvstats.com for more details on this process). The tolerance value is 1 minus the proportion of the variable's variance explained by the other independent variables. Thus, a high tolerance value indicates little collinearity, and tolerance values approaching zero indicate that the variable is almost totally accounted for by the other variables (high multicollinearity). The variance inflation factor is the reciprocal of the tolerance value; thus we look for small VIF values as indicative of low correlation among variables.

Diagnosing Multicollinearity. In our example, tolerance values for the variables in the equation range from .728 (X_6) to .347 (X_{12}), indicating a wide range of multicollinearity effects (see Table 11). Likewise, the VIF values range from 1.373 to 2.701. Even though none of these values indicate levels of multicollinearity that should seriously distort the regression variate, we must be careful even with these levels to understand their effects, especially on the stepwise estimation process. The following section will detail some of these effects on both the estimation and interpretation process.