

# Advanced Diagnostics for Multiple Regression:

*A Supplement to Multivariate Data Analysis*

*Multivariate Data Analysis*

Pearson Prentice Hall Publishing

# Table of Contents

|   |           |
|---|-----------|
| <b>LEARNING OBJECTIVES.....</b>   | <b>3</b>  |
| <b>PREVIEW .....</b>  | <b>3</b>  |
| <b>KEY TERMS .....</b>  | <b>3</b>  |
| <b>Assessing Multicollinearity .....</b>  | <b>6</b>  |
| <i>A Two-Part Process.....</i>  | <i>6</i>  |
| <b>Identifying Influential Observations .....</b>                                 | <b>7</b>  |
| <i>Step 1: Examining Residuals.....</i>   | <i>7</i>  |
| Analysis of Residuals.....  | 7         |
| Cases Used In Calculating The Residual .....                                      | 8         |
| Partial Regression Plots .....  | 9         |
| <i>Step 2: Identifying Leverage Points from the Predictors.....</i>               | <i>9</i>  |
| Hat Matrix.....   | 9         |
| Mahalanobis Distance.....   | 10        |
| <i>Step 3: Single-Case Diagnostics Identifying Influential Observations .....</i> | <i>10</i> |
| Influences on Individual Coefficients .....                                       | 11        |
| Overall Influence Measures .....  | 11        |
| <i>Step 4: Selecting and Accommodating Influential Observations .....</i>         | <i>12</i> |
| Overview.....   | 12        |
| <b>Summary.....</b>   | <b>12</b> |
| <b>Questions.....</b>   | <b>13</b> |
| <b>References.....</b>  | <b>13</b> |

# Advanced Diagnostics for Multiple Regression Analysis

## LEARNING OBJECTIVES

After reading our discussion of these techniques, you should be able to do the following:

1. Understand how the condition index and regression coefficient variance–decomposition matrix isolate the effects, if any, of multicollinearity on the estimated regression coefficients.
2. Identify those variables with unacceptable levels of collinearity or multicollinearity.
3. Identify the observations with a disproportionate impact on the multiple regression results.
4. Isolate influential observations and assess the relationships when the influential observations are deleted.

## PREVIEW

Multiple regression is perhaps the most widely used statistical technique, and it has led the movement toward increased usage of other multivariate techniques. In moving from simple to multiple regression, the increased analytical power of the multivariate model requires additional diagnostics to deal with the correlations between variables and those observations with substantial impact on the results. This appendix describes advanced diagnostic techniques for assessing (1) the impact of multicollinearity and (2) the identity of influential observations and their impact on multiple regression analysis. The chapter on multiple regression dealt with the basic diagnoses for these issues; here we discuss more sensitive procedures that have recently been proposed specifically for multivariate situations. These procedures are not refinements to the estimation procedures but instead address questions in interpreting the results that occur in the presence of multicollinearity and influential observations.

## KEY TERMS

Before reading this appendix, review the key terms to develop an understanding of the concepts and terminology used. Throughout the appendix the key terms appear in **boldface**. Other points of emphasis in the appendix are *italicized*. Also, cross-references in the Key Terms appear in *italics*.

**Collinearity** Relationship between two (collinearity) or more (*multicollinearity*) variables. Variables exhibit complete collinearity if their correlation coefficient is 1 and a complete lack of collinearity if their correlation coefficient is 0.

**Condition index** Measure of the relative amount of variance associated with an eigenvalue so that a large condition index indicates a high degree of *collinearity*.

**Cook's distance ( $D_i$ )** Summary measure of the influence of a single case (observation) based on the total changes in all other residuals when the case is deleted from the estimation process. Large values (usually greater than 1) indicate substantial influence by the case in affecting the estimated regression coefficients.

**COVRATIO** Measure of the influence of a single observation on the entire set of estimated regression coefficients. A value close to 1 indicates little influence. If the COVRATIO value minus 1 is greater than  $\pm 3p/n$  (where  $p$  is the number of independent variables + 1, and  $n$  is the sample size), the observation is deemed to be influential based on this measure.

**Deleted residual** Process of calculating *residuals* in which the influence of each observation is removed when calculating its residual. This is accomplished by omitting the  $i$ th observation from the regression equation used to calculate its predicted value.

**DFBETA** Measure of the change in a regression coefficient when an observation is omitted from the regression analysis. The value of DFBETA is in terms of the coefficient itself; a standardized form (SDFBETA) is also available. No threshold limit can be established for DFBETA, although the researcher can look for values substantially different from the remaining observations to assess potential influence. The SDFBETA values are scaled by their standard errors, thus supporting the rationale for cutoffs of 1 or 2, corresponding to confidence levels of .10 or .05, respectively.

**DFFIT** Measure of an observation's impact on the overall model fit, which also has a standardized version (SDFFIT). The best rule of thumb is to classify as influential any standardized values (SDFFIT) that exceed  $2/\sqrt{p/n}$ , where  $p$  is the number of independent variables + 1 and  $n$  is the sample size. There is no threshold value for the DFFIT measure.

**Eigenvalue** Measure of the amount of variance contained in the correlation matrix so that the sum of the eigenvalues is equal to the number of variables. Also known as the latent root or characteristic root.

**Hat matrix** Matrix that contains values for each observation on the diagonal, known as *hat values*, which represent the impact of the observed dependent variable on its predicted value. If all cases have equal influence, each would have a value of  $p/n$ , where  $p$  equals the number of independent variables + 1, and  $n$  is the number of cases. If a case has no influence, its value would be  $-1 \div n$ , whereas total domination by a single case would result in a value of  $(n - 1)/n$ . Values exceeding  $2p/n$  for larger samples, or  $3p/n$  for smaller samples ( $n \leq 30$ ), are candidates for classification as influential observations.

**Hat value** See *hat matrix*.

**Influential observation** Observation with a disproportionate influence on one or more aspects of the regression estimates. This influence may have as its basis (1) substantial differences from other cases on the set of independent variables, (2) extreme (either high or low) observed values for the criterion variables, or (3) a combination of these effects. Influential observations can either be “good,” by reinforcing the pattern of the remaining data, or “bad,” when a single or small set of cases unduly affects (biases) the regression estimates.

**Leverage point** An observation that has substantial impact on the regression results due to its differences from other observations on one or more of the independent variables. The most common measure of a leverage point is the *hat value*, contained in the *hat matrix*.

**Mahalanobis distance ( $D^2$ )** Measure of the uniqueness of a single observation based on differences between the observation’s values and the mean values for all other cases across all independent variables. The source of influence on regression results is for the case to be quite different on one or more predictor variables, thus causing a shift of the entire regression equation.

**Multicollinearity** See *collinearity*.

**Outlier** In strict terms, an observation that has a substantial difference between its actual and predicted values of the dependent variable (a large *residual*) or between its independent variable values and those of other observations. The objective of denoting outliers is to identify observations that are inappropriate representations of the population from which the sample is drawn, so that they may be discounted or even eliminated from the analysis as unrepresentative.

**Regression coefficient variance–decomposition matrix** Method of determining the relative contribution of each *eigenvalue* to each estimated coefficient. If two or more coefficients are highly associated with a single eigenvalue (*condition index*), an unacceptable level of *multicollinearity* is indicated.

**Residual** Measure of the predictive fit for a single observation, calculated as the difference between the actual and predicted values of the dependent variable. Residuals are assumed to have a mean of zero and a constant variance. They not only play a key role in determining if the underlying assumptions of regression have been met, but also serve as a diagnostic tool in identifying *outliers* and *influential observations*.

**SDFBETA** See *DFBETA*.

**SDFFIT** See *DFFIT*.

**Standardized residual** Rescaling of the *residual* to a common basis by dividing each residual by the standard deviation of the residuals. Thus, standardized residuals have a mean of 0 and standard deviation of 1. Each standardized residual value can now be viewed in terms of standard errors in middle to large sample sizes. This provides a direct means of identifying outliers as those with values above 1 or 2 for confidence levels of .10 and .05, respectively.

**Studentized residual** Most commonly used form of *standardized residual*. It differs from other standardization methods in calculating the standard deviation employed.

To minimize the effect of a single *outlier*, the standard deviation of residuals used to standardize the *i*th residual is computed from regression estimates omitting the *i*th observation. This is done repeatedly for each observation, each time omitting that observation from the calculations. This approach is similar to the deleted residual, although in this situation the observation is omitted from the calculation of the standard deviation.

**Tolerance** Commonly used measure of *collinearity* and *multicollinearity*. The tolerance of variable *i* ( $TOL_i$ ) is  $1 - R_i^{*2}$ , where  $R_i^{*2}$  is the coefficient of determination for the prediction of variable *i* by the other predictor variables. Tolerance values approaching zero indicate that the variable is highly predicted (collinear) with the other predictor variables.

**Variance inflation factor (VIF)** Measure of the effect of other predictor variables on a regression coefficient. VIF is inversely related to the *tolerance* value ( $VIF_i = 1 \div TOL_i$ ). The  $\sqrt{VIF}$  reflects the extent to which the standard error of the regression coefficient is increased due to multicollinearity. Large VIF values (a usual threshold is 10.0, which corresponds to a tolerance of .10) indicate a high degree of collinearity or multicollinearity among the independent variables, although values of as high as four have been considered problematic.

## ASSESSING MULTICOLLINEARITY

As discussed in the chapter on multiple regression, **collinearity** and **multicollinearity** can have several harmful effects on multiple regression, both in the interpretation of the results and in how they are obtained, such as stepwise regression. The use of several variables as predictors makes the assessment of multiple correlation between the independent variables necessary to identify multicollinearity. But this is not possible by examining only the bivariate (also known as the zero order) correlation matrix which shows only simple correlations between two variables. We now discuss a method developed specifically to diagnose the amount of multicollinearity present and the variables exhibiting the high multicollinearity. All major statistical programs have analyses providing these collinearity diagnostics.

### A TWO-PART PROCESS

The method has two components which each depict both the overall level of multicollinearity as well as its presence across the independent variables:

1. **Condition index** -- represents the collinearity of combinations of variables in the data set (actually the relative size of the **eigenvalues** of the matrix).
2. **Regression coefficient variance-decomposition matrix** -- shows the proportion of variance for each regression coefficient (and its associated independent variable) attributable to each condition index (eigenvalue).

We combine these in a two-step procedure:

1. Identify all condition indices above a threshold value. The threshold value usually is in a range of 15 to 30, with 30 the most commonly used value.
2. For all condition indices exceeding the threshold, identify variables with variance proportions above 90 percent. A collinearity problem is indicated when a condition index identified in step 1 as above the threshold value accounts for a substantial proportion of variance (.90 or above) for *two or more* coefficients.

## IDENTIFYING INFLUENTIAL OBSERVATIONS

In the chapter on multiple regression, we examined one approach to identifying **influential observations**, that being the use of studentized residuals to identify outliers. As noted then, however, observations may be classified as influential even though they are not recognized as outliers. In fact, many times an influential observation will not be identified as an outlier because it has influenced the regression estimation to such a degree as to make its residual negligible. Thus, we need to examine more specific procedures to measure an observation's influence in several aspects of multiple regression [2]. In the following discussion, we discuss a four-step process of identifying outliers, leverage points, and influential observations. As noted before, an observation may fall into one or more of these classes, and the course of action to be taken depends on the judgment of the researcher, based on the best available evidence.

### STEP 1: EXAMINING RESIDUALS

Residuals are instrumental in detecting violations of model assumptions, and they also play a role in identifying observations that are **outliers** on the dependent variable. We employ two methods of detection: the analysis of residuals and partial regression plots.

#### Analysis of Residuals

The **residual** is the primary means of classifying an observation as an outlier. The residual for the  $i$ th observation is calculated as the actual minus predicted values of the dependent variable, or:

$$\text{Residual}_i = Y_i - \hat{Y}_i$$

The residual can actually take many forms based on the results of two procedures: the cases used for calculating the predicted value, and the use (or nonuse) of some form of standardization. We will discuss each procedure in the following sections and then discuss how they are “combined” to derive specific types of residuals.

**Cases Used In Calculating The Residual** We have already seen how we calculate the residual using all of the, but a second form, the deleted residual, differs from the normal residual in that the  $i$ th observation is omitted when estimating the regression equation used to calculate the predicted value for that observation. Thus, each observation has no impact on its own predicted value in the deleted residual. The deleted residual is less commonly used, although it has the benefit of reducing the influence of the observation on its calculation.

**Standardizing The Residual** The second procedure in defining a residual involves whether to standardize the residuals. Residuals that are not standardized are in the scale of the dependent variable, which is useful in interpretation but gives no insight as to what is too large or small enough not to consider. **Standardized residuals** are the result of a process of creating a common scale by dividing each residual by the standard deviation of residuals. After standardization, the residuals have a mean of 0 and a standard deviation of 1. With a fairly large sample size (50 or above), standardized residuals approximately follow the  $t$  distribution, such that residuals exceeding a threshold such as 1.96 (the critical  $t$  value at the .05 confidence level) can be deemed statistically significant. Observations falling outside the threshold are statistically significant in their difference from 0 and can be considered outliers. This means that the predicted value is also significantly different from the actual value at the .05 level. A stricter test of significance has also been proposed, which accounts for multiple comparisons being made across various sample sizes [4].

A special form of standardized residual is the **studentized residual**. It is similar in concept to the deleted residual, but in this case the  $i$ th observation is eliminated when deriving the standard deviation used to standardize the  $i$ th residual. The rationale is that if an observation is extremely influential, it may not be identified by the normal standardized residuals because of its impact on the estimated regression model. The studentized residual eliminates the case’s impact on the standardization process and offers a “less influenced” residual measure. It can be evaluated by the same criteria as the standardized residual.

The five types of residuals typically calculated by combining the options for calculation and standardization are (1) the normal residual, (2) the deleted residual, (3) the standardized residual, (4) the studentized residual, and (5) the studentized deleted residual. Each type of residual offers unique perspectives on

both the predictive accuracy of the regression equation by its designation of outliers and the possible influences of the observation on the overall results.

### **Partial Regression Plots**

To graphically portray the impact of individual cases, the partial regression plot is most effective. Because the slope of the regression line of the partial regression plot is equal to the variable's coefficient in the regression equation, an outlying case's impact on the regression slope (and the corresponding regression equation coefficient) can be readily seen. The effects of outlying cases on individual regression coefficients are portrayed visually. Again, most computer packages have the option of plotting the partial regression plot, so the researcher need look only for outlying cases separated from the main body of observations. A visual comparison of the partial regression plots with and without the observation(s) deemed influential can illustrate their impact.

### **STEP 2: IDENTIFYING LEVERAGE POINTS FROM THE PREDICTORS**

Our next step is finding those observations that are substantially different from the remaining observations on one or more independent variables. These cases are termed **leverage points** in that they may "lever" the relationship in their direction because of their difference from the other observations (see the chapter on multiple regression for a general description of leverage points).

### **Hat Matrix**

When only two predictor variables are involved, plotting each variable on an axis of a two-dimensional plot will show those observations substantially different from the others. Yet, when a larger number of predictor variables are included in the regression equation, the task quickly becomes impossible through univariate methods. However, we are able to use a special matrix, the **hat matrix**, which contains values (**hat values**) for each observation that indicate leverage. The hat values represent the combined effects of all independent variables for each case.

Hat values (found on the diagonal of the hat matrix) measure two aspects of influence. First, for each observation, the hat value is a measure of the distance of the observation from the mean center of all other observations on the independent variables (similar to the Mahalanobis distance discussed next). Second, large diagonal values also indicate that the observation carries a disproportionate weight in determining its predicted dependent variable value, thus minimizing its residual.

This is an indication of influence, because the regression line must be closer to this observation (i.e., strongly influenced) for the small residual to occur. This is not necessarily “bad,” as illustrated in the text, when the influential observations fall in the general pattern of the remaining observations.

What is a large hat value? The range of possible values are between 0 and 1, and the average value is  $p/n$ , where  $p$  is the number of predictors (the number of coefficients plus one for the constant) and  $n$  is the sample size. The rule of thumb for situations in which  $p$  is greater than 10 and the sample size exceeds 50 is to select observations with a leverage value greater than twice the average ( $2p/n$ ). When the number of predictors is less than 10 or the sample size is less than 50, use of three times the average ( $3p/n$ ) is suggested. The more widely used computer programs all have options for calculating and printing the leverage values for each observation. The analyst must then select the appropriate threshold value ( $2p/n$  or  $3p/n$ ) and identify observations with values larger than the threshold.

### **Mahalanobis Distance**

A measure comparable to the hat value is the **Mahalanobis distance ( $D^2$ )**, which considers only the distance of an observation from the mean values of the independent variables and not the impact on the predicted value. The Mahalanobis distance is another means of identifying outliers. It is limited in this purpose because threshold values depend on a number of factors, and a rule of thumb threshold value is not possible. It is possible, however, to determine statistical significance of the Mahalanobis distance from published tables [1]. Yet even without the published tables, the researcher can look at the values and identify any observations with substantially higher values than the remaining observations. For example, a small set of observations with the highest Mahalanobis values that are two to three times the next highest value would constitute a substantial break in the distribution and another indication of possible leverage.

### **STEP 3: SINGLE-CASE DIAGNOSTICS IDENTIFYING INFLUENTIAL OBSERVATIONS**

Up to now we have found outlying points on the predictor and criterion variables but have not formally estimated the influence of a single observation on the results. In this third step, all the methods rely on a common proposition: the most direct measure of influence involves deleting one or more observations and observing the changes in the regression results in terms of the residuals, individual coefficients, or overall model fit. The researcher then needs only to examine the values and select those observations that exceed the specified value. We have already discussed one such measure, the studentized deleted residual, but will now explore several other measures appropriate for diagnosing individual cases.

## **Influences on Individual Coefficients**

The impact of deleting a single observation on each regression coefficient is shown by the **DFBETA** and its standardized version the **SDFBETA**. Calculated as the change in the coefficient when the observation is deleted, DFBETA is the relative effect of an observation on each coefficient. Guidelines for identifying particularly high values of SDFBETA suggest that a threshold of  $\pm 1.0$  or  $\pm 2.0$  be applied to small sample sizes, whereas  $\pm 2\sqrt{n}$  should be used for medium and larger data sets.

## **Overall Influence Measures**

**Cook's distance** ( $D_i$ ) is considered the single most representative measure of influence on overall fit. It captures the impact of an observation from two sources: the size of changes in the predicted values when the case is omitted (outlying studentized residuals) as well as the observation's distance from the other observations (leverage). A rule of thumb is to identify observations with a Cook's distance of 1.0 or greater, although the threshold of  $4/(n - k - 1)$ , where  $n$  is the sample size and  $k$  is the number of independent variables, is suggested as a more conservative measure in small samples or for use with larger data sets. Even if no observations exceed this threshold, however, additional attention is dictated if a small set of observations has substantially higher values than the remaining observations.

A similar measure is the **COVRATIO**, which estimates the effect of the observation on the efficiency of the estimation process. Specifically, COVRATIO represents the degree to which an observation impacts the standard errors of the regression coefficients. It differs from the DFBETA and SDFBETA in that it considers all coefficients collectively rather than each coefficient individually. A threshold can be established at  $1 \pm 3p/n$ . Values above the threshold of  $1 + 3p/n$  make the estimation process more efficient, whereas those less than  $1 - 3p/n$  detract from the estimation efficiency. This allows the COVRATIO to act as another indicator of observations that have a substantial influence both positively and negatively on the set of coefficients.

A third measure is **SDFFIT**, the degree to which the fitted values change when the case is deleted. A cutoff value  $2\sqrt{(k+1)/(n-k-1)}$  has been suggested to detect substantial influence. Even though both Cook's distance and SDFFIT are measures of overall fit, they must be complemented by the measures of steps 1 and 2 to enable us to determine whether influence arises from the residuals, leverage, or both. An unstandardized version (**DFFIT**) is also available.

#### **STEP 4: SELECTING AND ACCOMMODATING INFLUENTIAL OBSERVATIONS**

The identification of influential observations is more a process of convergence by multiple methods than a reliance on a single measure. Because no single measure totally represents all dimensions of influence, it is a matter of interpretation, although these measures typically identify a small set of observations. In selecting the observations with large values on the diagnostic measures, the researcher should first identify all observations exceeding the threshold values, and then examine the range of values in the data set being analyzed and look for large gaps between the highest values and the remaining data. Some additional observations will be detected that should be classified as influential.

After identification, several courses of action are possible. First, if the number of observations is small and a justifiable argument can be made for their exclusion, the cases should be deleted and the new regression equation estimated. If, however, deletion cannot be justified, several more “robust” estimation techniques are available, among them robust regression [3]. Whatever action is taken, it should meet our original objective of making the data set most representative of the actual population, to ensure validity and generalizability.

#### **OVERVIEW**

The identification of influential cases is an essential step in interpreting the results of regression analysis. The analyst must be careful, however, to use discretion in the elimination of cases identified as influential. There are always outliers in any population, and the researcher must be careful not to trim the data set so that good results are almost guaranteed. Yet, one must also attempt to best represent the relationships in the sample, and the influence of just a few cases may distort or completely inhibit achieving this objective. Thus, we recommend that one use these techniques wisely and with care, as they represent both potential benefit and harm.

#### **SUMMARY**

As the applications of multiple regression analysis increase in both scope and complexity, it becomes essential to explore the issues addressed in this appendix. Both multicollinearity and influential observations can have substantial impacts on the results and their interpretation. However, recent advances in diagnostic techniques, such as those described earlier, now provide the researcher with a simplified method of performing analyses that will identify problems in each of these areas. Whenever the regression analysis encounters either multicollinearity or influential observations, the researcher is encouraged to investigate the issues raised here and to take the appropriate remedies if needed.

## QUESTIONS

1. Describe the reasons for not relying solely on the bivariate correlation matrix for diagnosing multicollinearity.
2. In what instances does the detection of outliers potentially miss other influential observations?
3. Describe the differences in using residuals (including the studentized residual) versus the single-case diagnostics of DFBETA and DFFIT.
4. What criteria would you suggest for determining whether an observation was to be deleted from the analysis?

## REFERENCES

1. Barnett, V., and T. Lewis (1984), *Outliers in Statistical Data*, 2d ed. New York: Wiley.
2. Belsley, D. A., E. Kuh, and R. E. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
3. Rousseeuw, P. J., and A. M. Leroy (1987), *Robust Regression and Outlier Detection*. New York: Wiley.
4. Weisberg, S. (1985), *Applied Linear Regression*. New York: Wiley.