MULTIPLE REGRESSION

UNDELSTANDING MULTIPLE REGRESSION

Multiple regression analysis (MRA) is any of several related statistical methods for evaluating the effects of more than one independent (or predictor) variable on a dependent (or outcome) variable. Since MRA can handle all ANOVA problems (but the reverse is not true), some researchers prefer to use MRA exclusively. MRA answers two main questions: (1) What is the effect (as measured by a regression coefficient) on a dependent variable (DV) of a one-unit change in an independent variable (IV), while controlling for the effects of all the other IVs? (2) What is the total effect (as measured by the $R^2$) on the DV of all the IVs taken together?

Recall that simple linear (bivariate) regression, primarily used for prediction, has a single independent (predictor) variable ($X$) and a single dependent (criterion) variable ($Y$). In bivariate linear regression, a straight line was fitted to the scatterplot of points; the equation of this line, called the regression equation, has the form $\hat{Y} = bX + a$, where $\hat{Y}$ (Y hat) is the predicted value of the criterion variable for a given $X$ value on the predictor variable. The regression coefficient ($b$) is the slope of the line. The regression constant ($a$) is the $Y$ intercept.

In multiple linear regression, scores on the criterion variable ($Y$) are predicted using multiple predictor variables ($X_1, X_2, \ldots, X_k$). In multiple linear regression, we again have a single criterion variable ($Y$), but we have $K$ predictor variables ($k \geq 2$). These predictor variables are combined into an equation, called the multiple regression equation, which can be used to predict scores on the criterion variable ($\hat{Y}$) from scores on the predictor variables ($X_i$s). The general form of this equation is: $\hat{Y} = b_1X_1 + b_2X_2 + \ldots + b_kX_k + a$ (where the $b$s are the regression coefficients for the respective predictor variables and $a$ is the regression constant).

SELECTING PREDICTOR VARIABLES

Predictor variables need to serve a specific purpose on the regression analysis. That is, their selection should be supported in the literature. While researcher curiosity may play a part in selecting predictor variables, a rational (justification) of inclusion should be provided, with the best rational being supported in the applicable literature on the topic area.

In selecting predictor variables to use in multiple regression, we want to maximize the amount of explained variance ($R^2$) and reduce error (improve accuracy). Our predictor variables should be chosen carefully. The predictor variables ($X$s) should be highly correlated with the criterion variable ($Y$). However, the predictor variables should have a low correlation with other predictor variables (see multicollinearity). A suppressor variable is one that changes the $R^2$ but has a low correlation with the criterion variable and a high correlation with other predictor variables.

NUMBER OF PREDICTOR VARIABLES

Including more than 5 or 6 predictor variables in a regression analysis rarely produces a substantial increase in the multiple $R$. An adjusted $R^2$ is used when there are numerous predictor variables and a small number of observations ($n$). In studies where the number of observations is large ($n > 100$) and there are only 5 or 6 predictor variables – this is not a major concern.
SCREENING DATA PRIOR TO ANALYSIS

As with any statistical analysis, one should always conduct an initial screening of their data. One of the more common concerns with data error(s) is simply data entry errors. For example, making an incorrect entry of a variable value or omitting a value that should have been entered. The researcher should always double-check their data entry to avoid unnecessary (and completely avoidable) errors. Dealing with missing or potentially influential data can be handled in multiple ways. For example, the case(s) can be deleted (typically only if they account for less than 5% of the total sample) transformed or substituted using one of many options (see for example Tabachnick & Fidell, 2001).

Data can be screened as ungrouped or grouped. With ungrouped data screening, we typically conduct frequency distributions and include histograms (with a normal curve superimposed), measures of central tendency, and measures of dispersion. This screening process will allow us to check for potential outliers as well as data entry errors. We can also create scatter plots to check for linearity and homoscedasticity. Detection of multivariate outliers is typically done through regression – checking for mahalanobis distance values of concern and conducting a collinearity diagnosis (discussed in more detail below). Data can also be screened as grouped data. These procedures are similar to those for ungrouped data, with the exception that each group is analyzed separately. A split file option is conducted for separate analysis. Here again, descriptive statistics are examined, as well as histograms and plots.

MULTICOLLINEARITY

When there are moderate to high intercorrelations (e.g., $r > \pm .90$) among the predictors, the problem is referred to as multicollinearity. The extreme form of multicollinearity is singularity, when a perfect linear relationship exists between variables, or, in terms of correlation, when $r = \pm 1.00$. Multicollinearity poses a real problem for the researcher using multiple regression for three reasons:

1. It severely limits the size of $R$, because the predictors are going after much of the same variance on $y$.
2. Multicollinearity makes determining the importance of a given predictor difficult because the effects of the predictors are confounded due to the correlation among them.
3. Multicollinearity increases the variances of the regression coefficients. The greater these variances, the more unstable the prediction equation will be.

The following are three common methods for diagnosing multicollinearity:

1. Examine the variance inflation factors (VIF) for the predictors.

   The quantity $\frac{1}{(1-R_j^2)}$ is called the $j$th variance inflation factor, where $R_j^2$ is the squared multiple correlation for predicting the $j$th predictor from all other predictors.

   The VIF for a predictor indicates whether there is a strong linear association between it and all the remaining predictors. It is distinctly possible for a predictor to have only moderate and/or relatively weak associations with the other predictors in terms of simple correlations, and yet to have a quite high $R$ when regressed on all the other predictors.
When is the value for the variance inflation factor large enough to cause concern? As indicated by Stevens (2002): While there is no set rule of thumb on numerical values to compare the VIF, it is generally believed that if any VIF exceeds 10, there is reason for at least some concern. In this case, one should consider variable deletion or an alternative to least squares estimation. The variance inflation factors are easily obtained from SPSS or other statistical packages.

2. Examine the **tolerance** value, which refers to the degree to which one predictor can itself be predicted by the other predictors in the model. Tolerance is defined as $1 - R^2$. The higher the value of tolerance, the less overlap there is with other variables. The higher the tolerance value, the more useful the predictor is to the analysis; the smaller the tolerance value, the higher the degree of collinearity. While it is largely debated on target values, a tolerance value of .50 or higher is generally considered acceptable (Tabachnick & Fidell, 2001). Some statisticians accept a value as low as .20 before being concerned.

Tolerance tells us two things. First, it tells us the degree of overlap among the predictors, helping us to see which predictors have information in common and which are relatively independent. Just because two variables substantially overlap in their information is not reason enough to eliminate one of them. Second, the tolerance statistic alerts us to the potential problems of instability in our model. With very low levels of tolerance, the stability of the model and sometimes even the accuracy of the arithmetic can be in danger. In the extreme case where one predictor can be perfectly predicted from the others, we will have what is called a singular covariance (or correlation) matrix and most programs will stop without generating a model. If you see a statement in your SPSS printout that says that the matrix is singular or “not positive-definite,” the most likely explanation is that one predictor has a tolerance of 0.00 and is perfectly correlated with other variables. In this case, you will have to drop at least one predictor to break up that relationship. Such a relationship most frequently occurs when one predictor is the simple sum or average of the others.

Most multiple-regression programs (e.g., SPSS) have default values for tolerance (1 – SMC, where SMC is the squared multiple correlation, or $R^2$) that protect the user against inclusion of multicollinear IVs (predictors). If the default values for the programs are in place, the IVs that are very highly correlated with other IVs already in the equation are not entered. This makes sense both statistically and logically because the IVs threaten the analysis due to inflation of regression coefficients and because they are not needed due to their correlation with other IVs.

3. Examine the **condition index**, which is a measure of tightness or dependency of one variable on the others. The condition index is monotonic with SMS, but not linear with it. A high condition index (values of 30 or greater) is associated with variance inflation in the standard error of the parameter estimate of the variable. When its standard error becomes very large, the parameter estimate is highly uncertain.

If we identify a concern from the condition index, Tabachnick and Fidell (2001) suggest looking at the **variance proportions**. We do not want two variance proportions to be greater than .50 for each item. This consideration does not include the column labeled constant, which is the Y intercept (the value of the dependent variable when the independent variables are all zero).
There are several ways of combating multicollinearity.

1. Combine predictors that are highly correlated.

2. If one has initially a fairly large set of predictors – consider doing a principal components analysis (a type of factor analysis) to reduce to a much smaller set of predictors.

3. Consider deleting (omitting) a variable that is highly correlated with another variable or variables.

4. Consider a technique called ridge regression, which is an alternative to OLS (ordinary least squares) methods of estimating regression coefficients that is intended to reduce the problems in regression analysis associated with multicollinearity.

**Checking Assumptions for the Regression Model**

Recall that in the linear regression model it is assumed that the errors are independent and follow a normal distribution with constant variance. The normality assumption can be checked through the use of the histogram of the standardized or studentized residuals.

The independence assumption implies that the subjects are responding independently of one another. This is an important assumption in that it will directly affect the Type I error rate.

**Residual Plots**

There are various types of plots that are available for assessing potential problems with the regression model. One of the most useful graphs is the standardized residuals \( r_i \) versus the predicted values \( \hat{y}_i \). If the assumptions of the linear regression model are tenable, then the standardized residuals should scatter randomly about a horizontal line defined by \( r_i = 0 \). Any systematic pattern or clustering of the residuals suggests a model violation(s). What we are looking for is a distinctive pattern and/or a curvilinear relationship.

If nonlinearity or nonconstant variances are found, there are various remedies.

- For non-linearity, perhaps a polynomial model is needed. Or sometimes a transformation of the data will enable a nonlinear model to be approximated by a linear one.

- For non-constant variance, weighted least squares is one possibility, or more commonly, a variance stabilizing transformation (such as square root or log) may be used.

**Outliers and Influential Data Points**

Since multiple regression is a mathematical maximization procedure, it can be very sensitive to data points within “split off” or are different from the rest of the points, that is, to outliers. Just one or two such points can affect the interpretation of results, and it is certainly moot as to whether one or two points should be permitted to have such a profound influence. Therefore, it is important to be able to detect outliers and influential data points. There is a distinction between the two because a point that is an outlier (either on \( y \) or for the predictors) will not necessarily be influential in affecting the regression equation.
Data Editing

Outliers and influential cases can occur because of recording errors. Consequently, researchers should give more consideration to the data-editing phase of the data analysis process (i.e., always listing the data and examining the list for possible errors). There are many possible sources of error from the initial data collection to the final data entry. Take care to review your data for incorrect data entry (e.g., missing or excessive numbers, transposed numbers, etc.).

Measuring Outliers on y

For finding subjects whose predicted scores are quite different from their actual y scores (i.e., they do not fit the model well), the standardized residuals ($r_i$) can be used. If the model is correct, then they have a normal distribution with a mean of 0 and a standard deviation of 1. Thus, about 95% of the $r_i$ should lie within two standard deviations of the mean and about 99% within three standard deviations. Therefore, any standardized residual greater than about 3 in absolute value is unusual and should be carefully examined (Stevens, 2002). Tabachnick and Fidell (2001) suggest measuring standardized residuals at an alpha level of .001, which in turn would give us a critical value of ±3.29.

Measuring Outliers on Set of Predictors

The hat elements ($h_{ii}$) can be used here. It can be shown that the hat elements lie between 0 and 1, and that the average hat elements is $\frac{p}{n}$, where $p = k + 1$. Because of this, Stevens suggests that $\frac{3p}{n}$ be used. Any hat element (also called leverage) greater than $\frac{3p}{n}$ should be carefully examined. This is a very simple and useful rule of thumb for quickly identifying subjects who are very different from the rest of the sample on the set of predictors.

Measuring Influential Data Points

An influential data point is one that when deleted produces a substantial change in at least one of the regression coefficients. That is, the prediction equations with and without the influential point are quite different. Cook’s distance is very useful for identifying influential data points. It measures the combined influence of the case being an outlier on y and on the set of predictors. As indicated by Stevens (2002), a Cook’s distance > 1 would generally be considered large. This provides a “red flag,” when examining the computer printout, for identifying influential points.

Mahalanobis’ Distance

Mahalanobis’ distance ($D^2$) indicates how far the case is from the centroid of all cases for the predictor variables. A large distance indicates an observation that is an outlier for the predictors. How large must $D^2$ be before one can say that a case is significantly separated from the rest of the data at either the .01 or .05 level of significance? If it is tenable that the predictors came from a multivariate normal population, then the critical values can be compared to an established table of values. If $n$ is moderately large (50 or more), then $D^2$ is approximately proportional to $h_{ii}$. Therefore, $D^2 = (n - 1) h_{ii}$ can be used.
Tabachnick and Fidell (2001) suggest using the Chi-Square critical values table as a means of detecting if a variable is a multivariate outlier. Using a criterion of $\alpha = .001$ with $df$ equal to the number of independent variables to identify the critical value for which the Mahalanobis distance must be greater than. According to Tabachnick and Fidell, we are not using $N-1$ for $df$ because “the Mahalanobis distance is evaluated with degrees of freedom equal to the number of variable” (p. 93). Thus, all Mahalanobis distance values must be examined to see if the value exceeds the established critical value.

**Options for dealing with outliers**

1. Delete a variable that may be responsible for many outliers especially if it is highly correlated with other variables in the analysis. If you decide that cases with extreme scores are not part of the population you sampled then delete them. If the case(s) represent(s) 5% or less of the total sample size Tabachnick and Fidell (2001) suggest that deletion becomes a viable option.

2. If cases with extreme scores are considered part of the population you sampled then a way to reduce the influence of a univariate outlier is to transform the variable to change the shape of the distribution to be more normal. Tukey said you are merely reexpressing what the data have to say in other terms (Howell, 2002).

3. Another strategy for dealing with a univariate outlier is to assign the outlying case(s) a raw score on the offending variable that is one unit larger (or smaller) than the next most extreme score in the distribution” (Tabachnick & Fidell, 2001, p. 71).

4. Univariate transformations and score alterations often help reduce the impact of multivariate outliers but they can still be problems. These cases are usually deleted (Tabachnick & Fidell, 2001). **All transformations, changes to scores, and deletions are reported in the results section with the rationale and with citations.**

**Summary**

In summarizing – the examination of standardized residuals will detect $y$ outliers, and the hat elements or the Mahalanobis distances will detect outliers on the predictors. Such outliers will not necessarily be influential data points. To determine which outliers are influential data points, find those points whose Cook distances are $> 1$. Those points that are flagged as influential by Cook’s distance need to be examined carefully to determine whether they should be deleted from the analysis. If there is a reason to believe that these cases arise from a process different from that for the rest of the data, then the cases should be deleted. If a data point is a significant outlier on $y$, but its Cook distance is $< 1$, there is no real need to delete the point since it does not have a large effect on the regression analysis. However, one should still be interested in studying such points further to understand why they did not fit the model.

**References**


### Critical Values for an Outlier on the Predictors as Judged by Mahalanobis $D^2$

<table>
<thead>
<tr>
<th>Number of Predictors</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
<th>$k = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>5%</td>
<td>1%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>5</td>
<td>3.17</td>
<td>3.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4.00</td>
<td>4.11</td>
<td>4.14</td>
<td>4.16</td>
</tr>
<tr>
<td>7</td>
<td>4.71</td>
<td>4.95</td>
<td>5.01</td>
<td>5.10</td>
</tr>
<tr>
<td>8</td>
<td>5.32</td>
<td>5.70</td>
<td>5.77</td>
<td>5.97</td>
</tr>
<tr>
<td>9</td>
<td>5.85</td>
<td>6.37</td>
<td>6.43</td>
<td>6.76</td>
</tr>
<tr>
<td>10</td>
<td>6.32</td>
<td>6.97</td>
<td>7.01</td>
<td>7.47</td>
</tr>
<tr>
<td>12</td>
<td>7.10</td>
<td>8.00</td>
<td>7.99</td>
<td>8.70</td>
</tr>
<tr>
<td>14</td>
<td>7.74</td>
<td>8.84</td>
<td>8.78</td>
<td>9.71</td>
</tr>
<tr>
<td>16</td>
<td>8.27</td>
<td>9.54</td>
<td>9.44</td>
<td>10.56</td>
</tr>
<tr>
<td>18</td>
<td>8.73</td>
<td>10.15</td>
<td>10.00</td>
<td>11.28</td>
</tr>
<tr>
<td>35</td>
<td>11.10</td>
<td>13.20</td>
<td>12.85</td>
<td>14.92</td>
</tr>
<tr>
<td>40</td>
<td>11.53</td>
<td>13.74</td>
<td>13.36</td>
<td>15.56</td>
</tr>
<tr>
<td>50</td>
<td>12.33</td>
<td>14.60</td>
<td>14.18</td>
<td>16.56</td>
</tr>
<tr>
<td>100</td>
<td>14.22</td>
<td>16.95</td>
<td>16.45</td>
<td>19.26</td>
</tr>
<tr>
<td>200</td>
<td>15.99</td>
<td>18.94</td>
<td>18.42</td>
<td>21.47</td>
</tr>
<tr>
<td>500</td>
<td>18.12</td>
<td>21.22</td>
<td>20.75</td>
<td>23.95</td>
</tr>
</tbody>
</table>

*Stevens (p. 133, 2002)*