HBAT Multiple Regression (August 2019)

Jason Huggy, Student, Lewis University

I. INTRODUCTION

The purpose of this study was to utilize multiple regression to analyze the HBAT dataset found in *Multivariate Data Analysis* [1]. The data set includes 99 lines of unlabeled data, and 23 variables. Multiple regression analyzes the relationship between multiple independent variable and one dependent variable. The key is to understand the bivariate relationship between the dependent variable and each independent variable to help build the model. While the mathematics behind this algorithm are simple with one variable, in a multivariate situation it can become overwhelming. Therefore, SAS Enterprise Guide (EG) is used to perform all analysis on the multiple regression model.

II. METHODOLOGY

To conduct this study, SAS Enterprise Guide is used to produce all figures and data tables. To begin, the HBAT data is imported into EG and verified to be free of missing values or errors. A hasty examination of the data shows that the first five columns of data, as well as the last, are nonmetric features. The rest of the features are all metric. For this study only 14 features are utilized. X variables 6 through 18 of the HBAT data serve as the independent variables, while variable 19 serves as the dependent variable. Each column contains metric data, so there is no need to alter the data any more at this point.

A. Data Analysis

After importing the HBAT data into EG, a data characterization report is created. This is used to determine if there are any extreme outliers in the data, or errors that may need to be fixed. To check normality, histograms are plotted for every variable. While data transformations are needed for multiple variables to fix issues with normality, this study uses multiple regression on the variables without transformations to look at the impact of leaving the variables as they are.

B. Building the Model

To build the regression model, the linear regression function is used in SAS Enterprise Guide. Variable 19 is first selected as the dependent variable, and then then the other 13 variables are designated as the independent variables. To conduct the estimation process of selecting the right number of variables to use for the best model, the stepwise option is selected under the model attributes. In addition, a variety of analysis tools are printed along with the model in order to assess the fit.

III. RESULTS AND DISCUSSION

After running the multiple regression function in SAS Enterprise Guide, the resulting information was analyzed to determine the model's performance. The stepwise estimation process finished in six iterations and ended with only variables 6, 7, 9, 11, 12, and 16. The current R-Squared at this point was 0.79 and the adjusted R-Squared was 0.78. This shows that the model had some success as far as accuracy but has a ways to go to be totally accurate. When evaluating the correlation matrix with the remaining variables, it can be seen that several of the variables are correlated above 0.5 (or -0.5), which means that there are cases of multicollinearity that may need to be removed for further accuracy. The idea is to try to use variables that do not correlate with each other, while also explaining as much of the variance in the data as possible.

When looking at Figure 1, the data appears to be mostly random, with a few of the data points stretching a little far from a majority of the body of points. These points may need to be assessed to see if they are throwing off the effectiveness of the model. Figure 2 shows the relationship between each variable and the

dependent variable, at these plots show much of the same.

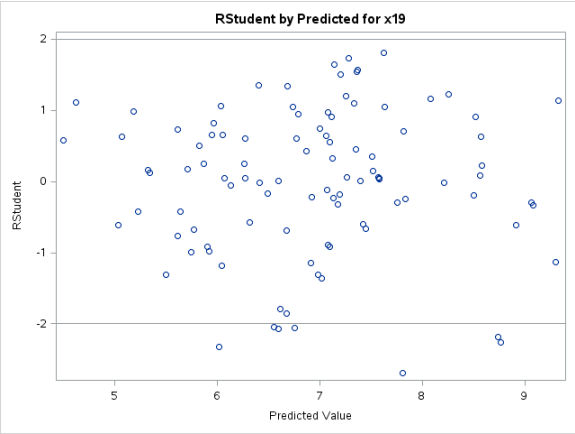

Figure 1. Analysis of Standardized Residuals

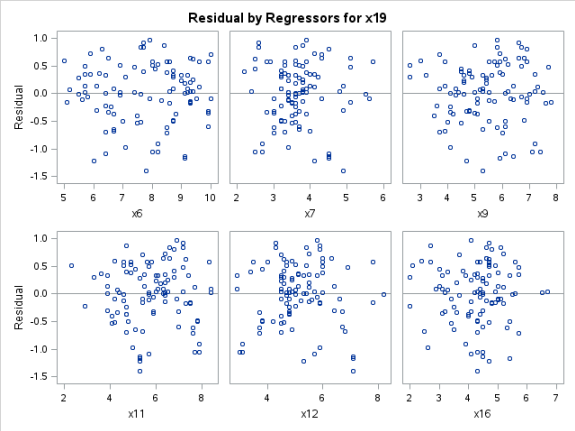

Figure 2. Standardized Partial Regression Plots

Figure 3 depicts the normal probability plot of the model. While most of the data aligns with the line shown, there are points at the beginning and the end of the plot that show outliers. This is consistent with the outliers seen in Figure 1.

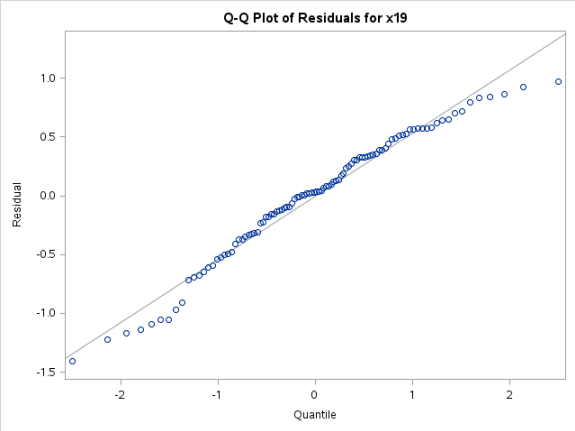

Figure 3. Normal Probability Plot

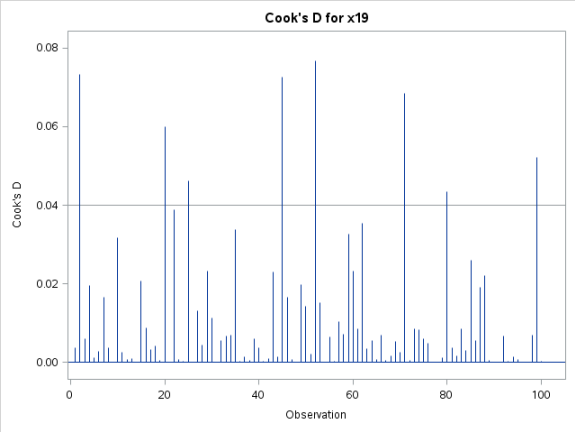

Figure 4. Cook's D Plot for HBAT Data

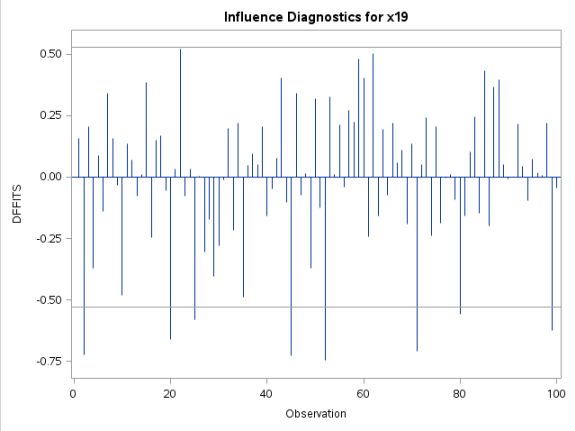

Figure 5. DFFITS Plot for HBAT Data

Figures 4 and 5 show the major influencers in the data. These influential points are consistent in both plots. These influencers need to be accessed to see if they are affecting the accuracy of the multiple regression model. If these points are outliers, then there is a need to determine if they are errors that need to be removed. If they are not errors, then it would benefit the researcher to try to further understand the relationship behind the outlier. Regardless, these influencers need action. Either the points must be removed, or more data may be needed to further study their impact on the model.

IV. CONCLUSION

Overall, the multiple regression model created in this study was accurate but could be better. First, data transformations need to be applied to several of the variables in order to fix their normality. This alone could have a significant impact on improving the model. Next, there are several variables, even after the

stepwise estimation process, where multicollinearity occurs. Some variables may need to be removed in order to reduce the amount of correlation between independent variables. Lastly, there is a presence of many outliers and influencers. More in-depth analysis is required to determine if these data points are limiting the effectiveness of the model, and if they need to be removed or further backed up with more data. Regardless, the overall model explains a good portion of the variation in data, and even an adjusted R-Squared of 0.78 is pretty good for the beginning stages of the process. The predictive capabilities of this model could be useful still, even if further tweaks to the model are unable to be made.

REFERENCES

[1] J. Hair, W. Black, B. Babin, R. Anderson, *Multivariate Data Analysis: Pearson New International Edition*. 2013.