Option #1: Food & Beverage BI Use Case: Wine Recognition Using SAS

**Scott Miner**

Colorado State University – Global Campus

**Option #1: Food & Beverage BI Use Case: Wine Recognition Using SAS**

The assignment asks the student to choose two variables at random from the results of a chemical analysis performed on wines grown in the same region in Italy, from three different cultivars or plant varieties. The Wine Recognition dataset comprises 178 observations, with 59 belonging to the first cultivar, 71 to the second, and 48 to the third. Thirteen constituents comprise the cultivars, represented in the dataset by numeric, continuous variables. “Cultivar” is the only categorical variable, assuming one of three numeric values: 1, 2, or 3. After using SAS to plot the variables, we calculate measures for the separation and cohesion. We then use multinomial logistic regression (MLR) to determine the variables most relevant for classification. These variables are plotted on a scatterplot, and we calculate measures for the separation and cohesion in this context and discuss our findings.

**Calculating Cohesion & Separation**

I chose first to plot proanthocyanidins by magnesium between cultivars. Proanthocyanidins are a type of phenol, or chemical, that affects the taste, color, and texture of a wine. Magnesium is a metal, affecting wine flavor (Homework 3: Part II, n.d.). Figure 1 shows these variables plotted on a scatterplot. We see a wide range of dispersion among the different cultivars; the classes do not appear very separate or cohesive, as the three colors are seen overlapping each other. We use measures of dispersion, including separation and cohesion, to assess the validity or separateness of our clusters. We aim to create clusters that are highly separated and cohesive. We calculate cohesion, also called the Sum of Squared Error (SSE), by finding the *Within* Cluster Sum of Squares (WCSS): . What this is saying is that for every data point within each cluster, we subtract the coordinate values from the mean value for the cluster (e.g., plant variety), square the results to eliminate negative values, and sum these values first amongst themselves and then amongst all three levels of cultivar to calculate the WCSS or cohesion. We calculate separation by finding the *Between* Cluster Sum of Squares (BCSS) = , where is the size of cluster *i,* and *m* is the centroid of the entire data set. What this is saying is that for every data point (e.g., the size of the cluster or 59 observations in the case of class 1), we subtract the coordinate measurements of the centroid of a given cluster from the coordinate measurements of the grand mean, or the mean of the means, then square and add the results as above, to determine the BCSS or separation. The sum of these two measurements creates the Total Cluster Sum of Squares (TCSS): Then, using the ratios allowed by these statistics (e.g., , ), we can determine how much of the data dispersion is due to cohesion and how much is due to separation (Gao, n.d.). To determine these measurements for the Wine Recognition data set, I used Excel to perform the calculations described above on the given data points. Figure 2 shows the results of these calculations. On the left-hand side of Figure 2, highlighted in red, we see the results of our calculations performed on the magnesium and proanthocyanidin variables. We can thus state 96.4% of dispersion is due to separation and 3.5% due to cohesion. These seem like adequate measurements, though compare them to the measurements on the right side, the separation and cohesion measurements for two variables (e.g., alcohol and flavonoids) found to most heavily influence the output variable (e.g., cultivar) as the result of MLR analysis. The calculations performed on these variables show the separation accounts for 99% of the variability between categories, while the cohesion accounts for less than 1% of the overall variation within categories. We see more tightly coupled clusters in our second model, discussed shortly.

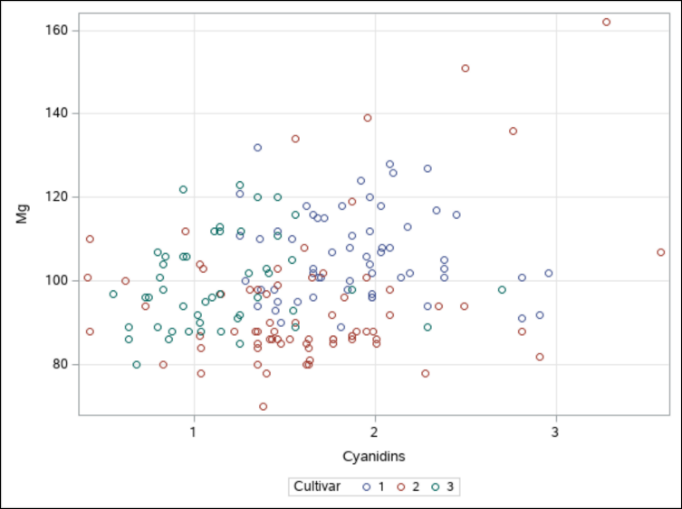


Figure . Scatterplot of Mg by Cyanidins in three cultivars

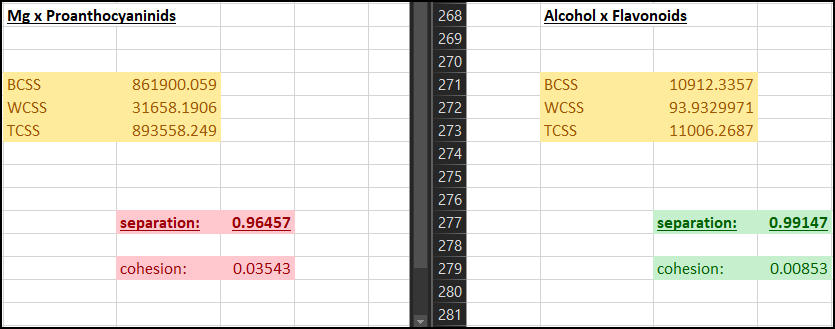


Figure 2. Calculating the Sum of Squares, Separation, and Cohesion for chemical

attributes across cultivars

How can we determine whether we are on the right track? The Sum of Squares (SS) is closely related to another measurement of dispersion, variance. We define variance as: , where *n* represents the number of observations. Given the variance, we can thus solve for the SS by multiplying the variance by (Z-5: Sum of Squares, Variance, and the Standard Error of the Mean—Westgard, n.d.). Figure 3 shows the variances for four variables in the Wine Recognition dataset: magnesium, proanthocyanidins, alcohol, and flavonoids. Magnesium and proanthocyanidins we already discussed. We measure alcohol content in units of ABV (alcohol by volume). Flavonoids are another phenol found in wine (Homework 3: Part II, n.d.). Using our calculator, we solve for the SS given the variances and number of observations per class. Figure 4 displays the results of our verifications. The final calculation in this result (31658.19057), equals the calculation performed above by Excel for the WCSS statistic in Figure 2. In effect, the two methods verify each other and confirm our results.

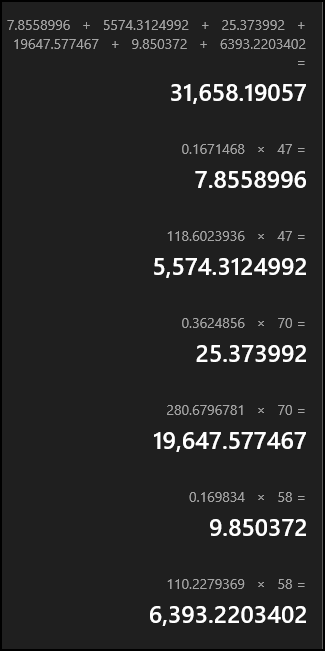


Figure 4. Verifying the WCSS

**Multinomial Logistic Regression (MLR)**

How were we able to determine which predictor variables within the dataset were most likely to influence the outcome variable? To perform this analysis, we turn our attention to MLR. Monyai et al. (2016) describe MLR as being used when the dependent variable has two or more categories, as was the case in our situation. El-Habil (2012) says many statisticians consider MLR to be one of the most important analyses applied to categorical data. Denham (2010) describes the MLR model as an extension of binary logistic regression, like discriminant analysis, though perhaps more useful. We thus implement MLR analysis via SAS to determine the chemical attributes most relevant in determining cultivars.

The logistic procedure in SAS provides four automatic model selection techniques: forward selection, backward elimination, stepwise selection, and the best subset selection. The model presented in this paper makes use of stepwise selection, which Shtatland et al. (2001) describes as being a “more flexible and sophisticated selection procedure” (p. 22) than the others, incorporating elements of both the forward selection and backward elimination models. Using a stepwise selection technique allows us to create a parsimonious model, capable of getting the most from the fewest number of factors. Figure 5 displays the Analysis of Effects output from our MLR model. We determine there is a significant relationship between the predictor variables, alcohol and flavonoids, and the outcome variable, cultivar. These variables are said to be the most effective indicators of the outcome variable. As indicated on the right-hand side of Figure 2, we see the separation accounts for 99% of the variability within this model, and the cohesion accounts for < 1%. In effect, we have a highly separated, tightly coupled model. Figure 6 shows this model plotted on a scatterplot. In this scatterplot, we see the separability among classes to be more distinct.

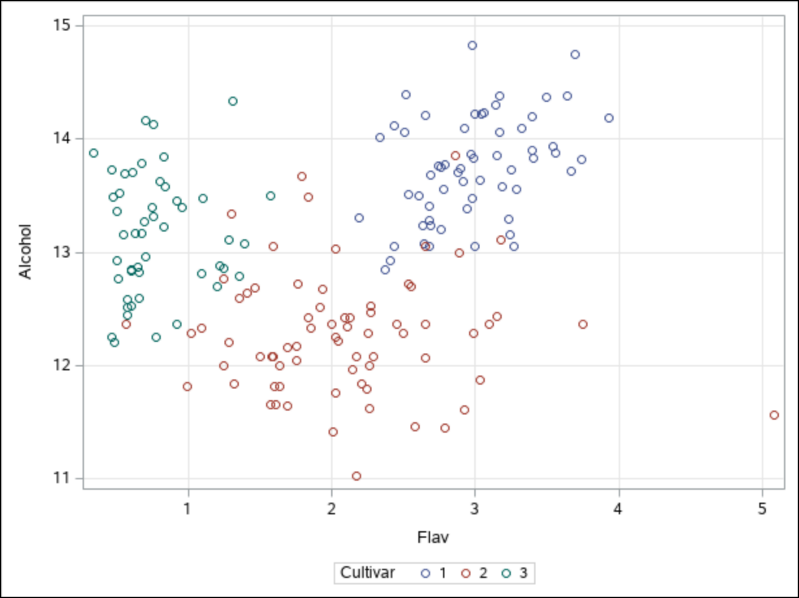


Figure 6. Scatterplot of Alcohol by Flavonoid in three cultivars

**Conclusion**

In closing, we calculated the sum of squares for four attributes found in the Wine Recognition dataset, mg, cyanidins, alcohols, and flavonoids, using SAS in combination with Excel and verified our results. We used MLR to determine the factors most influential in predicting plant variety: alcohol and flavonoids. We calculated measures of cohesion and separation for these models, concluding with a model where separation accounts for 99% of the variability, and cohesion the remaining 1%. We used SAS to plot the scatterplots of these variables to visualize the results. We can see, by comparison, how the results in Figure 6 are more tightly coupled than those in Figure 2.

References

Denham, Bryan E. (2010). Measurement of risk perceptions in social research: A comparative analysis of ordinary least squares, ordinal, and multinomial logistic regression models. Journal of Risk Research, 13(5), 571–589. https://doi.org/10.1080/13669870903172386

El-Habil, A. M. (2012). An Application on Multinomial Logistic Regression Model. Pakistan Journal of Statistics & Operation Research, 8(2), 271–291. https://doi.org/10.18187/pjsor.v8i2.234

Gao, J. (n.d.). Clustering Lecture 1: Basics. 62.

Homework 3: Part II. (n.d.). Retrieved January 11, 2020, from <https://cs.brown.edu/courses/cs100/homework/homework-3-part-ii/>

Monyai, S., Lesaoana, ’Maseka, Darikwa, T., & Nyamugure, P. (2016). Application of multinomial logistic regression to educational factors of the 2009 General Household Survey in South Africa. Journal of Applied Statistics, 43(1), 128–139. <https://doi.org/10.1080/02664763.2015.1077941>

Shtatland, E. S., Cain, E. M., & Barton, M. B. (2001). THE PERILS OF STEPWISE LOGISTIC REGRESSION AND HOW TO ESCAPE THEM USING INFORMATION CRITERIA AND THE OUTPUT DELIVERY SYSTEM.

Z-5: Sum of Squares, Variance, and the Standard Error of the Mean—Westgard. (n.d.). Retrieved January 12, 2020, from https://www.westgard.com/lesson35.htm