**Module 8: Portfolio Project Option 2**

**Build a Wine Quality Multiple Linear Regression Model Using SAS Studio**

Joshua Humphreys

Colorado State University Global

Course Code:MIS500

Dr. Osama Morad

February 13, 2022

**Module 8: Porfolio Project**

For my portfolio project, I selected option 2 and built a wine quality multiple linear regression model. Below you will find screenshots documenting successful completion of each assignment requirement. Following the screenshots, I have included an explanation and interpretation of my work.

**Figure 1.**

*Screenshot of importing winequality-red.csv data into Wine\_quality\_red dataset*

Graphical user interface, text, email

Description automatically generated

*Note.* current date in lower right of screenshot.

**Figure 2.**

*Screenshot 1 of 2 of data inputs*

Graphical user interface, application

Description automatically generated

**Figure 3.**

*Screenshot 2 of 2 of data inputs*

Graphical user interface, application

Description automatically generated

**Figure 4.**

*Screenshot of Model Effects Input*

Graphical user interface, application

Description automatically generated

**Figure 5.**

*Screenshot of statistics selection*

Graphical user interface, text, application

Description automatically generated

**Figure 6.**

*Screenshot of model selection input using default (none) selection*

Graphical user interface, text, application

Description automatically generated

**Figure 6.**

*Screenshot of model selection input using stepwise selection*

Graphical user interface, application

Description automatically generated

**Figure 7.**

*Output of Linear Regression Model using default (none) selection*

Table

Description automatically generated

**Chart

Description automatically generated**

**A picture containing diagram

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Figure 8.**

*Output of Linear Regression Model using stepwise selection*

Table

Description automatically generated

Table

Description automatically generated

**Chart

Description automatically generated**

Table

Description automatically generated

Table

Description automatically generated

**Chart

Description automatically generated**

**A picture containing diagram

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Explanation**

I started this project by loading the dataset in to SAS studio. I initially completed this step when I did my univariate and multivariate analysis. The task of loading the data did end up being a little involved. The original CSV used semi colons as delimiters instead of commas which SAS didn’t like. In order to resolve the issue I had to download the original CSV, open it in Excel, and then convert it to a CSV using commas as delimiters. After I loaded the data in SAS, I imported the dataset as demonstrated in Figure 1.

Next came the process of building the linear regression model. I selected *quality* as the dependent variable. I then selected all other variables in the *continuous variables* field. In the *Model* tab I added all variables to the model effects field. In the *Options* tab I did change the *Display statistics* field from the default setting to *Default and selected statistics* and selected *Variance inflation factors.* During my previous analysis I discovered correlation between multiple variables. At the recommendation of Cody (2019), I elected to include this statistic in order to assess for potential collinearity.

Up to this point, the process for creating both multiple linear regression models was the same. Therefore, I only included one set of screenshots to demonstrate completion of these tasks. For the first model, under the *Selection* tab, I left the *Model Selection* field set to *none.* By leaving the selection method set to *none* “the regression task will enter all of the predictor variables that you have selected, either as Classification variables or Continuous Variables” (Cody, 2019, p. 157).

For the second model, I decided to use stepwise selection. I used the adjusted R-squared statistic for my selection criteria. According to Cody (2019) forward selection “first chooses the ‘best’ variable depending on which of the selection criteria you have chosen” (p. 157). In this case, it selected the variable with the highest adjusted R-square value. Then forward selection continues to select variables based on their adjusted R-square values until the p-value exceeds a default value. In this case I left the *default value* field alone. Stepwise selection does the same thing as forward selection, but “variables already selected can be removed if adding other variables causes a particular variable to no longer satisfy the selection criterion” (cody, 2019, p. 157).

**Interpretation**

For my interpretation I’m going to attempt to work through both models somewhat simultaneously. That is, not all of the outputs are the same, so a true side-by-side comparison isn’t reasonable. However, I think we glean the most information from this assignment by comparing the two as best as possible. The first and most obvious difference between the two models can be visualized in the *Stepwise Selection Summary* (Figure 8). We can see that not all variables were selected in the model. Importantly, we see that *fixed acidity, density,* and *residual sugar* were not included in the model. Recall from my multivariate analysis that none of these variables demonstrated correlation with quality.

Moving down the output from the stepwise selection model, we see the *fit criteria for quality* matrix. This matrix shows four different methods for selecting variables. Interestingly, using Adjusted R-square selected the most variables. *Free sulfur dioxide* and *citric acid* were not selected by SBC. *Citric acid* was only included when using the adjusted R-square criteria.

The *Analysis of Variance* tables shows that the mean square due to the model is much larger than the means square due to error in both models yeilding large F values for both models. However, we do see a larger F value in the stepwise selection model. The adjusted R-square for the stepwise model was 0.3567 while the adjusted R-square for the default model was .3561.

The *Parameter* *Estimates* tables include the *variance inflation* statistic. In the default model we see high values of variance inflation for *fixed acidity* and *density.* Recall that neither of these variables were included in the stepwise selection model. We also see somewhat higher values (>3) for citric acid and pH in the default model. The stepwise selection model also produced somewhat higher (>2) values for citric acid. Recall from my multivariate analysis that citric acid has a strong relationship with volatile acidity. Furthermore pH did not demonstrate a relationship with quality. Based on these values and analysis, the model could benefit from removing both of these variables. Recall that the adjusted R-square selection criteria was the only selection criteria evaluated in the *Fit Criteria for quality* matrix that included citric acid in the model.

Examination of both *Fit Diagnostics for quality* tables reveals that the residuals for both models are approximately normal. Also, the *Residual by Regressors for quality* plots for both models do not demonstrate any clear patterns.

Finally, I’d like to talk about the *Observed by Predicted for quality* plots. This plot, for both models, clearly demonstrates that *quality* is not truly a continuous variable. Of course we knew that from our preliminary analysis. As the description of the data set suggests, it would certainly be appropriate to apply a classification task as opposed to regression. However, I don’t think that means that regression is useless. I think the approach to practical deployment of this model would be to simply round the model output to the nearest whole number*.* The *observed by Predicted for quality* plot also allows us to visualize how effective this model is at predicting wine quality. Recall that the adjusted R-square value for both models was approximately .35. That is, 35% of the variance in quality is explained by the predictors. The model appears to be pretty good at predicting quality for “middle of the road” wines. However, the model doesn’t do a very good job at capturing either high quality or low quality wines.

**Reflection**

Overall, this project was enlightening. The task of attempting to predict a subjective value like quality using objective measurements was intriguing. Though I suppose it does make sense. That is, our senses are little more than chemical measurement devices. The problem inherent in a task such as this is that human perception is filtered through our minds. Furthermore, there are an unknown number of confounding factors that have nothing to do with chemical characteristics that influence individual taste. With that being said it’s quite impressive that we are capable of making any sort of prediction like this. In that way, this project highlights the power of data analytics.

**References**

Cody, R. (2019). *A gentle Introduction to Statistics Using SAS Studio.* SAS Institute Inc.

Dietrich, D., Heller, B., & Yang, B. (2015). *Data science & big data analytics: Discovering, analyzing, visualizing and presenting data.* John Wiley & Sons.