**Analyses One: Stepwise Binary Logistic Regression**

Thus far, all assumptions for a Logistic Regression have been tested.

The following assumptions passed:

* sample size
* outliers
* independence of errors

The following assumptions did not pass:

* multicollinearity
* logit linearity

**Addressing Multicollinearity:**

Q - Can Multicollinearity be Corrected?

A - Yes.

Q - Is it Necessary to Correct for Multicollinearity? -

A - That’s debatable.

Background / Research on the effects of Multicollinearity:

<https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>

“The need to reduce multicollinearity depends on its severity and your primary goal for your regression model. Keep the following three points in mind:

* The severity of the problems increases with the degree of the multicollinearity. Therefore, if you have only moderate multicollinearity, you may not need to resolve it.
* Multicollinearity affects only the specific independent variables that are correlated. Therefore, if multicollinearity is not present for the independent variables that you are particularly interested in, you may not need to resolve it. Suppose your model contains the experimental variables of interest and some control variables. If high multicollinearity exists for the control variables but not the experimental variables, then you can interpret the experimental variables without problems.
* Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If your primary goal is to make predictions, and you don’t need to understand the role of each independent variable, you don’t need to reduce severe multicollinearity.”

“VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.”

**For our Project, The VIF between Fixed Acidity & Density is -5.61.**

**This is the only variable pair that’s problematic.**

Q - How do we correct for Multicollinearity?

A -

“The potential solutions include the following:

* Remove some of the highly correlated independent variables.
* Linearly combine the independent variables, such as adding them together.
* Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.
* LASSO and Ridge regression are advanced forms of regression analysis that can handle multicollinearity. If you know how to perform linear least squares regression, you’ll be able to handle these analyses with just a little additional study.

As you consider a solution, remember that all of these have downsides. If you can accept less precise coefficients, or a regression model with a high R-squared but hardly any statistically significant variables, then not doing anything about the multicollinearity might be the best solution.”

**Q - Do we as a team want to correct for Multicollinearity or proceed without correction?**

**Addressing Logit Linearity:**

Q - Can Logit Linearity be Corrected?

A - Yes

Q - Is it Necessary to Correct for Logit Linearity?

A - That depends on how strong you want your model to be and how much work you’re prepared to do to.

Q - How can Logit Linearity be corrected?

A - <https://www.researchgate.net/post/What_options_do_I_have_if_the_assumption_of_the_linearity_of_the_logit_is_not_met_in_a_binary_logistic_regression_model>

* “You can allow the curve to be more flexible, and then based on cross-validation or some other method decide how flexible the curve should be. See generalized additive models or GAMs.”
* “You could try transformation of your IV(s). This might reduce the magnitude of the interaction terms to ignorable levels. A polynomial approach might work as well (e.g., include X, and X-squared--after centering--as model IVs.”
* “Yet another approach would be bootstrap model estimation. Here's a simple link showing how to elicit in r: <https://www.r-bloggers.com/2015/08/bootstrap-with-logistic-regression/>"

**Q - Do we as a team want to correct for Logit Linearity or proceed without correction?**

final / finishing tasks:

**Running the Logistic Regression**

**Interpreting the Results in Plain Language**