**WIND DATA ANALYSIS**

# **Objective Statement**

# **Central Figure**

# **Introduction**

This project investigated a “wine review” dataset to understand patterns, if any, that link wine varieties, their prices, and the no. of points they receive upon being rated (superior = points or not) on multiple characteristics (taste, smell).

The dataset is fairly big with 2500 data points, but the distribution of those data points among the various wine varieties is highly uneven. For instance, there are only 3 data points for the "Petit Manseng" variety of wine while there are 417 data points corresponding to the “Chardonnay” variety. In such a case wherein multiple groups of interest are underrepresented in the data set; frequentist methods may be unreliable with the model being prone to overfitting. Overfitting here, refers to the model fitting the data corresponding to varieties with highest no. of data points very well but failing to fit underrepresented varieties well enough due to a lack of data points, to ultimately result in high variance between estimates w.r.t inputs belonging to different groups. Thus, adopting the Bayesian framework was deemed a good idea because by incorporating prior information and considering hierarchical structures, it would provide more robust estimates through partial pooling that leverages the idea of borrowing strength across varieties to better estimate parameters.

The decision to use a Hierarchical Logistic Regression model to fit the data was made because such a model would be able to capture the discrete binary response variable in light of a Bernoulli distribution (if wine received superior rating then 1 = success and 0 = failure otherwise) through a linear combination of predictor variables (, , , , , , , , , , , ), while also taking into account the grouping structure (wine ) present in the data such that group parameters are pooled together to share information (partial pooling), thereby allowing for underrepresented groups to draw statistical support from others and also adding a regularization effect which should reduce over-fitting.

Subscripts

Give below is some important notation to bear in mind w.r.t subscripts in order to understand the model definition that follows. Because of the hierarchical nature of this model which also involves multiple predictor variables in a linear combination to explain each response variable, the subscripts are especially important, but may get confusing. Thus, this prelude before diving into expressions that define the model exists to make the purpose of each subscript clear.

* refer to groups that are present in the data set and are being considered in the model. Here, because the ***grouping variable*** is wine and there are different varieties of wine in the data set, .
* refer to data points that are present within a group. For example, for the “Chardonnay” variety of wine in the data set, there are 417 data points (wines). Thus, w.r.t the “Chardonnay” group, .
* refer to predictor variables used to estimate each instance of the response variable. In the model considered here, the 12 ***predictor variables*** are , , , and . Thus, .

The Model

The ***response variable*** (the variable we want to model) is . This is a binary variable such that when it is 1, this means that the corresponding wine was assigned a rating of points. Consequently, implies that the wine received a rating points.

Since here, the response variable , meaning that outcomes are discrete and binary, logistic regression would be a better choice for a data model than linear regression because the latter predicts continuous values while the former deals specifically with binary outcomes. Moreover, a linear regression model would try to fit a straight line to the data, which likely wouldn’t accurately represent the binary nature of the response variable (superior rating or not). In contrast, a logistic regression model uses a sigmoid function to transform the linear combination of predictors into a probability between 0 and 1 which can then be interpreted as the chance of observing a superior rating () for a particular wine given its characteristics .

The data model is assumed to be wherein, a wine’s rating being superior is a success, and it receiving a lower rating is seen as failure.

The log-odds of a success (a wine being deemed superior), is modelled as a linear function of the predictor variables and the grouping term to obtain a logistic regression model as follows wherein are the coefficients associated with each predictor variable that explains the corresponding response variable instance.

Intercepts are special because they represent how, assuming a certain global baseline log odds of wines being rated as superior, it varies from group to group as captured by the Random Effects (RE) intercept .

The link function here, is . This link function computes so that the resulting values are valid probabilities and do not fall outside the [0, 1] range [1]. Thus, given covariates and an associated vector of coefficients , can be obtained as , meaning that can be used to predict the value of .

The RE intercept is assumed to be normally distributed as follows. It is the inclusion of this RE term, that makes this a special kind of hierarchical model called a Random Effects Hierarchical Model. Here, is the difference between baseline log-odds and that of each group. Thus, because the RE intercept is simply variance from a fixed term, the mean of the normal distribution that is used to model it, is set as 0. [1]

Further, priors must also be chosen for population level coefficients which, in accordance with known conjugate priors, can be assumed to be a normal distribution. This may be expressed as follows where subscript refers to each value .

Under this set up, there are 4 hyperparameter’s whose value is to be determined. These are TO DO …

The R package was used to fit the model. This required defining the model using the syntax as given blow. Here, the term indicates that a random hierarchical intercept term should be included in the model and that the intercepts should grouped at the wine level.

A weakly informative prior ) was chosen to model population parameters .

# **Data Description**

The dataset comprised several binary indicator variables that specifies multiple characteristics of wine. A mosaic plot was generated that visualizes proportion of datapoints that had each characteristic/not and were rated superior/not (Figure 1).

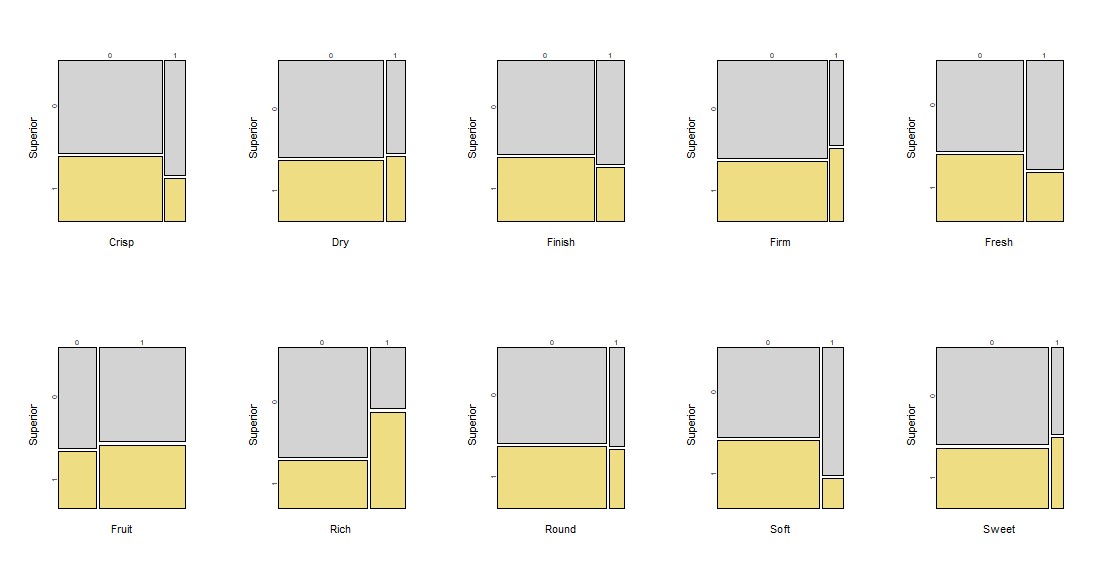


Figure . Mosaic plot of v/s wine characteristics.

Observations made were as follows.

* The most common characteristic is "Fruity" but that doesn't tell us much about the superiority of the wine as both superior and non-superior rated wines have a fruity flavor.
* Rich, although small in proportion, seem more likely to be rated highly.
* Soft wines, although small in proportion, seem less likely to be rated highly. Another seemingly significant indicator of non-superiority, is crispness of wine.

# **Analysis**

Model Output

# **Conclusions**