**WIND DATA ANALYSIS**

**Note:** Kindly find the codebase related to this project on my GitHub repository here please.

# **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 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 , , , , the interaction between , and between . 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 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.

The Random Effects (RE) intercept is special because it represents how, assuming a certain population baseline log odds of wines being rated as superior, this varies from group to group. It is the inclusion of this RE term, that makes this a special kind of hierarchical model called a Random Effects Hierarchical Model. W.r.t this dataset, represents difference between the population baseline log-odds of wine being ranked as superior and the log-odds for the same per group (wine variety). [1]

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 [2]. Thus, given covariates and an associated vector of coefficients , can be obtained as , meaning that can be used to predict the value of .

By default, the package that was used to build the model, assumes that the RE intercept follows a Student T distribution as follows wherein the mean is 0 (to reflect no prior belief about the average value of the intercept across varieties), standard deviation is 2.5 (to imply a weak prior that allows for some variability without asserting a strong influence) and the degrees of freedom is set to 3.

Generally, the bigger the degrees of freedom, the closer to a normal distribution, the t distribution is [3]. So, with low degrees of freedom here, the tails of the T distribution will be fat, meaning that more extreme values are allowed and thus a weaker prior. The package packages has set this as the default to keep the prior weakly informative while also “providing at least some regularization to considerably improve convergence and sampling efficiency” [4]. These conditions are favorable for our use case here as well. Hence, we stick to this default prior definition.

Since population level parameter coefficients are unknown, priors must be chosen for them as well, which, in accordance with prior conjugacy, can be assumed to be a normal distribution. This may be expressed as follows where subscript refers to each parameter with index value . Due to a lack of domain knowledge regarding wines and the absence of an acquaintance who is a domain expert, here the choice was made to assume a non-informative prior such that mean is and standard deviation is .

The function that comes with the R package was used to define the model as follows using the syntax. 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.

fit **<-** brm**(**

superior\_rating **~** 1 **+** price\_log10 **+** tannin **+** alcohol **+** Rich **+** price\_log10**\***body **+**

price\_log10**\***tfidf\_tsne\_1\_norm **+** **(**1**|**variety**)**,

family **=** bernoulli**(**logit**)**, data **=** wine\_reviews, prior **=** prior**(**normal**(**0, 10**)**, class**=**b**)**

**)**

*Note:**Please find reasons for choice of selected predictor variables in section 4 below.*

# **Data Description**

Initial data processing was done using Python. There is a

# **Analysis**

Model Output

# **Conclusions**