Wines - Final Report

Richard Hardis, Paulina Jane LoCicero, James Trawick

12/1/2019

Table of Contents

# ISYE 6414 Final Project: Regression Analysis of Wines

## 1. Abstract

Our goal is to evaluate factors that influence wine quality and price, based on a dataset collected from <https://www.kaggle.com/zynicide/wine-reviews>, which was scraped from WineEnthusiast in June of 2017, then again in November 2017 (<https://www.winemag.com/?s=&drink_type=wine>). We augmented this dataset with several other geographical and textual predictors to help further explore the drivers of wine price and quality. Winery latitude, longitude, and elevation were gathered from the the mapquest Open APIs. Temperature and Precipitation data for countries was gathered from the WorldBank’s open climate API. The foundational winemag.com data set contains information for 130,000 different wines, some of the factors include: country, description, and designation among others. Length of review was also calculated for each wine to use as a potentially significant predictor. Due to API limitations, we are randomly sampling 2000 wines from the original 130,000 to use for data cleaning and then build our model. With this data, we hope to answer the following questions: Under which conditions does higher wine rating correspond with higher reviews? Specifically, with which predicting variables included in a model does that model show rating being a statistically significant predictor of price? How accurate is a multiple linear regression model built off of the winemag.com 2017 dataset at predicting the prices of the top 20 wines of 2019 according to totalwine.com? Do some regions produce wines with higher quality (how does region influence points awarded?). We intend to run an initial simple linear regression, a multiple linear regression, an ANOVA test, and will evaluate other models which may fit the data well based on our initial findings. Finally, our team will also explore robust regression methodologies (lasso, elastic net, etc…) in order to identify important attributes and reduce overfitting in for our predictive models.

## 2. Background/Introduction

### Study Motivations

We want to know more about wine…

### Study Expectations

Wine price and quality are related and can be explained better with the help of other variables

## 3.1 Data

### Raw Data and Collection

Our raw data started from a root dataset of wines taken from <https://www.kaggle.com/zynicide/wine-reviews> as part of their machine learning competition. This dataset contains wine data for over 130,000 distinct wines and was scraped from the search results pages of winemag.com in June and November of 2017. The original dataset contained country, description, designation, points (quality score), price, province (state in the US or province in Europe), region\_1, region\_2 (subregion of the province or region\_1), taster\_name, taster\_twitter\_handle, title, variety, and winery. This dataset had many data points, but did not have many useful predictors for regression analysis, so we chose to scrape further webpages to gather more factors. Two datasets were developed with additional factors. The first large dataset has predictors abv (alcohol by volume), bottle\_size, category (red, white, rose), and importer scraped from the individual wine page on winemag.com. This large dataset has 73,257 individual wines. The second, smaller dataset contains all of the features of the large dataset and also includes the year produced, average country temperature, standard deviation of country temperature, average country precipitation, standard deviation of country precipitation, lattitude, longitude, elevation, and review\_length (word count). There are 1123 unique wines in this smaller dataset. Winery latitude, longitude, and elevation were gathered from the the mapquest Open APIs. Temperature and Precipitation data for countries was gathered from the WorldBank’s open climate API. The large quantity of data requires detailed exploration to ensure its quality and suitability for regression analysis.

### Data Exploration

The first step in the data exploration is cleaning. This involved removing any wines that did not have data for price, quality, year,temperature, precipitation, lattitude, longitude, elevation, review length, abv, bottle size, or importer. Because of the large quantity of data available, this approach of removing wines with missing data is preferable to imputing using averages or other means. Next, data exploration of the data can commence.

## 3.2 Modeling Analyses Methods

### Model Approach 1: Multiple Linear Regression

MLR with variable selection and metric analyses (r^2, adj. r^2, assumptions, etc.)

### Model Approach 2: Poisson Regression

Poisson model with variable selection and metric analyses (statistical significance of the model and predictive power, GOF, assumptions, etc.)

In this section, we attempt to fit a Poisson Logistic Regression model. We wish to see if the points awarded per bottle of wine can be modeled as a rate, and we test different predicting variables for different possible models.

First, we create a full model, which includes the following predicting variables:

#### Statistical Significance and Predictive Power

We compare the fitted model to the null model using a chi-square test. Our p-value for this test is **FILL HERE**. This high p-value indicates that our model is not significant overall, and thus has poor predictive power. We still check goodness of fit to test if the Poisson Regression is appropriate for modeling the response variable, points.

#### Goodness of Fit

We take a look at the Deviance and Pearson Residuals. We find the following.

CONFIRM VALUES WITH FINAL DATASET

Deviance Residual is approximately **1**.

Pearson Residual is approximately **1**.

Furthermore, we check the dispersion parameter, and find it to be **FILL HERE**.

High p-values for the Goodness of Fit Tests, and a dispersion parameter of less than 2, suggests that we do not reject the null hypothesis that the Poisson Regression Model is a good fit for this data. The dispersion parameter suggests that we do not see more variance with this data than we would expect to see if it were distributed with a Poisson rate.

Thus, while our model may not have very good predictive power, the Poisson Regression may be a good model fit. We will check assumptions after we select our final model.

#### Variable Selection

To help narrow down our final model, we will use three methods for variable selection: Forward Stepwise, Lasso, and Elastic Net. We compare each model with the full model we began with in the previous section.

**Forward Stepwise Regression for Variable Selection**

The regression selects only price as a predicting variable.

**Lasso Regression for Variable Selection**

**Elastic Net Regression for Variable Selection**

**Comparing and Selecting Models**

#### Checking Assumptions for Selected Model

## 4. Results

## 5. Conclusions

### Implications

What did we find?

### Further Questions

What’s next?