The Role of Residual Sugar in Wine Quality

W203 Team November: Cindy Xun, Jing An, Yiying Liu, Beijing Wu

December, 2021

Contents

Introduction and Research Question

As a consultant for a local Napa winery and in hopes of understanding what makes a bottle of wine great, we want to evaluate the role of sweetness in expert rated wine quality. A recent study indicated that after surveying 40,000 American households, it was found that sales of sweet wine increased by 40.1% during the COVID-19 pandemic. Generally, dry wines have almost no residual sugar (0 - 4 grams per liter, g/L) and sweet wines have the most residual sugar (> 120 g/L), as shown in **Figure 1**. Based on a wine consumer report of 2018, more consumers prefer sweet or semi-sweet wine than dry wine with no sugar.

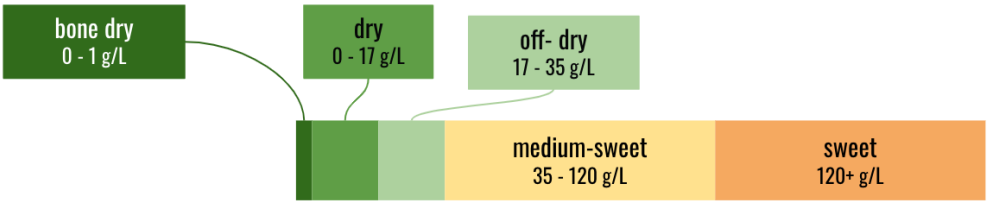

Figure 1: Residual Sugar Conent in Wine

In the past, however, it was not uncommon for sweet wines to be perceived as cheap or poor quality. With time, this misconception has been cleared by many winemakers and experts. And customers no longer believe that sweet wines are unsophisticated due to their highly sweet or unbalanced flavor. Today, many young consumers have been shown to prefer sweeter or non-traditional wines, and this demographic may help explain the growth in sales. With changes in perceptions and market preferences, it is therefore critical for the industry to adapt, innovate and produce wines suited to consumer preferences that they enjoy.

By better understanding how sweetness affects the quality perception of wine experts, wineries may be able to tailor their wines better based on popular customer tastes and increase sales. Our project, therefore, aims to study the difference in the level of residual sugar, the main chemical component that adds sweetness, and its effects on wine quality. Residual sugar is the amount of sugar from natural grapes that's leftover in a bottle of wine after the fermentation process is completed. During the fermentation process, sugar is converted to alcohol with the help of yeast, and winemakers can choose at which point to terminate the fermentation process before all the sugar gets consumed. This decision could affect both the levels of residual sugar and the alcohol concentration.

Therefore, our research question is

> *Does sweeter non-sparkling wine lead to better quality review by wine experts?*

We hypothesize the effect of the causal relationship between levels of residual sugar and wine quality would be that sweeter wine will lead to a better quality review. Based on industry research on sales and consumer preferences, we believe that as residual sugar concentration increases in wine, it will receive better quality ratings from wine experts. Although there are a lot of factors that could lead to the success of a type of wine, our research focuses specifically on residual sugar level because this is a chemical property that can be tightly controlled by winemakers. We specifically chose Likert scale rating on wine quality as the final response variable because as a consumer product, wines with higher quality ratings are more likely to lead to more sales or a better reputation. Therefore, these two variables specifically operationalize a controllable wine feature and potential wine success that could help us find if there are any causality present. Finally, we would hope to apply anything we learned from this project to help advance the sales and reputation for Napa wineries.

Data and Theoretical Causal Model

Our Dataset

Our research leverages data from the UC Irvine Machine Learning Repository (http://www3.dsi.uminho.pt/pcortez/wine/), which contains 4,898 records of exported white and 1,599 records of red wines from the *Vinho Verde* region located in northern Portugal. The dataset includes data for common physicochemical properties recorded by machines and also sensory assessment evaluated by humans. The wine quality assessments were recorded on a Likert scale of 0 to 10 with 0 being very bad and 10 being excellent. Since the assessment process was different for white and red wine, recorded data were kept separately in two datasets. Both the physiochemical properties and the sensory assessment were conducted during the wine certification process, which is a common process for export goods in order to help countries stratify wines into different categories for ease of pricing.

Revision of Research Question - Why White Wine Only?

We choose to not combine the white and red wine datasets, and evaluate only the white wine dataset, for two reasons detailed below.

First, there exists an inherent taste difference between red and white wine samples. Furthermore, within our dataset, white and red wines' physicochemical signals and qualities were collected in separate processes. Therefore, differences in the quality metric or calibration of machines could give contradictory results for red and white wines, and we felt it would be inconsistent to combine the two datasets and analyze them together.

Secondly, we find that the range in residual sugar in the red wine dataset is quite limited. With the maximum residual sugar concentration for the red wine dataset being 15.5 g/L, these red wines can at most be categorized as dry wines based on **Figure 1**. Therefore, this dataset is not suited for us to extrapolate any of our analysis into the sweet or medium sweet red wine categories because such red wine samples do not exist in our dataset.

As such, we decide to examine causal relationships between levels of residual sugar in non-sparkling white wine and wine quality, and specifically, we tailored our research question to the below:

> *Does sweeter non-sparkling white wines lead to better quality review by wine experts?*

Literature Review and Variables

Our white wine dataset (hereinafter referred to as our dataset) contains eleven (11) physicochemical parameters. **Table 1** presents the description of each physicochemical parameter, along with its descriptive statistics.

Theoretical Causal Theory

Based on our literature review, we have created a causal pathway below that details the relationship of all of our 11 physicochemical signals. The direction of effect between two variables is presented by the directional arrows of the arrow and each two signals' relationship is denoted by either an increasing effect correlation (+) or a decreasing effect (-).

In the absence of oxygen, through the process of fermentation, yeast converts the natural sugar in grapes into alcohol and carbon dioxide (CO_2). The more sugar there is in the grapes, the higher the potential alcohol level of wine (i.e., `alcohol`) if the yeast are allowed to carry out fermentation to completion. And

Table 1: Physicochemical Parameters for White Wines

Variable	Measure	Unit	Description	Range	Stdev
fixed acidity	Tartaric acid conc.	g/L	One of the three main acids found in wine grapes. Plays prominent role in maintaining chemical stability of the wine, wine's color, and influencing the taste of the finished product.	3.8 - 14.2	0.844
volatile acidity	Acetic acid conc.	g/L	Associated with "acetification" of wine and is responsible for the sour taste of vinegar.	0.08 - 1.10	0.101
citric acid	Citric acid conc.	g/L	Add "freshness" and flavor to wine.	0.00 - 1.66	0.121
residual sugar	Residual sugar conc.	g/L	From natural grape sugars leftover in wine after the alcoholic fermentation finishes. 10 g/L ~ 1% sweetness.	0.6 - 65.8	5.072
chlorides	Sodium chloride conc.	g/L	Reflects local soil and water conditions (i.e., distance between vineyard and coast)	0.009 - 0.346	0.022
free sulfur dioxide	Free SO2 conc.	mg/L	antioxidant and antimicrobial, making it an effective preservative for wine. Amount depends on pH. Higher pH, less free SO2.	2 - 289	17.007
total sulfur dioxide	Total SO2 conc.	mg/L	Free SO2 and those Abound to other chemicals in wine.	9 - 440	42.498
density	density	g/L	Determined by concentration of alcohol, sugar, glycerol, and other dissolved solids.	0.9871 - 1.0390	0.003
pH	pH	NA	A measure of acidity. Lower pH indicates higher acid levels.	2.720 - 3.820	0.151
sulphates	Potassium sulphate conc.	g/L	Acts as an antioxidant, removing oxygen suspended in the wine, which slows down aging.	0.22 - 1.08	0.114
alcohol	% vol. alcohol	%	Alcohol content in wine only changes during fermentation.	8.0 - 14.2	1.231

this, in turn, means that the less residual sugar is left in wine. In other words, the earlier the winemakers stop fermentation, the more residual sugar will be left, and hence the sweeter the wine is.

Opposite to the relationship between fermentation duration and `residual.sugar.conc`, usually the longer that fermentation goes on, the more sugar is converted into alcohol, resulting in higher alcohol content in wine. Therefore, there exists a 2-way relationship between `residual.sugar.conc` and `alcohol` in wine, that is, the higher the alcohol content, the lower the `residual.sugar.conc`. Since the goal of our project is to examine the causal relationship between `residual.sugar.conc` and `quality`, we think that `alcohol` needs to be excluded from our models as it will mask the effect of `residual.sugar.conc` on `quality`.

In addition, both alcohol content and level of residual sugar affect wine density. As alcohol has a lower density than water, the higher the alcohol content, the lower the wine density. The higher the `residual.sugar.conc`, the higher the wine `density`. Because of this correlation, `density` is excluded from our model building process.

During the process of fermentation, acetic acid, which is considered the main component of volatile acidity in wine, is produced as a byproduct through oxidation, and wines that have long fermentation periods generally accumulate higher levels of acetic acid. Excess acetic acid can contribute to the wine fault, giving the wine a "nail polish remover" smell and vinegar taste. Therefore, high `acetic.acid.conc` is considered to have a negative effect on wine quality.

Citric acid is another acid that has many uses in wine production. Citric acid is often added to the finished wines to give a "fresh" flavor, and hence has a positive effect on wine quality. However, citric-sugar co-

metabolism can increase the formation of volatile acid (i.e., acetic acid) in wine, which can affect the wine aroma and wine quality negatively. We hypothesize that there should be an optimal `citric.acid.conc` in wine, where it will add freshness to wine without giving a vinegar taste.

Sulfur Dioxide is a common chemical compound used in winemaking as a preservative and an antibacterial agent. `total.sulfur.dioxide.conc` is the portion of sulfur dioxide that is free in the wine (i.e., `free.sulfur.dioxide.conc`), plus the portion that is bound to other chemicals in the wine. `free.sulfur.dioxide.conc` and `pH` of the wine determines how much sulfur dioxide is available in the active form to help protect the wine from oxidation and spoilage. Sulfur dioxide's antimicrobial and antioxidant efficiency in wine depends on a wide range of factors, of which the most important ones are `pH`, `residual.sugar.conc`, and `alcohol`. The lower the pH, the less sulfur dioxide needed; the higher the alcohol content, the less sulfur dioxide; the lower the level of residual sugar, the less sulfur dioxide needed. Because sulfur dioxide also affects wine quality, residual sugar is a cofounder in this relationship. To simplify our model building process, sulfur dioxide is left out.

`pH` is a method of determining the strength of acid in a solution. The effect of `pH` on wine quality can be explained by a combination of all acids. `pH` is not independent from acid concentrations, and if we include pH in our model, it will mask other effects of various acids on quality. Therefore, we decided to exclude `pH` in our model building process.

Wine grapes from cooler climates have a higher level of `tartaric.acid.conc`, typically 6 g/L and higher. Wine grapes from warmer climates generally have a lower level of `tartaric.acid.conc`, around 1 - 2 g/L. Therefore, we transform the variable `tartaric.acid.conc`, which is a continuous measure of tartaric acid concentration, to a categorical control variable `climates` to depict the climate the wine grapes are from. `climates` is set to be equal to `cooler` if `tartaric.acid.conc` is greater or equal to 6 g/L and `warmer` if it is less than 6 g/L. Similarly, the `sodium.chloride.conc` is highly correlated with the soil and the water condition of certain geographical regions. We have added `sodium.chloride.conc` as a control variable as well into our hypothesized models.

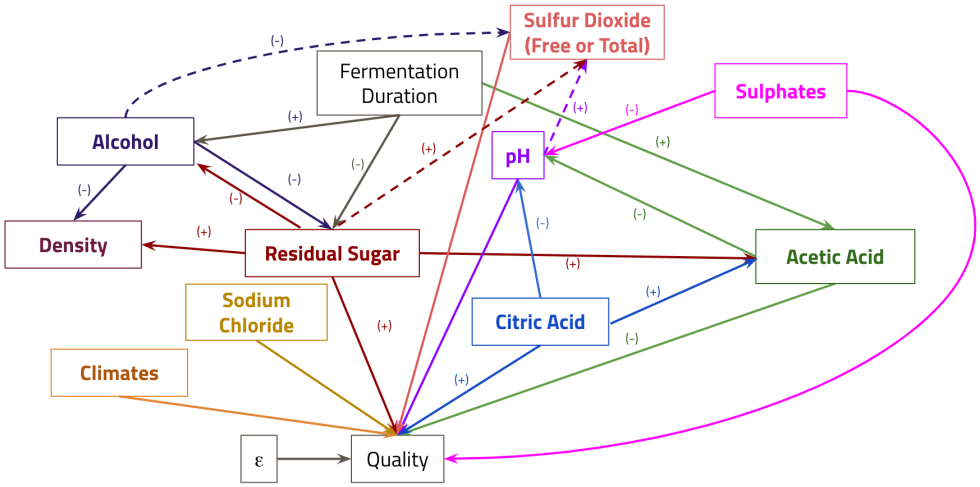

Figure 2: Theoretical Causal Theory

The error term, epsilon, is everything else that could affect wine quality and that is independent of any of the other variables in our model. This could include factors such as oak and barrel aging process, the harvest of the grapes, capping method, wine storage, and transportation. We also assume that there are no factors in the error term that directly affect both the outcome variable and the explanatory variables.

Rejection Criteria

To interpret the coefficient estimate between physicochemical signals and the wine quality, our rejection criteria will be set at 0.05. If the significance level is less than 0.05, we will reject the null hypothesis and know that the true effect size between one variable and the outcome variable is not zero. Otherwise, we don't have sufficient evidence to show that the coefficient estimate is not zero.

Research Design

Our research design involves dividing the whole dataset into a training and a testing set. The training set will be used for descriptive analysis in order to evaluate each physicochemical signal and identify if there are any transformations we need to perform. We have evaluated how each physicochemical signal interacts with each other in order to identify control variables. Then, we can leverage the causal theory to identify a one-equation structural model and assess if there are any violations for our causal graph (**Figure 2**). There could exist an optimal region for various physicochemical signals to achieve the highest quality. Therefore, we will also look at adding quadratic terms to our model. In the end, we will use linear regression to model the relationship between `residual.sugar.conc` and `quality`.

After we have finished all exploratory research using the training dataset, we will use the testing dataset to evaluate the causality theory we have hypothesized and evaluate its statistical and practical significance.

The Model Building Process

Train vs. Test

We separated the entire white wine dataset of 4,898 samples into 30%, or 1,469 records to be used for training purposes and 70%, or 3,429 records to be used for testing purposes.

Correlation & Distribution

We conducted a complete correlation analysis for all physicochemical signals for white wines (**Figure 3**) and did not find any two features with perfect collinearity. The highest correlation we observed was 0.839, which is between `density` and `residual.sugar.conc`. However, as we described in our theoretical causal theory, `density` will not be included in our model building process, and hence this high correlation will not affect our interpretation of the models. All variables that are potentially be used in the model building process do not have high correlation.

Transformation

Because we have over 100 samples for both our training or testing datasets, we can use the large sample model to evaluate coefficient estimates. The two assumptions that need to be met are IID samples and unique BLP exists. Therefore, whether each variable follows a normal distribution is not a necessary condition for us to leverage the large sample model.

The variable `tartaric.acid.conc` is transformed into a categorical variable `climates`, where `tartaric.acid.conc` is greater or equal to 6 g/L, climate is set to be `cooler`. When `tartaric.acid.conc` is less than 6 g/L, climate is set to be `warmer`. This variable `climates` is used as a control variable in our models to take into account the condition in which the wine grapes come from.

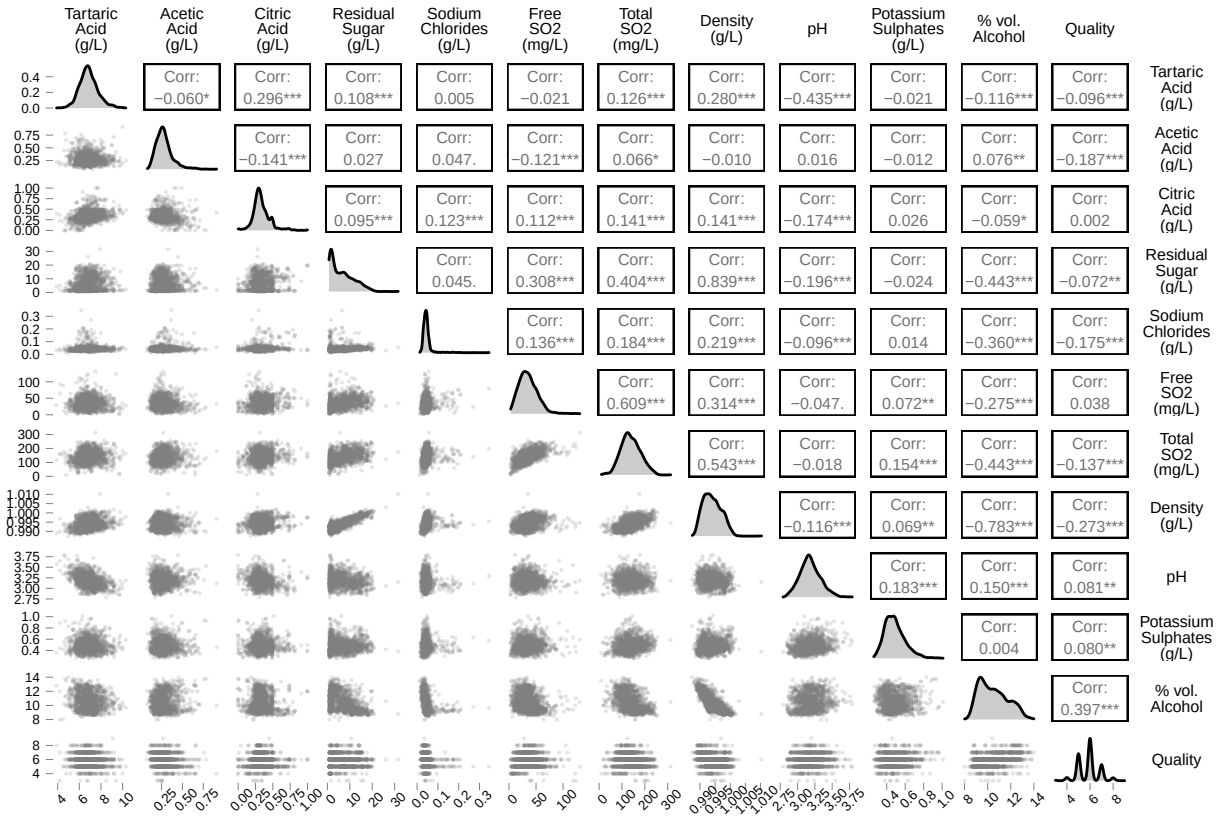

Figure 3: Correlogram of Variables

Models

We have implemented 7 models in incremental stages to identify the relationship among different variables and quality ratings. Each subsequent model is built upon the previous one in order to potentially show the consistent effect of the coefficient estimate of residual sugar on wine quality.

model_base and model_base_quad

`model_base` and `model_base_quad` use variables `residual.sugar.conc` and `quality`. In these two models, we explore the relationship between residual sugar and quality. We would like to understand how `residual.sugar.conc` affects `quality`, or how the sweetness can influence the wine quality. As the most basic and simplified model, we would like to have a short model that contains `residual.sugar.conc` as the only variable. In addition, we wondered about the possibility whether there exists an optimal range of `residual.sugar.conc` that received the highest quality. Therefore, we added a quadratic term of `residual.sugar.conc` to our linear regression model (i.e., `model_base_quad`).

```
model_base <- lm(quality ~ residual.sugar.conc, data = train)
model_base_quad <- lm(quality ~ residual.sugar.conc
                      + I(residual.sugar.conc^2), data = train)

stargazer(
   model_base,
```

```
  model_base_quad,
  type = "text", header = FALSE,
  star.cutoffs = c(0.05, 0.01, 0.001),
  font.size = "small",
  align = TRUE,
  no.space = TRUE
)
```

```
##
## ========================================================================
##                                      Dependent variable:
##                          ---------------------------------------------
##                                            quality
##                                (1)                      (2)
## ------------------------------------------------------------------------
## residual.sugar.conc           -0.012**                  0.007
##                               (0.004)                  (0.014)
## I(residual.sugar.conc2)                                 -0.001
##                                                         (0.001)
## Constant                      5.944***                 5.897***
##                               (0.037)                  (0.050)
## ------------------------------------------------------------------------
## Observations                   1,469                    1,469
## R2                             0.005                    0.007
## Adjusted R2                    0.004                    0.005
## Residual Std. Error       0.867 (df = 1467)        0.867 (df = 1466)
## F Statistic             7.608** (df = 1; 1467) 4.798** (df = 2; 1466)
## ========================================================================
## Note:                                      *p<0.05; **p<0.01; ***p<0.001
```

From `model_base`, we saw that the coefficient estimate for residual sugar and quality has a negative correlation (coefficient estimate -0.012) and achieved statistical significance ($p < 0.01$). However, both the linear and quadratic coefficients became insignificant in `model_base_quad`, indicating that the behavior of residual sugar on quality cannot be confidently modeled by a quadratic term. Hence, we remove the quadratic `residual.sugar.conc` term out of our subsequent models.

The residual plot of `model_base` (**Figure 4, left**) shows quite a few data points on both sides of 0 that have large residuals, indicating that although our base model is accurate in explaining the relationship between residual sugar and wine quality, it is pretty off at times. Normal Q-Q plot (**Figure 4, right**) shows the residuals for `model_base` are normally distributed.

model_control_citric and model_control_citric_quad

`model_control_citric` and `model_control_citric_quad` use variables `residual.sugar.conc`, `citric.acid.conc`, and `quality`. Besides the sweetness of a wine, the sourness also affects its quality. Therefore, we included citric acid as a feature since it is often derived from fruits and gives wine a fruity and fresh flavor. However, increasing `citric.acid.conc` will also increase the density of diacetyl. At concentrations greater than 5 mg/L, diacetyl can be overpowering, resulting in distinct buttery or nutty flavor. We believe that there could be an optimal concentration for citric acid, where when it is in excess, it could affect wine quality negatively.

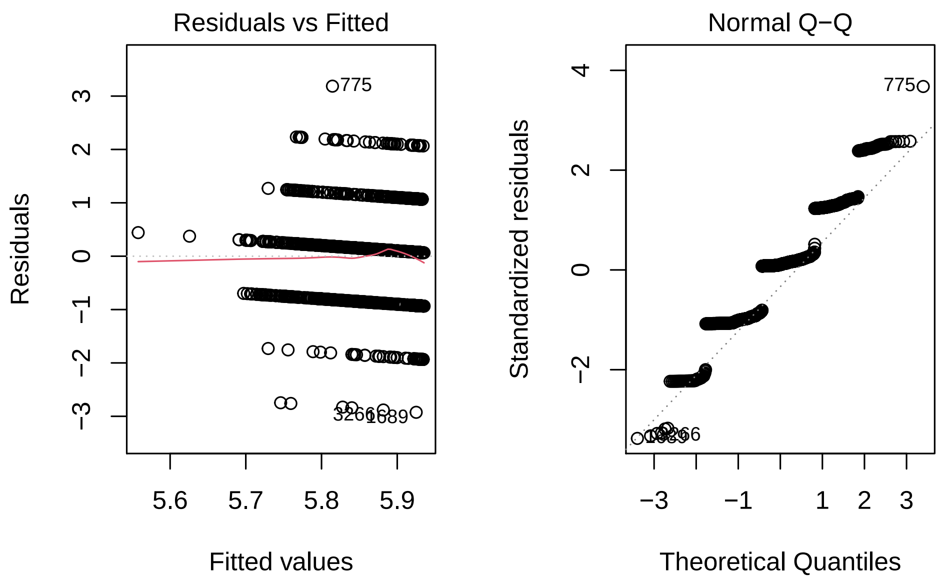

Figure 4: Residual Plots for Base Model

```
model_control_citric <- lm(quality ~ residual.sugar.conc
                    + citric.acid.conc,
                    data = train)
model_control_citric_quad <- lm(quality ~ residual.sugar.conc
                    + citric.acid.conc
                    + I(citric.acid.conc^2),
                    data = train)

stargazer(
   model_base,
   model_control_citric,
   model_control_citric_quad,
   type = "text", header = FALSE,
   star.cutoffs = c(0.05, 0.01, 0.001),
   font.size = "small",
   align = TRUE,
   no.space = TRUE
)
```

```
##
## ===========================================================================================
##                                               Dependent variable:
##                          ------------------------------------------------------------------
##                                                    quality
##                                  (1)                  (2)                   (3)
## -------------------------------------------------------------------------------------------
## residual.sugar.conc           -0.012**             -0.012**              -0.011*
##                               (0.004)              (0.004)               (0.004)
```

```
## citric.acid.conc                                  0.066                3.347***
##                                                  (0.185)              (0.603)
## I(citric.acid.conc2)                                                  -4.087***
##                                                                       (0.716)
## Constant                      5.944***            5.923***             5.337***
##                               (0.037)             (0.070)              (0.123)
## ------------------------------------------------------------------------------------------
## Observations                   1,469               1,469                1,469
## R2                             0.005               0.005                0.027
## Adjusted R2                    0.004               0.004                0.025
## Residual Std. Error    0.867 (df = 1467)   0.867 (df = 1466)    0.858 (df = 1465)
## F Statistic          7.608** (df = 1; 1467) 3.865* (df = 2; 1466) 13.505*** (df = 3; 1465)
## ==========================================================================================
## Note:                                                      *p<0.05; **p<0.01; ***p<0.001
```

`model_control_citric` shows that the effect of citric acid on wine quality is insignificant, but when we include the quadratic term of `citric.acid.conc`, the coefficient estimates become significant. The coefficient for the linear term is positive and the coefficient for the quadratic form is negative, both with p value less than 0.001. This result agrees with our research that citric acid has an optimal concentration in wine that adding just the right amount could give the wine a fresh flavor, but too much would affect wine quality negatively. In addition, the coefficient estimate for `residual.sugar.conc` in `model_control_citric_quad` only decreases by 0.001 compared to our `model_base`, showing consistent relationship between `residual.sugar.conc` and `quality`.

model_control_climate

`model_control_climate` uses variables `residual.sugar.conc`, `citric.acid.conc`, `climates`, and `quality`.

```
model_control_climate <- lm(quality ~ residual.sugar.conc
                    + citric.acid.conc
                    + I(citric.acid.conc^2)
                    + factor(climates),
                    data = train)
stargazer(
   model_base,
   model_control_climate,
   type = "text", header = FALSE,
   star.cutoffs = c(0.05, 0.01, 0.001),
   font.size = "small",
   align = TRUE,
   no.space = TRUE
)
```

```
##
## ========================================================================
##                                      Dependent variable:
##                          ---------------------------------------------
##                                             quality
##                                (1)                       (2)
## ------------------------------------------------------------------------
## residual.sugar.conc           -0.012**                  -0.010*
##                               (0.004)                   (0.004)
```

```
## citric.acid.conc                                        3.698***
##                                                         (0.616)
## I(citric.acid.conc2)                                   -4.390***
##                                                         (0.723)
## factor(climates)warmer                                  0.190**
##                                                         (0.071)
## Constant                          5.944***              5.226***
##                                   (0.037)               (0.130)
## ------------------------------------------------------------------------
## Observations                       1,469                 1,469
## R2                                 0.005                 0.032
## Adjusted R2                        0.004                 0.029
## Residual Std. Error      0.867 (df = 1467)       0.856 (df = 1464)
## F Statistic            7.608** (df = 1; 1467) 11.963*** (df = 4; 1464)
## ========================================================================
## Note:                                      *p<0.05; **p<0.01; ***p<0.001
```

As discussed in previous sections of this report, we transform the continuous variable `tartaric.acid.conc` in wine to a categorical variable `climates` in order to take into account the climates in which the wine grapes are grown. When we include `climates` as a control variable in our `model_control_climate`, the coefficient estimate for warmer climate is 0.190 with p value less than 0.01, suggesting that wine grapes grown in warmer climates affect wine quality positively.

model_control_NaCl

`model_control_NaCl` uses variables `residual.sugar.conc`, `citric.acid.conc`, `climates`, `sodium.chloride.conc`, and `quality`.

```
model_control_NaCl <- lm(quality ~ residual.sugar.conc
                     + citric.acid.conc
                     + I(citric.acid.conc^2)
                     + factor(climates)
                     + sodium.chloride.conc,
                     data = train)
stargazer(
   model_base,
   model_control_NaCl,
   type = "text", header = FALSE,
   star.cutoffs = c(0.05, 0.01, 0.001),
   font.size = "small",
   align = TRUE,
   no.space = TRUE
)
```

```
##
## ========================================================================
##                                      Dependent variable:
##                          ---------------------------------------------
##                                             quality
##                                (1)                       (2)
## ------------------------------------------------------------------------
## residual.sugar.conc           -0.012**                  -0.009*
##                               (0.004)                   (0.004)
```

```
## citric.acid.conc                                    3.476***
##                                                     (0.609)
## I(citric.acid.conc2)                               -3.974***
##                                                     (0.717)
## factor(climates)warmer                              0.150*
##                                                     (0.070)
## sodium.chloride.conc                               -5.962***
##                                                     (0.959)
## Constant                        5.944***            5.524***
##                                 (0.037)             (0.137)
## ------------------------------------------------------------------------
## Observations                     1,469               1,469
## R2                               0.005               0.057
## Adjusted R2                      0.004               0.053
## Residual Std. Error      0.867 (df = 1467)   0.845 (df = 1463)
## F Statistic           7.608** (df = 1; 1467) 17.545*** (df = 5; 1463)
## ========================================================================
## Note:                                  *p<0.05; **p<0.01; ***p<0.001
```

Besides climate, we further include the `sodium.chloride.conc` as a control variable due its correlation with the soil and water condition of different geographic locations. While the coefficient estimate for residual sugar on quality is consistent, the coefficient for climate changed slightly. The coefficient estimate changes from -0.012 to -0.009, representing more than 20% decrease. This tells us that since both `climates` and `sodium.chloride.conc` paint a similar picture of where the grapes are growing from, there is most likely a covariate relationship between these two variables.

model_full

`model_full` uses variables `residual.sugar.conc`, `citric.acid.conc`, `climates`, `sodium.chloride.conc`, `acetic.acid.conc`, and `quality`.

```
model_full <- lm(quality ~ residual.sugar.conc
                    + citric.acid.conc
                    + I(citric.acid.conc^2)
                    + factor(climates)
                    + sodium.chloride.conc
                    + acetic.acid.conc,
                    data = train)
stargazer(
   model_base,
   model_full,
   type = "text", header = FALSE,
   star.cutoffs = c(0.05, 0.01, 0.001),
   font.size = "small",
   align = TRUE,
   no.space = TRUE
)
```

```
##
## ========================================================================
##                                      Dependent variable:
##                          ---------------------------------------------
##                                            quality
```

```
##                                 (1)                    (2)
## ------------------------------------------------------------------------
## residual.sugar.conc           -0.012**                -0.009*
##                               (0.004)                 (0.004)
## citric.acid.conc                                      2.316***
##                                                       (0.637)
## I(citric.acid.conc2)                                 -2.728***
##                                                       (0.743)
## factor(climates)warmer                                0.147*
##                                                       (0.070)
## sodium.chloride.conc                                 -5.741***
##                                                       (0.950)
## acetic.acid.conc                                     -1.298***
##                                                       (0.231)
## Constant                      5.944***                6.103***
##                               (0.037)                 (0.170)
## ------------------------------------------------------------------------
## Observations                   1,469                   1,469
## R2                             0.005                   0.077
## Adjusted R2                    0.004                   0.073
## Residual Std. Error      0.867 (df = 1467)       0.837 (df = 1462)
## F Statistic            7.608** (df = 1; 1467) 20.197*** (df = 6; 1462)
## ========================================================================
## Note:                                      *p<0.05; **p<0.01; ***p<0.001
```

For our last exploratory model, we include an additional variable, `acetic.acid.conc`, that could affect wine quality. Because acetic acid is the main compound found in vinegar, having too much of acetic acid will give the wine a vinegar taste. However, acetic acid has a complicated relationship with both residual sugar and citric acid. Residual sugar and citric acid co-metabolism can produce acetic acid. With any chemical reaction, there could be an equilibrium and the direction of effect could be reversed as well. Upon the addition of acetic acid, even though the coefficient estimate for residual sugar on quality stays consistent, the coefficient estimate for other acids have been altered.

Results

Test Dataset Coefficient Estimate Validation

After exploring our dataset using the 30% testing data through a total of 7 models, we decide to choose the `model_control_NaCl` model as our final model.

$$\text{quality} = \beta_0 + \beta_1(\text{residual sugar}) + \beta_2(\text{citric acid}) + \beta_3(\text{citric acid}^2) + \beta_4(\text{climates}) + \beta_5(\text{NaCl}) + \epsilon$$

The reason we chose this model and not the `model_full` was because we are not entirely confident with the directional effect between acetic acid and residual sugar. Like every chemical reaction, there could exist co-equilibrium. Though `model_full` gives us a better R^2, the coefficient estimate for residual sugar on quality is the same at -0.009 with $p < 0.05$.

```
test_model_control_NaCl <- lm(quality ~ residual.sugar.conc
                              + citric.acid.conc
                              + I(citric.acid.conc^2)
                              + factor(climates)
                              + sodium.chloride.conc,
```

```
                              data = test)
stargazer(
  test_model_control_NaCl,
  type = "text", header = FALSE,
  star.cutoffs = c(0.05, 0.01, 0.001),
  font.size = "small",
  align = TRUE,
  no.space = TRUE
)
```

```
##
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                   quality
## ------------------------------------------------
## residual.sugar.conc              -0.014***
##                                   (0.003)
## citric.acid.conc                 2.405***
##                                   (0.337)
## I(citric.acid.conc2)             -2.630***
##                                   (0.373)
## factor(climates)warmer           0.182***
##                                   (0.048)
## sodium.chloride.conc             -8.816***
##                                   (0.704)
## Constant                         5.883***
##                                   (0.081)
## ------------------------------------------------
## Observations                       3,429
## R2                                 0.074
## Adjusted R2                        0.073
## Residual Std. Error          0.860 (df = 3423)
## F Statistic              55.085*** (df = 5; 3423)
## ================================================
## Note:                  *p<0.05; **p<0.01; ***p<0.001
```

All of the variables we added to the right hand side of the linear regression model are all significant with a significance level below 0.001. This means that we can reject the null hypothesis and conclude that these estimates are all nonzero. Specifically, the effect of residual sugar on quality is at -0.0014. This means with every unit (g/L) increase in residual sugar, the wine quality rating will decrease by 0.014.

Similarly, we have found a negative correlation between the quadratic citric acid term and quality. We looked at this relationship in conjunction with the positive relationship between citric acid and quality. These indicate that some of the behavior in quality rating can be explained by an optimal range in citric acid. The climate categorical variable has a positive correlation with quality and shows that a wine from warmer climate will lead to an increase of 0.182 in quality rating. The coefficient estimate of negative 8.816 for sodium chloride concentration on quality seems large in comparison with other coefficient estimates. But this is due to the low sodium chloride concentration presented in our dataset, from 0.009 g/L to 0.346 g/L.

Statistical vs Practical Significance

From using our test dataset, we obtained the coefficient estimate of the effect of residual sugar on quality to be -0.014 for the base model with a p-value less than 0.001. This means that the null hypothesis that the

coefficient estimate is zero can be rejected. The base model indicates that for every unit increase in g/L of residual sugar, there will be an approximate decrease of 0.014 in expert rated wine quality.

It is also important to evaluate the practical significance of our results. Given that with a large sample size of over 100 samples in both the training and testing dataset, our coefficient estimate is 0.014, which is one hundredth of a unit in Likert scale. Based on our linear regression model, a wine with no residual sugar will receive a 5.883 grade of rating by wine experts. With any unit increase of g/L for `residual.sugar.conc`, our model tells us that we will observe a -0.014 decrease in rating. Based on the typical range of residual sugar in wine that ranges from 0 g/L to 120 g/L, when winemakers increase the leftover residual sugar concentration from the minimum of 0 g/L to 120 g/L, the wine quality effect due to residual sugar would be a decrease of 1.68 in ratings. The drastic changes in the manufacturing process to achieve a 120 g/L to 0 g/L reduction in residual sugar will at most lead to a 1.68 degree of increase in wine quality rating. This leads us to conclude that although our model was able to establish a significant causal relationship between residual sugar and wine quality, the effect size is really small, hence no practical significance can be evaluated.

Furthermore, we understand that it is easier to achieve statistical significant results given our large sample and small standard error. Therefore, the reason that our coefficient estimate is significant could simply be due to having a large sample size.

Discussion

Structural Limitations

Our model contains several limitations, the first being that the data for quality is on a Likert scale, which is not a continuous variable. It has a limited range and is discrete and ordinal. This may somewhat limit the insights we can gather from our dataset. In addition, we do not have any insight into the actual expert review process. Whether it was the same or different experts who reviewed all of the wines is unknown to us. In the case that different experts reviewed different wines, it is possible that due to their unique opinions and preferences, the same wine may receive different ratings from different reviewers. With any Likert scale variable, we are not certain that the difference in quality is the same between two different scores. Therefore, this poses a limitation in how confident we are at the quality information we can extract from our dataset.

With many physicochemical variables present in our dataset, our team believes that there are still omitted variables of our causal model that could potentially affect our analysis and interpretation. We have identified fermentation duration as one such omitted variable. As mentioned above, residual sugar concentration can be controlled by the fermentation duration. Increasing fermentation duration would result in less residual sugar and greater percentage alcohol. On the contrary, decreasing fermentation time would result in more residual sugar and a lower percentage alcohol. This could in turn, also influence the flavor and quality of the wine along with other characteristics. Since we don't have fermentation duration information present in our dataset, we thought that we could use acetic acid as an intermediate proxy to hypothesize the effect of fermentation duration on quality.

To find the omitted variable bias, we need the direction of correlation between 1) fermentation duration and `quality` ($\hat{\beta}_1$) and between 2) fermentation duration and `residual.sugar.conc` (σ_1). For σ_1, we know that there is a negative correlation between fermentation duration and `residual.sugar.conc`. For β_k, we think the correlation would also be negative because with longer fermentation duration, more acetic acid will be produced. Hence, wine quality will suffer due to vinegar taste. Therefore, according to the omitted bias equation

$$\hat{\beta}_1 - \beta_1 = \sigma_1 * \beta_k$$

we found that our omitted variable bias would be away from zero.

In addition, we have identified additional control variables that are missing from our causal theory. The first two are grape types and wineries, both are categorical variables and could affect the quality of wine. Some grapes tend to provide more complexity in their flavors than others and some wineries may be more

experienced at producing high quality wines than other wineries. The last variable is the age of the barrel. Studies have shown that new barrels tend to impart a lot of its flavors onto wines while more aged barrels tend to provide neutral flavors.

Statistical limitations

Unfortunately our samples do not satisfy the requirement to be independently and identically distributed variables. The wines evaluated were from a 4 year period between 2004 and 2007, all from the same region of Portugal, and of the same wine type. This makes us believe that a clustering effect could exist within our dataset. The climate, soil and winemaking process for all of the wines may therefore be similar. Therefore, the various wines reviewed may only represent that particular climate, soil and wine type and do not represent wines in general in diverse regions and types, and our model does not satisfy the IID assumption. Trying to design a model that satisfies the IID assumption would likely need to utilize data from a variety of regions, climates and wine types.

Conclusion

Despite our proposed causal theory that sweeter non-sparkling wines lead to better quality review by wine experts, analysis of our dataset showed the opposite. In fact, `residual.sugar.conc` has a negative and significant correlation (β_1 = -0.019, $p < 0.01$) effect on wine `quality` as reviewed by wine experts when we fitted our testing dataset against the `model_control_NaCl`. Our results disprove the theory that simply adding sugar will make the wine quality score higher. There could exist additional complex relationships among many of the variables that we have not found. Thus, increasing the residual sugar may inadvertently affect other variables in the winemaking process, and therefore negatively affect the quality.

The combination of a statistical significant result and a small effect size could indicate that the quality of wine may be driven by a myriad of other multiple characteristics and the sweetness of white wine may not be a significantly important factor. In our residuals vs fitted plot (**Figure 4, left**), values appear to be in multiple parallel lines, some of which are far from 0, indicating that there are other variables that have an effect, and that the model could not explain the relationship fully. Additionally, our model has the adjusted R^2 value of 0.073. This indicates that only 7.3% of the wine quality variation can be attributed to the model. As this value is relatively small, it does not adequately explain the relationship.

It was also determined that different types of wine have different intended levels or ranges of residual sugar. This means that various types have been evaluated over many years and a certain recommended balance of the variables involved in the winemaking have been achieved with a recommended amount of residual sugar. It is possible that many consumers prefer types of wine that are sweeter in general, but as our study involved only white wine, this does not mean that only increasing the residual sugar in that type of wine will increase the quality. Furthermore, since our dataset was collected more than a decade ago and changes in consumer preferences could have happened more recently. Therefore, we could conduct follow up analysis on looking at the causal relationship between residual sugar and quality for more recent wines produced from U.S. wineries.

Therefore, our recommendation is for winemakers to refrain from simply increasing residual sugar in the hopes of making the wine receive a higher quality rating, but rather engage in research on their specific type of wine, and a large number of variables that may affect the flavor to achieve a superior and more desirable product. Increasing the residual sugar alone, is not recommended to achieve a more desirable and higher quality wine.

Reference

1. A snapshot of the American wine consumer in 2018. (2018, December 10). Retrieved December 8, 2021, from https://www.winebusiness.com/news/?go=getArticle&dataId=207060.

2. Application note #105 – chloride in wine by titration. (2017, July). Retrieved December 11, 2021, from https://mantech-inc.com/wp-content/uploads/2014/07/105-Chloride-in-Wine-by-Titration.pdf.

3. Barrel fermentation versus barrel aging – what's the difference?: Hawk Haven Vineyard & Winery. Hawk Haven Vineyard & Winery Experience Handcrafted Wines from Cape May County. (2020, March 11). Retrieved December 11, 2021, from https://hawkhavenvineyard.com/root-stock-fridays/.

4. Can you explain the terms "residual sugar," "Brix," "total acidity" and "ph" as they relate to wine? Wine of the Month Club. (n.d.). Retrieved December 8, 2021, from https://www.wineofthemonthclub.com/category/RS-brix-acidity-pH.

5. CASERTANO, R. O. N. (2017, March 13). Ph & Wine Quality. Consumer Fresh Winemakers. Retrieved December 10, 2021, from https://www.cfpwinemakers.com/blog/post/?pH-Wine-Quality-30.

6. Cortez, P., Cerdeira Antonio, Almeida, F., Matos, T., & Reis Jose. (2009, May 22). Modeling wine preferences by data mining from ... - uminho. Retrieved December 8, 2021, from http://www3.dsi.uminho.pt/pcortez/wine5.pdf.

7. Dr. Peter Costello, Magali Deleris-Bou, Dr. Richard Descenzo, Dr. Nichola Hall, Dr. Sibylle Krieger, Prof. Dr. Aline Lonvaud-Funel Piet Loubser, Jose Maria Heras, Shirley Molinari, Dr. Rich Morenzoni, Anthony Silvano, Gordon Specht, Francine Vidal, Dr. Caroline Wilde, MALOLACTIC FERMENTATION- IMPORTANCE OF WINE LACTIC ACID BACTERIA IN WINEMAKING from https://www.lallemandwine.com/wp-content/uploads/2015/10/Lallemand-Malolactic-Fermentation.pdf.

8. Hakim, S. (2018, April 3). Citric acid. Viticulture and Enology. Retrieved December 11, 2021, from https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid#:~:text=Citric%20acid%20is%20often%20added,give%20a%20%E2%80%9Cfresh%E2%80%9D%20flavor.&text=Since%20bacteria%20use%20citric%20acid,the%20growth%20of%20unwanted%20microbes.

9. Hale, N. (2021, August 25). What is acidity in wine? Wine Enthusiast. Retrieved December 11, 2021, from https://www.winemag.com/2019/06/19/what-is-acidity-in-wine/.

10. Liz Thach, M. W. (2021, May 20). Why wineries should consider making more sweet wine, according to new E&J Gallo Study. Forbes. Retrieved December 8, 2021, from https://www.forbes.com/sites/lizthach/2021/05/20/why-wineries-should-consider-making-more-sweet-wine-according-to-new-ej-gallo-study/?sh=512303679839.

11. McIntyre, D. (2018, January 11). There's nothing wrong with being a little sweet on wine. The Washington Post. Retrieved December 8, 2021, from https://www.washingtonpost.com/lifestyle/food/theres-nothing-wrong-with-being-a-little-sweet-on-wine/2018/01/11/4a4ffc92-f70a-11e7-a9e3-ab18ce41436a_story.html.

12. Miller, M. (n.d.). Monitoring acids and ph in winemaking - gencowinemakers.com. Retrieved December 10, 2021, from https://www.gencowinemakers.com/docs/Acids%20Presentation.pdf.

13. Monaco, S. M. D., Curilen, Y., CarmenMaturano, R. D., Bravo, S. M. E., & Adriana Beatriz Simesand Adriana Catalina Caballero. (2016, October 19). The use of indigenous yeast to develop high-quality Patagonian wines. IntechOpen. Retrieved December 8, 2021, from https://www.intechopen.com/chapters/52025.

14. Moroney, M. (2018, February 27). Total sulfur dioxide – why it matters, too Midwest Grape and Wine Industry Institute. Retrieved December 11, 2021, from https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/#:~:text=Simply%20put%2C%20Total%20Sulfur%20Dioxide, aldehydes%2C%20pigments%2C%20or%20sugars.

15. Mowery, L. (2021, March 16). Sugar high: A quick guide to sweet red wine. Wine Enthusiast. Retrieved December 8, 2021, from https://www.winemag.com/2020/08/27/good-sweet-red-wine-beginners/.

16. Puckette, M. (2020, February 20). Weird wine flavors and the science behind them. Wine Folly. Retrieved December 11, 2021, from https://winefolly.com/deep-dive/weird-wine-flavors-and-the-science-behind-them/.

17. Puckette, M. (2019, September 20). What is residual sugar in wine? Wine Folly. Retrieved December 11, 2021, from https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/.

18. Revisiting sulfur dioxide use. The Australian Wine Research Institute. (2011, June 13). Retrieved December 11, 2021, from https://www.awri.com.au/industry_support/winemaking_resources/fining-stabilities/microbiological/avoidance/sulfur_dioxide/.

19. Sturm, J. (2020, April 21). Preserving your wine with potassium metabisulfite. Preserving your Wine with Potassium Metabisulfite. Retrieved December 11, 2021, from https://www.vynova-group.com/blog/preserving-wine-with-potassium-metabisulfate.

20. Sylvia Wu July 16, & Wu, S. (2020, July 16). What is residual sugar in wine? – ask decanter. Decanter. Retrieved December 8, 2021, from https://www.decanter.com/learn/residual-sugar-46007/.

21. Understanding tartrates crystals in wine and the effects of cold stabilization. Stonestreet Alexander Mountain Estate. (n.d.). Retrieved December 11, 2021, from https://www.stonestreetwines.com/tartrates-crystals.

22. WEATHERWAX, J. O. S. H. U. A. (n.d.). Wine alcohol content: How much alcohol is in wine? Retrieved December 11, 2021, from https://home.binwise.com/blog/wine-alcohol-content.

23. Why is testing for acetic acid important in winemaking? Randox Food. (2019, May 21). Retrieved December 11, 2021, from https://www.randoxfood.com/why-is-testing-for-acetic-acid-important-in-winemaking/#:~:text=Acetic%20acid%20is%20a%20two,small%20amount%20of%20acetic%20acid.

24. Why is testing for tartaric acid important in wine making? Randox Food. (2020, January 22). Retrieved December 8, 2021, from https://www.randoxfood.com/why-is-testing-for-tartaric-acid-important-in-wine-making/.

25. Wikimedia Foundation. (2021, November 9). Diacetyl. Wikipedia. Retrieved December 8, 2021, from https://en.wikipedia.org/wiki/Diacetyl.

26. Williams, S. (2020, April 27). How much sugar does wine contain, and how is it controlled during the winemaking process? Wiens Family Cellars. Retrieved December 11, 2021, from https://www.wienscellars.com/wine-101-blog/how-much-sugar-does-wine-contain-and-how-is-it-controlled-during-the-winemaking-process.

27. Wine analysis. ETS Laboratories. (n.d.). Retrieved December 11, 2021, from https://www.etslabs.com/analyses/DEN.