Wine taste and quality prediction by GLM

tom

2017-12-11

Introduction

Ubiquitos era is blurring out physical barrier. People from different countries consume different goods from all over the world. the Wine market is no exception of this phenomenon. Deviating the Europe continent, the wine products are getting consumed by more and more areas in the world.

Meanwhile, the wine production areas is currently enlarging. Not only old wine regions in Europe, but also new wine regions such as China, America continents, and Africa yield large amount of wine products. As the quality and taste of the wine heavily relies on the climate and environment of the region, the diversity of wines grow rapidly.

In this situation, traditional role of oenologist reach its limit. To match a wine product to the exact world market, we need another way of wine analyzing. GLM can be a way to conduct that task. By analyzing physicochemical ingredients of the wine which people in a region prefer the best, we can distribute wine products in very effecient way. In order to do that, we can conduct regression model analyzing to figure out the portion of the components of wine products.

This project aims at analyzing the relationship between the taste preference of traditional market and physicochemical ingredients of wine produced in traditonal region. This method can be used further to analyze the relationships between new wine market and new wine regions. As the result, we can easily match the demand and consumption in a easier way.

Data Discription

2. Understand data. Describe how the data is collected and what the data is. -> data description

The wine from Portugal, vinho verde is the subject of this project which is a unique product from the Minho (northwest) region of Portugal. The data were collected from May/2004 to February/2007 with random sample method. The data was collected in two ways: physicochemical ingredients data and sensory data. Official certification entity (CVRVV), collected basic physicochemical statistics of red and white wine. For the sensory data, three sensory assessors evaluted each sample in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations.

Original dataset contains red wine data and White wine data. In this project, we will consider only red wine. In the dataset, there are 11 variables of physicochemical statistics and 1 variable from sensory score, quality.

There are 1599 samples.

Attribute (units) Red wine	Min	Max	Mean	Median
Fixed acidity (g(tartaric acid)/dm3)	4.6	15.9	8.3	7.9
Volatile acidity (g(acetic acid)/dm3)	0.1	1.6	0.5	0.52
Citric acid (g/dm3)	0.0	1.0	0.3	0.26
Residual sugar (g/dm3)	0.9	15.5	2.5	2.2
Chlorides (g(sodium chloride)/dm3)	0.01	0.61	0.08	0.079
Free sulfur dioxide (mg/dm3)	1	72	14	14
Total sulfur dioxide (mg/dm3)	6	289	46	38
Density (g/cm3)	0.990	1.004	0.996	0.99
pH	2.7	4.0	3.3	3.31
Sulphates (g(potassium sulphate)/dm3)	0.3	2.0	0.7	0.62

Attribute (units) Red wine	Min	Max	Mean	Median
Alcohol (vol.%)	8.4	14.9	10.4	10.2
quality	3	8	5.6	6

The histogram of quality seems to follow normal distribution.

3. Exploratory Analysis

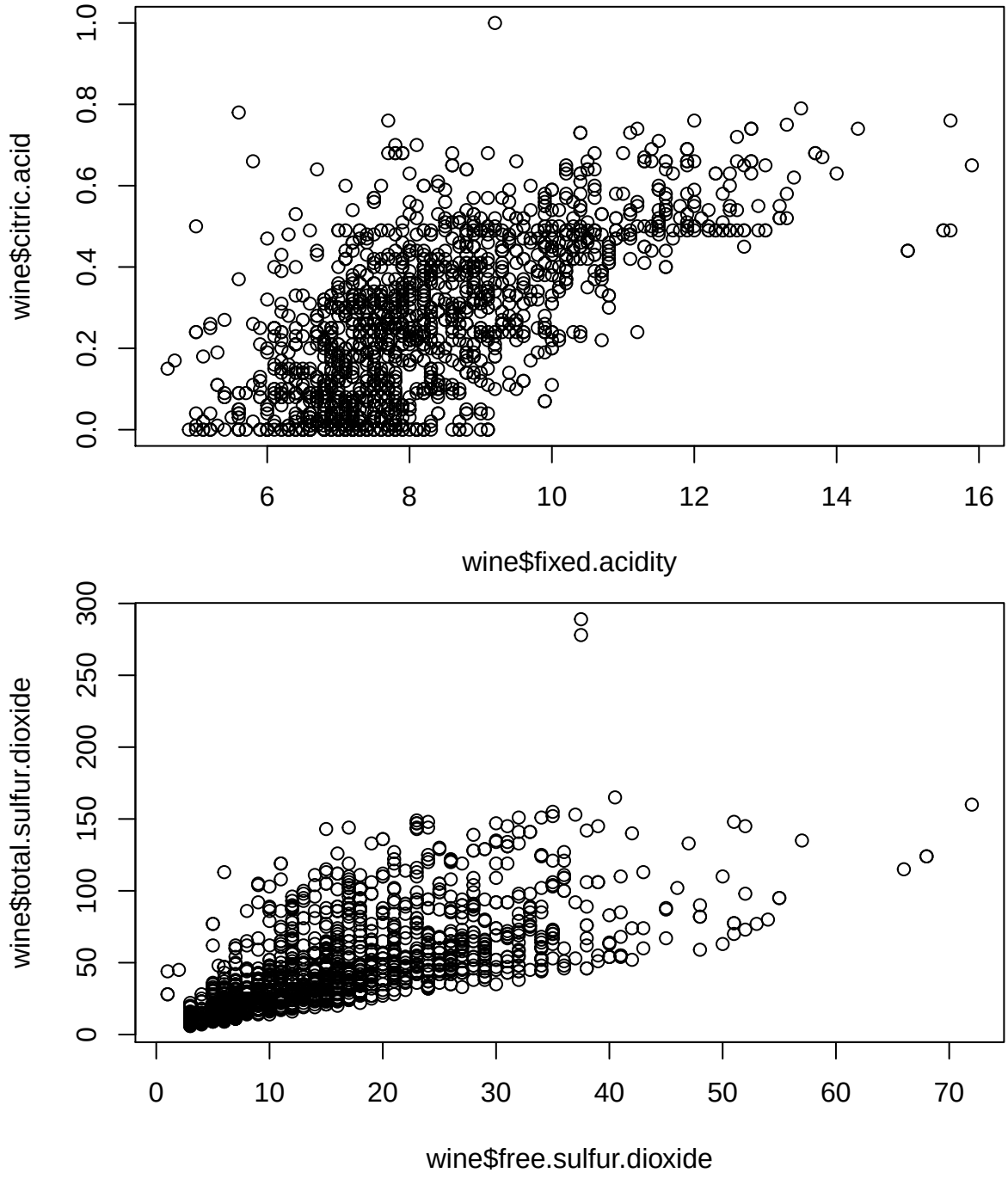
3. Use summary statistics / exploratory data analysis to understand the data better or to identify any obvious patterns. -> exploratory data analysis

Observing correlation matrix, there exists weak relationship between two predictor, alchole 0.47 and volatile.acidity -0.39 for quality. As the correlation was not strong, we can see that the quality of the wine is not dependent on any single covariate.

```
str(wine)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
```

```
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```




```
## [1] 0.6717034
```

In this stage, there are many correlation between covariates found.

Variables	Correlation coefficient
citric.acid, fixed.acidity	0.67
total.sulfur.dioxide, free.sulfur.dioxide	0.67
density, fixed acidity	0.67
pH, fixed acidity	-0.68

Four pairs of covariates has high correlations. However, since all of the kinds of acidity hugely affects the tastes of the wines, we do not delete any of acid covariates. It is obvious that density affects taste, and pH is affected by the level of acid. so we do not delete density nor pH.

Dioxide is used for presevation of wine. free sulfur dioxide is to prevents microbial growth and the oxidation of wine. Total sulfur dioxide accounts for amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine. As total sulfur dioxide is more affective to the taste, we delete free.sulfur.dioxide from the dataset.

4. Confirmatory data analysis

We conduct Generalized Linear Model and linear regression model. We first conduct Poisson model and see which variables are adequate for the modeling. Then we try other models: logit model and linear regression model.

- is -> seems to
- check the model's validity by using other models, as we did in assignment2

5. Interpret the results. -> results/discussions
6. Conclude. -> conclusion
7. Now finish up intro and finally summary/abstract.

LRT test -> cook's distance

two sample T test