红酒质量分析报告

回归分析小组

2024 年 1 月 4 日

目录

1 问题背景

随着酒类产品在市场上的广泛受欢迎，对于酒的质量和特征的深入了解变得至关重要。为了更好地了解酒的品质，我们进行了一项回归分析，重点关注了酒的酸度，二氧化硫（SO_2）含量等特征。这些特征在很大程度上影响了酒的口感、风味和保存能力。

2 数据说明

我们通过 UC Irvine仓库下载了红酒数据，数据集中的特征包含酒的各类化学成分指标以及酒的品质评价。其中，成分指标，如 pH 值、SO_2 含量、残糖量等通过物理化学检测得出；酒的品质由专业品酒师做出评价。（每个样本由三个品酒师做出评价，每个人的评分为 0（差）到 10（好）的一个整数，最终评价取三人的中位数）。红酒共 1599 条数据，其中不含缺失值，共记录了 12 个特征，特征说明如 1 所示。

表 1: 酒的特征说明

变量名	中文含义	变量类型	单位
fixed acidity	固定酸度	连续型变量	g/L
Volatile Acidity	挥发性酸度	连续型变量	g/L
Citric Acid	柠檬酸	连续型变量	g/L
Residual Sugar	残糖	连续型变量	g/L
Chlorides	氯化物	连续型变量	g/L
Free Sulfur Dioxide	游离二氧化硫	连续型变量	g/L
Total Sulfur Dioxide	总二氧化硫	连续型变量	g/L
Density	密度	连续型变量	g/mL
pH	葡萄酒的 pH 值	连续型变量	
Sulphates	硫酸盐	连续型变量	g/L
Alcohol	醇度	连续型变量	%
Quality	酒品	离散型变量	

3 描述性统计

3.1 数值特征

对数据的初步描述如表 2，表3和表4 所示，包含平均值、最小值、最大值、中位数。

表 2: 数据描述 1-4

fixed.acidity	volatile.acidity	citric.acid	residual.sugar
Min. : 4.60	Min. :0.1200	Min. :0.000	Min. : 0.900
1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.090	1st Qu.: 1.900
Median : 7.90	Median :0.5200	Median :0.260	Median : 2.200
Mean : 8.32	Mean :0.5278	Mean :0.271	Mean : 2.539
3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.420	3rd Qu.: 2.600
Max. :15.90	Max. :1.5800	Max. :1.000	Max. :15.500

表 3: 数据描述 5-8

chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
Min. :0.01200	Min. : 1.00	Min. : 6.00	Min. :0.9901
1st Qu.:0.07000	1st Qu.: 7.00	1st Qu.: 22.00	1st Qu.:0.9956
Median :0.07900	Median :14.00	Median : 38.00	Median :0.9968
Mean :0.08747	Mean :15.87	Mean : 46.47	Mean :0.9967
3rd Qu.:0.09000	3rd Qu.:21.00	3rd Qu.: 62.00	3rd Qu.:0.9978
Max. :0.61100	Max. :72.00	Max. :289.00	Max. :1.0037

表 4: 数据描述 9-12

pH	sulphates	alcohol	quality
Min. :2.740	Min. :0.3300	Min. : 8.40	Min. :3.000
1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000
Median :3.310	Median :0.6200	Median :10.20	Median :6.000
Mean :3.311	Mean :0.6581	Mean :10.42	Mean :5.636
3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000
Max. :4.010	Max. :2.0000	Max. :14.90	Max. :8.000

3.2 因变量描述

首先通过直方图观察酒品分布情况，可以发现，数据主要集中在 $5, 6$。

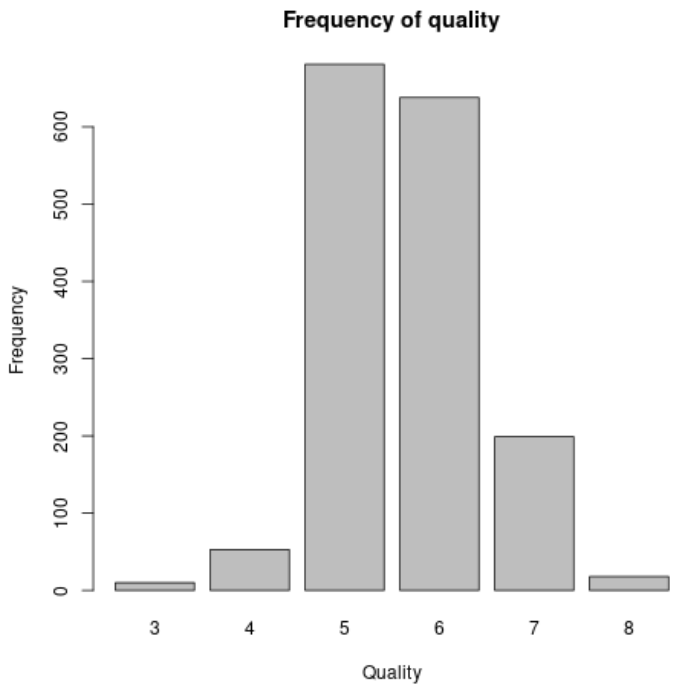

图 1: 酒品数据分布

3.3 自变量描述

我们通过自变量的分布图和自变量与酒品的箱线图对其进行描述。下面展示部分具有代表性的数据。

3.3.1 硫酸盐（Sulphates）

硫酸盐的分布图和关于酒品的箱线图如图所示。

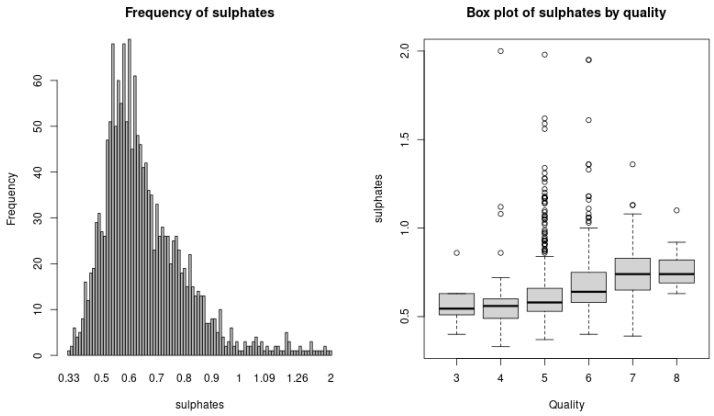

图 2: 硫酸盐分布及箱线图

3.3.2 酒精（Alcohol）

酒精含量的分布图和关于酒品的箱线图如图所示。

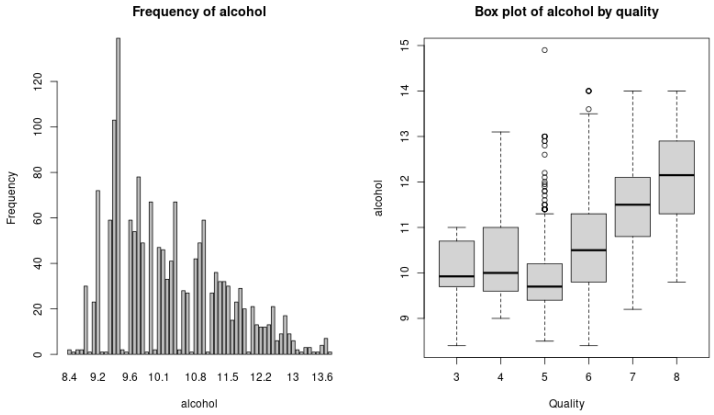

图 3: 酒精分布及箱线图

3.3.3 总二氧化硫（Total Sulfur Dioxide ）

总二氧化硫含量的分布图和关于酒品的箱线图如图所示。

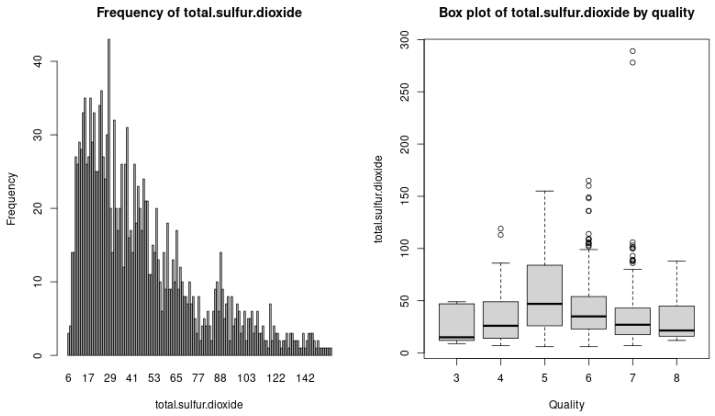

图 4: 总二氧化硫分布及箱线图

3.3.4 氯化物（Chlorides）

氯化物含量的分布图和关于酒品的箱线图如图所示。

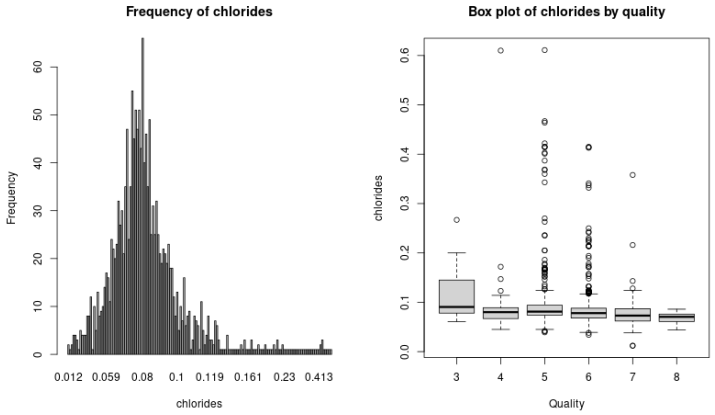

图 5: 氯化物分布及箱线图

3.4 自变量相关性分析

我们对自变量相关性进行分析，结果如图所示。可以发现，与体现酒的酸碱性的自变量相关性相对较大，需要通过后续处理消除它们的复共线性。

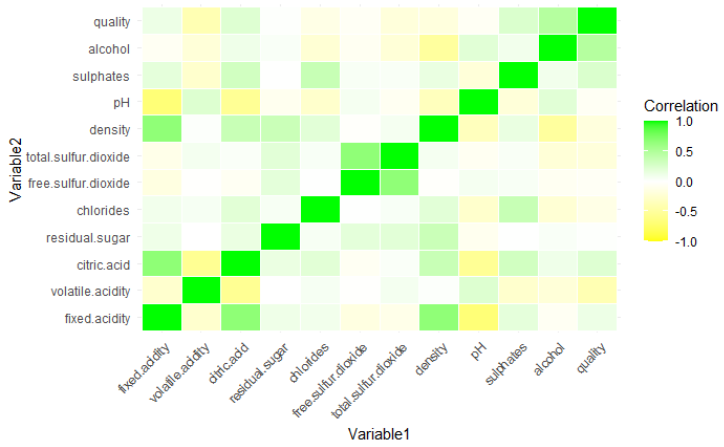

图 6: 自变量相关性热力图

4 数据建模

我们首先使用全模型进行建模，通过红酒的化学成分特征，对酒品进行回归分析。然后使用选模型对变量进行选择，最后使用 Box-cox 对 quality 进行变换。

4.1 全模型

首先，我们对数据进行最小二乘回归，得到结果如下:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.9652	21.1946	1.04	0.3002
fixed.acidity	0.0250	0.0259	0.96	0.3357
volatile.acidity	-1.0836	0.1211	-8.95	0.0000 ***
citric.acid	-0.1826	0.1472	-1.24	0.2150
residual.sugar	0.0163	0.0150	1.09	0.2765
chlorides	-1.8742	0.4193	-4.47	0.0000 ***
free.sulfur.dioxide	0.0044	0.0022	2.01	0.0447 *
total.sulfur.dioxide	-0.0033	0.0007	-4.48	0.0000 ***
density	-17.8812	21.6331	-0.83	0.4086
pH	-0.4137	0.1916	-2.16	0.0310 *
sulphates	0.9163	0.1143	8.01	0.0000 ***
alcohol	0.2762	0.0265	10.43	0.0000 ***
Residual standard error: 0.648 on 1587 degrees of freedom				
Multiple R-squared: 0.3606, Adjusted R-squared: 0.3561				
F-statistic: 81.35 on 11 and 1587 DF, p-value: $<$ 2.2e-16				

结果显示，在 0.05 的置信水平下，挥发性酸度、氯化物、总二氧化硫、硫酸盐和醇度与酒品呈正相关。

然后我们对模型进行回归诊断，结果如下图：

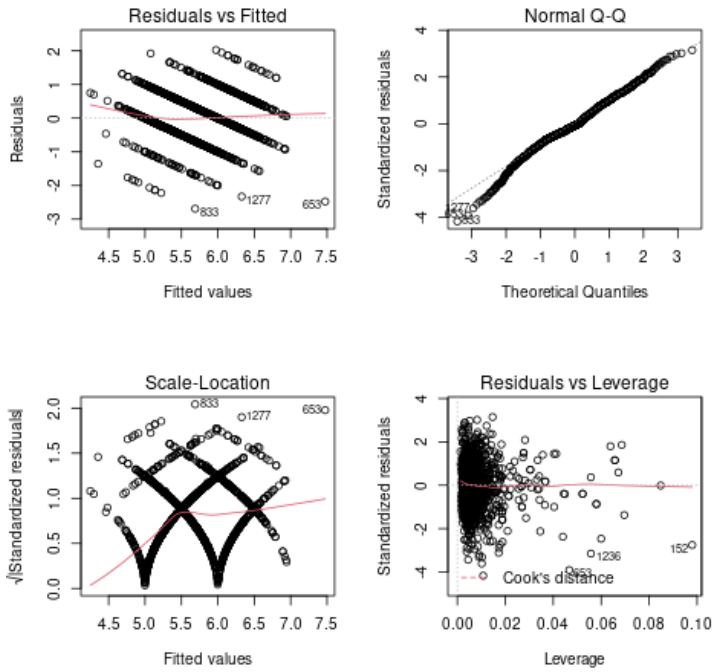

图 7: 全模型诊断图

观察残差图，可以发现残差的分布近似一个椭圆，说明残差可能不满足方差的假设条件。

4.1.1 模型检验

4.2 选模型

通过 AIC 准则和 BIC 准则选取变量。

4.3 Box-cox 变换

4.4 主成份分析

5 结论及建议