

红酒质量分析报告

回归分析小组

2024 年 1 月 5 日

目录

1	问题背景	1
2	数据说明	1
3	描述性统计	2
3.1	数值特征	2
3.2	因变量描述	4
3.3	自变量描述	4
3.3.1	固定酸度 (Fixed Acidity)	5
3.3.2	挥发性酸度 (Volatile Acidity)	5
3.3.3	游离二氧化硫 (Free Sulfur Dioxide)	6
3.3.4	醇度 (Alcohol)	6
3.4	自变量相关性分析	7
4	数据建模	7
4.1	全模型	8
4.1.1	异常值检验	10
4.1.2	异方差检验	10
4.1.3	自相关性检验	11
4.1.4	多重共线性检验	11
4.2	选模型	11
4.2.1	AIC 准则	12
4.2.2	BIC 准则	12
4.3	Box-cox 变换	12
4.4	主成份分析	13
5	结论及建议	13

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 599 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. : 6.557	Min. :0.0052	Min. :-0.1331	Min. :1.223
1st Qu.: 7.857	1st Qu.:0.3278	1st Qu.: 0.3126	1st Qu.:1.784
Median : 8.622	Median :0.4581	Median : 0.4563	Median :1.977
Mean : 8.859	Mean :0.4808	Mean : 0.4535	Mean :2.407
3rd Qu.: 9.309	3rd Qu.:0.6300	3rd Qu.: 0.5803	3rd Qu.:2.457
Max. :12.642	Max. :1.0396	Max. : 1.1742	Max. :6.780

表 3: 数据描述 5-8

chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
Min. :-0.17986	Min. : 2.872	Min. : 9.874	Min. :0.7226
1st Qu.: 0.07193	1st Qu.: 8.041	1st Qu.: 24.112	1st Qu.:0.9328
Median : 0.16598	Median :15.930	Median : 56.041	Median :0.9960
Mean : 0.21012	Mean :16.475	Mean : 63.791	Mean :1.0014
3rd Qu.: 0.35014	3rd Qu.:22.089	3rd Qu.:102.992	3rd Qu.:1.0719
Max. : 0.76331	Max. :39.160	Max. :151.197	Max. :1.2901

表 4: 数据描述 9-12

pH	sulphates	alcohol	quality
Min. :2.611	Min. :-0.09304	Min. : 8.768	Min. :4.000
1st Qu.:3.017	1st Qu.: 0.34213	1st Qu.: 9.273	1st Qu.:5.000
Median :3.134	Median : 0.50100	Median : 9.498	Median :5.000
Mean :3.129	Mean : 0.64124	Mean : 9.888	Mean :5.581
3rd Qu.:3.256	3rd Qu.: 0.72854	3rd Qu.:10.346	3rd Qu.:6.000
Max. :3.582	Max. : 2.84062	Max. :12.465	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, 7。

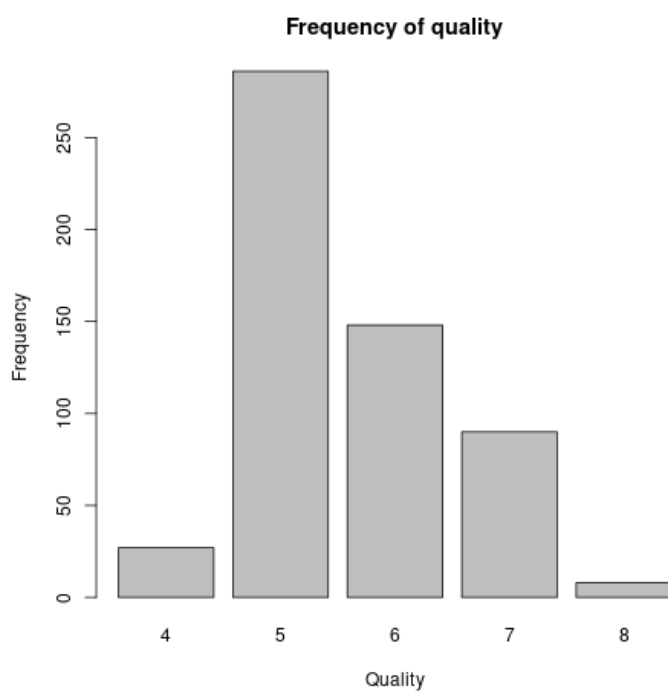


图 1: 酒品数据分布

3.3 自变量描述

我们通过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 固定酸度 (Fixed Acidity)

固定酸度关于酒品的箱线图如图所示。

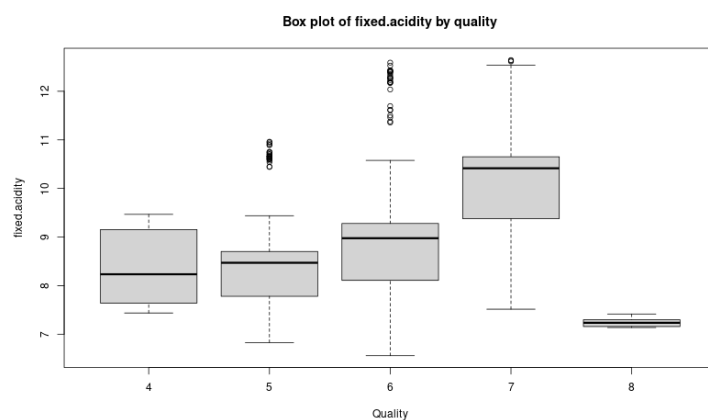


图 2: 固定酸度箱线图

3.3.2 挥发性酸度 (Volatile Acidity)

挥发性酸度关于酒品的箱线图如图所示。

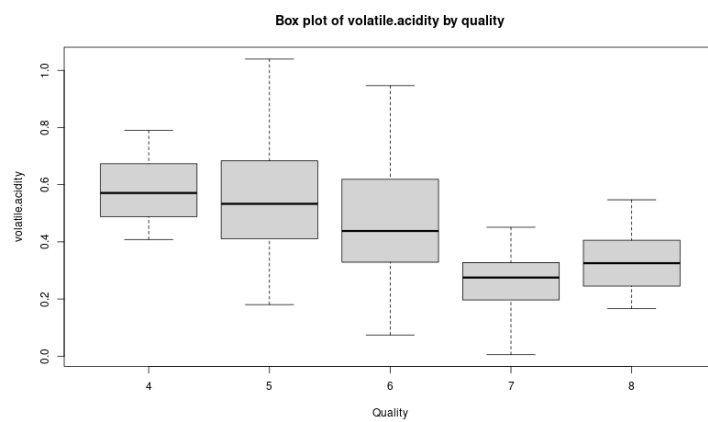


图 3: 挥发性酸度分布箱线图

3.3.3 游离二氧化硫 (Free Sulfur Dioxide)

游离二氧化硫含量关于酒品的箱线图如图所示。

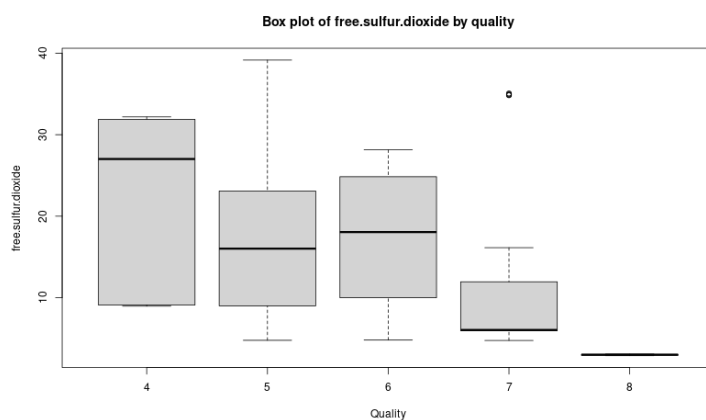


图 4: 游离二氧化硫箱线图

3.3.4 醇度 (Alcohol)

酒精含量关于酒品的箱线图如图所示。

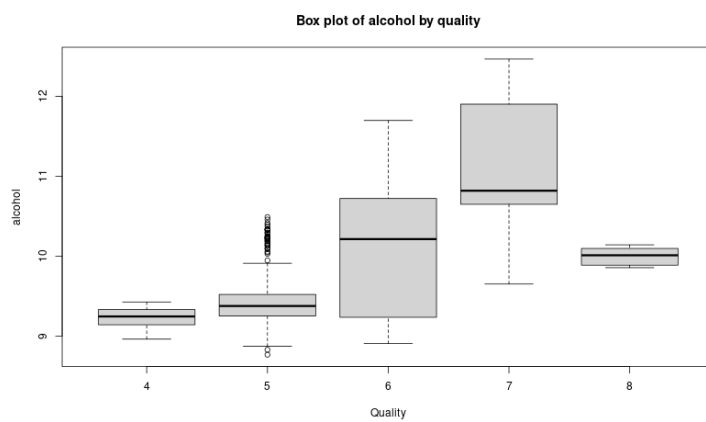


图 5: 醇度箱线图

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。

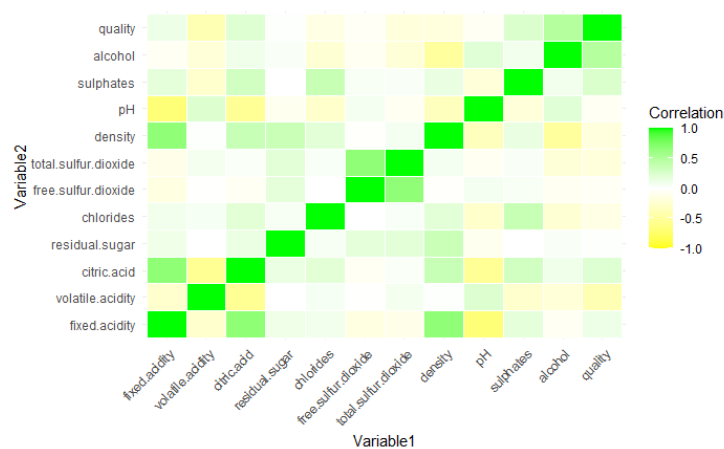


图 6: 自变量相关性热力图

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，我们对数据进行全模型最小二乘回归，得到结果如下：

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.5577	0.8651	1.80	0.0723 .
fixed.acidity	0.0771	0.0220	3.51	0.0005 ***
volatile.acidity	-1.0448	0.1477	-7.07	0.0000 ***
citric.acid	0.0886	0.1538	0.58	0.5648
residual.sugar	0.0100	0.0277	0.36	0.7174
chlorides	-0.2167	0.1891	-1.15	0.2523
free.sulfur.dioxide	-0.0181	0.0044	-4.10	0.0000 ***
total.sulfur.dioxide	-0.0008	0.0012	-0.65	0.5157
density	0.1462	0.2437	0.60	0.5488
pH	-0.2349	0.2072	-1.13	0.2574
sulphates	-0.0626	0.0550	-1.14	0.2556
alcohol	0.4856	0.0376	12.91	0.0000 ***
Residual standard error: 0.5762 on 547 degrees of freedom				
Multiple R-squared: 0.5656, Adjusted R-squared: 0.5569				
F-statistic: 64.76 on 11 and 547 DF, p-value: < 2.2e-16				

结果显示，在 0.05 的置信水平下，固定酸度、挥发性酸度、游离二氧化氮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：

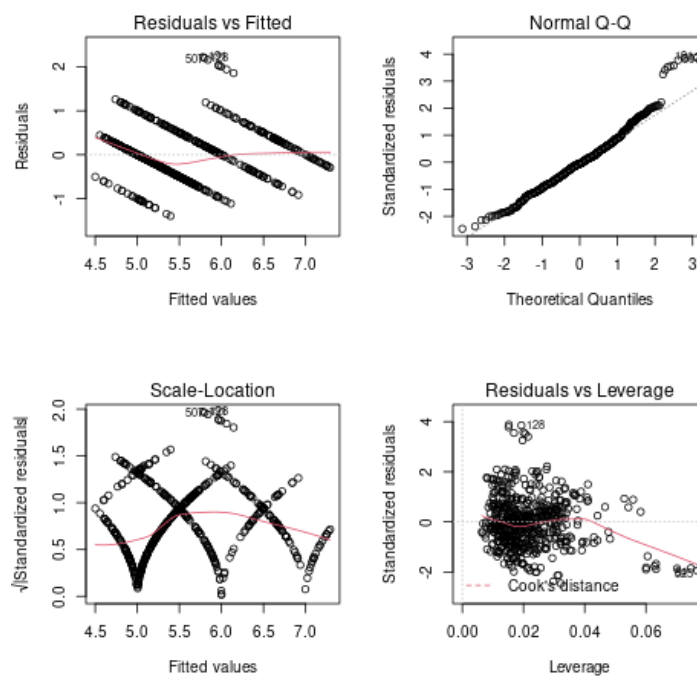


图 7: 全模型诊断图

在左上角残差图中，有一部分离群点；在右上角 Q-Q 图中，有一部分数据发生了偏离；因此有必要对数据进一步处理。

在进一步处理之前，我们先对模型进行一系列检验。

4.1.1 异常值检验

利用 Cook 距离检验数据集中的异常值点，结果如下：

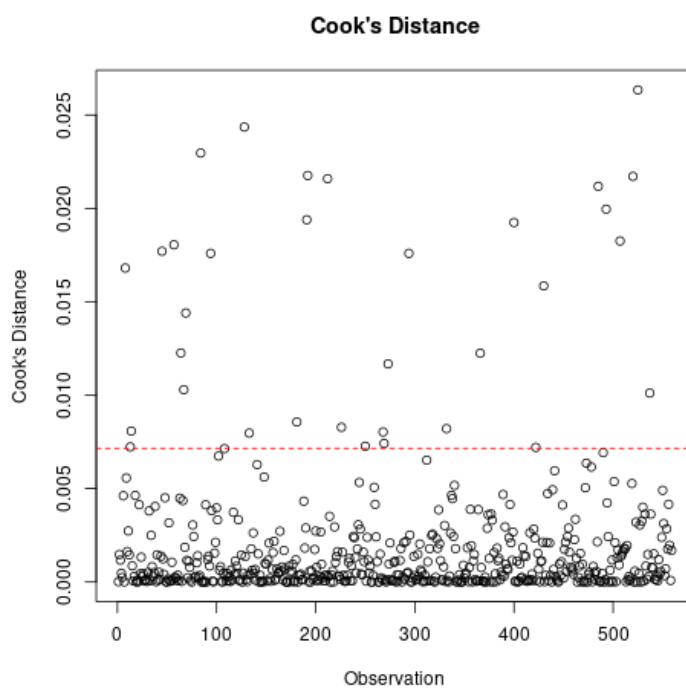


图 8: Cook 距离

可以看到，所有数据点的 Cook 距离都接近 0。

4.1.2 异方差检验

利用 `ncvTest` 检验模型异方差性，结果如下：

表 5: 异方差检验

Chisquare	Df	p
0.1033325	1	0.74787

检验得到 p 值为 0.74787，因此接受原假设，认为模型不具有异方差性。

4.1.3 自相关性检验

利用 `dwtest` 对误差自相关性进行检验，结果如下：

表 6: 自相关性检验

DW	p
2.0775	0.8176

检验得到 p 值为 0.8176，因此接受原假设，认为模型不具有自相关性。

4.1.4 多重共线性检验

利用 VIF 对自变量共线性进行检验，结果如下：

表 7: 多重共线性

	VIF
fixed.acidity	1.50
volatile.acidity	1.45
citric.acid	1.62
residual.sugar	1.75
chlorides	2.08
free.sulfur.dioxide	2.94
total.sulfur.dioxide	4.12
density	1.04
pH	2.08
sulphates	1.49
alcohol	1.62

若 VIF 小于 1，表示自变量不存在多重共线性的问题；若 VIF 在 1 到 5 之间，表示存在轻微的多重共线性问题；若 VIF 大于 5，表示存在较强的多重共线性。可以看到，总二氧化硫可能与其他变量的相关性相对较强。

4.2 选模型

为了解决多重共线性的问题，我们通过 AIC 准则和 BIC 准则选取变量。

4.2.1 AIC 准则

使用 AIC 准则选择变量，结果如下：

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.6495	0.3696	1.76	0.0794 .
fixed.acidity	0.0952	0.0190	5.01	7.46e-07 ***
volatile.acidity	-1.0721	0.1330	-8.06	4.78e-15 ***
free.sulfur.dioxide	-0.0201	0.0026	-7.66	8.48e-14 ***
alcohol	0.4991	0.0329	15.16	< 2e-16 ***
Residual standard error: 0.5743 on 554 degrees of freedom				
Multiple R-squared: 0.563, Adjusted R-squared: 0.5598				
F-statistic: 178.4 on 4 and 554 DF, p-value: < 2.2e-16				

4.2.2 BIC 准则

使用 BIC 准则选择变量，结果如下：

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.6495	0.3696	1.76	0.0794 .
fixed.acidity	0.0952	0.0190	5.01	7.46e-07 ***
volatile.acidity	-1.0721	0.1330	-8.06	4.78e-15 ***
free.sulfur.dioxide	-0.0201	0.0026	-7.66	8.48e-14 ***
alcohol	0.4991	0.0329	15.16	< 2e-16 ***
Residual standard error: 0.5743 on 554 degrees of freedom				
Multiple R-squared: 0.563, Adjusted R-squared: 0.5598				
F-statistic: 178.4 on 4 and 554 DF, p-value: < 2.2e-16				

使用 AIC 准则和 BIC 准则筛选出来的变量相同。

4.3 Box-cox 变换

为了改善数据的非正态性与异方差性，我们对 Quality 进行 Box-cox 变换。

4.4 主成份分析

5 结论及建议