红葡萄酒成分与质量的探索性数据分析

摘要

利用 *R* 语言作为研究对象，分析红葡萄酒数据中质量和红葡萄酒的物理化学指标变量的主要影响因素。由于变量之间可能具有相关性，所以要先对红葡萄酒的物理化学指标变量之间做双因素相关性分析，**定性分析**相关性较强的变量。接着对变量预处理，建立**多元线性回归**模型，但是由于质量只存在 *3*、*4*、*5*、*6*、*7*、*8*，六种评分数据，就使得逐步线性回归建立的模型的误差分析图呈现多条倾斜直线的异常情况。进一步分析，产生这种异常情况的原因可能是由于质量数据是分类变量，但其本质上仍然是有大小关系，不同于其他分组变量。所以要对质量做**加噪声处理**，即添加一个均值为 0，方差为 1 的正态分布噪声，使其变为连续性变量，进而建立线性回归模型。但由于正态分布噪声的随机性，进而导致了线性回归模型系数的随机性，所以就需要做**重复试验**取平均处理，减低模型的偶然性。并通过**交叉检验**，判断利用 *AIC* 和 *BIC* 准则建立的模型哪个误差较小。

数据来自 UCI 机器学习数据库的葡萄酒质量数据，关于红葡萄酒样品的 *11* 种物理化学指标和质量等级。

一、 问题背景

葡萄酒是一种成分复杂的酒精饮料，不同产地、年份和品种的葡萄酒成分不同，这也是导致质量差异过大的重要因素。至今，质量评价主要还是依靠专家的感官。味道是最难理解的一种感官，因此用味蕾评价葡萄酒也就成为一件艰巨的任务。为了评估葡萄酒的质量，我们提出的方法就是根据酒的物理化学性质与质量的关系，找出高品质的葡萄酒具体与什么性质密切相关，这些性质又是如何影响葡萄酒质量的，是本节课题所研究的问题。

二、 数据集介绍

红葡萄酒数据集，提供了红葡萄酒的样本数据，并由专家做质量评估(*0-10* 分不等)，并进行了理化指标的检验，其中包含以下 *11* 种理化指标：

表 变量名称

理化指标	含义
fixed acidity	非挥发性酸
volatile acidity	挥发性酸
citric acid	柠檬酸
residual sugar	残糖
chlorides	氯化物
free sulfur dioxide	游离二氧化硫
total sulfur dioxide	总二氧化硫
density	密度
pH	酸碱度
sulphates	硫酸盐
alcohol	酒精含量

三、 变量相关性分析

3.1 双变量相关性分析

由于变量之间可能存在相关性，所以首先对变量进行相关性分析，计算各个变量之间的相关系数。

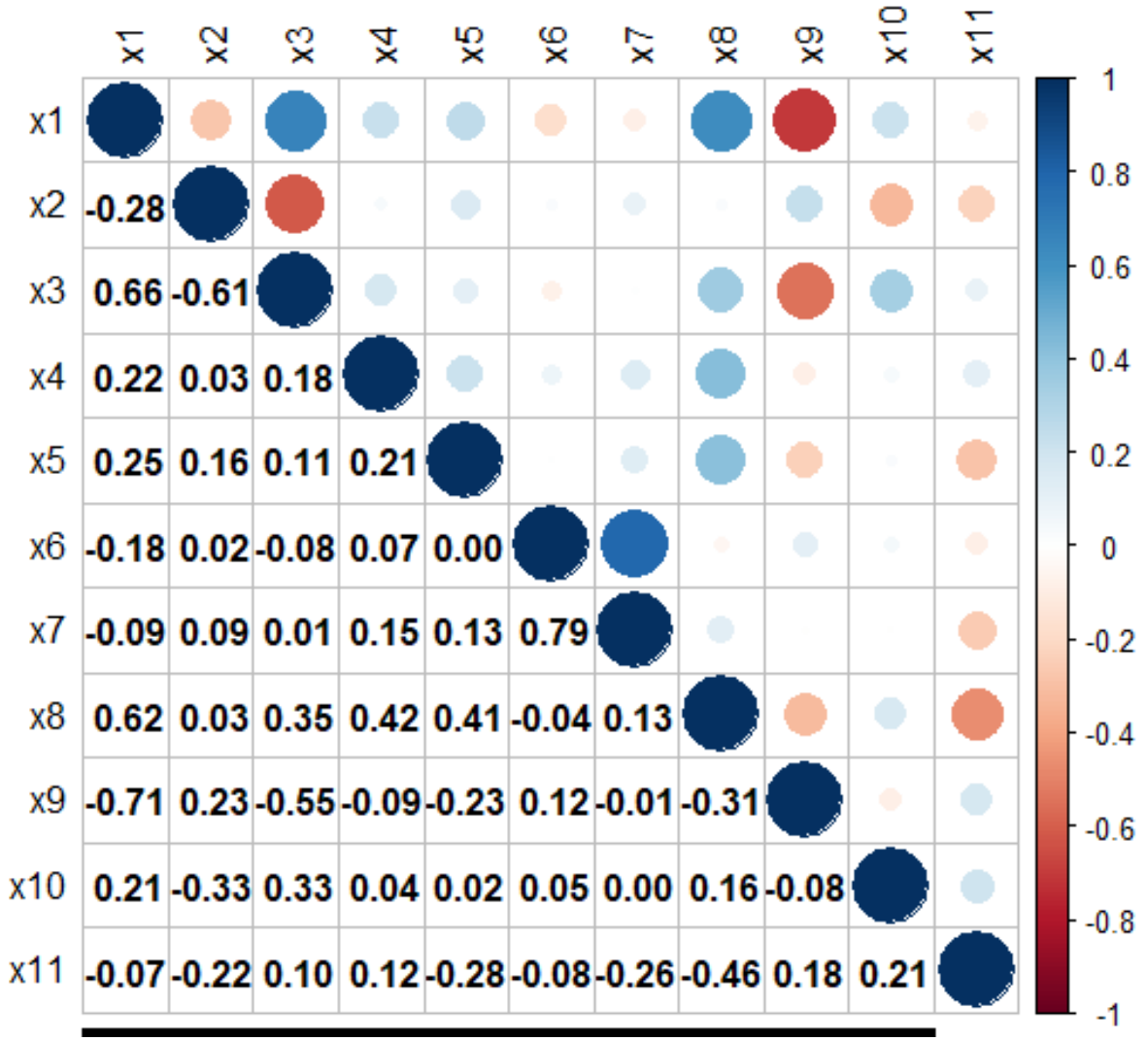


图 变量相关性热图

由相关系数矩阵可知：

free sulfur dioxide 和 *total sulfur dioxide*、*fixed acidity* 和 *citric acid*、*fixed acidity* 和 *density*、*residual sugar* 和 *density*、*chlorides* 和 *density* 有较强的正相关性；

fixed acidity 和 *pH*、*volatile acidity* 和 *citric acid*、*citric acid* 和 *pH*、*density* 和 *alcohol* 有较强的负相关性。

其中 *fixed acidity* 和 *pH*、*fixed acidity* 和 *citric acid*、*free sulfur dioxide* 和 *total sulfur dioxide*、*density* 和 *alcohol* 的散点图分别如下：

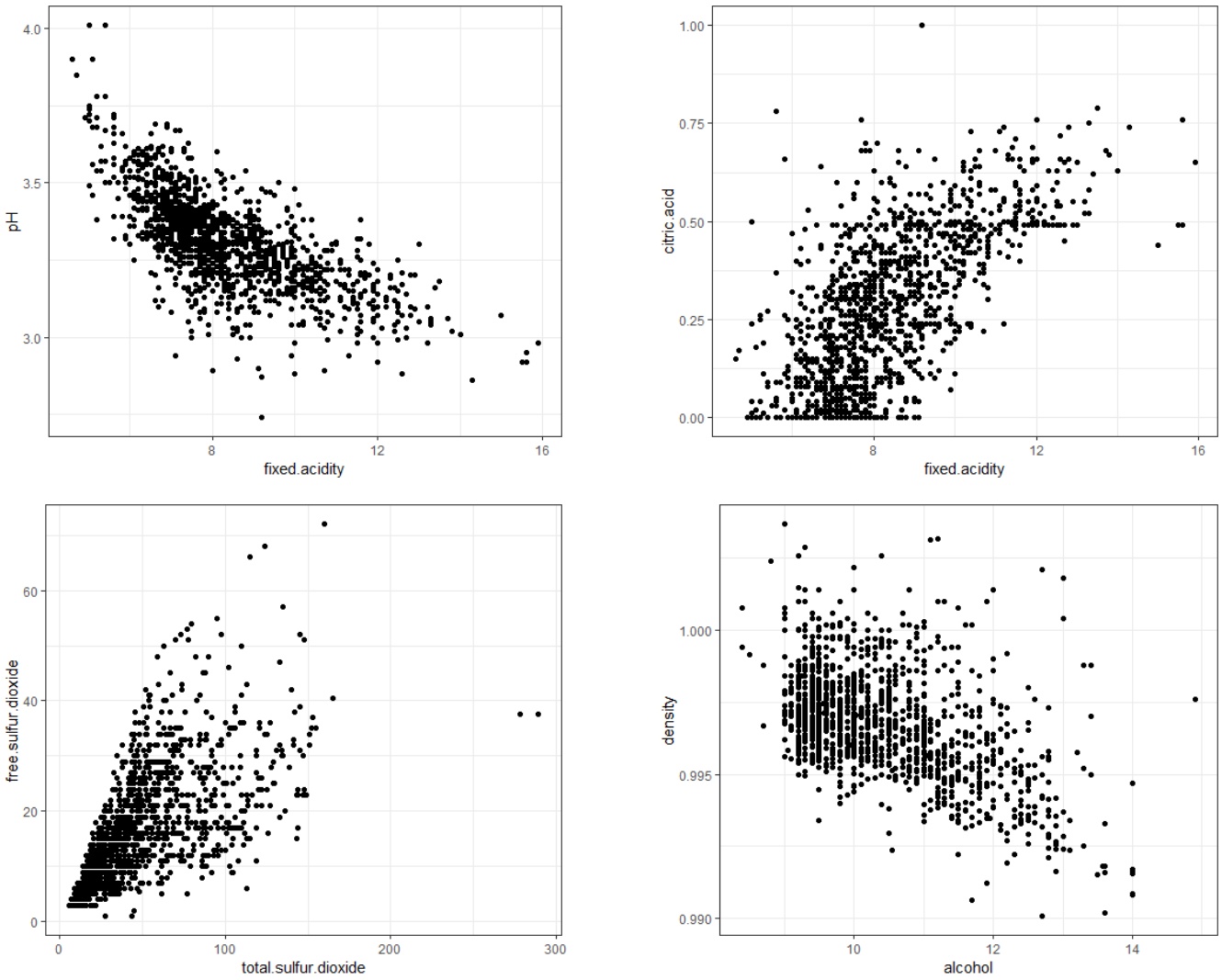


图 部分相关性较强变量的散点图

3.2 单变量相关性分析

由上述的双变量相关性分析可知，部分变量之间存在较强的相关性。接下来，需要分析 *11* 种变量对质量评估的相关性。由于质量评估仅仅只有六种，故我们以两种方式更加直观的定性分析变量对质量之间的相关性。

表 变量对质量之间的相关系数

quality		***quality***	
fixed acidity	*0.11*	*total sulfur dioxide*	*-0.20*
volatile acidity	*-0.38*	*density*	*-0.18*
citric acid	*0.21*	*pH*	*-0.04*
residual sugar	*0.03*	*sulphates*	*0.38*
chlorides	*-0.19*	*alcohol*	*0.48*
free sulfur dioxide	*-0.06*		

由于数据集中质量评估的评分结果仅仅只有六种，接下来用箱线图的形式更加直观的定性分析各个变量对质量的相关性。

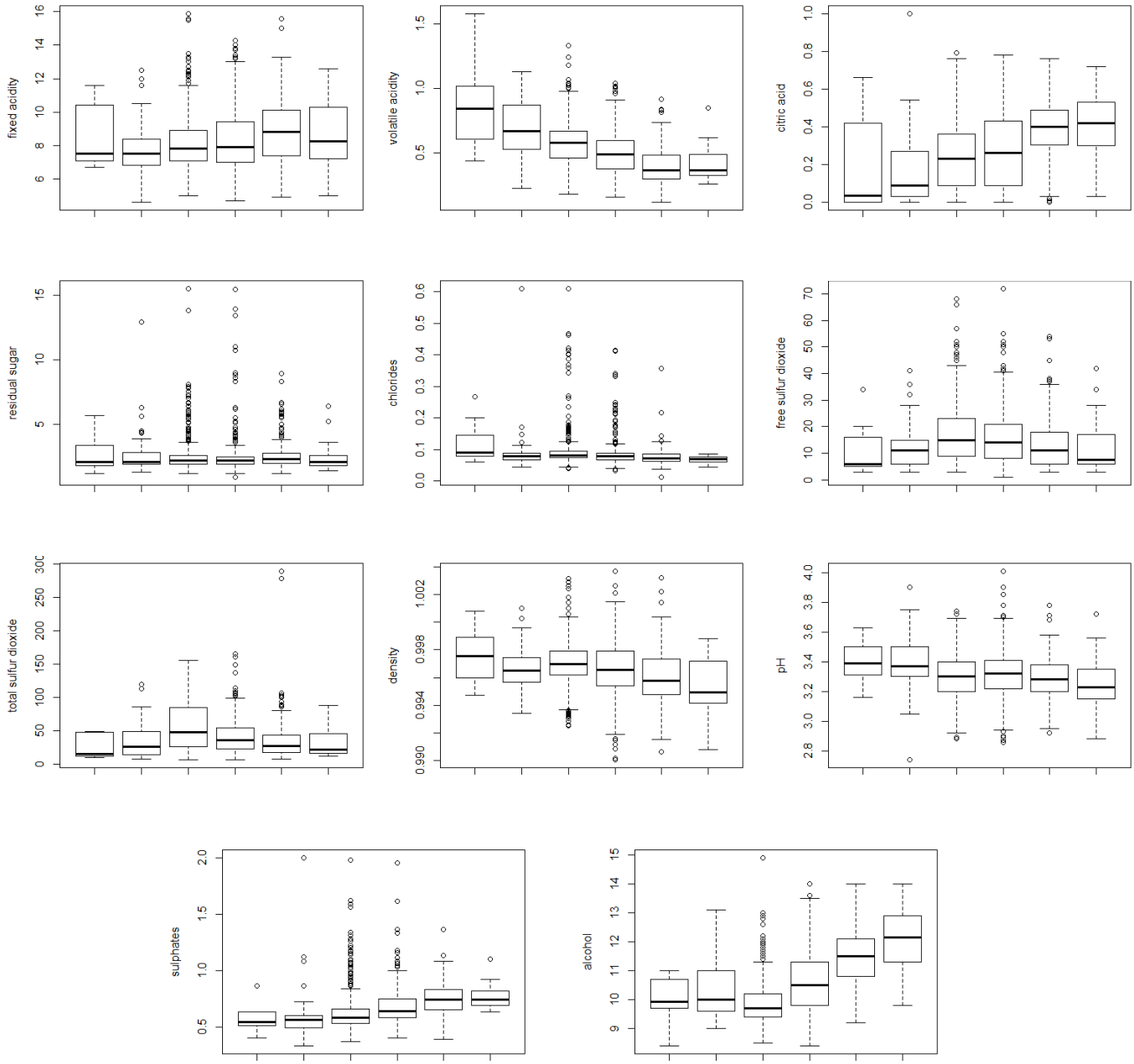


图 各个变量对质量的相关性分析

结果表明：

volatile acidity、*citric acid*、*alcohol* 对 *quality* 的波动较大；

fixed acidity、*free sulfur dioxide*、*total sulfur dioxide*、*density*、*pH*、*sulphates* 对 *quality* 的波动较小；

residual sugar、*chlorides* 对 *quality* 基本不波动。

四、 模型建立

4.1 线性回归分析原理

假定因变量y与自变量$x_1,x_2,\cdots,x_p$有以下形式的线性关系：

$$y=\beta_0+\beta_1x_1+\cdots+\beta_nx_n+\varepsilon$$

其中ε可看成是随机误差，一般来讲它服从$N(0,\sigma^2)$分布是合理的。

4.2 未加噪的质量评估线性回归

由于部分变量之间存在极强的相关性，所以我们需要找到影响较大的因素建模，这里用逐步回归方法和 *AIC* 准则筛选变量，进而建立线性回归模型。

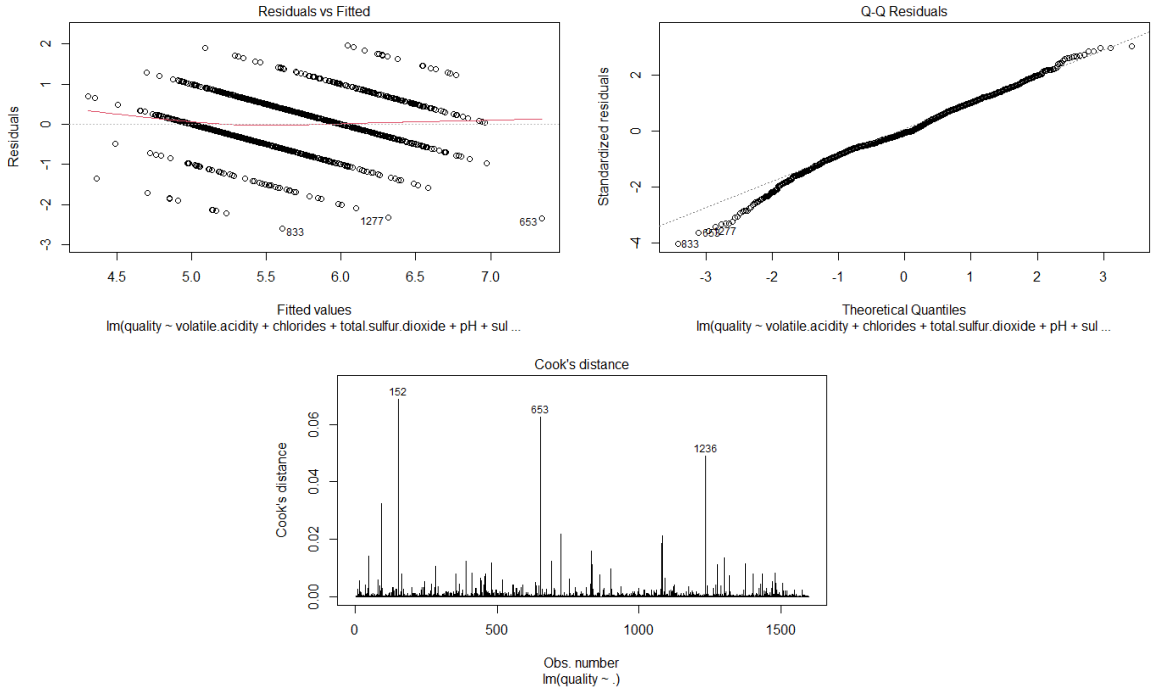
质量评估结果的数据特征：

质量评分	3	4	5	6	7	8
频数	10	53	681	638	199	18

利用 *AIC* 准则逐步回归筛选出的变量及显著性：

变量	系数
Intercept	5.64^{***}
volatile acidity	-0.18^{***}
chlorides	-0.09^{***}
free sulfur dioxide	0.05^{*}
total sulfur dioxide	-0.11^{***}
pH	-0.07^{***}
sulphates	0.15^{***}
alcohol	0.31^{***}

未加噪线性回归模型检验可视化：



结果分析：

Residuals vs Fitted 图呈现多条倾斜直线的情况，结果异常，说明模型存在问题；

Q-Q Residuals 图中，基本残差分布在*-3* 到 *2* 之间，并且基本服从正态分布，没有出现异常；

Cook's distance 图说明存在三个强影响点，分别为 *152*、*653*、*1236*。

4. 3 加噪的质量评估线性回归

分析上述结果产生原因，可能是因为质量评分太离散导致，解决办法是对质量评分进行加噪声处理，即对评分加上一个均值为 *0*，方差为 *1* 的正态噪声，使质量评分变成一个连续型变量，进而建立线性回归模型。

利用 *AIC* 准则逐步回归筛选变量及显著性：

变量	系数
Intercept	5.68^{***}
volatile acidity	-0.21^{***}
citric acid	$-0.09^{\cdot}$
residual sugar	-0.05
density	0.14^{**}
pH	-0.13^{***}
sulphates	0.08^{**}
alcohol	0.49^{***}

单个变量对质量评分的一元线性回归系数及显著性：

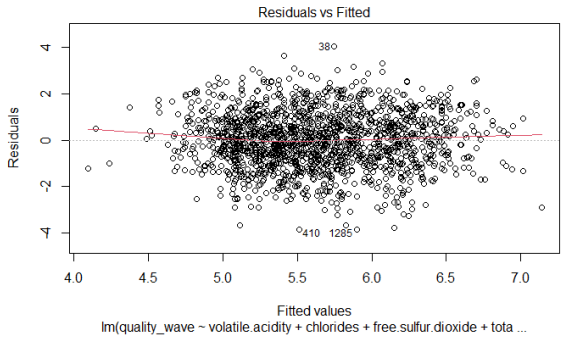
变量	常数	系数	变量	常数	系数
fixed acidity	0.15^{***}	5.68^{***}	*total sulfur dioxide*	-0.13^{***}	5.68^{***}
volatile acidity	-0.31^{***}	5.68^{***}	*density*	-0.11^{**}	5.68^{***}
citric acid	0.22^{***}	5.68^{***}	*pH*	-0.09^{**}	5.68^{***}
residual sugar	0.02	5.68^{***}	*sulphates*	0.20^{***}	5.68^{***}
chlorides	-0.08^{*}	5.68^{***}	*alcohol*	0.44^{***}	5.68^{***}
free sulfur dioxide	-0.06^{*}	5.68^{***}			

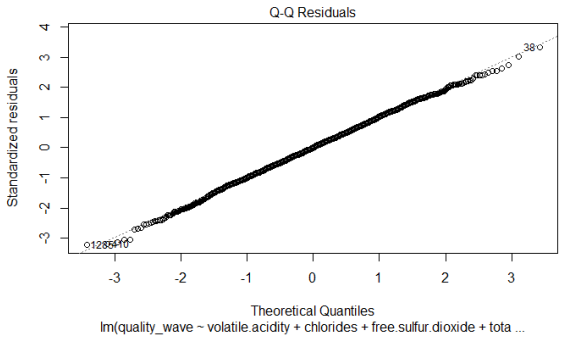
全部变量对质量评分的十一元线性回归系数及显著性：

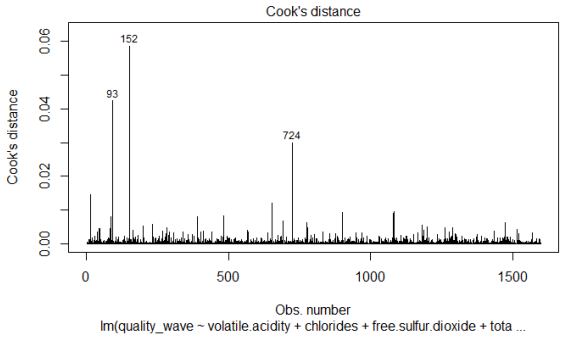
变量	系数	变量	系数
fixed acidity	-0.01	*total sulfur dioxide*	-0.02^{**}
volatile acidity	-0.20^{***}	*density*	0.13
citric acid	-0.07	*pH*	0.13
residual sugar	-0.04	*sulphates*	0.10
chlorides	-0.04	*alcohol*	0.47^{***}
free sulfur dioxide	-0.01	*Intercept*	5.68^{***}

结果表明：前后 *alcohol* 系数显著性和大小基本没有差异，*volatile acidity* 显著性没有差异，大小存在细微差异，其余变量显著性存在较大差异。

加噪线性回归模型检验可视化：







结果分析：

三个图均没有异常，模型合理。

五、 交叉检验及调参

对质量评分进行一次加噪声处理，可能存在偶然性，故建立的线性回归模型也会存在偶然性，所以要重复操作调整参数。

由于逐步回归中，有 *AIC* 准则和 *BIC* 准则，其表达式分别为：

$$AIC = 2k - 2\ln(L)$$

$$BIC = \ln(n)k - 2\ln(L)$$

其中，*L* 是似然函数，*n* 是样本大小，*k* 是参数数量。

很明显，*BIC* 的惩罚比 *AIC* 大，考虑了样本数量。

设置训练集和测试集，利用训练集调参，建立根据 *AIC* 和 *BIC* 准则的线性回归模型，最后利用测试集计算误差，选择误差小的作为最后建立的回归模型。

训练集和测试集数据特征：

	训练集(*80*%)	**测试集(*20*%)**
样本量	*1281*	*318*

AIC、*BIC* 回归模型调参：

变量	***AIC***	***BIC***	**变量**	***AIC***	***BIC***
fixed acidity	*0.01*	*0.00*	*total sulfur dioxide*	*-0.09*	*-0.03*
volatile acidity	*-0.20*	*-0.21*	*density*	*-0.02*	*0.00*
citric acid	*-0.02*	*0.00*	*pH*	*-0.06*	*-0.01*
residual sugar	*0.02*	*0.00*	*sulphates*	*0.16*	*0.13*
chlorides	*-0.07*	*-0.02*	*alcohol*	*0.29*	*0.31*
free sulfur dioxide	*0.02*	*0.00*	*Intercept*	*5.64*	*5.64*

误差公式：$error = \sum_{i=1}^{318} (y - \hat{y})^2$

两种模型的误差分别为：

AIC	***BIC***
139.55	*141.71*

所以，利用 *AIC* 建立的模型，误差较小更精准。

根据 *AIC* 建立的模型，可判断 *volatile acidity*、*sulphates* 和 *alcohol* 系数较大，说明这三种指标对质量评分影响较大。

我们将质量评分进行分组，将红葡萄酒品质分为低、中、高三种

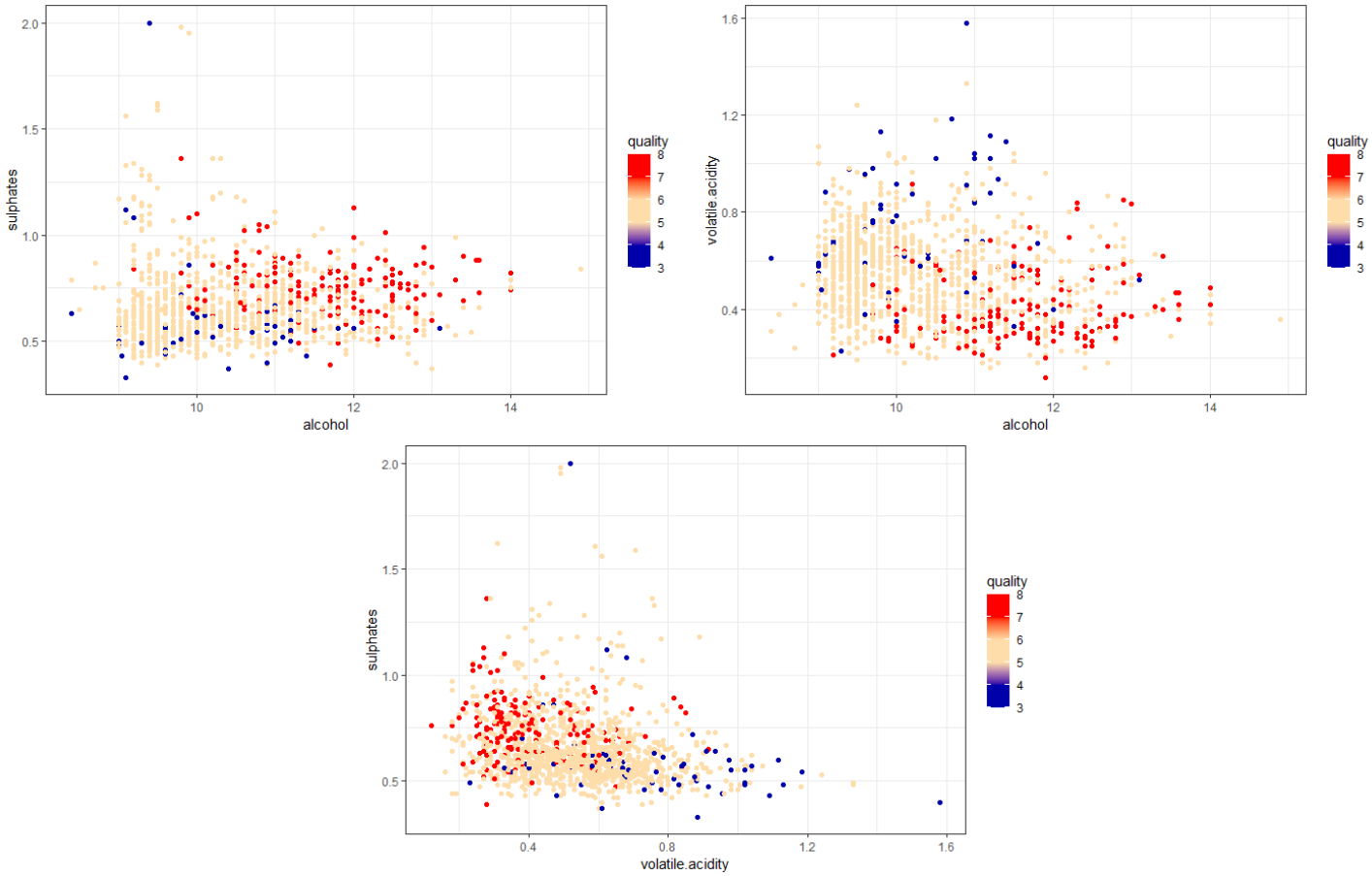


图 三种指标两两之间的散点图

结果表明：

alcohol 度在[*10*,*14*]之间，*sulphates* 浓度在[*0.5*,*1*]之间，*volatile acidity* 浓度在[*0.2*，*0.4*]之间，更容易产出高品质的红葡萄酒。