Business Statistics

“Vinho Verde”红葡萄酒酒精度数影响因素研究

2018级数据科学4班 刘慧怡 320180940030

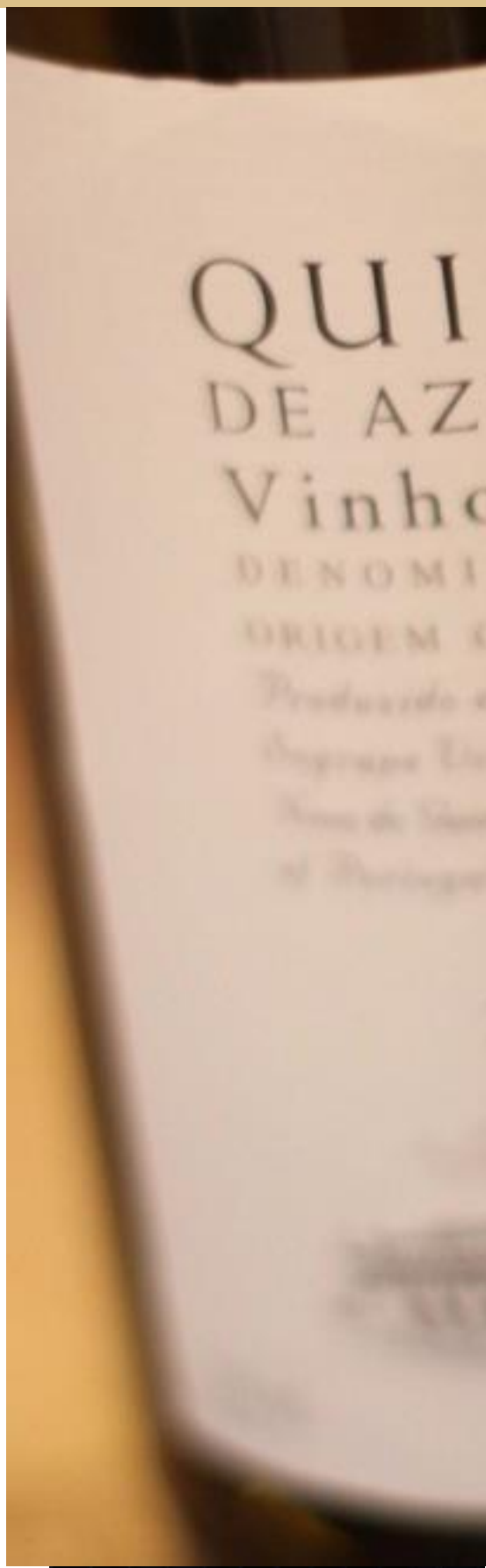
PRAFACE

前 言

Vinho Verde——葡萄牙青酒，产自葡萄牙北部，主要产自米尼奥（Minho）地区，在葡萄牙的受欢迎程度仅次于波尔图葡萄酒。这种葡萄酒，该酒有独特的口感和特征，一般口感清爽，味道淡雅，酸度高，清新、芳香。特点是清淡，酒精含量也很低，清爽甘冽，为夏季饮用佳品。

本研究报告采用多元线性回归分析方法，研究Vinho Verde的物理化学参数和感官参数与其酒精度数的关联关系，旨在为酿酒厂提供一定的参考价值，通过对化学成分的控制和感官参数的测量，实现对酒精度数的调控与估计，以呈现出符合预期的酒精度数和最优质的口感。

数据集收录于网站*https://archive.ics.uci.edu/ml/datasets/wine+quality*，由Vinho Verde产区葡萄栽培委员会(CVRVV)发布，具有较高的真实性与可信度。



CONTENTS

目 录

一、数据说明

数据集概况

数据集共有1599个观测值，12个特征。由于隐私和逻辑问题，只有物理化学和感官变量可用（例如，没有葡萄类型、葡萄酒品牌、葡萄酒销售价格等数据）。

```
'data.frame':   1599 obs. of  12 variables:
$ fixedacidity      : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
$ volatileacidity   : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
$ citricacid        : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
$ residualsugar     : 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 ...
$ freesulfurdioxide : num  11 25 15 17 11 13 15 15 9 17 ...
$ totalsulfurdioxide: 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 ...
```

从上面的数据探查可以发现，该数据的变量大多数是数值型变量，变量quality（感官参数综合评分）是整数型变量。

参数说明

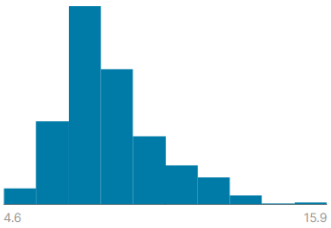
根据研究目标，确定变量alcohol作为因变量，其余变量作为解释变量。

变量编号	变量名	变量说明
X_1	fixed acidity	与酒有关的，不易挥发的酸含量
X_2	volatile acidity	醋酸含量
X_3	citric acid	柠檬酸含量
X_4	residual sugar	糖分含量
X_5	chlorides	盐含量
X_6	free sulfur dioxide	游离态二氧化硫含量
X_7	total sulfur dioxide	总二氧化硫含量
X_8	density	酒的密度
X_9	pH	pH值，用于描述酒的酸性或碱性
X_{10}	sulphates	硫酸盐，一种葡萄酒添加剂
X_{11}	quality	质量评级（基于感官数据综合得出）
Y	**alcohol**	**酒精含量**

各变量具体总览情况如下：

fixed acidity

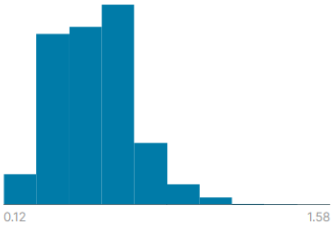
most acids involved with wine or fixed or nonvolatile (do not evaporate readily)



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	8.32	
Std. Deviation	1.74	
Quantiles	4.6	Min
	7.1	25%
	7.9	50%
	9.2	75%
	15.9	Max

volatile acidity

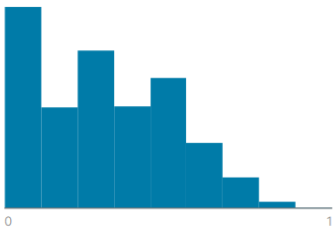
the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.53	
Std. Deviation	0.18	
Quantiles	0.12	Min
	0.39	25%
	0.52	50%
	0.64	75%
	1.58	Max

citric acid

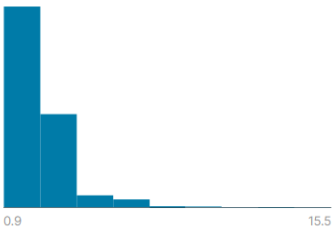
found in small quantities, citric acid can add 'freshness' and flavor to wines



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.27	
Std. Deviation	0.19	
Quantiles	0	Min
	0.09	25%
	0.26	50%
	0.42	75%
	1	Max

residual sugar

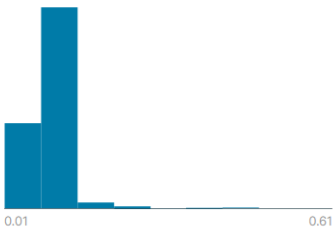
the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	2.54	
Std. Deviation	1.41	
Quantiles	0.9	Min
	1.9	25%
	2.2	50%
	2.6	75%
	15.5	Max

chlorides

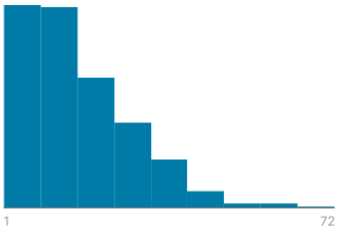
the amount of salt in the wine



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.09	
Std. Deviation	0.05	
Quantiles	0.01	Min
	0.07	25%
	0.08	50%
	0.09	75%
	0.61	Max

free sulfur dioxide

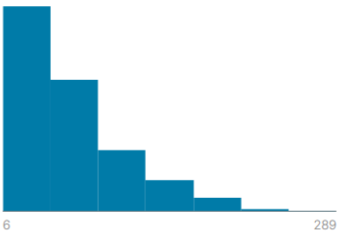
the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	15.9	
Std. Deviation	10.5	
Quantiles	1	Min
	7	25%
	14	50%
	21	75%
	72	Max

total sulfur dioxide

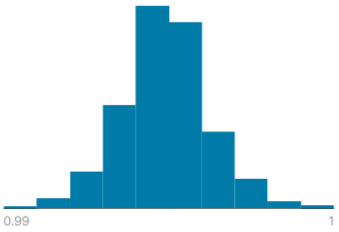
amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	46.5	
Std. Deviation	32.9	
Quantiles	6	Min
	22	25%
	38	50%
	62	75%
	289	Max

density

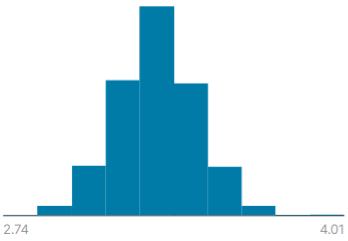
the density of water is close to that of water depending on the percent alcohol and sugar content



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	1	
Std. Deviation	0	
Quantiles	0.99	Min
	1	25%
	1	50%
	1	75%
	1	Max

pH

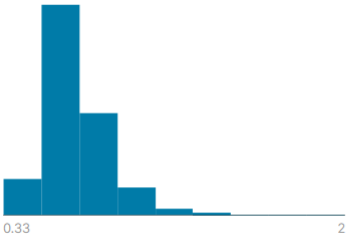
describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale



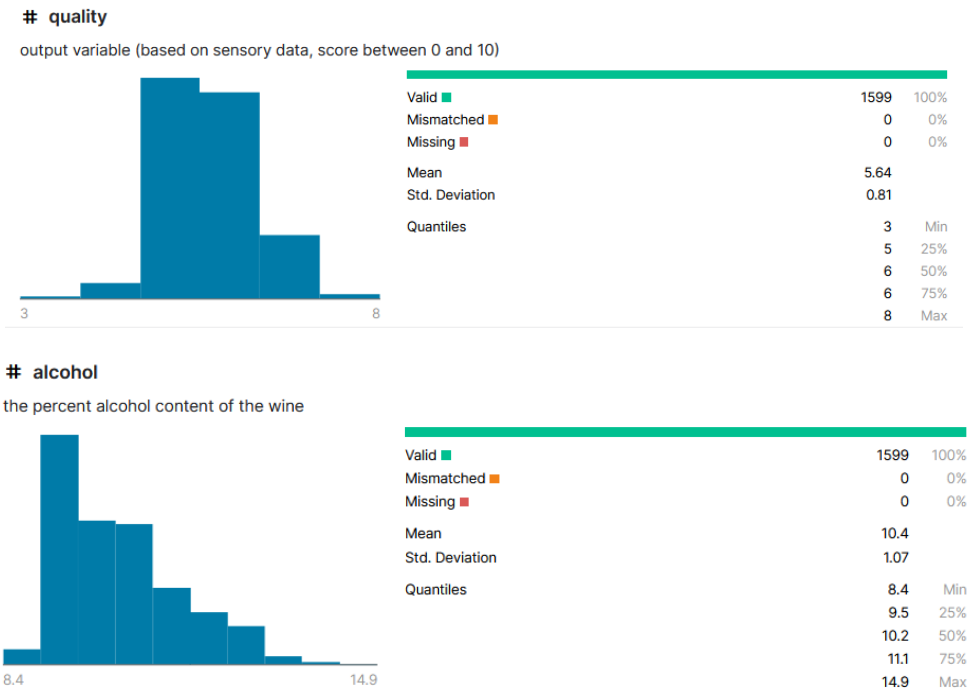
Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	3.31	
Std. Deviation	0.15	
Quantiles	2.74	Min
	3.21	25%
	3.31	50%
	3.4	75%
	4.01	Max

sulphates

a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant



Valid	1599	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.66	
Std. Deviation	0.17	
Quantiles	0.33	Min
	0.55	25%
	0.62	50%
	0.73	75%
	2	Max



各个变量的数据较为标准，不存在缺失值或不合法数据，因此无需进行初步清洗。

二、模型假设

响应变量alcohol为数值型数据即定量指标，解释变量基本上绝大部分也是定量指标（除quality为定性等级指标，已转换为整型数值），由于响应变量可能受到其他多个解释变量的影响，采用多元线性回归分析。

设多元线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

其中，$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$为偏回归系数，表示在其他自变量保持不变时，$X_j$中增加或减少一个单位时Y的期望值的平均变化量。$\varepsilon$代表随机误差，本次分析中假设随机误差是随机的、独立的、服从正态分布的。

应用于本数据集中，回归方程可以写成更直观的形式：

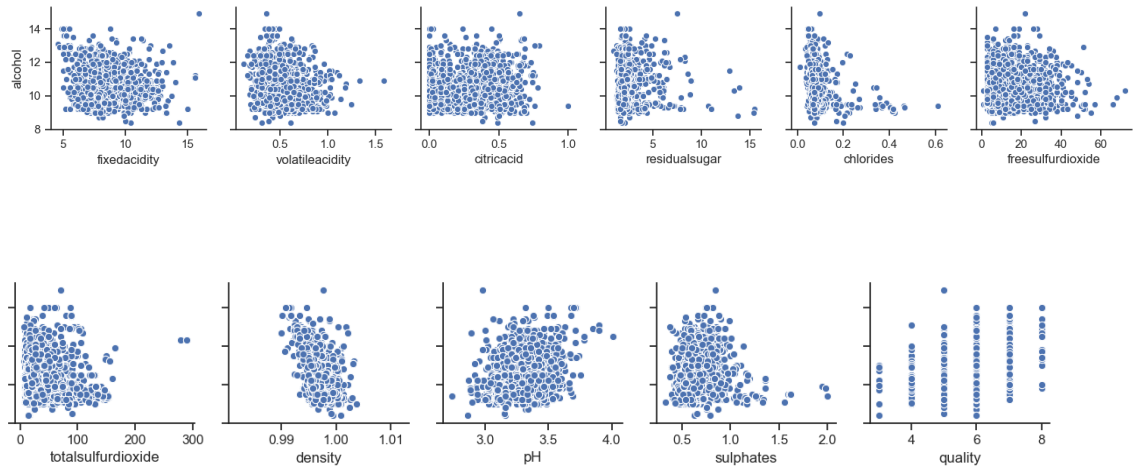
$$\text{alcohol} = \beta_0 + \beta_1 \text{fixedacidity} + \beta_2 \text{volatileacidity} + \beta_3 \text{citricacid} + \beta_4 \text{residualsugar} + \beta_5 \text{chlorides} + \beta_6 \text{freesulfurdioxide} + \beta_7 \text{totalsulfurdioxide} + \beta_8 \text{density} + \beta_9 \text{pH} + \beta_{10} \text{sulphates} + \beta_{11} \text{quality} + \varepsilon$$

三、相关分析

各变量的相关系数表如下：

相关系数	fixedacidity	volatileacidity	citricacid	residualsugar	chlorides	freesulfurdioxide	totalsulfurdioxide	density	pH	sulphates	alcohol	quality
fixedacidity	1.000	-0.256	0.672	0.115	0.094	-0.154	-0.113	0.668	-0.683	0.183	-0.062	0.124
volatileacidity	-0.256	1.000	-0.552	0.002	0.061	-0.011	0.076	0.022	0.235	-0.261	-0.202	-0.391
citricacid	0.672	-0.552	1.000	0.144	0.204	-0.061	0.036	0.365	-0.542	0.313	0.110	0.226
residualsugar	0.115	0.002	0.144	1.000	0.056	0.187	0.203	0.355	-0.086	0.006	0.042	0.014
chlorides	0.094	0.061	0.204	0.056	1.000	0.006	0.047	0.201	-0.265	0.371	-0.221	-0.129
freesulfurdioxide	-0.154	-0.011	-0.061	0.187	0.006	1.000	0.668	-0.022	0.070	0.052	-0.069	-0.051
totalsulfurdioxide	-0.113	0.076	0.036	0.203	0.047	0.668	1.000	0.071	-0.066	0.043	-0.206	-0.185
density	0.668	0.022	0.365	0.355	0.201	-0.022	0.071	1.000	-0.342	0.149	-0.496	-0.175
pH	-0.683	0.235	-0.542	-0.086	-0.265	0.070	-0.066	-0.342	1.000	-0.197	0.206	-0.058
sulphates	0.183	-0.261	0.313	0.006	0.371	0.052	0.043	0.149	-0.197	1.000	0.094	0.251
alcohol	-0.062	-0.202	0.110	0.042	-0.221	-0.069	-0.206	-0.496	0.206	0.094	1.000	0.476
quality	0.124	-0.391	0.226	0.014	-0.129	-0.051	-0.185	-0.175	-0.058	0.251	0.476	1.000

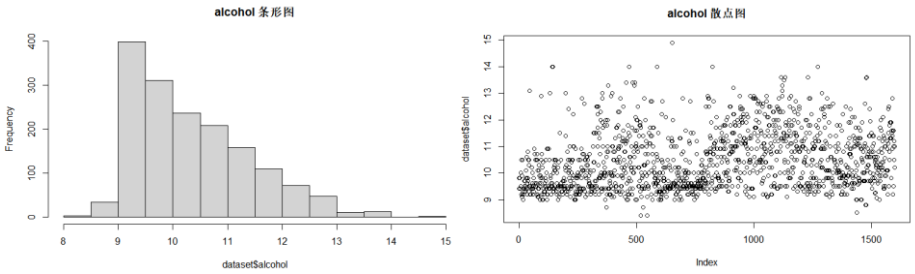
做出每一个自变量与因变量的散点图如下：



从以上散点图和相关系数结果表可以看出，各个自变量与因变量的关系中较为明显的是密度density和质量评级quality。alcohol与density呈负相关；与quality呈弱正相关。散点图粗略的直观反映出了变量间的关系。

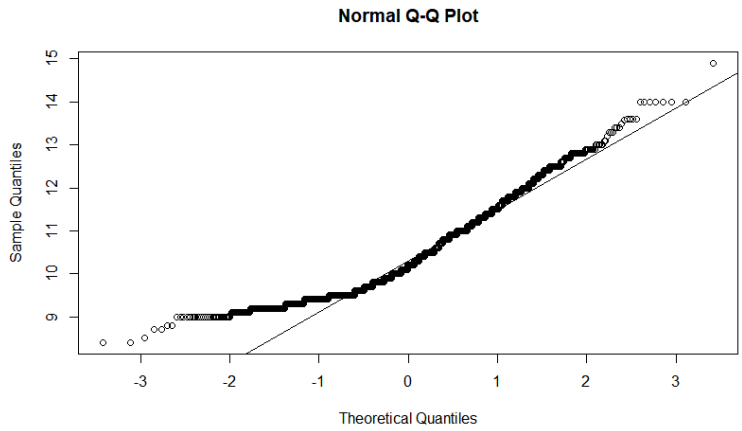
四、正态性检验

条形图和散点图



对因变量alcohol做直方图和散点图，根据直方图可以粗略看出大致符合正态分布。

QQ图



正态性检验主要采用QQ图，即分位数图示法（Quantile Quantile Plot，简称 Q-Q 图）。其原理是先将数据集排序，再计算每个数字对应的累计分布值（i-0.5)/n（i为n中的第i个值)。累计分布值表示某个特定值以下的值所占数据的比例。其后画出标准正态分布曲线（平均值=0，标准方差=1）。最后根据以上所得数据画QQ图，第一个点是第一个累计分布值对应的数值。横坐标是正态分布，纵坐标是数据集。

酒精含量的QQ图中x和y构成的点总体上分布在一条直线上，证明样本数据与正态分布存在线性相关性，即服从正态分布。

综上，我们可以近似的认为该变量是服从正态分布的。

五、模型建立

初步回归

从下面的回归结果我们可以看到，模型的R方等于0.6913，调整的R方为0.6891，总体来说，模型的效果较好，但仍存在其他影响酒精含量的因素未被发现。 从各个自变量的显著性来看，除了变量freesulfurdioxide系数检验不显著，其余变量在显著性水平0.05下均表现显著。

```
Call:
lm(formula = alcohol ~ ., data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15912 -0.36953 -0.04903  0.34211  2.51405

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.632e+02  1.334e+01  42.201  < 2e-16 ***
fixedacidity         4.925e-01  2.034e-02  24.215  < 2e-16 ***
volatileacidity      5.893e-01  1.128e-01   5.223 2.00e-07 ***
citricacid           8.197e-01  1.334e-01   6.143 1.02e-09 ***
residualsugar        2.624e-01  1.208e-02  21.715  < 2e-16 ***
chlorides           -9.330e-01  3.862e-01  -2.416   0.0158 *
freesulfurdioxide   -3.019e-03  1.992e-03  -1.515   0.1299
totalsulfurdioxide  -1.390e-03  6.715e-04  -2.071   0.0386 *
density             -5.736e+02  1.365e+01 -42.027  < 2e-16 ***
pH                   3.617e+00  1.507e-01  24.001  < 2e-16 ***
sulphates            9.540e-01  1.042e-01   9.154  < 2e-16 ***
quality              2.322e-01  2.227e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5942 on 1587 degrees of freedom
Multiple R-squared:  0.6913,    Adjusted R-squared:  0.6891
F-statistic:   323 on 11 and 1587 DF,  p-value: < 2.2e-16
```

逐步回归——选择最优特征和模型

为了选择最优的模型，我们利用逐步回归，以AIC信息统计量为准则，通过选择最小的AIC信息统计量，来达到删除或增加变量的目的。AIC全称是最小化信息量准则（Akaike Information Criterion）。它是拟合精度和参数个数的加权函数： AIC=2（模型参数的个数）-2ln（模型的极大似然函数）。

下面是逐步回归的结果：

```
Start:  AIC=-1652.8
alcohol ~ fixedacidity + volatileacidity + citricacid + residualsugar +
    chlorides + freesulfurdioxide + totalsulfurdioxide + density +
    pH + sulphates + quality

                    Df Sum of Sq     RSS      AIC
<none>                           560.30 -1652.80
- freesulfurdioxide   1      0.81  561.11 -1652.49
- totalsulfurdioxide  1      1.51  561.82 -1650.49
- chlorides           1      2.06  562.36 -1648.93
- volatileacidity     1      9.63  569.93 -1627.55
- citricacid          1     13.32  573.63 -1617.22
- sulphates           1     29.58  589.88 -1572.53
- quality             1     38.40  598.70 -1548.81
- residualsugar       1    166.48  726.79 -1238.81
- pH                  1    203.38  763.68 -1159.63
- fixedacidity        1    207.02  767.33 -1152.02
- density             1    623.59 1183.90  -458.61
```

根据AIC的公式，增加自由变量的数目提高了拟合的优良性，AIC鼓励数据拟合的优良性但是尽量避免出现过拟合的情况。所以优先考虑的模型应该是AIC值较小的那一个。根据AIC

准则可以找出最好的解释数据但是包含最少自由参数的模型。

从上面的逐步回归结果可以看出，初始模型的AIC是1652.80.随着变量的删除或者增加，模型的AIC相比起始模型都是增加的，说明起始模型就是当前最优的模型。

六、模型检验

多重共线性检验

多元线性回归的前提是自变量相互独立，不存在线性关系，因此我们对自变量进行多重共线性检验，同时也是对自变量进行二次筛选。

①相关系数矩阵

相关系数	fixedacidity	volatileacidity	citricacid	residualsugar	chlorides	freesulfurdioxide	totalsulfurdioxide	density	pH	sulphates	alcohol	quality
fixedacidity	1.000	-0.256	0.672	0.115	0.094	-0.154	-0.113	0.668	-0.683	0.183	-0.062	0.124
volatileacidity	-0.256	1.000	-0.552	0.002	0.061	-0.011	0.076	0.022	0.235	-0.261	-0.202	-0.391
citricacid	0.672	-0.552	1.000	0.144	0.204	-0.061	0.036	0.365	-0.542	0.313	0.110	0.226
residualsugar	0.115	0.002	0.144	1.000	0.056	0.187	0.203	0.355	-0.086	0.006	0.042	0.014
chlorides	0.094	0.061	0.204	0.056	1.000	0.006	0.047	0.201	-0.265	0.371	-0.221	-0.129
freesulfurdioxide	-0.154	-0.011	-0.061	0.187	0.006	1.000	0.668	-0.022	0.070	0.052	-0.069	-0.051
totalsulfurdioxide	-0.113	0.076	0.036	0.203	0.047	0.668	1.000	0.071	-0.066	0.043	-0.206	-0.185
density	0.668	0.022	0.365	0.355	0.201	-0.022	0.071	1.000	-0.342	0.149	-0.496	-0.175
pH	-0.683	0.235	-0.542	-0.086	-0.265	0.070	-0.066	-0.342	1.000	-0.197	0.206	-0.058
sulphates	0.183	-0.261	0.313	0.006	0.371	0.052	0.043	0.149	-0.197	1.000	0.094	0.251
alcohol	-0.062	-0.202	0.110	0.042	-0.221	-0.069	-0.206	-0.496	0.206	0.094	1.000	0.476
quality	0.124	-0.391	0.226	0.014	-0.129	-0.051	-0.185	-0.175	-0.058	0.251	0.476	1.000

根据 Klein's Rule，只有当预测因子的两两相关性超过多元相关系数R时，我们才应该担心回归系数估计值的稳定性。根据之前分析可得，模型的多元相关系数大于表格中任意两个自变量之间的相关系数，因此粗略估计不存在多重共线性。

②方差膨胀因子（VIF）

为了进一步更全面的验证各个自变量之间是否存在共线性问题，此处利用方差膨胀因子进行判断。方差膨胀因子VIF是指回归系数的估计量由于自变量共线性使得方差增加的一个相对度量。一般建议，如VIF>10，表明模型中有很强的共线性问题。VIF的计算公式如下：

$$VIF_j = \frac{1}{1 - R_j^2}$$

各自变量的VIF值如下所示：

```
> vif(model, digits = 3)
       fixedacidity    volatileacidity         citricacid      residualsugar          chlorides
              5.675              1.848              3.058              1.314              1.495
freesulfurdioxide totalsulfurdioxide            density                 pH          sulphates
              1.965              2.209              3.004              2.450              1.413
            quality
              1.464
```

从上面的结果可以看出，各变量的方差膨胀因子最大是5.7<10,因此所有变量之间不存在共线性问题。

显著性检验

① F检验

首先对回归方程整体进行显著性检验，即F检验。检验做如下假设：

H_0: 所有的系数都为0 ($\beta_1 = \beta_2 = \ldots = \beta_k = 0$)

H_1: 至少一个系数不为0

回归结果显示（如下图），由于f检验的p值小于2.2e-16，约等于0，不仅满足p<0.05，甚至满足p<o.o1，具有极其显著的统计学差异。因此拒绝原假设，接受备择假设。该线性回归方程的整体方程系数显著异于零，至少有一个解释变量对因变量有影响。

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5942 on 1587 degrees of freedom
Multiple R-squared:  0.6913,    Adjusted R-squared:  0.6891
F-statistic:   323 on 11 and 1587 DF,  p-value: < 2.2e-16
```

② t检验

分别检验回归模型中各个回归系数是否具有显著性，即检验解释变量的系数是否在规定的显著性水平上显著。对于回归系数$\beta_i(1 \leqslant i \leqslant 11)$，作如下假设：

H_0: $\beta_i=0$

H_1: $\beta_i \neq 0$

根据下表可以看出，除游离态二氧化硫含量freesulfurdioxide之外，其他解释变量的p值均大于0.05，拒绝H_0，接受H_1，系数显著不为0。意味着除了freesulfurdioxide外每一个自变量都对因变量酒精含量产生影响。

```
Call:
lm(formula = alcohol ~ ., data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15912 -0.36953 -0.04903  0.34211  2.51405

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.632e+02  1.334e+01  42.201  < 2e-16 ***
fixedacidity         4.925e-01  2.034e-02  24.215  < 2e-16 ***
volatileacidity      5.893e-01  1.128e-01   5.223 2.00e-07 ***
citricacid           8.197e-01  1.334e-01   6.143 1.02e-09 ***
residualsugar        2.624e-01  1.208e-02  21.715  < 2e-16 ***
chlorides           -9.330e-01  3.862e-01  -2.416   0.0158 *
freesulfurdioxide   -3.019e-03  1.992e-03  -1.515   0.1299
totalsulfurdioxide  -1.390e-03  6.715e-04  -2.071   0.0386 *
density             -5.736e+02  1.365e+01 -42.027  < 2e-16 ***
pH                   3.617e+00  1.507e-01  24.001  < 2e-16 ***
sulphates            9.540e-01  1.042e-01   9.154  < 2e-16 ***
quality              2.322e-01  2.227e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5942 on 1587 degrees of freedom
Multiple R-squared:  0.6913,    Adjusted R-squared:  0.6891
F-statistic:   323 on 11 and 1587 DF,  p-value: < 2.2e-16
```

由于解释变量freesulfurdioxide的p值大于0.05，差异不显著，无法有效拒绝原假设，因此删除该变量做第二次线性回归，结果如下：

Regression Analysis

R?	0.691
Adjusted R?	0.689
R	0.831
Std. Error	0.594
n	1599
k	10
Dep. Var.	**alcohol**

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	1,253.6509	10	125.3651	354.79	0.00E+00
Residual	561.1136	1588	0.3533		
Total	1,814.7645	1598			

Regression output					*confidence interval*	
variables	*coefficients*	*std. error*	*t (df=1588)*	*p-value*	*95% lower*	*95% upper*
Intercept	561.5896					
fixedacidity	0.4893	0.0202	24.179	3.12E-110	0.4496	0.5290
volatileacidity	0.6161	0.1115	5.525	3.84E-08	0.3974	0.8348
citricacid	0.8538	0.1316	6.489	1.15E-10	0.5957	1.1119
residualsugar	0.2601	0.0120	21.683	1.57E-91	0.2366	0.2837
chlorides	-0.9717	0.3855	-2.521	.0118	-1.7278	-0.2156
totalsulfurdioxide	-0.0021	0.0005	-4.134	3.76E-05	-0.0031	-0.0011
density	-571.9288	13.6083	-42.028	3.83E-260	-598.6209	-545.2367
pH	3.5849	0.1493	24.017	5.47E-109	3.2921	3.8777
sulphates	0.9485	0.1042	9.102	2.58E-19	0.7441	1.1528
quality	0.2308	0.0223	10.370	2.00E-24	0.1871	0.2745

与原先的回归结果对比可以发现，去除该解释变量后，回归分析模型中R^2 adj 与R^2 的变化可

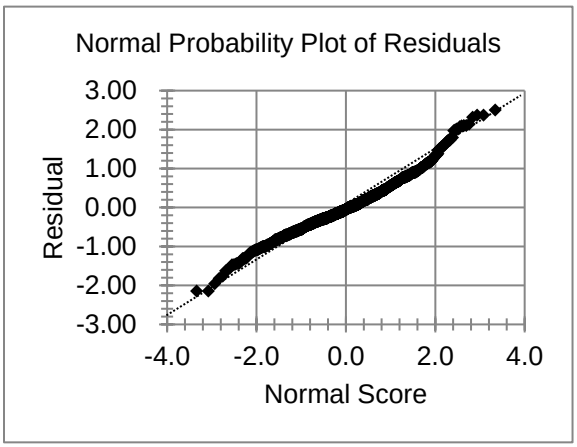
以忽略不计，说明去除的变量对模型影响不大。ANOVA表中，f检验的p值仍然约等于0，说明该模型的回归系数仍显著不为0。回归输出表的结果中，各个解释变量的的p值均减小，10个解释变量中9个变量的p值接近于0，各个解释变量的p值均小于0.05，因此可以都拒绝原假设，各个回归系数显著不为0。

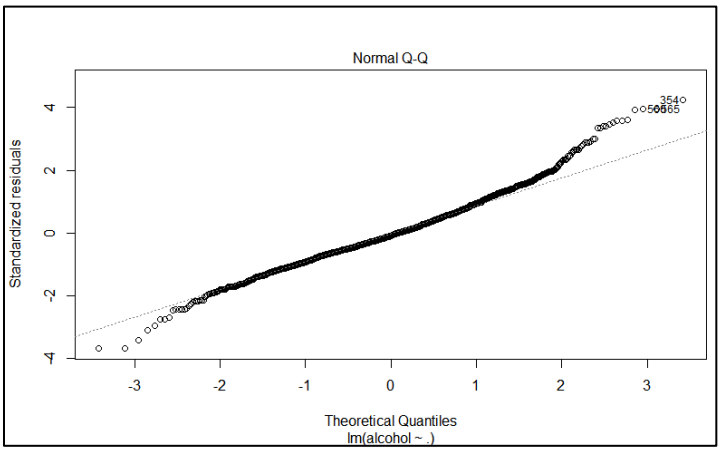
七、残差分析

在多元线性回归模型的构建中，我们对残差做了同方差、独立、服从正态分布的假设，多元线性回归建立在该假设的基础上。因此，应对该假设进行验证。

残差序列的正态性检验

非正态的残差会造成参数的置信区间不可靠，因为它们是使用正态性假设来构造的。因此有必要对残差进行正态性检验。绘制标准化残差的累积概率图如下。概率图的上下两端与线性略有偏差，但总体上符合正态假设。





绘制残差的QQ图，由于近似一条直线，同样可以说明总体上符合正态假设。

残差序列的独立性检验

误差存在自相关时，模型中的系数用最小二乘估计计算会不准确，往往会算出的系数的真实方差值和误差项的方差值会偏小。为了检验得到的方程的准确性，我们进行自相关检验。由于样本数量较大，选择根据DW值进行检验而非直观化图形检验。DW检验的原假设和备择假设分别为：

H_0: ρ=0（u不存在自相关）

H_1: ρ≠0 (u存在一阶自相关)

DW的计算公式为：

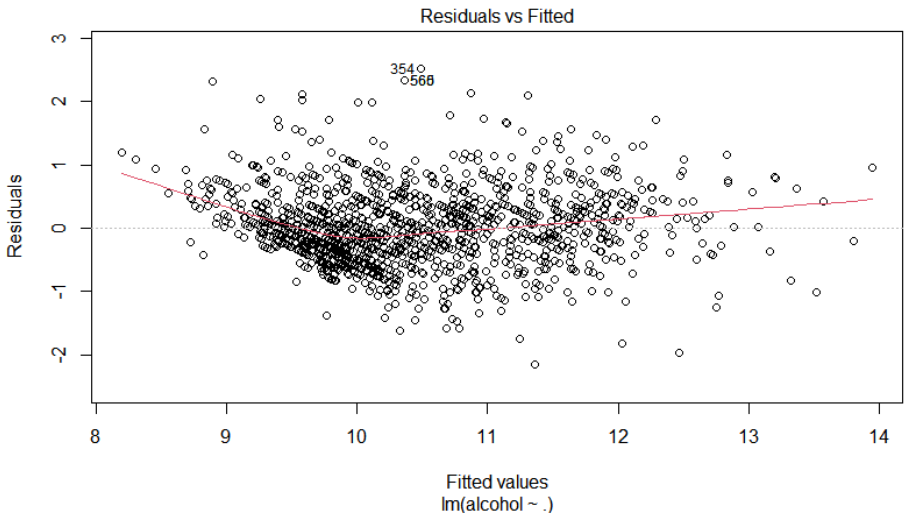
$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

利用MegaStat 计算出DW值为1.86411026933906≈2，若DW值小于2证明有正自相关性，大于2时则说明有负相关性。此处DW值接近2，因而说明模型较为理想，不存在自相关性。

残差序列的同方差检验

假设中残差具有同方差性。由于异方差的存在使得最小二乘估计量不再是最好的线性无偏估计量，会导致模型的残差不再是同方差的，所以要对模型进行异方差检验。

绘制残差和预测值的散点图，可以直观地看出残差与样本数据是否有明显关系。若残差随着样本数据的变化而变化，那么就存在异方差性。由于图中各点随机分布在过0点的直线两侧，没有固定的分布模式，所以不需要考虑异方差。



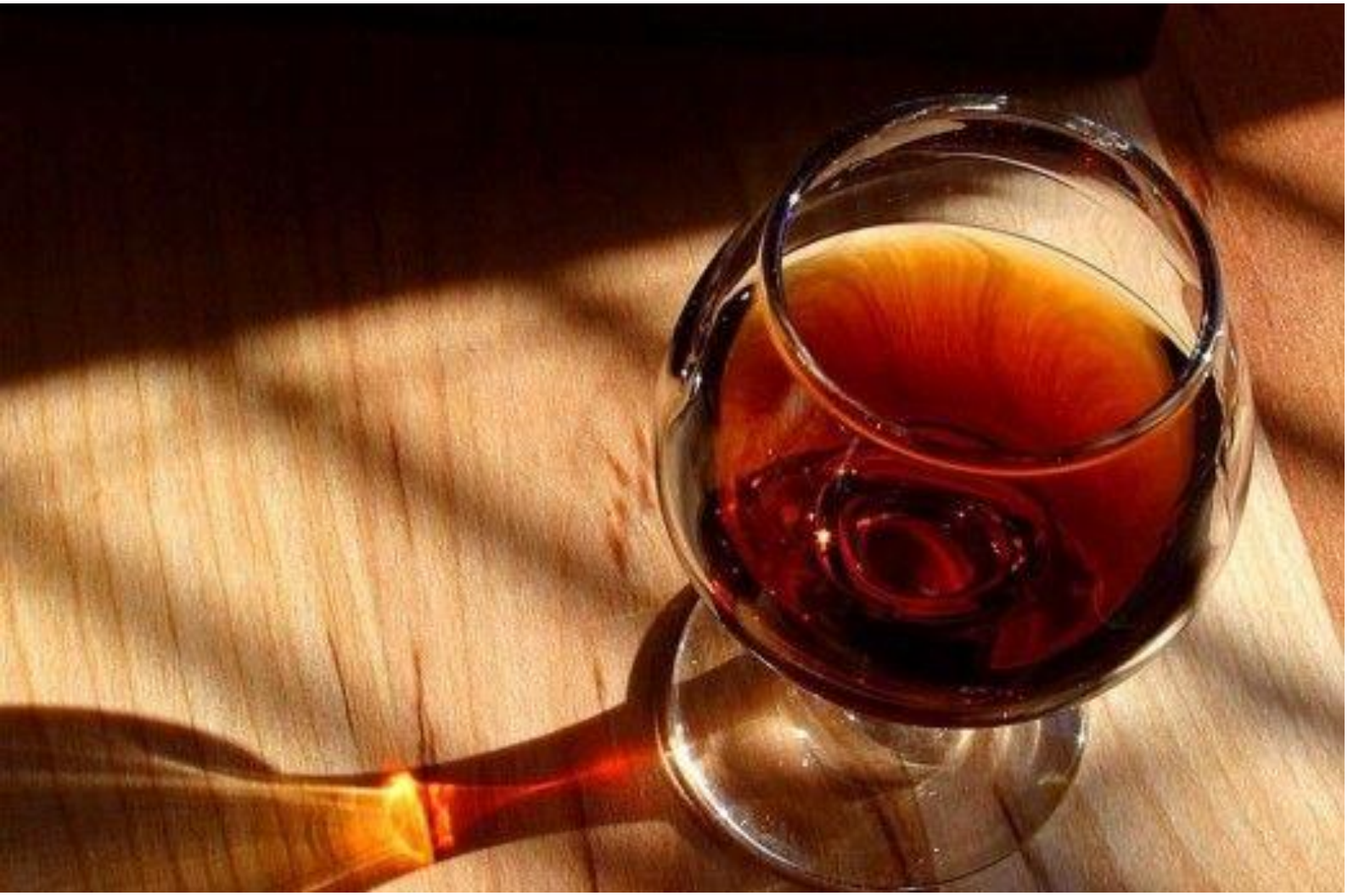
八、异常值检测

异常值指在所获得的数据中相对误差较大的观测值，也称离群值。

使用Megastat获取学生化外残差，其计算方式计算相当于重新进行n次回归，依次省略每个观测，并重新计算研究的残差。下图展示了部分异常观测值。数据样本总数为1599，其中

有87个观测值被标记，因为他们属于高杠杆点，杠杆值超过杠杆比率2(k+1)/n=2(10+1)/1599≈0.0138。说明这些记录的的一个或多个解释变量与该变量的平均值相差很大，但只有通过检查各个数据列，我们才能确定具体是哪些解释变量。有48条记录的残差值异常，即预测的酒精含量与实际的酒精含量之差的绝对值超过标准差的2倍。这些异常观测值可能对回归模型造成影响。下图是部分异常值的截取图像，由于数据条数过多，此处不做全部展示。

Observation	alcohol	Predicted	Residual	Leverage	Studentized Residual	Studentized Deleted Residual
14	9.10	10.15	-1.05	0.028	-1.798	-1.799
18	9.30	9.52	-0.22	0.028	-0.382	-0.382
20	9.20	9.26	-0.06	0.022	-0.102	-0.102
21	9.40	10.70	-1.30	0.005	-2.194	-2.196
34	9.40	10.70	-1.30	0.024	-2.213	-2.215
39	9.80	10.76	-0.96	0.016	-1.634	-1.635
43	10.50	9.70	0.80	0.022	1.354	1.355
44	10.30	10.63	-0.33	0.014	-0.557	-0.557
46	13.10	11.93	1.17	0.018	1.987	1.989
80	9.10	9.26	-0.16	0.016	-0.270	-0.270
82	9.40	9.32	0.08	0.044	0.137	0.137
84	9.40	9.18	0.22	0.031	0.381	0.381
127	10.90	11.16	-0.26	0.018	-0.447	-0.447
128	10.90	11.17	-0.27	0.018	-0.451	-0.451
132	13.00	11.50	1.50	0.008	2.535	2.539
133	13.00	11.50	1.50	0.008	2.535	2.539
143	14.00	12.83	1.17	0.013	1.974	1.975
145	14.00	12.83	1.17	0.013	1.974	1.975
152	9.40	8.36	1.04	0.097	1.847	1.848
162	9.20	9.36	-0.16	0.018	-0.264	-0.264
170	9.50	9.43	0.07	0.037	0.114	0.114
199	13.00	12.53	0.47	0.014	0.801	0.801
218	9.10	10.28	-1.18	0.002	-1.994	-1.996
227	9.50	10.70	-1.20	0.032	-2.051	-2.053
235	9.00	10.43	-1.43	0.007	-2.422	-2.426
240	9.00	10.43	-1.43	0.007	-2.422	-2.426



九、总结

经过以上模型建立及验证过程，最终的多元线性回归方程为：

alcohol=561.5896 + 0.4893*fixedacidity + 0.6161*volatileacidity + 0.8538*criticacid + 0.2601*residualsugar-0.9717*chlorides-0.0021*totalsulfurdioxide -571.9288* density + 3.5849*pH + 0.9485*sulphates + 0.2308*quality

模型R方值为0.691，意味着fixedacidity, volatileacidity, residualsugar, chlorides, totalsulfurdioxide, density, pH, sulphates, quality, citricacid解释变量可以解释alcohol的69.1%变化原因。模型截距无意义。