2012年电子科技大学中山学院优秀论文

葡萄酒等级划分体系模型的探究

摘　要

针对目前葡萄酒评价体系不完善的现状，本文对葡萄酒评价体系作出探究。

对于问题一，运用单因素方差分析法，利用 Matlab 软件，以 Anoval 函数求解。求出 p-value，显著性水平取 0.05 作为标准来判断那组有显著性，以及通过比较方差来判断那组数据更加可信。

对于问题二，在问题一中得到第二组评分更可信，因此根据该组的评分进行分级，通过用 Matlab 软件的 Corrcoef 和 Regress 函数对该组成分进行相关性验证和用 EXCEL 画出图表进行分析，找出影响葡萄酒分级的成分，然后在酿酒葡萄数据中找出与影响葡萄酒分级相同的成分，再结合葡萄酒评分对葡萄样品进行分级，得出葡萄样品成分的排列，结合成分的量和葡萄酒分级得出影响酿酒葡萄分级成分的范围。

对于问题三，通过问题二的解答，可以知道葡萄酒和酿酒葡萄的划分级别，利用附件二的资料，对每一种理化指标的数据，根据对应的含量建立模型，运用 matlab 软件拟合数据，作出拟合线性图，并采用多元回归分析法进行回归分析，最后根据拟合线性图和回归系数来分析两类理化指标之间的关系。

对于问题四，分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。结合题目给出芬香物质的数据，对感官指标和理化指标进行综合分析，用 MABTLE 拟合感官指标和理化指标的数据，得出结论：需要结合葡萄酒的理化指标和感官指标对葡萄酒的质量进行综合评价。

关键字：方差分析法　　分级　　理化指标　　线性相关　　回归分析

一、问题的重述

随着我国经济的快速发展，葡萄酒市场竞争也异常激烈和无序 “三精一水”、假年份、假产地酒、假酒庄，影响消费者的健康，虽然我国的 GB15037-2006《葡萄酒》国家标准对葡萄酒的质量作了规定，但由于相应规范的制定工作限制，我国关于葡萄酒质量等级分划的标准还未完善，国家迫切需要制定统一的质量等级制度。

确定葡萄酒质量时一般是通过聘请一批有资质的评酒员进行品评。每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到其总分，从而确定葡萄酒的质量。酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系，葡萄酒和酿酒葡萄检测的理化指标会在一定程度上反映葡萄酒和葡萄的质量，文章给出了某一年份一些葡萄酒的评价结果及该年份这些葡萄酒的和酿酒葡萄的成分数据。本文尝试解决以下问题：

问题一： 由于评酒师对葡萄酒的评分存在主观性，需对评酒师的分数进行客观分析，分析两组评酒员的评价结果有无显著性差异，哪一组结果更可信？

问题二：葡萄酒的质量离不开原料酿酒葡萄的质量，所以酿酒葡萄的理化指标至关重要。需根据酿酒葡萄的理化指标和葡萄酒的质量对这些酿酒葡萄进行分级。

问题三：酿酒葡萄与葡萄酒的理化指标之间的联系可能影响着葡萄酒质量，所以需建立模型，酿酒葡萄与葡萄酒的理化指标之间的联系。

问题四：分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量，能否综合感官指标和理化指标，建立模型，来评价葡萄酒的质量是问题关键所在。

二、模型假设

1. 品酒员打分相互之间没有影响；
2. 品酒员对样品的给的总分是他对该样品所有方面评分的总和，并且该样品的最终得分可认为是 10 位品酒员打分的平均值；
3. 题目所给的数据真实可靠；
4. 酿酒方式及酿酒过程对葡萄酒的质量没有影响；
5. 不同种类葡萄酒的成份数据值统一标准没有差异 ；
6. 所有样品的酿造过程相同。

三、符号说明

n	测试数量
r	测试水平量
A	因素

SS	各类数据源的平方和
Df	各类数据相应的自由度
MS	各类的均方值
F	统计量
P	大于 F 的概率
S_A	各组均值对总方差的偏差平方和
S_E	各组数据对均值偏差平方和的总和

四、问题分析

问题一的分析

我们要根据附件 1 的数据可知：评酒员对红酒 27 组样品，和白酒 28 组样品进行评分，每件样品都进行了两次评分，即是有两组评分数据，题目要求分析两组评酒员的评分结果有无显著性差异，以及那一组数据更加可信，对于显著性的判断，我们采用单因素方差分析法（Analysis Of Variance）。对于每件样品，评酒员对外观，香气，口感，及其整体评价进行打分，每一组的每件样品都有十名品酒员进行评分，故求每个品酒员对样品酒的总分，之后求出这十名品酒员给的总分的平均分，此平均分就是该样品的总分，葡萄酒分为白酒和红酒，我们对第一组的红酒和第二组的红酒进行方差分析法，运用 matlab 软件中的 anoval 函数可得出 p-value，及 F 值，通过分析就可知道那组更加具有显著性。方差是考察数据的波动性的，方差小就说明数据比较稳定，方差大就是波动性比较大 ，故通过比较两组数据的方差大小，就知道那一组数据更加可信。

问题二的分析

根据问题一可知，第二组的评酒员的评酒分数更可靠，所以选择第二组葡萄酒的数据进行处理。从评酒员对葡萄酒评分的分数入手，用逆向思维反推葡萄的等级。首先将第一问中第二组的白葡萄酒和红葡萄酒的每一种样品的评分进行分等级，依次分为四个等级，然后用 EXCEL 将每个等级的样品酒的理化指标画成曲线图，忽略异常数据点，观察各等级间的理化指标有没有相关性，如果有相关性，找出影响葡萄酒质量的相关因素，跟酿酒葡萄的理化指标数据进行对照，得出酿酒葡萄的分级依据。

问题三的分析

结合葡萄酒和酿酒葡萄的理化指标，作出每两个理化指标间的直观趋势图，观察两者之间的大体关系，根据曲线拟合的方法得出两者间的函数关系。

问题四的分析

由第三问求解可得出酿酒葡萄与葡萄酒的理化指标之间是呈线性相关的，因此我们要证明酿酒葡萄和葡萄酒的理化指标对葡萄酒质量是有影响的，只需证明酿酒葡萄的理化指标对葡萄酒质量是有影响。在综合附录 3 给出的芬香物质，用 MABTLE 拟合出理化指标和感官指标的关系图呈相关性，所以要综合葡萄酒的理化指标和感官指标一起来评价葡萄酒的质量。

五、模型建立与求解

5.1 问题一的模型建立和求解

对于两组评酒员的评价结果有无显著性差异，我们采用单因素方差分析法去解决。

单因素方差分析法:

只考虑一个因素A 对所关心的指标的影响，A 取几个水平，在每个水平上作若干个试验，试验过程中除A 外其它影响指标的因素都保持不变（只有随机因素存在),我们的任务是从试验结果推断，因素A 对指标有无显著影响，即当A 取不同水平时指标有无显著差别。A 取某个水平下的指标视为随机变量,判断A 取不同水平时指标有无显著差别，相当于检验若干总体的均值是否相等。

设 A 取 n 个水平 $A_1, A_2, A_3, \ldots A_n$，在水平 A_i 下总体 x_i 服从正态分步 $N(u_i, \sigma^2)$,i=1,...,n,这里 u, σ^2 未知， u_i 可以互不相同，但假定 x_i 有相同的方差，又设在每个水平 A_i 下作了 n_i 次独立试验，即从中抽取容量为 n_i 的样本，记作 $x_{ij}, j = 1,..., n_j, x_{ij}$ 服从 $N(u_i, \sigma^2)$， i=1,…,n,j=1,…, n_i 且且相互独立。将这些数据列成表 1（单因素试验数据表）的形式。

表 5.1 单因素试验数据表

分值	第一组红酒	第二组红酒	第一组白酒	第二组白酒
A1	X12	X21	X12	X21
A2	X21	X22	X21	X22
.....				
A3	X31	X32	X31	X32

根据上述理论，首先我们对数据进行处理，附件 1 里有四组数据：红葡萄酒和白葡萄酒各有两组数据，每种酒都有两组人进行对其进行评分，每件样品酒有十名

品酒员号打分，采用单因素方差分析法，我们将样品酒的总分作为唯一考虑的因素 A，运用 matlab 软件编程求出品酒员对每组样品打的总分的平均分，见下表：

表 5.2 组样品红酒和白酒的总分

样品号	第一组红葡萄酒品尝综合得分评分	第一组白葡萄酒品尝综合得分评分	第二组红葡萄酒品尝综合得分评分	第二组白葡萄酒品尝综合得分评分
1	62.7	82	68.1	77.9
2	80.3	74.2	74	75.8
3	80.4	79.7	74.6	75.6
4	68.6	79.4	71.2	76.9
5	73.3	71	72.1	81.5
6	72.2	68.4	66.3	75.5
7	71.5	77.5	65.3	74.2
8	72.3	71.4	66	72.3
9	81.5	72.9	78.2	80.4
10	74.2	74.3	68.8	79.8
11	70.1	72.3	61.6	71.4
12	53.9	63.3	68.3	72.4
13	74.6	65.9	68.8	73.9
14	73	72	72.6	77.1
15	58.7	72.4	65.7	78.4
16	74.9	74	69.9	67.3
17	79.3	78.8	74.5	80.3
18	59.9	73.1	65.4	76.7
19	78.6	72.2	72.6	76.4
20	79.5	77.8	75.8	76.6
21	77.1	76.4	72.2	79.2
22	77.2	71	71.6	79.4
23	85.6	75.9	77.1	77.4
24	78	73.3	71.5	76.1
25	69.2	77.1	68.2	79.5
26	73.8	81.3	72	74.3
27	73	64.8	71.5	77
28		81.3		79.6

对这四组数据，我们将白酒和红酒分开来判断其有无显著性，即第一组红酒与第二组红酒，第一组白酒和第二组白酒比较。

运用matlab软件对数据处理编程得出以下结果，标准ANOVA表分析见下表：

表5.3白葡萄酒ANOVA表

Source	SS	df	MS	F	Prob>F
Columns	86.034	1	86.0339	5.11	0.0278
Error	909.11	54	16.8354		
Total	995.144	55			

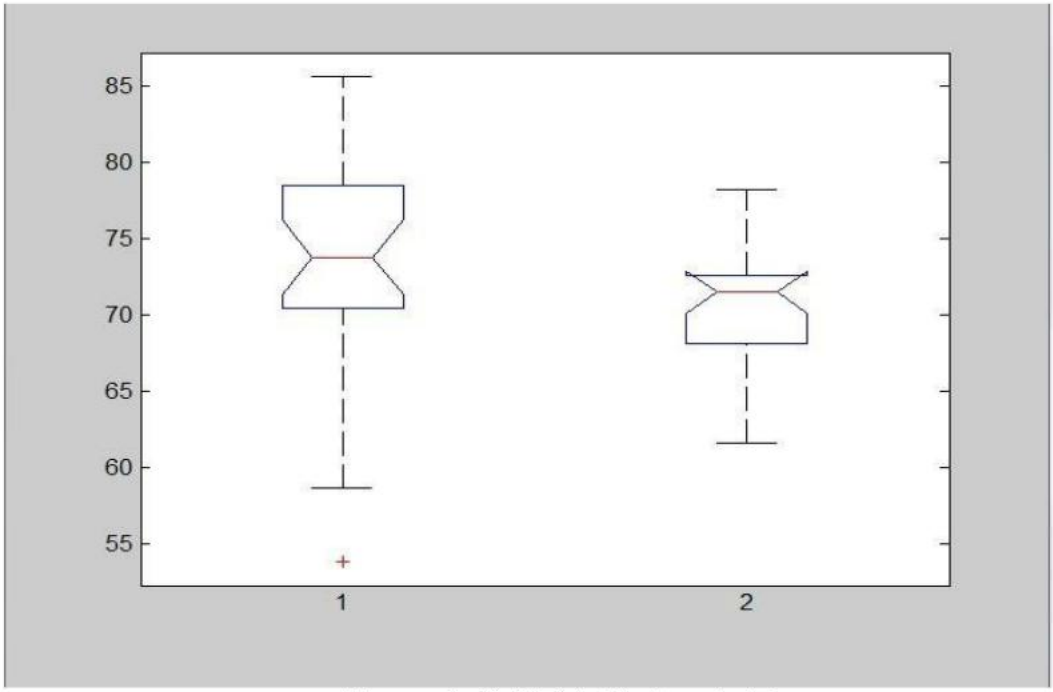


图5.1 白葡萄酒盒型（box）图

表 5.4 红葡萄酒 ANOVA 表

Source	SS	df	MS	F	Prob>F
Columns	88.94	1	88.935	2.55	0.1164
Error	1813.75	52	34.8799		
Total	1902.69	53			

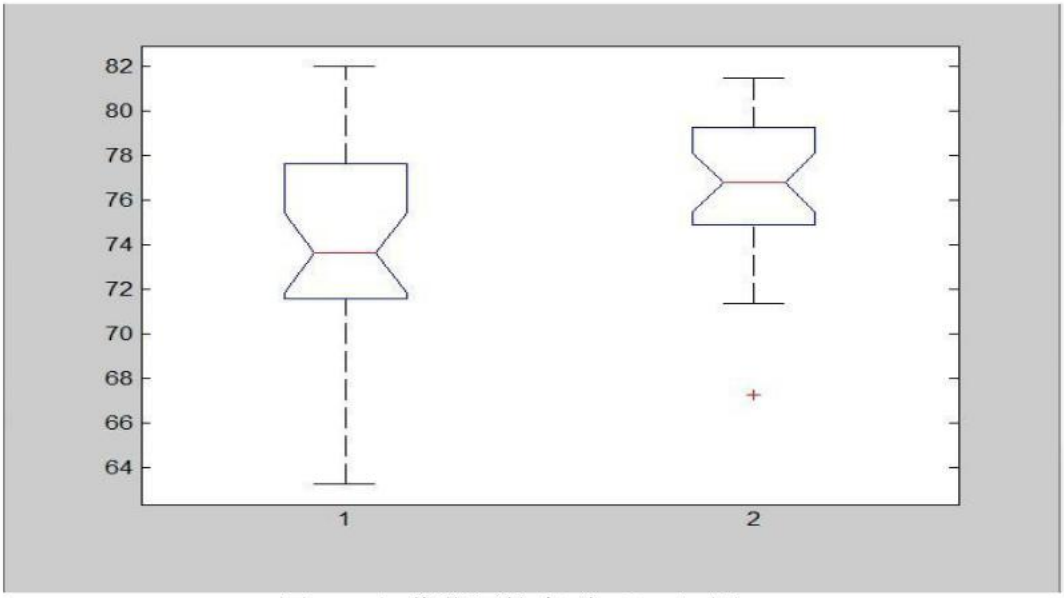


图5.2 红葡萄酒的盒型（box）图

表 5.5 方差分析表：

方差来源	平方和(SS)	自由度 (df)	均方（MS）	1-P 分数位 F	概率 p
因素 A	S_A	r-1	$\bar{S}_A = \frac{S_A}{r-1}$	$F_{1-p_r}(r-1, n-r)$	p_r
误差	S_E	n-r	$\bar{S}_E = \frac{S_E}{n-1}$		
总和	S_T				

通常情况下，实验结果 p 达到 0.05 水平或 0.01 水平，才可以说数据之间具备了差异显著或是极显著。在作结论时，应确实描述方向性（例如显著大于或显著小于）。sig 值通常用 P>0.05 表示差异性不显著。在此我们去 0.05 作为显著性水平标准，红酒中的 ANOVA 表中 Prob>F 栏 p 值为 0.0278 <0.05,故拒绝 Ho，且盒型图的中心线差差别不大，对应的 F 也很小，故可知品酒员对白酒的评分具有显著性。红葡萄酒酒中的 ANOVA 表中的 P>0.05，接受 Ho，故没有显著性。

对于那组数据更加可信，我们知道方差是考察数据的波动性的，方差小就说明数据比较稳定，方差大就是波动性比较大。故我们将红酒，白酒每组样品酒一一对应，第一组的红酒中样品一与第二组红酒中的样品一进行方差分析，以此类推，我们将所求到的方差用 matlab 进行画图。

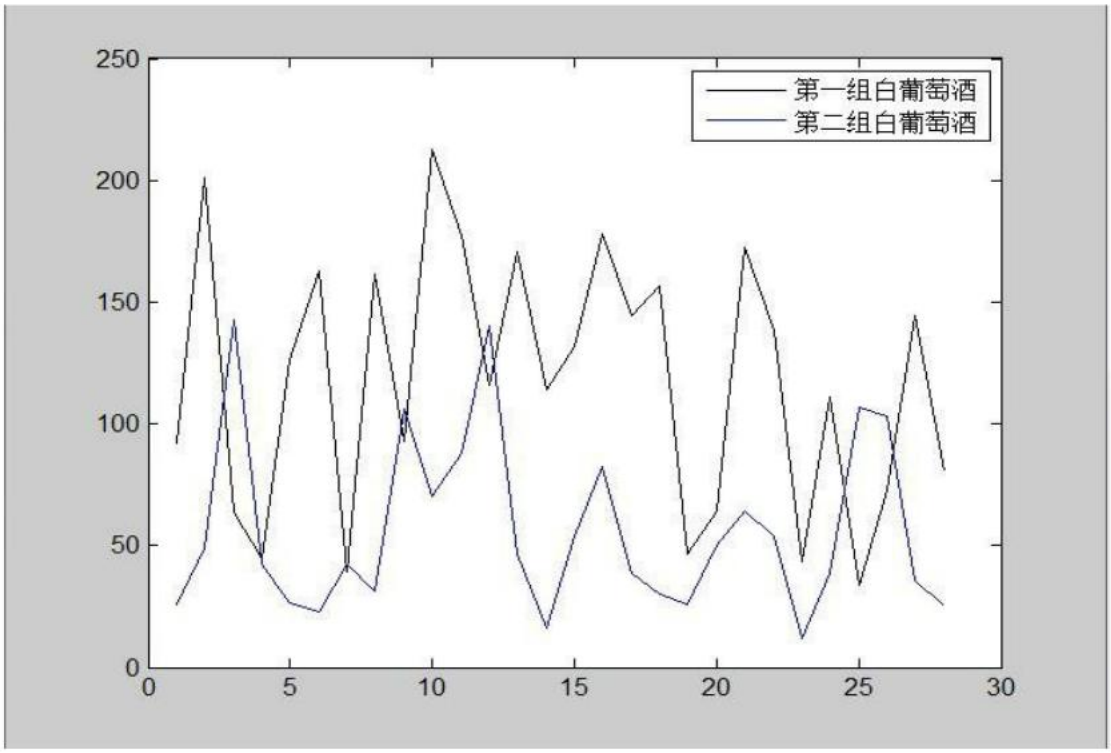


图 5.3 白葡萄酒的方差图

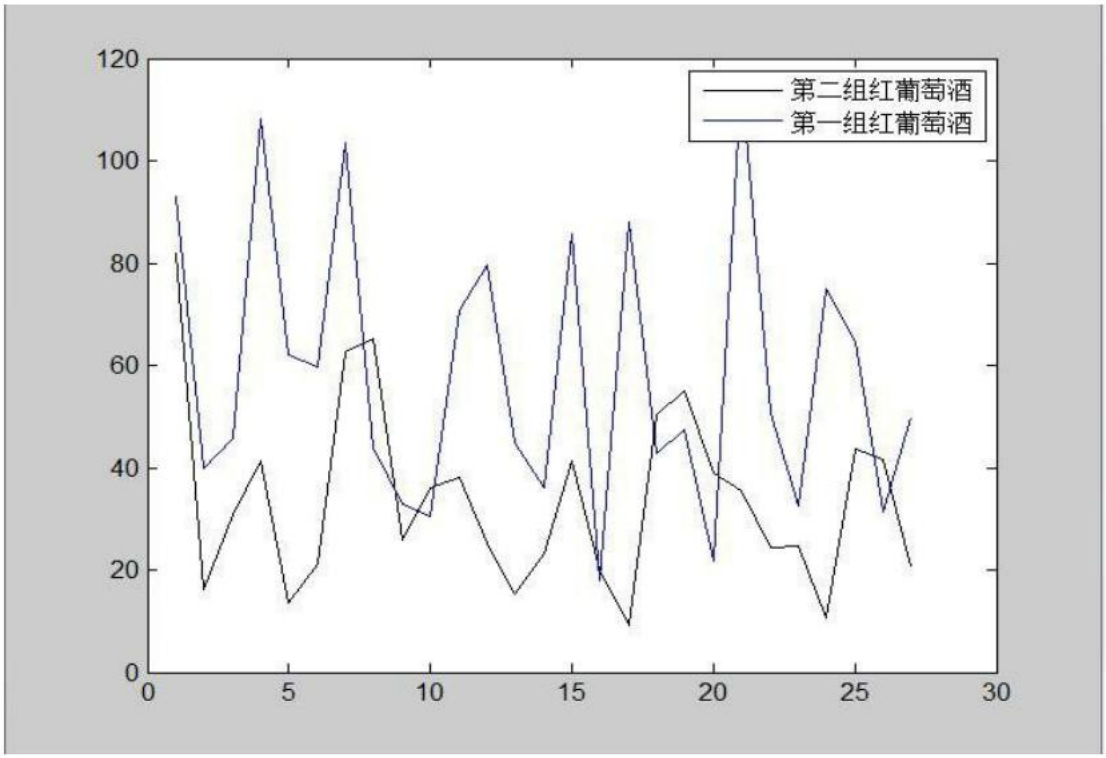


图 5.4 红葡萄酒的方差图

从两副图中，我们很明显的看到第二组数据的方差基本小于第一组数据，因此我们认为第二组数据更加可信。

5.2 酿酒葡萄的分级

5.2.1白葡萄酒的分级

通过统计第二组白葡萄酒的每个样品的分数，将白葡萄酒分为四个等级。

第一等级	（75，80]
第二等级	（70，75]
第三等级	（65，70]
第四等级	（0，65]

由分数等级标准可得到各个样品酒处于的那个等级段，表格如下：

第一等级	9, 23, 20
第二等级	3, 17, 2, 14, 11, 21, 5, 26, 22, 24, 27, 4
第三等级	16，10, 13, 12, 25，1, 6，8, 15, 18, 7
第四等级	11

上述各个等级的样品所对应的各个理化指标的关系图如下（其中每个图的横坐标不是样品号，而是依次每个等级中的样品，从左到右依次为一，二，三，四等级的样品）：

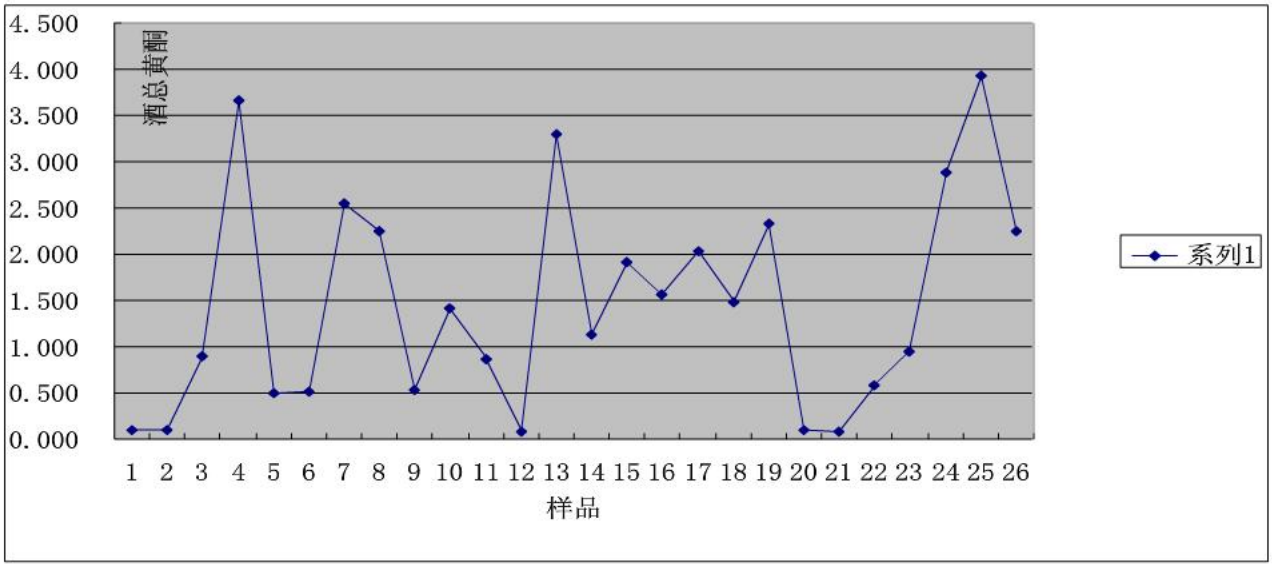


图 5.5 各个等级中各样品的酒总黄酮含量

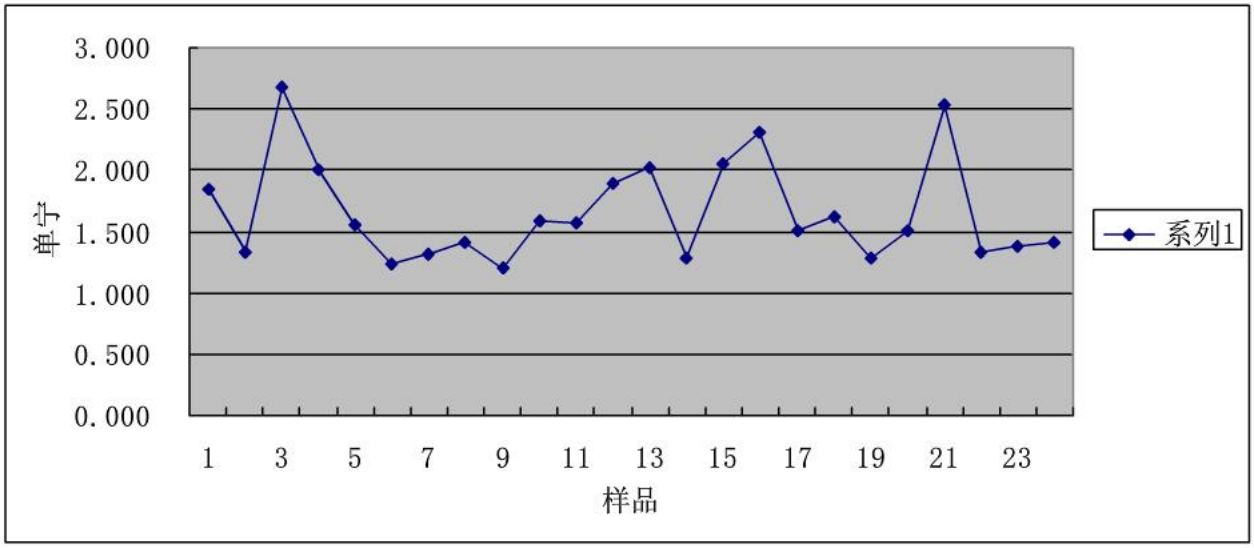


图 5.6 各个等级中各样品的单宁含量

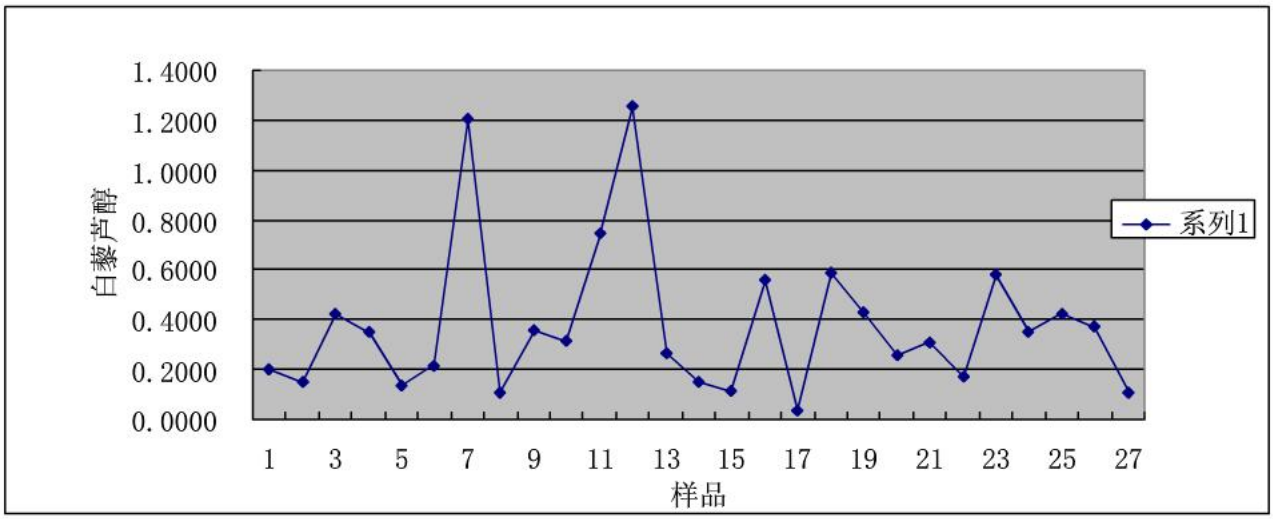


图 5.7 各个等级中各样品的白藜芦醇含量

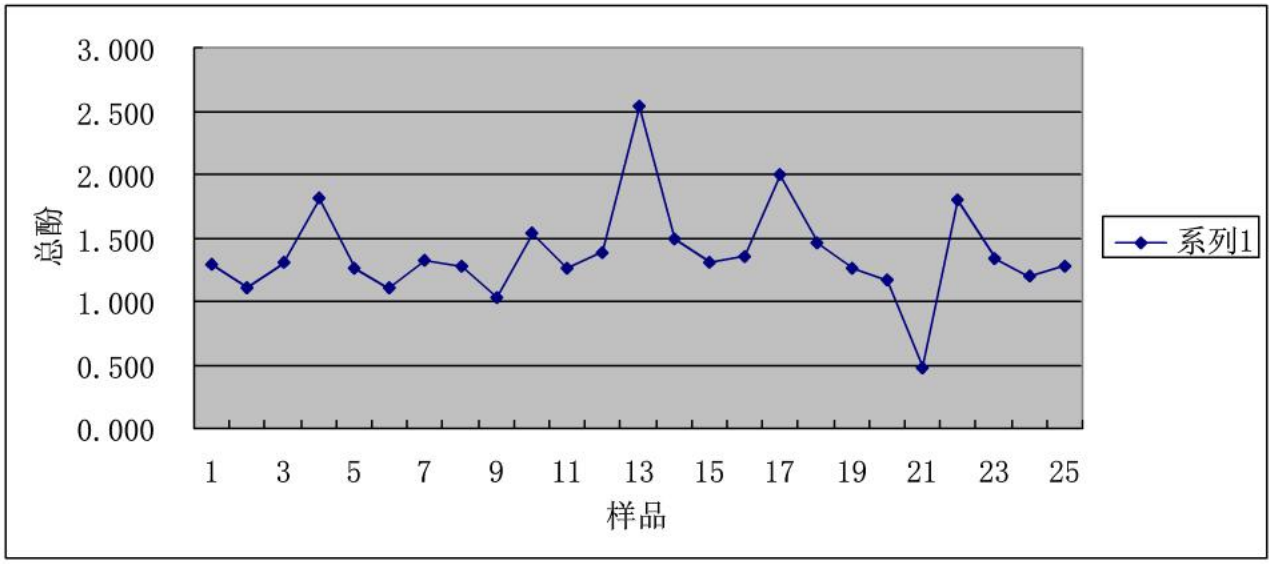


图 5.8 各个等级中各样品的总酚含量

用 MABTLE 软件对以上四幅图进行相关性分析，由 corrcoef 得出四幅图的结果的绝对值都非常不接近 1，且 regress 函数的出 stats 中的 p 远大于 0.05 故可知无相关性，其中酒总黄酮的相关系数为-0.0892，单宁的相关系数为-0.114，白藜

芦醇的相关系数为 0.2596，总酚的相关系数为-0.0391,四种因素与样品皆无相关性。
说明：相关系数的绝对值在 0 到 0.3 的呈无相关性，0.3 到 0.8 的呈弱相关性，0.8 到 1 呈强相关性）

5.2.2 红葡萄酒的分级

通过统计第二组红葡萄酒的每个样品的分数，将红葡萄酒分为四个等级。

第一等级	（74，78]
第二等级	（70，74]
第三等级	（66，70]
第四等级	（0，62]

由分数等级标准可得到各个样品酒处于的那个等级段，表格如下：

第一等级	9, 23，20, 3, 17
第二等级	2, 19, 14, 21, 5, 26, 22, 24, 27, 4
第三等级	16, 13, 10, 12, 25, 1, 6
第四等级	23, 15，18, 7, 11

上述各个等级的样品所对应的各个理化指标的关系图如下（其中每个图的横坐标不是样品号，而是依次每个等级中的样品，从左到右依次为一，二，三，四等级的样品，并且去掉异常数据）：

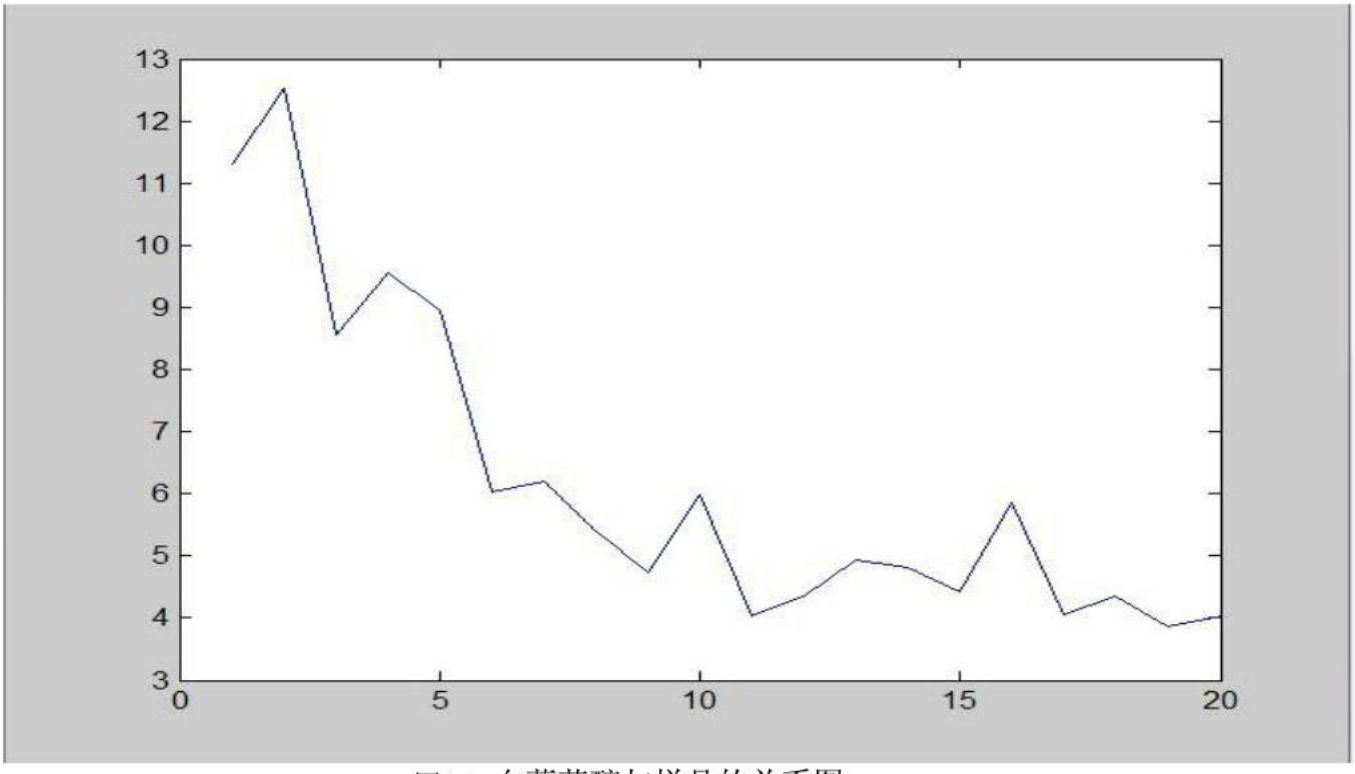


图5.9 白藜芦醇与样品的关系图

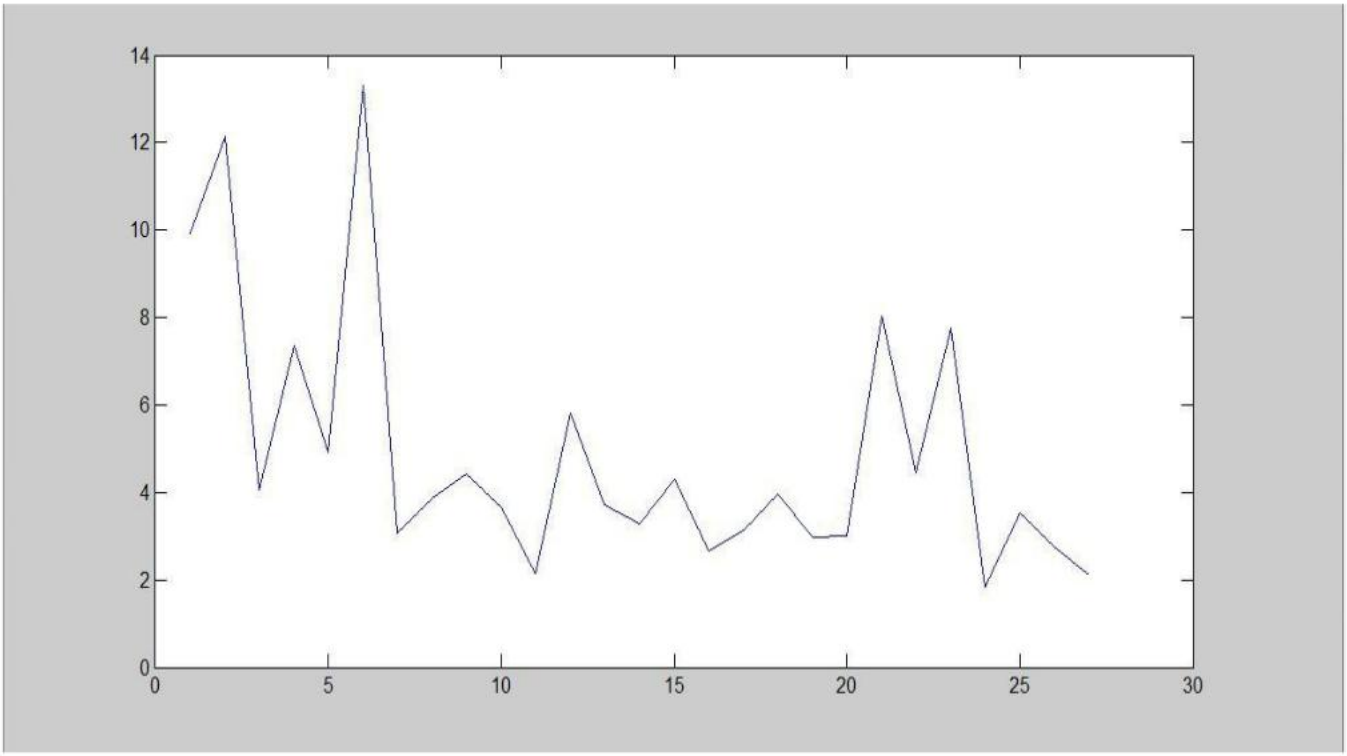


图5.10 黄酮与样品的关系图

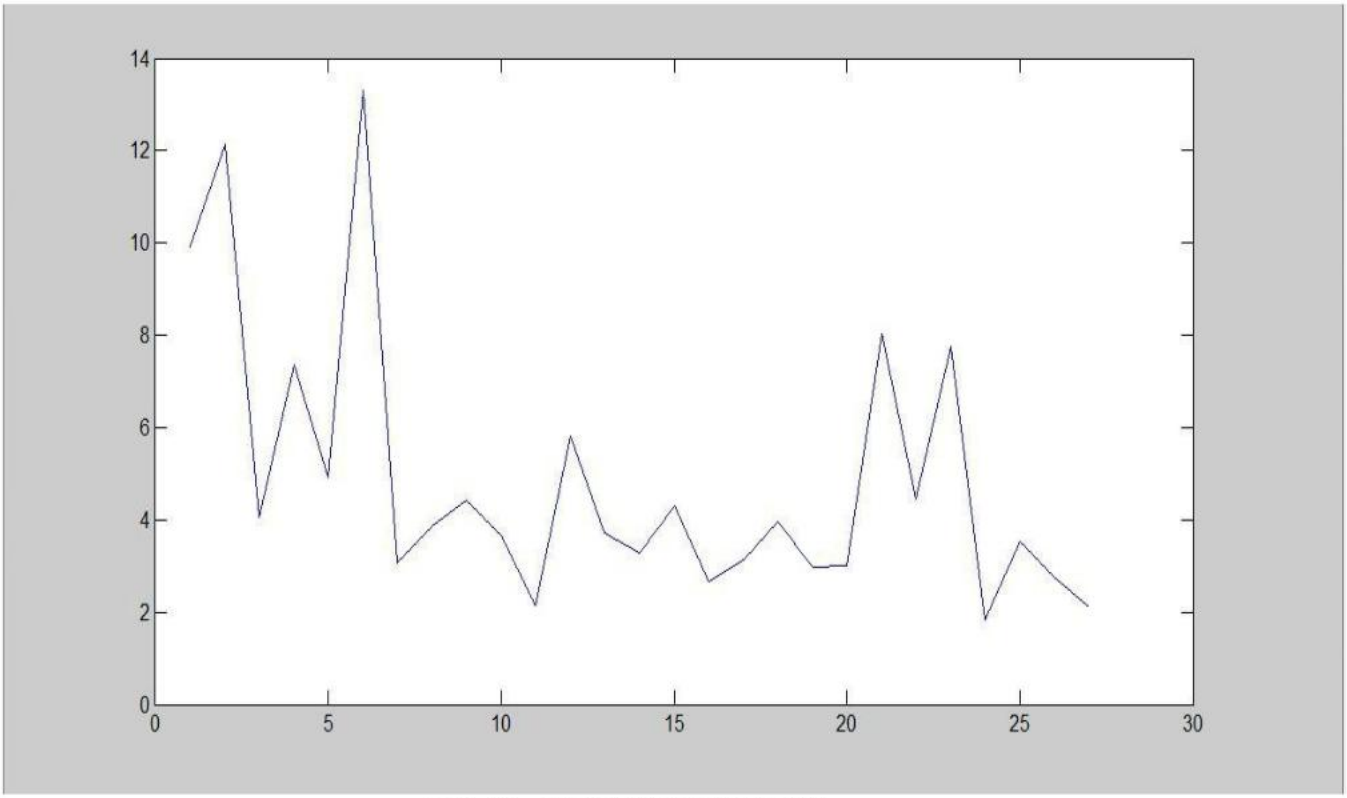


图5. 11 总酚与样品的关系图

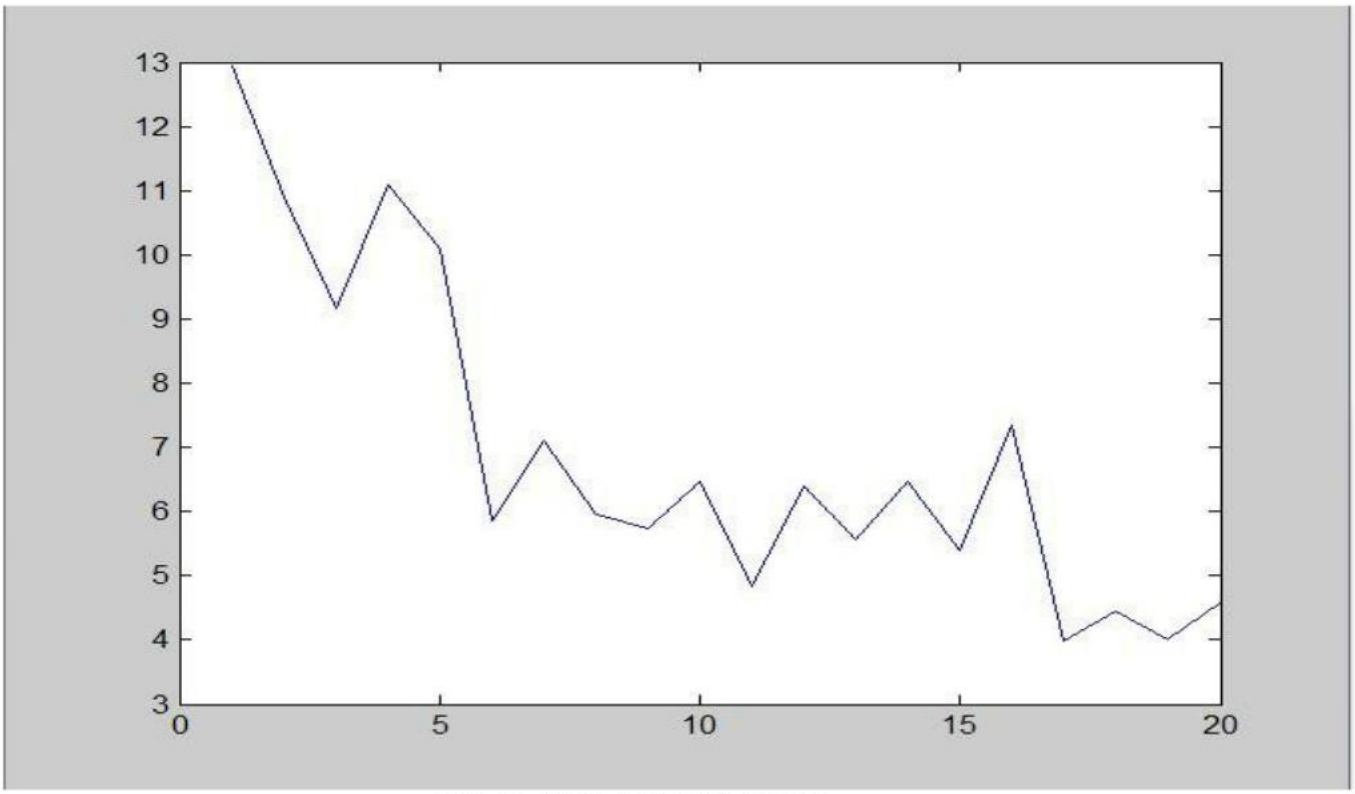


图 5.12 单宁与样品的关系图

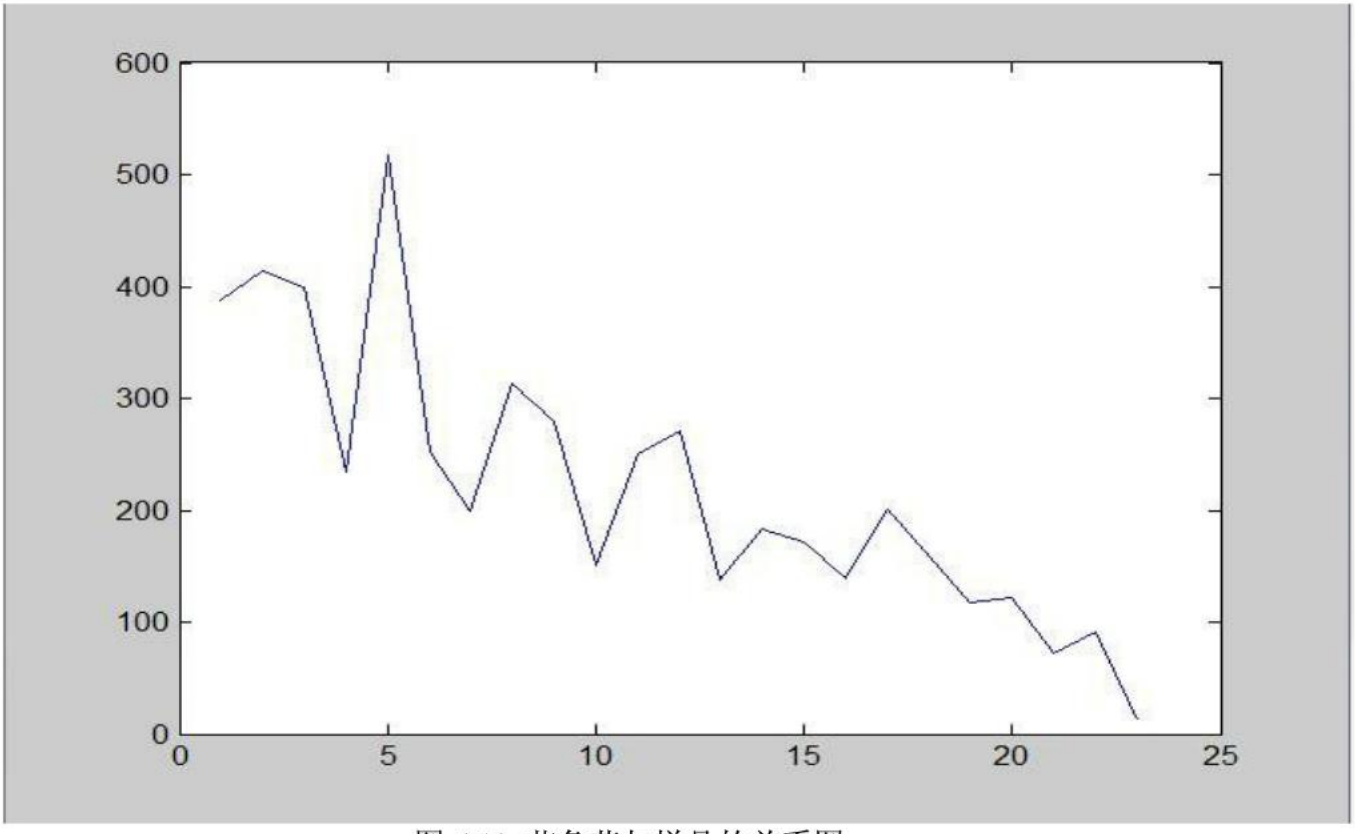


图 5.13 花色苷与样品的关系图

用 MABTLE 软件对以上四幅图进行相关性分析，由 corrcoef 得出四幅图的结果的绝对值都接近 1，且 regress 函数的出 stats 中的 p 小于 0.05 故可知有相关性，其中单宁的相关系数到达-0.8278，总酚的相关系数达到-0.8341，花色苷的相关系数达到-0.8533，呈强相关性。白藜芦醇的相关系数为-0.508，酒总黄酮的相关系数为-0.486，呈弱相关性。各图的代码如附录 2.

综上所述，影响红葡萄酒质量的等级的因素有单宁，总酚，花色苷，这三个因素直接影响了红葡萄酒的分级，但红葡萄酒的分级也直接影响了酿酒葡萄的质量分级。

下面通过红葡萄酒的理化指标结合酿酒葡萄的数据对酿酒葡萄进行分等级：

表 5.6 影响红葡萄酒分级的因素成分数据表

影响红葡萄酒分级的成分数据表				
红葡萄酒的样品	花色苷	总酚	单宁	所属的等级
葡萄样品9	240.843	80.741	12.933	第一等级
葡萄样品23	172.626	153.698	10.888	
葡萄样品20	23.523	58.922	5.864	
葡萄样品3	157.939	49.662	13.259	第二等级
葡萄样品17	59.424	67.062	9.170	
葡萄样品2	224.367	124.863	11.078	
葡萄样品14	140.257	108.190	6.073	
葡萄样品19	115.704	50.321	5.981	
葡萄样品21	89.282	238.064	10.090	
葡萄样品5	120.606	66.962	5.849	
葡萄样品26	58.469	63.033	3.615	
葡萄样品22	74.027	56.243	7.105	
葡萄样品27	34.190	34.694	5.961	
葡萄样品24	144.881	59.246	5.747	
葡萄样品4	79.685	116.270	6.477	
葡萄样品16	60.660	65.280	4.832	第三等级
葡萄样品13	65.324	52.845	6.385	
葡萄样品10	44.203	96.353	5.567	
葡萄样品12	32.343	72.813	6.458	
葡萄样品25	49.643	52.814	5.406	
葡萄样品1	408.028	159.522	11.030	
葡萄样品6	46.186	149.183	7.354	
葡萄样品8	241.397	94.152	12.028	
葡萄样品15	52.792	75.443	3.985	
葡萄样品18	40.228	56.502	4.447	
葡萄样品7	60.767	106.428	4.014	
葡萄样品11	7.787	65.235	4.588	第四等级

由上表的数据我们可把酿酒红葡萄进行划分等级，如下：

表 5.7 酿酒葡萄酒的等级划分表

酿酒红葡萄的类别	各成分的范围值（mg/100g 鲜重）			
酿酒红葡萄	花色苷 mg/100g 鲜重	总酚(mmol/kg)	单宁 (mmol/kg)	等级
	[70-240)	[150-200)	[11-15)	一级
	[60-170)	[100-150)	[7-11)	二级
	[20-60)	[50-100)	[4-7)	三级
	[0-20)	[0-50)	[0-4)	四级

5.3 酿酒葡萄和葡萄酒理化指标的关系

通过观察酿酒葡萄和理化指标的数据，用 MABTLE 将数据进行处理，将酿酒葡萄和葡萄酒的相同的理化指标的数据进行拟合，得到以下图形：

5.3.1 将酿酒葡萄和葡萄酒中的花色苷数据进行拟合，得到下图：

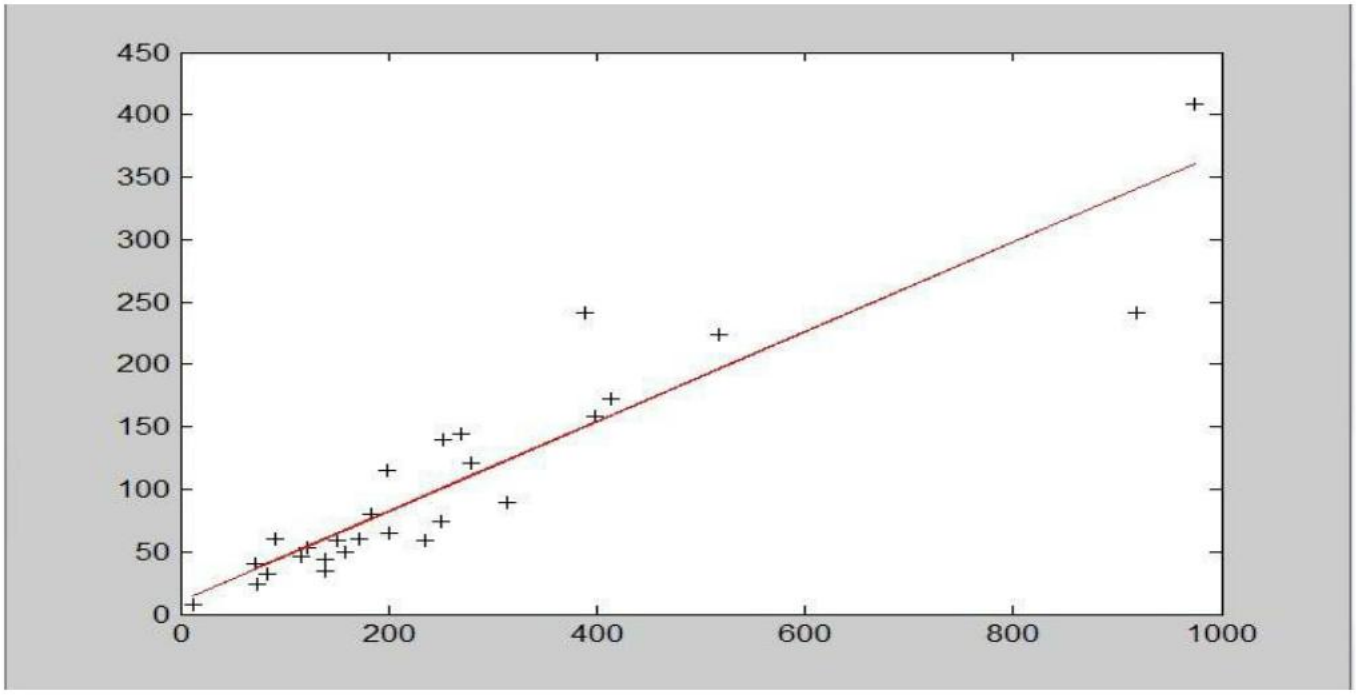


图 5.14 酿酒葡萄和葡萄酒中的花色苷数据拟合图

5.3.2 将酿酒葡萄和葡萄酒中的单宁数据进行拟合，得到下图：

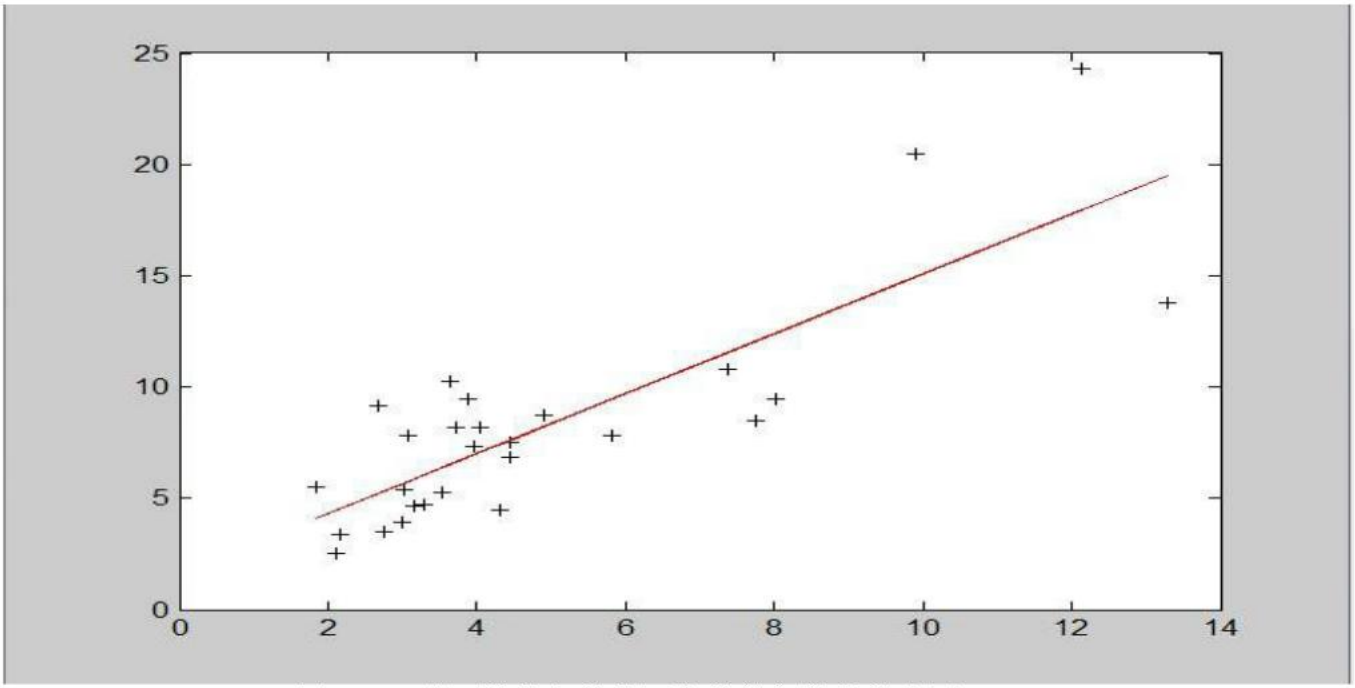


图 5.15 酿酒葡萄和葡萄酒中的单宁数据拟合图

5.3.3 将酿酒葡萄和葡萄酒中的葡萄总黄酮数据进行拟合，得到下图：

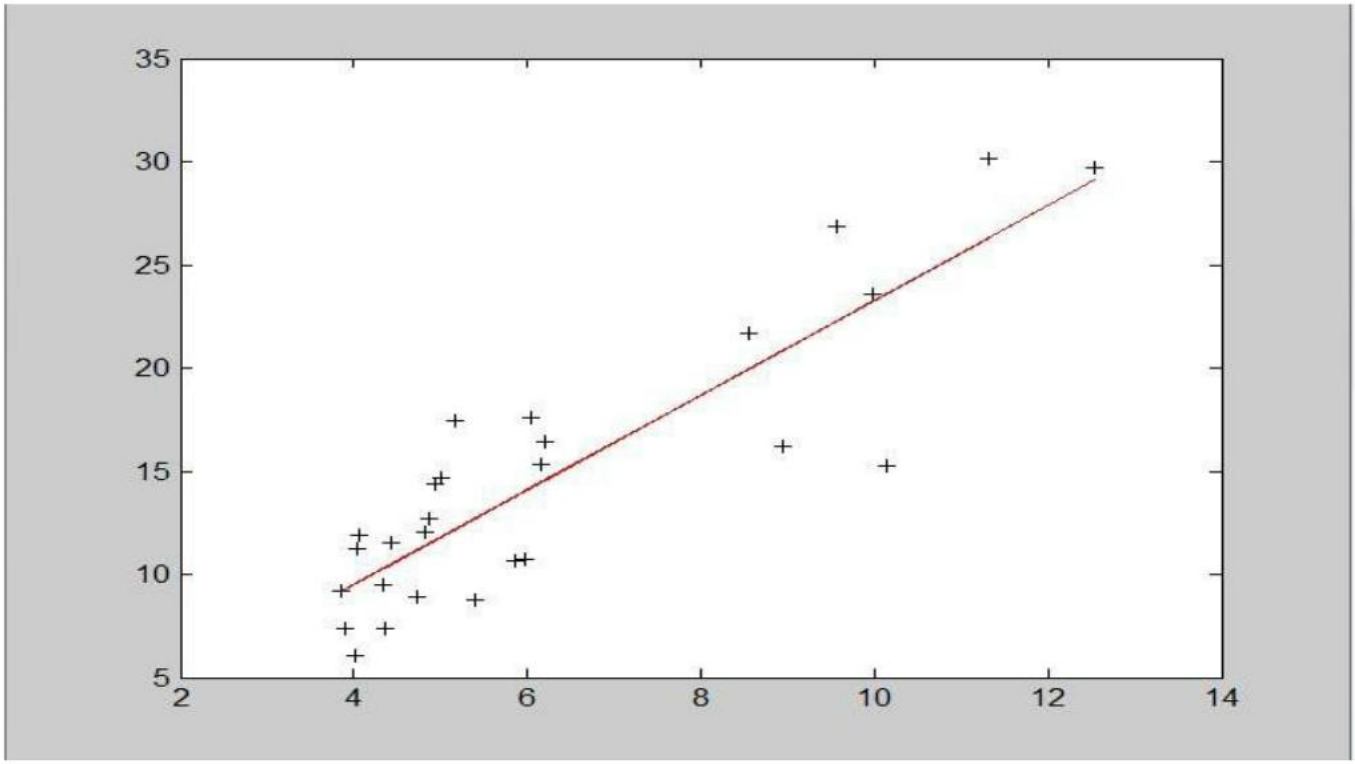


图 5.16 酿酒葡萄和葡萄酒中的葡萄总黄酮数据拟合图

5.3.4 将酿酒葡萄和葡萄酒中的白藜芦醇数据进行拟合，得到下图：

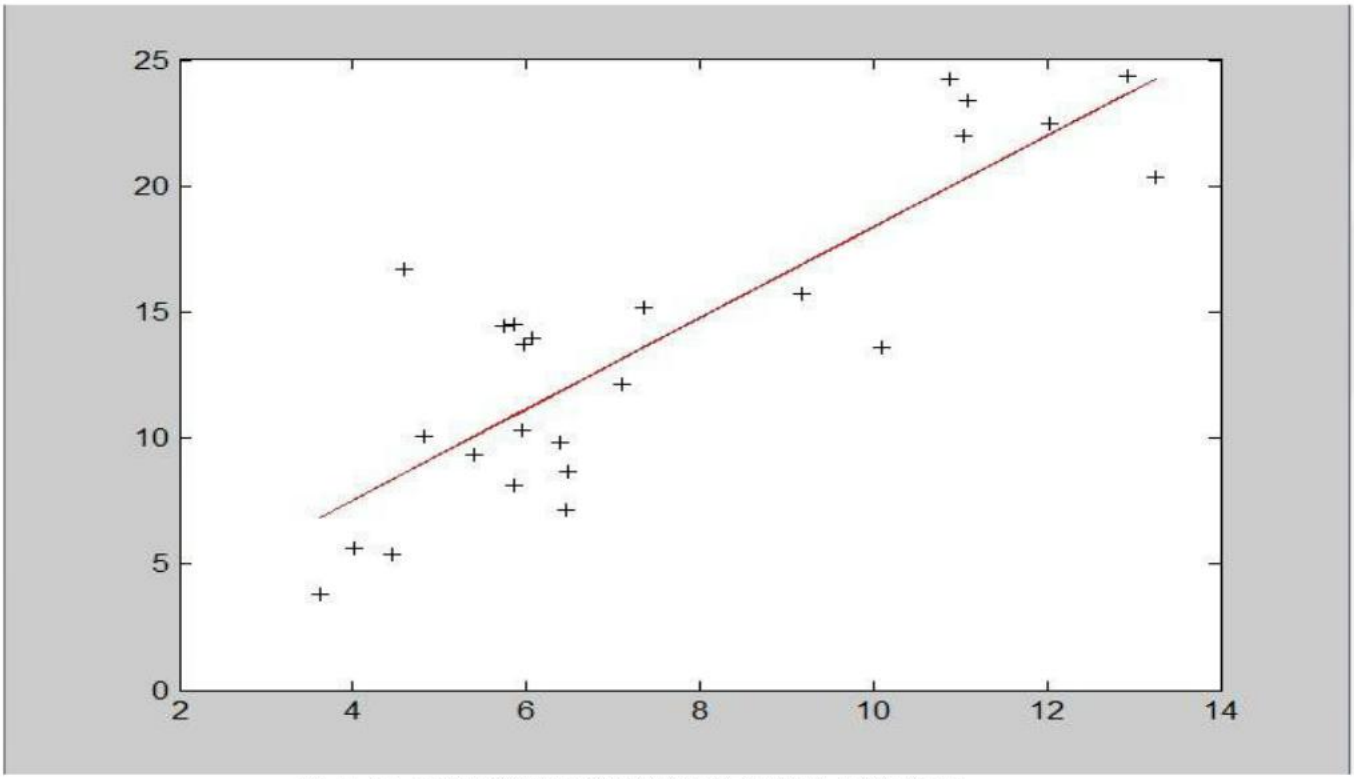


图 5.17 酿酒葡萄和葡萄酒中的总酚数据拟合图

5.3.5 将酿酒葡萄和葡萄酒中的白藜芦醇数据进行拟合，得到下图：

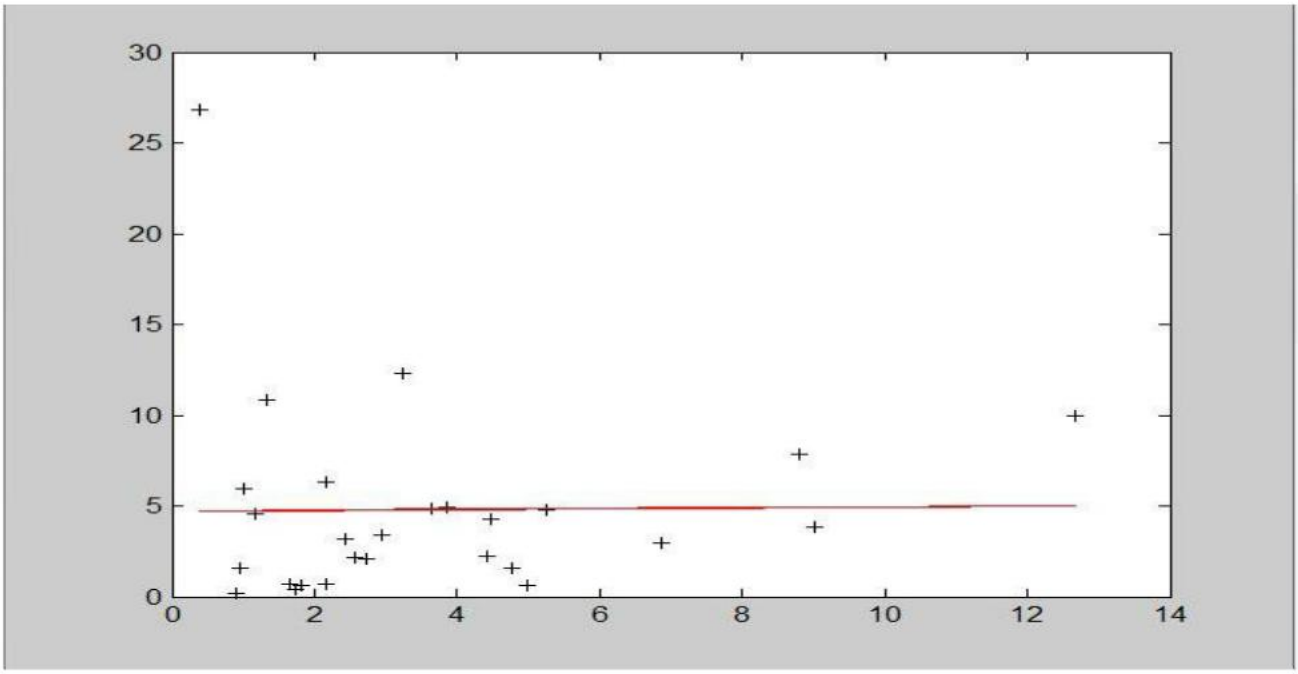


图 5.18 酿酒葡萄和葡萄酒中的白藜芦醇数据拟合图

由以上四幅图可知：
一．酿酒葡萄和葡萄酒中花色苷，单宁，葡萄总黄酮，总酚这四个因素呈线性关系，是正相关性。即酿酒葡萄中的花色苷含量，所酿出来的葡萄酒中花色苷含量越高；酿酒葡萄中的单宁含量，所酿出来的葡萄酒中单宁含量越高；酿酒葡萄中的葡萄总黄酮含量，所酿出来的葡萄酒中葡萄总黄酮含量越高；酿酒葡萄中的总酚含量，所酿出来的葡萄酒中总酚含量越高；
二．酿酒葡萄和葡萄酒中白藜芦醇对两者不影响

5.4 验证理化指标是否能成为葡萄酒等级评价依据

由第三问求解可得出酿酒葡萄与葡萄酒的理化指标之间是呈线性相关的，因此只需证明酿酒葡萄和葡萄酒的理化指标其中一种对葡萄酒质量是有影响，则可证明两者的理化指标是对葡萄酒质量是有影响的，因此我们用 Matlab 进行数据拟合来画其关系图，并用多元线性回归来判断其是否成线性关系。

我们先对葡萄酒的评分按照从低到高排列，相对应酿酒葡萄的各成分也得到相应的排列，横坐标是葡萄酒的评分，纵坐标是是酿酒葡萄的各成分，各图表如下：

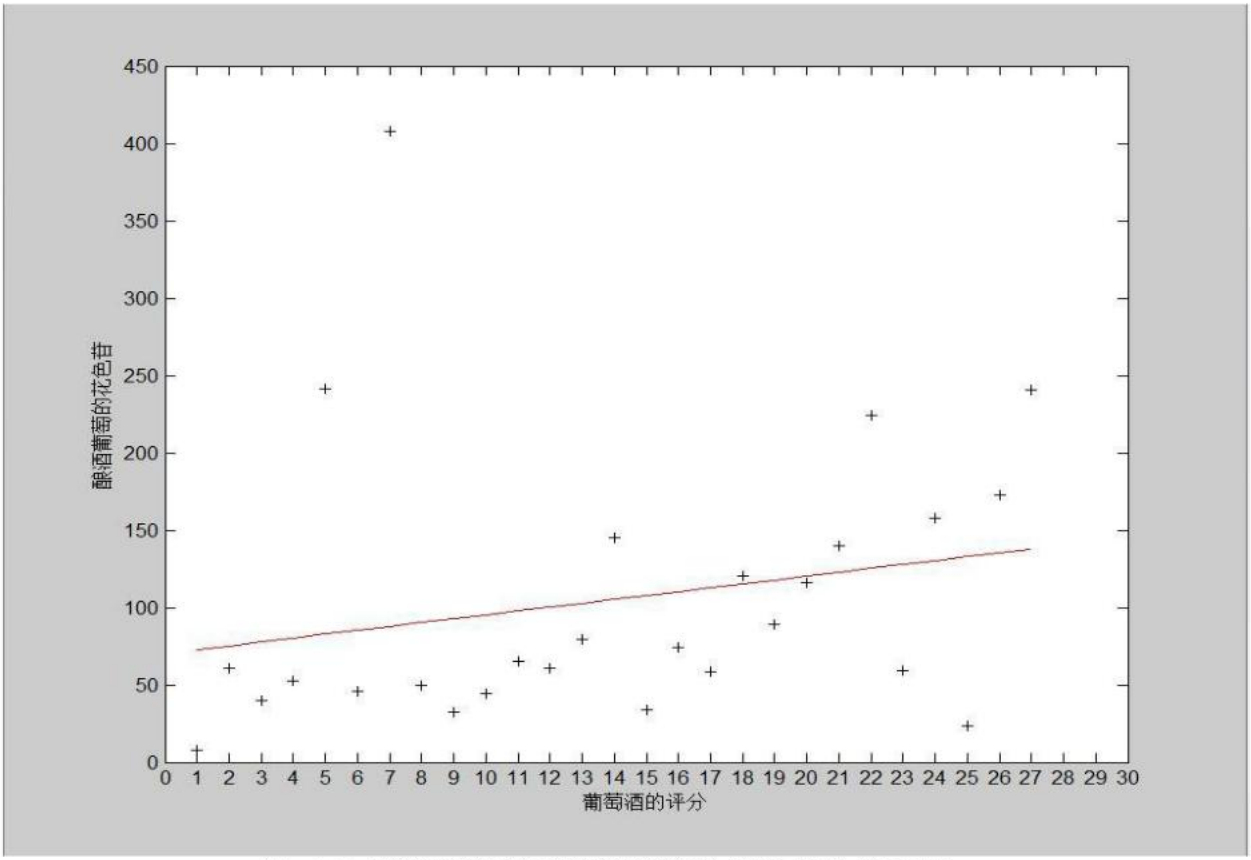


图 5.19 葡萄酒评分和酿酒葡萄的花色苷的关系图

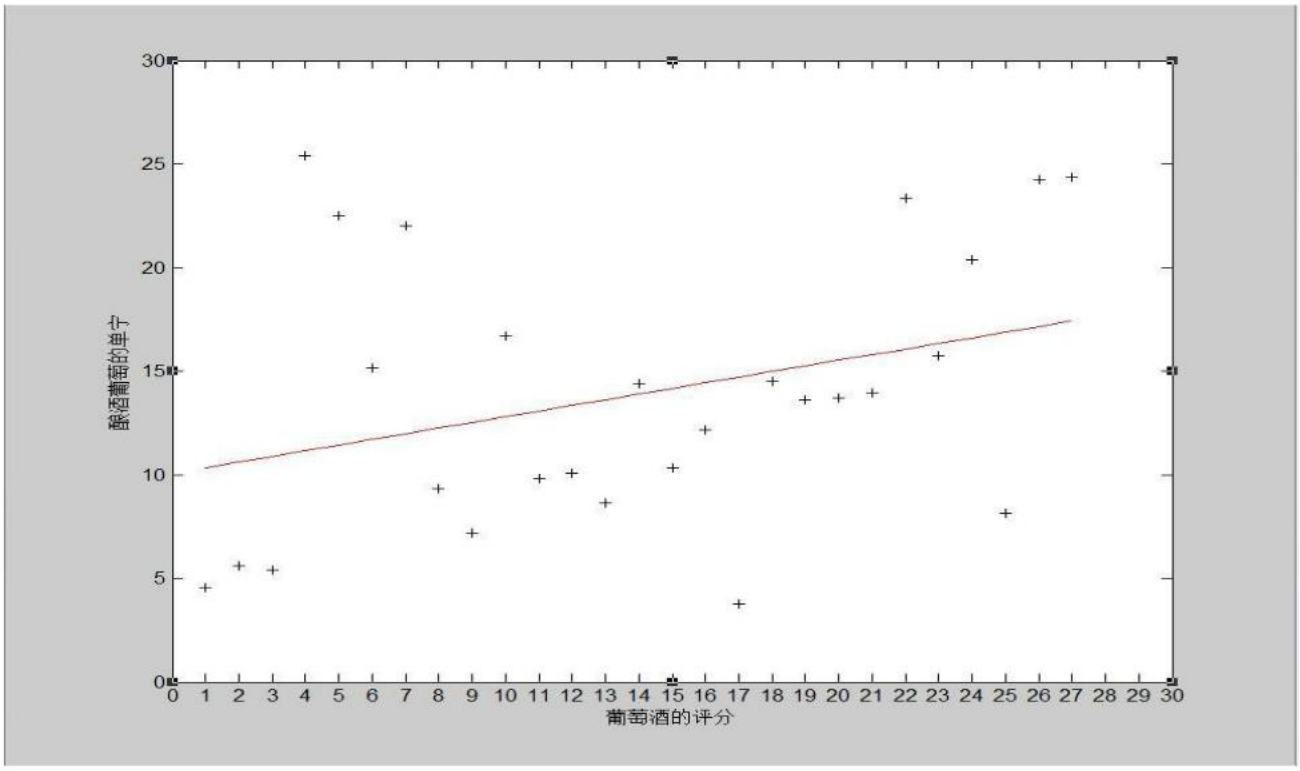


图 5.20 葡萄酒评分和酿酒葡萄的单宁的关系图

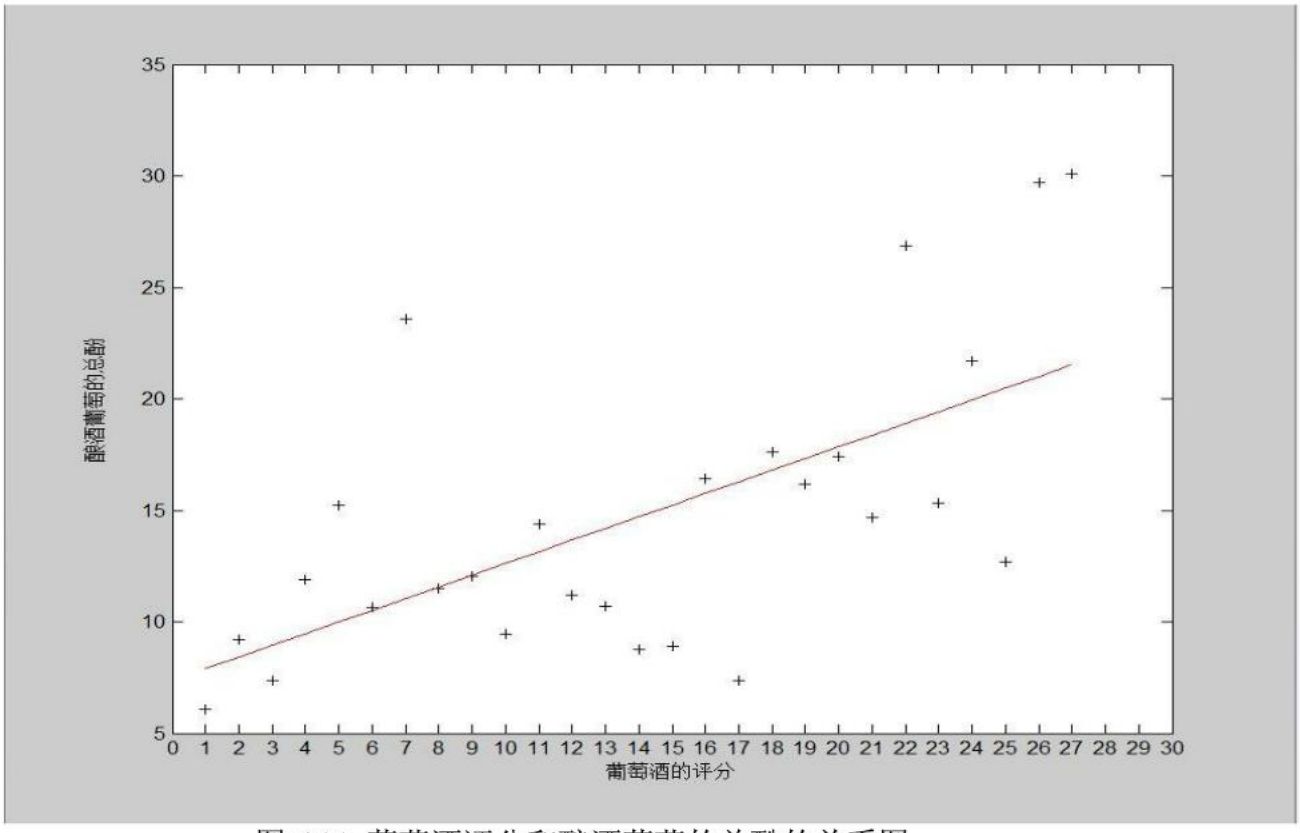


图 5.21 葡萄酒评分和酿酒葡萄的总酚的关系图

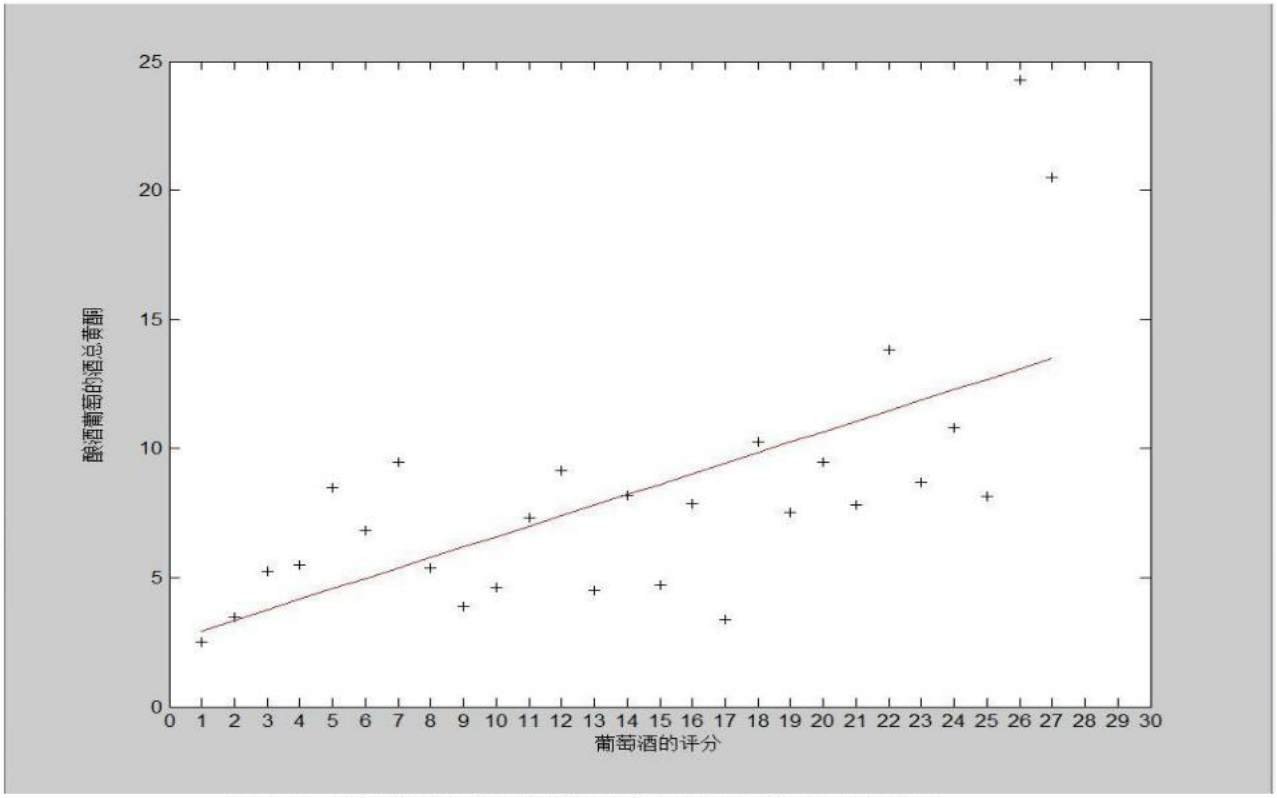


图 5.22 葡萄酒评分和酿酒葡萄的酒总黄酮的关系图

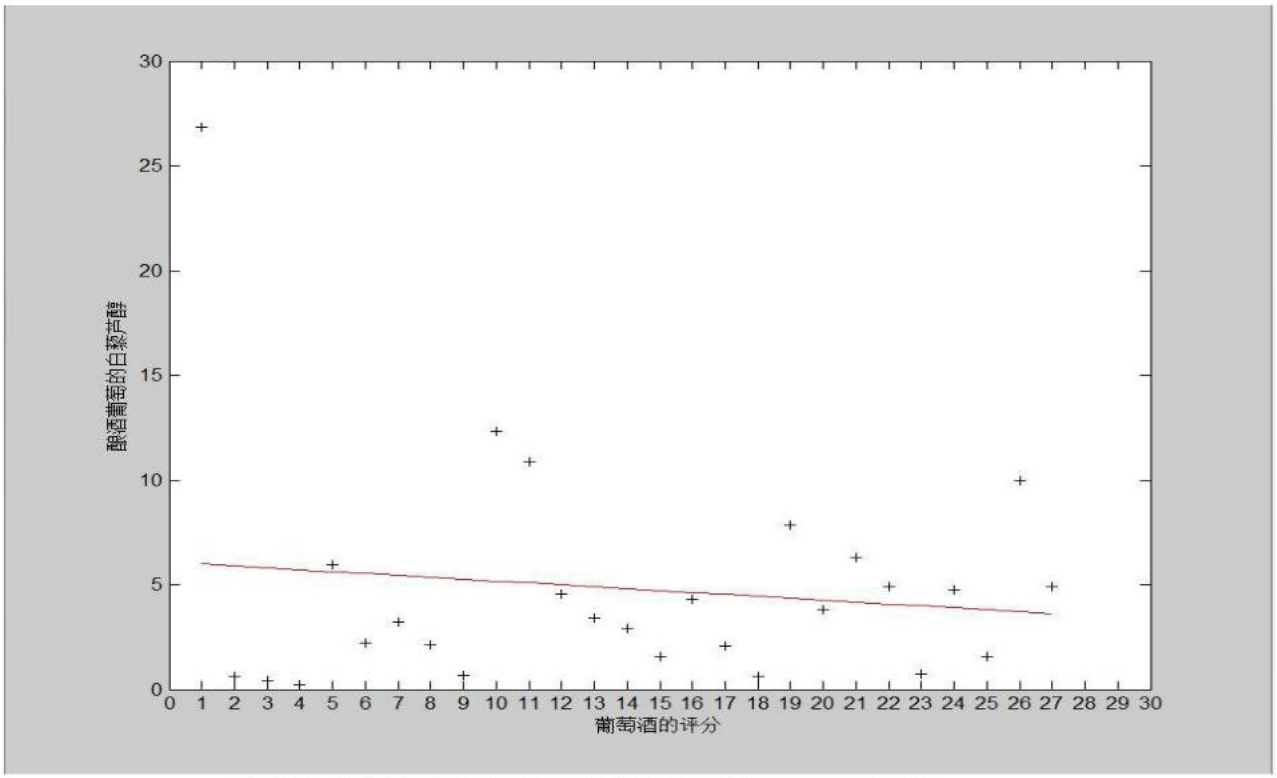
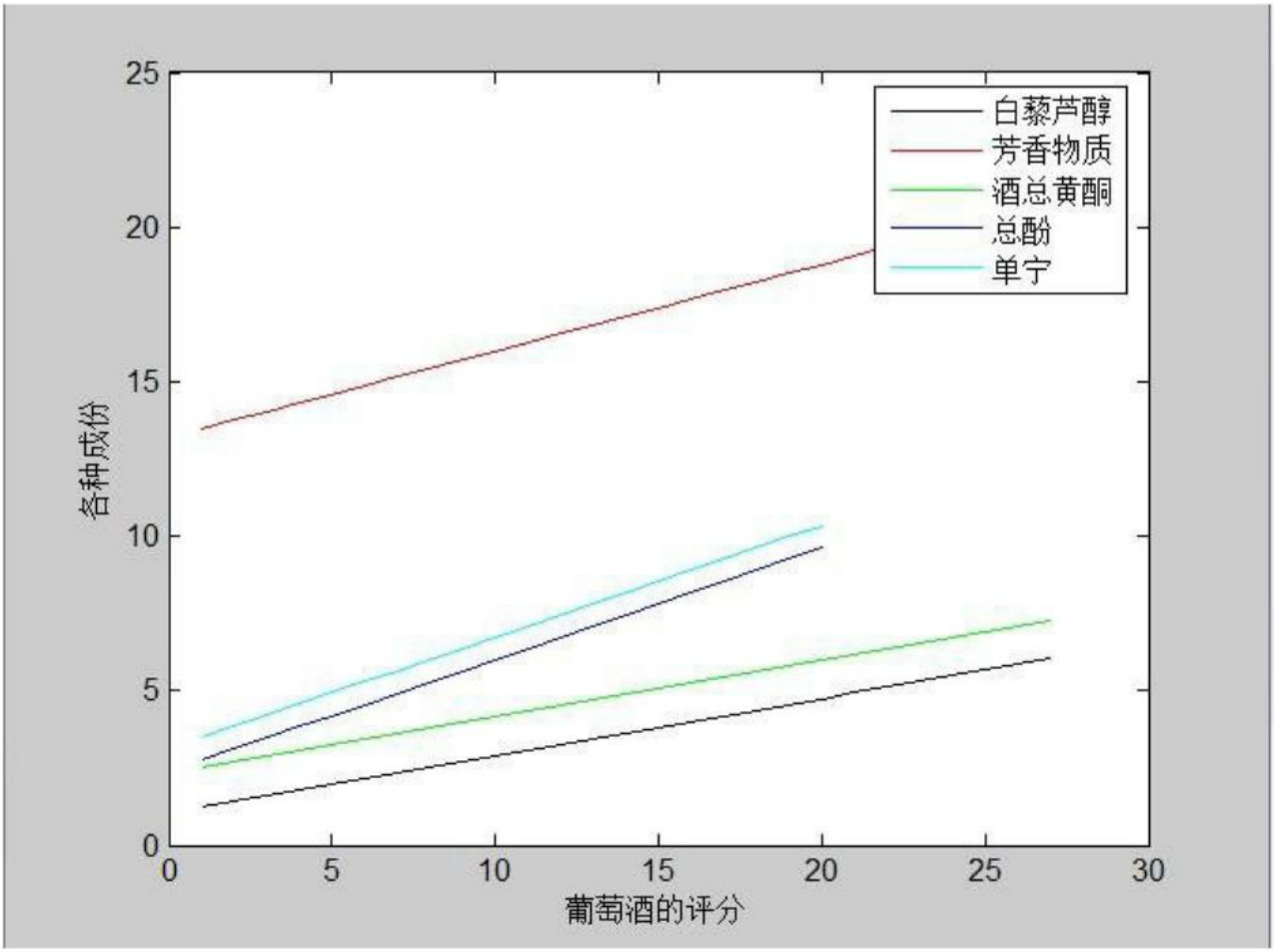


图 5.23 葡萄酒评分和酿酒葡萄的白藜芦醇的关系图

通过 Matlab 软件进行数据拟合并画出拟合线性图，得出酿酒葡萄各成分系数 p 的值分别为 0.0003，0.0963，0.0005，0.0002，0.5134（备注 p 越接近 0，该成分与葡萄酒的质量相关性越强），所以我们可得知酿酒葡萄的理化指标对葡萄酒质量是有影响，因为由第三问知酿酒葡萄与葡萄酒之间有相关性，所以也间接说明了酿酒葡萄和葡萄酒的理化指标对葡萄酒质量是有影响的。

但是结合附件 3 中各个样品中芬香物质的数据及样品的等级进行数据拟合，可得下图：



由图可知葡萄酒的感官指标中的芬香物质与葡萄酒的各个理化指标呈一致的相关性，所以不能只用葡萄酒的理化指标来评价葡萄酒的质量，还要加上葡萄酒的感官指标。即综合葡萄酒的理化指标和感官指标来进行对葡萄酒的质量进行评价。

六、模型的评价与优缺点

6.1 优点

（1）对数据进行合理的处理，采用拟合的方法对数据间的关系进行图像化，是

的其问题分析更加明显，简化。
（2）第二问中采用逆向思维的方法，反推出酿酒葡萄的等级分划。
（3）除去异常数据，使图像更加合理化，明显化。
（4）将每组数据进行有针对性对比，如第一组的白葡萄酒和第二组的白葡萄酒进行对比
（5）将每个小问题的分析串联起来，使其思路更加清晰

6.2 缺点

（1）数据量太庞大，导致我们忽略一些对题目影响不是很大的数据，使得模型存在误差。
（2）编程时数据量太大，存在很多困难。
（3）数据与数据将很难将其联系起来。
(4)对数据处理不大，没有整体进行求解，只针对某一种葡萄酒进行相关性验证，如第二问只对白葡萄酒进行分级。
（5）考虑问题不周全，如只对第四问验证酿酒葡萄与葡萄酒的理化指标之间的联系，而没有考虑进一步了解两者存在什么比例。

七、参考文献

[1] 姜启源，谢金鑫，叶俊.数学建模，北京：高等教育出版社，2004
[2] 单锋，朱丽梅 ，田贺民，数学建模，北京：国防工业出版社，2012.2
[3] （美）吉奥丹诺（Giordano，F.R）叶其孝，姜启源，数学建模（原书第三版）北京：机械工业出版社，2005,1
[4]冯杰，黄力伟，王勤，尹成义.数学建模原理与案例,北京：科学出版社，2007
[5] 宋叶志，贾东永，MATLAB 数值分析与运用，北京：机械工业出版社，2009.7