赛区评阅编号（由赛区组委会填写）：

2022 高教社杯全国大学生数学建模竞赛

承　诺　书

我们仔细阅读了《全国大学生数学建模竞赛章程》和《全国大学生数学建模竞赛参赛规则》（2019 年修订稿，以下简称为“竞赛章程和参赛规则”，可从全国大学生数学建模竞赛网站下载）。

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

我们完全清楚，抄袭别人的成果是违反竞赛章程和参赛规则的行为；如果引用别人的成果或资料（包括网上资料），必须按照规定的参考文献的表述方式列出，并在正文引用处予以标注。

我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们授权全国大学生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号（从 A/B/C/D/E 中选择一项填写）：C

我们的报名参赛队号（12 位数字全国统一编号）：1234

参赛学校（完整的学校全称，不含院系名）：桂林电子科技大学

参赛队员 (打印并签名)：1. 张湃

2. 江海瑶

3. 谭前程

指导教师或指导教师组负责人 (打印并签名)：覃义老师

（指导教师签名意味着对参赛队的行为和论文的真实性负责）

日期：2021 年 08 月 26 日

（请勿改动此页内容和格式。此承诺书打印签名后作为纸质论文的封面，注意电子版论文中不得出现此页。以上内容请仔细核对，如填写错误，论文可能被取消评奖资格。）

赛区评阅编号（由赛区组委会填写）：

2022 高教社杯全国大学生数学建模竞赛

编 号 专 用 页

赛区评阅记录（可供赛区评阅时使用）：

评阅人						
备注						

送全国评阅统一编号（由赛区组委会填写）：

全国评阅随机编号（由全国组委会填写）：

（请勿改动此页内容和格式。此编号专用页仅供赛区和全国评阅使用，参赛队打印后装订到纸质论文的第二页上。注意电子版论文中不得出现此页。）

葡萄酒的评价

摘要

葡萄酒是由新鲜葡萄经发酵而得到的一种成分复杂的酒精饮料，其质量由外观、香气和口感共同体现。不同品种的葡萄酒的成分有不同之处。一方面，葡萄酒中的氨基酸类、糖类、酸类和酚类化合物在不同的浓度下会呈现不同的风味；另一方面，葡萄酒中的大量挥发性物质，例如醇类、脂类和醛类化合物，都具有不同的香气。这些因素会综合作用在葡萄酒中，影响酒的质量。所以研究不同葡萄类别和其中成分，对分析和划分葡萄酒质量尤为关键。

针对问题一：我们首先对数据进行整理和分析，分别对两组评酒员对红、白葡萄酒的评分进行了 **Shapiro-Wilk 检验**和**平滑核直方图**的绘制，直观判断两组评酒员的评分差异。由于其中两组评酒员的评分不服从正态分布，我们使用了 **Mann-Whitney U 法**对评分中位数进行检验。综合两种检验的结果，我们认为第二组评酒员在红葡萄酒和白葡萄酒的评分上都更加可信。

针对问题二：为了研究葡萄的理化指标和葡萄酒的质量之间的关系，我们分别对红、白葡萄的理化指标与对应样本酿造的酒评分进行**相关性分析**，筛选出相关性较高的变量（分别有 10 个与 7 个），对筛选之后的变量进行 **K-Medoids 聚类**（分为三簇），对每一簇样本的酿造出的葡萄酒评分进行对应，发现均可以分为高、中、低三个档次。其中，中档占多数，这符合评分为类正态分布的结论，由此可以给出酿酒葡萄的分级。

针对问题三：建立红、白葡萄的理化指标与对应葡萄酒理化指标的**逐步回归模型**，通过逐步回归分别得出对红、白葡萄酒特定理化指标影响显著的葡萄理化指标因子，建立“最优”回归方程，并对回归方程进行显著性检验。结果均通过检验。

针对问题四：

关键字：Mann-Whitney U　ShapiroWilk 检验　K-Medoids 聚类　逐步回归模型

一、 问题重述

1.1 问题背景

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。

根据题中所给要求以及我们浏览的资料与文献，我们建立模型分别为以下几个问题进行求解：

1.2 需要解决的问题

1. 分析附件 1 中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?
2. 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。
3. 分析酿酒葡萄与葡萄酒的理化指标之间的联系。
4. 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?

二、 问题分析

2.1 问题一的分析

首先，对附表 1 中的评分数据进行分析和清洗。由于葡萄酒平常标准为百分制，其中外观分析 15 分、香气分析 30 分、口感分析 44 分、平衡/整体评价处理 11 分。因此，将每组中的各指标评分分数累加，对缺失数据和异常数据使用均值填补的方法填补或替换，得到评酒员对该葡萄酒的综合评分，并分别绘制红、白葡萄酒综合评分的平滑和直方图，与正态分布进行对比。

然后，对两组评酒员在红、白葡萄酒的综合评分进行单样本的 Shapiro-Wilk 检验，判断各小组综合评分是否满足正态分布。若数据满足正态分布，则对两组评酒员的评分数据进行 t 检验，判断两组之间是否有显著差异。若数据不服从正态分布，则使用两个独立样本的 Mann-Whitney U 法判断两组数据的中位数是否有差异。

最后，根据两种葡萄酒的综合评价均值（中位数）和方差判定评酒员评价的可信度。

2.2 问题二的分析

首先，对附表 2 和附表三中的理化数据进行分析和清洗。整合葡萄的理化指标并与酿成的葡萄酒进行对应，计算其相关性。得到红、白葡萄酒每个理化指标与相应总评分

相关性后，筛选出相关性较高的指标。

然后，对筛选出来的指标数据，预设 3 个簇进行 K-Medoids 聚类，得出聚类结果后，按照该结果计算各簇中葡萄酒样本的评分。

最后，按照分类结果和筛选出的理化数据，对葡萄的质量进行分级。

2.3 问题三的分析

本题需要分析酿酒葡萄和葡萄酒的理化指标之间的关系，于是考虑分别以红葡萄酒、白葡萄酒的各项理化指标作为因变量，以红、白葡萄的各项理化指标作为自变量。但由于本题所给数据的指标过多，所以对于因变量的选取，我们参考问题二所得到具有代表性的理化指标。另外，在本题中需要再众多葡萄酒的理化指标中用合适的方法筛选出影响作用显著的因子，从而得到清晰明了的因子与各项因变量之间的相关关系，从而可以得到红、白葡萄各项指标与葡萄酒各项指标之间的联系。

2.4 问题四的分析

三、 模型假设

1. 假设给出的的各项数据是真实可靠的;
2. 假设与总体评价极不和谐的异常数据很少，可以看做没有此现象;
3. 假设在短时间内，文中各个理化指标不会发生明显的变化;
4. 假设外界其他因素对模型所研究的方面影响很小, 不是决定因素;
5. 假设各评酒员对酒样的评分是相互独立的;
6. 假设各评酒员对酒的评价的客观公正的;
7. 假设评酒员在品尝完一种酒样后，不会对下一种或多种酒样的评价产生影响;
8. 假设葡萄酒质量一定能被某一确定分数量化，且评酒员在能力范围内会尽量接近该分数。

四、 符号说明

表 1 符号说明

符号	描述
A, B	两个总体
n_A, n_B	从 A, B 两个总体中随机抽取独立随机样本的容量
T_A, T_B	A, B 两个样本的秩和
U, U_α	Mann-Whitney U 检验量和临界值
p_j	第 j 个葡萄的理化指标
r	相关系数

五、 模型的建立与求解

5.1 问题一的模型建立与求解

5.1.1 模型的准备

1. 数据的分析、预处理

首先、对附件 1 中的数据进行直观分析，得出葡萄就品尝评分是按照百分制的标准执行，其中各类指标分数如下：外观分析 15 分（澄清度 5 分、色调 10 分）、香气分析 30 分（纯正度 6 分、浓度 8 分、质量 16 分）、口感分析 44 分（纯正度 6 分、浓度 8 分、持久性 8 分、质量 22 分）、平衡/整体评价处理 11 分。

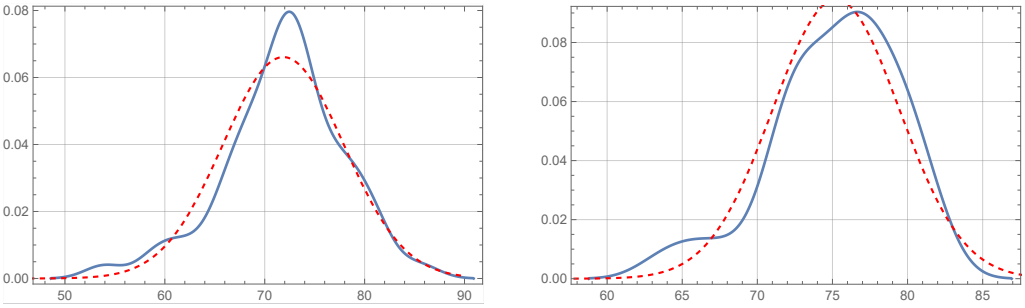
其次，为了综合评价葡萄酒，本文将两种葡萄酒中各指标的得分进行累加。同时，我们也对缺失和异常数据，使用均值填补法进行了填补。

其中，附件 1 第一组红葡萄酒 20 号样本中 4 号评酒员色调评分数据缺失；第一组白葡萄酒 3 号样本中 7 号评酒员口感分析持久性评分数据异常；第第一组白葡萄酒 8 号样本中 9 号评酒员口感分析持久性数据异常。

5.1.2 模型的建立与求解

1. 数据的正态性检验

对红、白葡萄酒每组的综合得分作出平滑核密度直方图如图 1，直观上可以认为两组数据服从正态分布.



(a) 红葡萄酒平滑核密度直方图与正态分布 **(b) 白葡萄酒平滑核密度直方图与正态分布**

图 1 红、白葡萄酒的平滑核密度直方图与对应的正态分布

2. Shapiro-Wilk 检验

使用单样本的 Shapiro-Wilk 检验方法检验各组红、白葡萄酒综合评分进行检验，结果如表 2

表 2 红、白葡萄酒单样本 S-W 检验结果

	组 1	组 2
红葡萄酒	0.0646029	0.906157
白葡萄酒	0.454315	0.194481

通过结果可以看出，在 90% 置信度下，组一红葡萄酒总评分不服从正态分布，故使用非参数的 Mann-Whitney U 法对两组间的评分中位数差异进行检验。

3. Mann-Whitney U 法流程与检验结果

Mann-Whitney U 检验假设两个样本分别来自除了总体均值以外完全相同的两个总体，目的是检验这两个总体的均值是否有显著的差别。其检验的流程如 algorithm 1.

Algorithm 1 Mann-Whitney U 检验

输入:

两个总体 A, B

1: 从两个总体 A, B 中随机抽取容量为 n_A, n_B 的两个独立随机样本，将 $(n_A + n_B)$ 个观察值按大小顺序排列，指定 i 为第 i 小 (或第 i 大) 观察值，若存在相同观察值，则使用位序的平均数;

2: 计算两个样本的秩和 T_A, T_B;

3: 对于总体 A, B, 计算检验量 $U_A = n_A n_B + n_A(n_A+1)/2 - T_A, U_B = n_A n_B + n_B(n_B+1)/2 - T_B$;

4: 选择其中较小的 U 值与 U 的临界值 U_α 比较，若 U 大于 U_α，接受原假设，反之则拒绝原假设;

输出:

返回: 检验结果 (U 或 p 值);

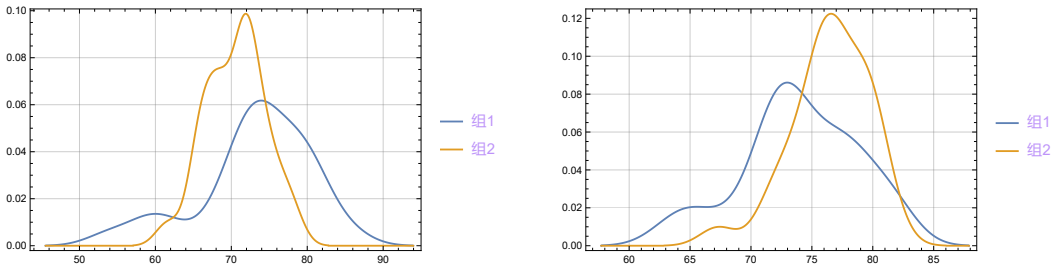
使用 algorithm 1对分组数据进行检验，得到的 p 值如表 3.

表 3 红、白葡萄酒组间 Mann-Whitney U 法检验结果

红葡萄酒	白葡萄酒
0.015	0.031

从结果可以看出，p 值均小于 0.05，说明在 95% 的置信度下，组 1 和组 2 评酒员对红、白葡萄酒的评分有显著差异。

从以上分析可以得出，不同组评酒员对同一葡萄酒的评价有差异，但由于我们假定葡萄酒的质量可以被决定，故方差越小的评分组，得出的分数越能接近真实的评分。而对于不同的组别，他们的平滑核密度直方图如图 2.



(a) 不同组红葡萄酒平滑核密度直方图 **(b) 不同组白葡萄酒平滑核密度直方图**

图 2 不同小组在红、白葡萄酒的平滑核密度直方图

所以，根据方差最小的原则，对于组 2 的评酒员在对红、白葡萄酒的评分上都更合理。

5.2 问题二的模型建立与求解

5.2.1 聚类指标的选取

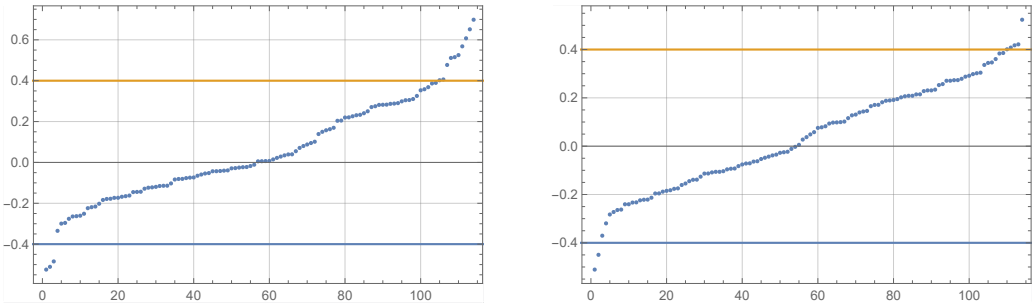
要对酿酒葡萄进行分级，首先需要确定适当的分级目标，一般来说，可以将葡萄的质量分为 3 个等级，分别为优秀、中等和较差。考虑到葡萄的理化指标与葡萄酒质量和评分之间有较多变量的影响，首先需要找出对葡萄酒质量影响较大的指标。由于葡萄的理化指标中，拥有二级指标的一级指标只需通过累加获得，为了更细致的研究不同理化指标对葡萄酒质量的影响，我们不考虑拥有二级指标的一级指标，然后按照剩余指标按照原有排序进行重新排序，共计 120 个指标，分别记为 $p_j, j = 1, 2, \ldots, 120$，同时对于其中第 i 个样本数据，记为 p_{ij}.

为了在不改变变量含义的情况下找出对葡萄酒质量影响较大的因素，本文选择分别对红、白葡萄的理化指标进行相似度度量，并通过计算相关系数衡量这些指标之间的相似程度。

指标 p_j 与 p_k 的样本相关系数为:

$$r_{jk} = \frac{\sum_{i=1}^{n}(p_{ij} - \mu_j)(p_{ik} - \mu_k)}{\sqrt{\sum_{i=1}^{n}(p_{ij} - \mu_j)^2 \sum_{i=1}^{n}(p_{ik} - \mu_k)^2}}$$

经过计算,得到的红、白葡萄理化指数与葡萄酒评分的相关系数经过排序后如图 3所示:



(a) 红葡萄理化指标与相应得分相关系数（排序后） **(b) 白葡萄理化指标与相应得分相关系数（排序后）**

图 3 红、白葡萄理化指标与相应得分相关系数（排序后）

从图中可以看出，只有很少的指标相关系数大于 0.4，而一般认为相关性小于 0.3 不存在相关性，数据的相似程度极低。所以，我们认为只有超过 ±0.4 的界限，才有变量之间较高的相似度。所以，以 ±0.4 为界限，筛选出红、白葡萄的正负指标，如表 4所示。

表 4 指标选取情况

	红葡萄	白葡萄
正指标	19,27,28,30,35,40,46,76,105,113	56,57,59
负指标	3,41,45,57,59	50,80

经过筛选后的指标不仅与葡萄酒的质量有了较大的相关性，同时也大幅度降低了指标的数量，减轻了样本数量过少带来的不确定性。

5.2.2 K-Medoids 聚类模型及其聚类结果

通常使用的聚类算法，例如 K-Means, 均值聚类等，受异常点的影响严重，且对样本数量较少的数据集效果不佳。所以我们选择能够有效削弱异常点影响的 K-Medoids 算法对筛选后的样本数据进行聚类，该算法的流程见 algorithm 2.

Algorithm 2 K-Medoids(Partitioning Around Medoids)

输入:

n 个数据点 p_n

1: **初始化:** 在 n 个点中随机选择 k 个作为中心 m_k，把剩余的每个点分配给离它最近的中心;

2: 随机地选择一个非中心点替换中心点，计算分配后的距离改进量;

3: 如果总的损失减少，则交换中心点和非中心点, 否则不进行交换;

4: 重复 2-3 的过程，直到所有的点不再发生变化，或已达到设定的最大迭代次数;

输出:

返回: 聚类结果 C;

使用此算法，对红、白葡萄进行聚类，得到的结果如表 5.

表 5 聚类结果

	红葡萄	白葡萄
第一类	9,23	4,5,14,21,23,28
第二类	1,2,3,5,8,10,13,14,16,17,19,20,21,22,24,25,26	2,3,9,10,20,24,25,26,27
第三类	4,6,7,11,12,15,18,27	1,6,7,8,11,12,13,15,16,17,18,19,22

分别对这三类葡萄发酵成的葡萄酒的平均分进行计算，结果如表 6.

表 6 红、白葡萄酒不同簇的平均分

	红葡萄酒	白葡萄酒
第一类	77.65	78.62
第二类	71.41	77.07
第三类	67.00	75.08

从评分表中可以看出，不同簇对应着不同质量的葡萄酒，其中第一类对应可以酿成优质葡萄酒的葡萄，第二类对应普通葡萄酿造的葡萄酒，而第三类对应较差葡萄酿造质量较低的葡萄酒。

由此，我们可以得出结论，对于红葡萄，提升葡萄酒质量的指标为较高的蛋白质、DPPH 自由基、总酚、PH 值等 10 个指标的含量，降低葡萄酒质量的指标为较高的果皮颜色；对于白葡萄，提升葡萄酒质量的指标为较高的苏氨酸、总糖、可溶性固形物、果皮颜色 b 和 c 含量，降低葡萄酒质量的指标为较高的果穗质量和 (E)-2-已烯醛的含量。

5.3 问题三的模型建立与求解

5.3.1 模型的建立

由于本题涉及的自变量过多，所以考虑使用逐步回归法，根据条件对不显著的变量以及相关性不强的变量进行剔除，使得回归方程中的每一个自变量的作用都具有显著意义，引入变量再次进行检验，对该过程进行若干次，直到没有变量可引入，也没有任何自变量可以剔除为止。

在逐步回归这个双向筛选的过程中，筛选的规则是基于对偏回归平方和 F 的检验，以下为 F 检验：

$$H_0 : \beta_j = 0, H_1 : \beta_j \neq 0$$

$$F = \frac{SS_1^{(l)}(X_j)}{SS_2^{(l)}(n-p-1)}$$

其中，p 为进行到第 l 步时方程中的自变量的个数，分子为第 l 步时 X_j 的偏回归平方和，分母为第 l 步时 X_j 的残差平方和。对于给定的检验水平 α, 当 $F \geq F_{a(1,n-p-1)}$ 时，可决定引入，否则，反之。

5.3.2 模型的求解

1. 回归方程的计算

下面对于红葡萄酒理化指标中的花色苷、单宁等八个理化指标与葡萄中的蛋白质、总酚等 20 个因子之间的关系进行逐步回归分析。

其中，以红葡萄酒中的花色苷 y_1 为例，取其作为因变量，选取与评分相关性高的 20 个因子作为自变量，利用软件对该方程进行回归分析，并建立逐步回归方程，经过调整后得到的回归方程系数表如表 7所示：

表 7 模型摘要

模型	R	R 方	调整后的 R	标准估算的错误
1	0.923	0.851	0.845	90.47914
2	0.973	0.947	0.942	55.24180

从表 7可知，在模型 2 中，逐步拟合的多元线性回归方程的因变量能够被自变量 (花色苷、反式白藜芦醇苷) 解释占 94.7%, 说明该模型效果较好。

表 8 逐步回归的参数估计

模型 2	系数	标准化系数	t	显著性
常量	7.578		0.456	0.653
花色苷	2.058	0.802	15.85	0.00
反式白藜芦醇苷	78.345	0.332	6.562	0.00

从表表 8可以看出逐步回归分析中的系数、标准化系数、t 检验两和显著性。从模型中各变量的 t 简言之及其显著水平可以发现，花色苷和反式白藜芦醇苷达到了高度显著水平 ($p<0.01$). 说明花色苷和反式白藜芦醇苷这两个因素对红葡萄酒花色苷指标影响是最显著的，由此，我们可以得到最优回归方程为

$$y_1 = 7.578 + 2.058x_{16} + 78.345x_{17}$$

1. 回归方程的显著性检验

逐步回归分析中的 F 检验是检验总体回归方程中的假定因变量 y 与自变量 $x_1, x_2, x_3 ... x_n$ 之间的线性关系是否显著，是否可以用线性模型来描述因变量和自变量之间的关系。

对逐步回归方程 (2) 进行 F 的检验，从表 9可知模型 2 中，偏回归平方和 (S_i) 为 1302576.536, 残差平方和为 73239.749, 自由度分别为 2,24,26, 残差均方差为 3051.656, F 检验量为 231.421, 显著性概率为 0.000<0.05, 所以，回归方程通过 F 检验，说明线性回归效果显著。

表 9 逐步回归的方差分析

模型 2	平方和	自由度	均方	F	显著性
回归	1302576.536	2	651288.268	213.421	0.000
残差	73239.749	24	3051.656		
总计	1375816.285	26			

1. 葡萄酒各指标的回归方程

利用类似的方法，求出红葡萄酒的其他各指标回归方程：

花色苷：$y_1 = 7.578 + 2.058x_{16} + 78.345x_{17}$

单宁：$y_2 = -15.808 + 0.082x_{18} + 0.546x_{17} + 18.212x_{11} + -0.091x_{14} + 0.003x_{12}$

总酚：$y_3 = 1.156 + 0.32x_2 + 0.796x_{17}$

白藜芦醇：$y_4 = 2.649 + 6.121x_{20}$

RWL：$y_5 = 50.417 + -0.147x_{16} + 6.443x_8 + -1.502x_3$

RWa：$y_6 = 69.167 + -0.94x_{16} + -4.818x_5$

RWH：$y_7 = 2.217 + -0.643x_6$

RWC：$y_8 = 167.812 + -8.539x_6 + -0.097x_{16} + -4.001x_4$

白葡萄酒的各指标回归方程：

单宁：$z_1 = 0.964 + 0.237x_1$

总酚：$z_2 = 0.920 + 0.142x_1 + 0.068x_7$

色泽 L：$z_4 = 102.416 + 0.024x_9 + -0.009x_{10} + -0.143x_3 + 0.007x_{14}$

色泽 a：$z_5 = -1.586 + -0.128x_{15} + 0.020x_9$

色泽 b：$z_6 = 8.804 + -0.136x_9 + 0.026x_{10} + -3.222x_{11}$

色泽 H：$z_7 = -22.861 + 3.052x_3 + -0.529x_5 + 0.134x_{14} + 0.076x_{10}$

色泽 C：$z_8 = 8.827 + -0.136x_9 + -0.026x_{10} + -3.272x_{11}$

5.4 问题四的模型建立与求解

1. 结果分析

根据上述模型的方法，可得红、白葡萄酒的理化指标和红葡萄的之间联系，结果如下表，表所示。

表 10 红葡萄与红葡萄酒理化指标的相关因子

红葡萄酒成分	相关因子
花色苷（y_1）	花色苷（x_{16}）、反式白藜芦醇苷（x_{17}）
单宁（y_2）	顺式白藜芦醇苷（x_{18}）、反式白藜芦醇苷（x_{17}）、DPPH 自由基 1/IC50 (x_{11})、多酚氧化酶活力 (x_{14})、褐变度（x_{12}）
总酚（y_3）	总酚（x_2）、反式白藜芦醇苷（x_{17}）
白藜芦醇（y_4）	顺式白藜芦醇（x_{20}）
RWL（y_5）	花色苷（x_{16}）、果皮颜色 C（x_8）、白藜芦醇（x_3）
RWa（y_6）	花色苷（x_{16}）、果皮颜色 a（x_5）
RWH（y_7）	果皮颜色 b（x_6）
RWC（y_8）	果皮颜色 b（x_6）、花色苷（x_{16}）、果皮颜色 L（x_4）