Statistics and Application 统计学与应用, 2017, 6(1), 104-110
Published Online March 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.61012

Hans汉斯

Wine Evaluation Model

Yuming Xu

Jiangxi University of Finance and Economics, Nanchang Jiangxi
Email: 474370777@qq.com

Received: Mar. 10th, 2017; accepted: Mar. 27th, 2017; published: Mar. 30th, 2017

Abstract

Based on large amounts of data which I get from the National Mathematical Modeling Contest, by using cluster analysis and principal component analysis, I established a variance model and regression model to find the relationship between wine grape and wine quality and what indicators can affect the quality of the wine. I get the relevant index system affect wine quality and the equations. The results obtained enable people to learn more about the relationship between wine grapes and wine, much quicker and easier to analyze and evaluate the quality of the wine.

Keywords

Variance Model, Clustering Analysis, Principal Component Analysis, Regression Model

葡萄酒评价模型

徐宇明

江西财经大学，江西 南昌
Email: 474370777@qq.com

收稿日期：2017年3月10日；录用日期：2017年3月27日；发布日期：2017年3月30日

摘　要

本文基于在参加全国大学生数学建模大赛时得到的大量数据，通过使用聚类分析和主成分分析等方法，建立了方差模型和回归模型，研究酿酒葡萄与葡萄酒质量的关系以及哪些指标能影响葡萄酒的质量，最后得出影响葡萄酒质量的相关指标体系以及其中的方程关系。得出的结果能使人们更多了解酿酒葡萄与葡萄酒的关系，更快捷简便的分析评价出葡萄酒的质量。

文章引用: 徐宇明. 葡萄酒评价模型[J]. 统计学与应用, 2017, 6(1): 104-110. https://doi.org/10.12677/sa.2017.61012

关键词

方差模型，聚类分析，主成份分析，回归模型

1. 问题的重述

确定葡萄酒的质量一般是由专业评酒员通过品尝对葡萄酒的口味各项指标进行打分，最后求和得出总分进行评判。由于不同评酒员对葡萄酒的偏好和侧重不尽相同，在打分上不可避免有着主观性和局限性，再加上葡萄酒口味与许多条件有关，包括葡萄、产地、年份等，因此一般还会通过对酒液中的理化指标进行测定分析，以得出葡萄酒内部理化指标与口味之间的联系，进一步的得出可信度较高的葡萄酒质量分类。本文数据建立来源于 2012 年全国大学生数学建模，通过数学模型讨论下列问题：

1) 分析两组评酒员的评价结果有无显著性差异？
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，并论证能否用酿酒葡萄和葡萄酒的理化指标来评价葡萄酒的质量？

2. 问题的假设

1) 假设所有品酒师对酒样品的评价都很可靠。
2) 假设两组品酒师品的样品酒都对应的完全相同。
3) 假设用仪器检测出的成分都真实有效。
4) 假设各样品的测量不考虑附件外的指标影响。

3. 符号说明

表 1 为相关符号说明。

4. 模型的建立和求解

4.1. 问题(1)的分析和求解

问题(1)的第一问是分析两组评酒员的评价结果有无显著性差异。通过建立协方差模型，有无显著性差异的分析可运用检验统计量 T^2 解决。公式如下：

对于红葡萄酒，设 (x_i, y_i)，$i=1,2\cdots$，令 $d=x_i-y_i, i=1,2\cdots$；又设 d_i 独立分布在 $N(\delta,\Sigma)$，其中 $\Sigma>0$，$\delta=\mu_1-\mu_2$。μ_1，μ_2 分别是总体 X 和总体 Y 的均值向量。其中可以通过检验假设：$H_0:\mu_1=\mu_2$ 等价于 $H_0:\delta=0$，从而将两个总体的均值比较检验的情形转化为一个总体的情形。当原假设 $H_0:\delta=0$ 为真时，统计量为：

$$\frac{i-p}{p(i-1)}T^2。其中 T^2=i(\bar{X}-\bar{Y})S^{-1}(\bar{X}-\bar{Y})'。$$

Table 1. Symbol description
表 1. 符号说明

i	表示某一种类葡萄酒的第 i 样品
$(j=1,2,\cdots,10)$	表示第 j 位品酒员
$(k=1,2,\cdots,10)$	表示第 k 指标的分数
$\bar{X}_{ijk}$	表示第 j 个品酒员对第 i 个某类葡萄酒样品的第 k 个项目的评价指数
n	表示葡萄酒的样品的总数
$\bar{X}_{ik}$	表示某类葡萄酒第 i 个样品各项指标分数的平均值
$\bar{S}_i$	表示某类葡萄酒第 i 个样品的方差
s	表示方差
S	表示两个样品的协方差矩阵
T_1^2	对于红葡萄酒两组间的检查统计量
T_2^2	对于白葡萄酒两组间的检查统计量
$\bar{X}$	表示第一组红葡萄酒 k 类指标所有分数取平均值的向量
$\bar{Y}$	表示第二组红葡萄酒 k 类指标所有分数取平均值的向量

服从自由度为 p 和 $i-p$ 的 F 分布，对给定的显著性水平 α，若 $T^2 \geq T_\alpha^2$，则有显著性差异；反之，则无显著性差异。

$$T_\alpha^2 = \frac{p(i-1)}{i-p} F_\alpha(p, i-p)$$

4.2. 问题(2)的分析和求解

问题(2)中要求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。以品酒师对葡萄酒的评分作为葡萄酒质量的标准，首先统计每种红葡萄酒所得的 A, B 组品酒师打出的总得分，后求其平均值。之后考虑红葡萄的分级，对酿酒葡萄的 29 项理化指标进行处理(其中果皮颜色分为颜色 a 和颜色 b)，对于多次测量的数据求平均值。导出这 28 个葡萄样品对应的 29 个指标的含量后用 SPSS [1]对这些数据进行聚类分析，分析出来后的结果图见附录。聚类后在得出每一类的得分后进行分级。分类情况如下表 2、表 3 所示。

我们同样的用 SPSS 对白葡萄酒进行同样的数据操作后得到下列表 4、表 5 分类情况：

根据表 2~5，可以看出将红葡萄分为三类或四类时葡萄酒的等级变化较小，因此聚类分析有一定的可行性。故可采用其结果，将葡萄酒分为四级。第一级最好的酒为 21 号，第二级较好的酒为 5、10、17、23、24 号，第三级一般的酒为 3、6、13、14、16、25、26、27 号，第四级较差的酒为 1、2、4、7、8、9、11、12、15、18、19、20、22。同时白葡萄酒也分为四类第一类最好的酒为 3、28 号；第二类较好的酒为 1、8、13、16、17、18、19、22 号；第三类一般的酒为 2、4、6、7、10、11、12、14、20、21、23、26 号；第四类较差的酒为 5、15、24、25、27 号。

4.3. 问题(3)的分析和求解

问题(3)中要求分析酿酒葡萄与葡萄酒的理化指标之间的联系，首先考虑红葡萄与红葡萄酒理化指标的联系。此处可以使用多元回归的方法求二者关系。但是由于红葡萄有 29 个指标，红葡萄酒有 6 个指标，

Table 2. Red Wine is divided into three classes
表 2. 红葡萄酒分为三类时

类别	葡萄样品编号	分数
第一类	3，21	76.1
第二类	5，6，10，13，14，16，17，23，24，25，26，27	73.1
第三类	1，2，4，7，8，9，11，12，15，18，19，20，22	70.0

Table 3. Red Wine is divided into four classes
表 3. 红葡萄酒分为四类时

类别	葡萄酒样品编号	分数
第一类	21	75.0
第二类	5，10，17，23，24	75.4
第三类	3，6，13，14，16，25，26，27	72.2
第四类	1，2，4，7，8，9，11，12，15，18，19，20，22	70.0

Table 4. White Wine is divided into three classes
表 4. 白葡萄酒分为三类时

类别	葡萄酒样品	分数
第一类	28	80.5
第二类	1，3，8，13，16，17，18，19，22	75.6
第三类	2，4，5，6，7，9，10，11，12，14，15，20，21，23，24，25，26，27	75.4

Table 5. White Wine is divided into four classes
表 5. 白葡萄酒分为四类时

类别	葡萄酒样品	分数
第一类	3，28	80.5
第二类	1，8，13，16，17，18，19，22	75.6
第三类	2，4，6，7，9，10，11，12，14，20，21，23，26	75.3
第四类	5，15，24，25，27	75.1

变量过多不易于处理。因此可以使用 excel 求出红葡萄酒的第 i 个指标和红葡萄第 j 个指标的相关系数 r_{ij}，第一次求出的相关相关系数情况。

根据得到的相关系数表，删除一些相关性都不好的噪音元素(即某一行或者某一列的 $|r_{ij}|<=0.3$)。根据上表可以得知在 $|r_{ij}|<=0.3$ 里的元素为红葡萄指标中的酒石酸、柠檬酸、白藜芦醇、总糖、还原糖、PH 值、可滴定酸、果穗质量、百粒质量、果皮颜色 a 和果皮颜色 b，这些都是弱相关指标，即噪音指标。删除掉这些元素后，剩下的都是相关系数处于区间 $0.3<|r_{ij}|<0.7$ 的指标，包括氨基酸、蛋白质、花色苷、苹果酸、多分氧化酶活力、褐变度、DPPH 自由基、总酚、单宁、葡萄总黄酮、黄酮醇、总糖、可溶性物质、固酸比、干物质含量、果梗比、出汁率、果皮质量。这些都是有相关性的指标。我们把删除后的数据进行第二次相关性分析，根据数据计算，可以得出处于大约 $|r_{ij}|>=0.7$ 里的有花色苷和总酚，表明它们对酿出的葡萄酒的理化指标有强相关性。处于 $0.3<|r_{ij}|<0.7$ 有氨基酸、蛋白质、苹果酸、多分氧化酶

活力、褐变度、DPPH 自由基、单宁、葡萄总黄酮、黄酮醇、总糖、果梗比、出汁率、果皮质量。其余的指标处于大约$|r_{ij}|<=0.3$的是可溶性物质，固酸比和干物质含量，这些指标都相关性不大。删除掉这些相关性弱的指标后进行最后的相关系数分析，由此可以比较出大约$|r_{ij}|>0.7$的有总酚、花色苷、单宁，说明它们与其它理化指标成强相关性。大约在$0.3<|r_{ij}|<0.7$有蛋白质、褐变度、DPPH 自由基、葡萄酒总黄酮、黄酮醇、果梗比和果皮质量，这些都与其它指标成相关性。它们的理化指标分布情况如下图 1 所示。

此时再考虑能否建立一个数学函数来表达它们之间的联系。分析后发现，可以建立回归模型。首先把红葡萄酒的 6 个成分定义为自变量用 SPSS 进行主成分分析，分析后然后用 MATLAB [2]对得到的主成份进行拟合得到的函数关系式为：

白葡萄的关系式为：

$$y=0.369x^4+0.1598*x^3-0.0568*x^2+0.1952*x+0.2272,$$

红葡萄关系式为：

$$y=0.844x^5+0.958x^4+0.988x^3+0.965x^2+0.956x-0.928.$$

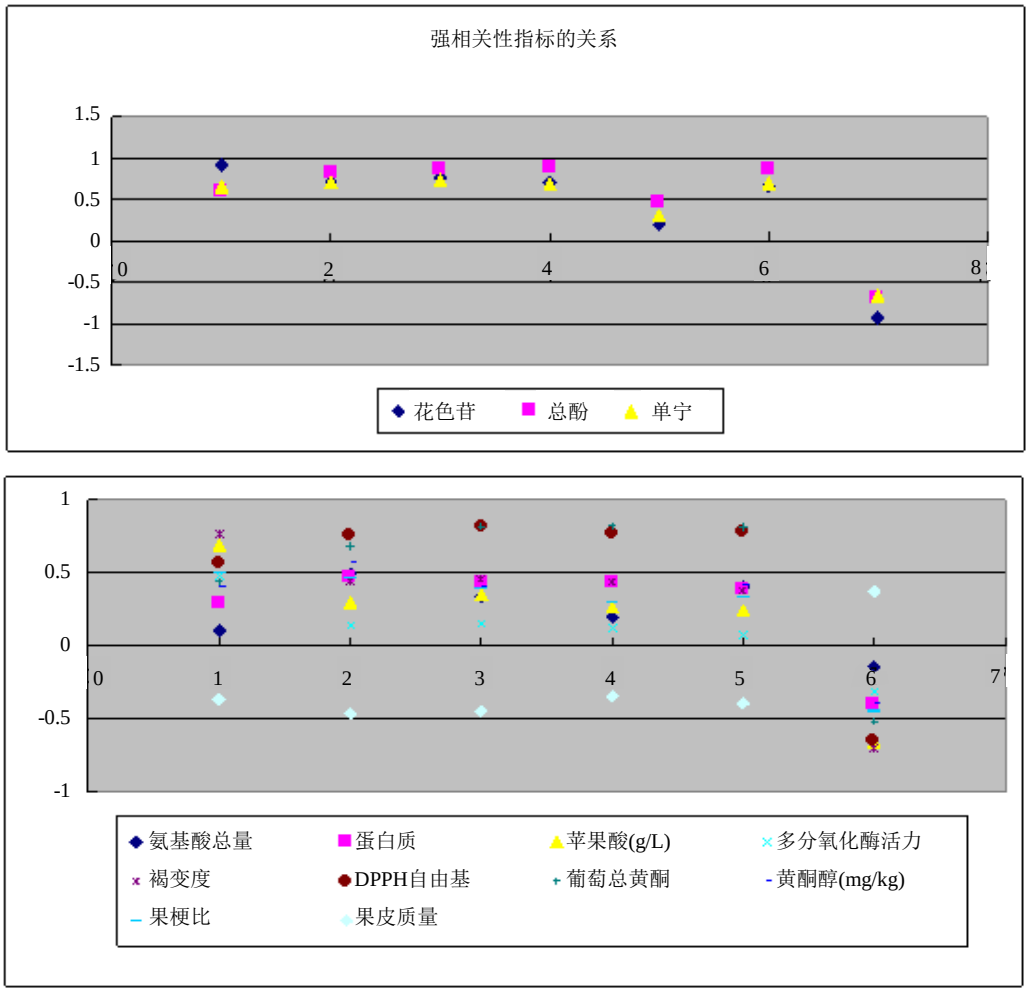

Figure 1. Relationship of strong correlation index
图 1. 强相关性指标的关系

4.4. 问题(4)的分析与求解

要分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，考虑到各理化指标对葡萄酒质量的不同方面的影响不同，因此将葡萄酒质量分为外观，口感，香气，整体 5 方面单独考虑[3]。又由于理化指标较多且相互联系紧密，信息重叠量大，所以先对理化指标用主成分分析法，并按各指标对主成分的载荷大小排序筛选出一些有代表性的指标。对每一方面而言，由于酿酒葡萄和葡萄酒的各理化指标对它的影响方式和影响大小未知，故选用多元线性回归分析法对数据进行拟合，确立影响较大的几个指标，再利用这些指标建立多元线性回归模型，从而得到理化指标对葡萄酒某方面质量的具体函数关系，得到酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响。

由于理化指标较多且相互联系紧密，信息重叠量大，所以先对理化指标用主成分分析法，并按各指标对主成分的载荷大小排序筛选出一些有代表性的指标。

选取葡萄和葡萄酒的理化指标共 38 个指标，以 27 种红葡萄和红葡萄酒的这 38 个指标的含量值为原始数据，利用 SPSS 软件求得其主成分矩阵，在 SPSS 软件中主成分分析分析后的分析结果中，“成分矩阵”反应的就是主成分载荷矩阵。在主成分载荷矩阵中对每一主成分下的各指标排序，按大小确定每一主成分主要代表的指标 i，从而得到少量的代表性的指标。由此，对上述 9 个主成分下指标分别排序，取每个主成分下载荷较大的几项构成新的指标，从而筛选出花色苷(酒)、单宁(酒)、总酚(酒)、白藜芦醇(酒)、DPPH(酒)果穗质量、A、干物质含量、pH、可溶性固形物、还原糖、总糖、总黄酮共 13 个指标。

又由参考文献[4]可知，芳香物质中可筛选出乳酸乙酯、己醇、庚醇 3 个影响葡萄酒香气的指标。综上，将葡萄与葡萄酒理化指标及芳香物质指标简化为 16 个影响指标。为进一步将这 16 个主要理化指标对应于葡萄酒不同评分项目得分的主要影响因素，因为第二组评价的分数较为可靠，故我们选择第二组评分为葡萄酒的分数。本文以各评分项目得分为因变量，以 16 个主要理化指标为自变量，初步建立全体主要理化指标对不同评分项目得分的多元线性回归模型。利用多元线性回归分析中的 Sig(显著性水平)大小比较来确定主要影响因素(理化指标)与主要理化指标的对应。

1) 多元线性回归分析分类筛选

下面以口感评价得分的主要影响因素筛选为例进行分析，各自变量 $(x_1, x_2, x_3 \cdots x_n)$ 及因变量 y 用 SPSS 进行回归分析。Sig 值要求小于给定的显著性水平，一般是 0.05、0.01 等，Sig 越接近于 0 越好。对所得各主要理化指标的 Sig 值进行大小比较，可知 PH、白藜芦醇、花色苷和单宁的 Sig 值较小于其他理化指标，因此这四种理化指标即可作为口感评价得分的主要影响因素。同理可得外观、香气及平衡/整体评价得分的主要影响因素。

2) 多元线性回归模型建立

以外观评价得分的多元线性回归模型的建立为例进行分析：由以上分类筛选已确定 PH、白藜芦醇、花色苷和单宁为口感评价得分的主要影响因素。下利用 SPSS 建立因变量口感评价得分与 4 个自变量之间的线性回归模型。将数据导入 SPSS，确定因变量和自变量，得到由此可得表 6：

由 Sig 值都很小，可知道该拟合良好。故其之间的方程关系式为：

$$y_1 = 19.329 - 0.004x_1 + 0.417x_2 + 2.81x_3 - 1.88x_4$$

同理可得出葡萄酒的香气质量可由葡萄酒的单宁及葡萄的还原糖、总黄酮和干物质总量评价，其评价模型为 $y_2 = 10.065 - 1.242x_1 - 0.532x_2 + 0.003x_3 + 0.147x_4$。外观质量可由葡萄酒的 a 指标、乳酸乙酯含量及白藜芦醇含量，葡萄的果穗质量评价，其评价模型为 $y_3 = 26.588 - 0.004x_1 + 0.456x_2 + 0.184x_3 + 0.029x_4 - 0.623x_5$。平衡/整体质量可由葡萄酒的乳酸乙酯、单宁、白藜芦醇以及葡萄的 PH 评价。其评价模型为 $y_4 = 6.616 + 0.04x_1 + 0.527x_2 - 0.028x_3 + 0.283x_4$。

Table 6. Results of SPSS analysis
表 6. SPSS 分析结果表

模型	*B*	标准误差	试用版	*t*	Sig
(常量)	19.329	2.961		6.528	0.000
花色苷	−0.004	0.001	−0.473	−2.630	0.015
单宁	0.417	0.111	0.694	3.753	0.001
PH	2.810	0.890	0.394	3.158	0.005
白藜芦醇	−0.188	0.037	−0.590	−5.069	0.000

5. 模型的评价与应用

第(2)问中由于葡萄的各种理化指标繁多，我们采用主成分分析法。主成分分析法作为统计上的一种多元分析方法[5]，我们是利用降维思想把主成分多指标转化为少数几个综合指标不仅能消除指标间信息的重叠，而且能根据指标所提供的原始信息生成客观的绩效得分。采用单因素方差分析法，主要缺点是如果假设不成立，只能获得 X 个相关变量具有差异性，而不能说明是哪几个具有差异，或者差异程度有多少。主成分分析法的主要缺点是当采样数据较多时，协方差阵的计算，求解特征值以及特征向量时的运算量会很大。

参考文献 (References)

[1] 张磊, 毕靖, 郭莲英. SPSS 实用教程[M]. 北京: 人民邮电出版社, 2008.
[2] 司守奎, 孙玺菁. 数学建模算法与应用[M]. 北京: 国防工业出版社, 2011.
[3] 李记明, 李华. 葡萄酒成分分析与质量研究[J]. 食品与发酵工业, 1994(2): 30-35.
[4] 李丽, 梁芳华, 孙爱东. 葡萄酒中的特征性香气成分的形成及其影响因素[J]. 饮料工业, 2009(5): 13-16.
[5] 王学民. 应用多元分析[M]. 上海: 上海财经大学出版社, 2009.