

**2020年兰州大学数学建模竞赛**

**中文标题**  混合物中特定成分的检测

队伍编号 1 1 1

学 院 信息科学与工程学院

专 业 数据科学

学生姓名 蔡琪琦

丁俊伟

曹延飞

**混合物中特定成分的检测**

**摘要**

本文针对混合物中特定成分有无的判定问题，运用数据挖掘的相关知识，基于Logistic回归分析建立了混合物中特定成分判定的数学模型，运用python、SPSS等软件得出合理的结论，并对测试数据做出了预测。

针对问题一要求给出判定特定成分存在的主要指标。我们首先对数据进行清洗，发现并没有明显的异常值，然后用Logistic回归方法构建数学模型。紧接着，我们考虑到了变量之间的多重共线性问题，尝试了主成分分析法和Logistic回归中“Forward：LR”方法消除多重共线性。根据SPSS的分析结果，我们对三个模型进行分析和比较，最终选用Logistic回归中“Forward：LR”方法。通过观察变量的优势比判断每一项指标对模型的贡献程度，优势比越大变量显著性越高，相关性高的变量即为判定特定成分存在的主要指标。我们得出结论：判定特定成分存在的主要指标为V1、V2、V7。

针对问题二要求判断是否存在模糊区域，若存在就估计出这一模糊区域。在Logistic回归模型中，当计算结果恰巧等于分类阈值时不好做出判断，因此本文选取分类阈值对应的区域为模糊区域。结果如下：-3.035+3.601v1+6.132v2-1.075v3-0.367v4+0.986v5-1.140v6+1.307v7=0七维空间中满足上述关系式的区域便是模糊区域。

针对问题三要求建立数学模型，判断测试数据中每一个混合物是否含有特定指标。在分析问题一、二的基础上，我们运用Logistic回归方法建立出数学模型，并借助准确度、精确度、灵敏度、特异度等指标对模型做出了分析和检测，发现模型是比较好的，而且准确率高达约92%。紧接着我们根据构建的模型预测出前十个样本是否含有特定成分，其中1表示含有，0表示不含有。预测结果为：前十个混合物中，第五个混合物中不含特定成分，其余九个混合物中含有特定成分。

本文主要应用SPSS软件和python对相关模型进行编程求解，计算方便、快速且较为准确，整篇文章采取图文并茂的效果。文章最后对建立的模型进行了客观的分析与评价，并提出了改进想法，使模型更加客观且具有现实意义。

**关键字：**特定成分；模糊区域；Logistic回归；

目录

[1. 问题重述 4](#_Toc45988533)

[2. 数据准备 4](#_Toc45988534)

[3. 模型假设 4](#_Toc45988535)

[4. 通用符号说明 4](#_Toc45988536)

[5. 问题一：给出判定特定成分存在的主要指标 4](#_Toc45988537)

[5.1 问题分析 4](#_Toc45988538)

[6. 问题二：成分存在与否的判定是否存在模糊区域 4](#_Toc45988539)

[6.1 问题分析 4](#_Toc45988540)

[7. 问题三：建立数学模型，判断测试数据 5](#_Toc45988541)

[7.1 问题分析 5](#_Toc45988542)

[8. 模型分析与评价 5](#_Toc45988543)

[8.1 模型优点 5](#_Toc45988544)

[8.2 模型改进 5](#_Toc45988545)

# 问题重述

为了检测某种混合物中是否含有特定成分，对该混合物的25000个样本进行了7项指标(记为V1,V2,…,V7)的检测。检测结果见附件“Data.xlsx”。数据文件包含两个表单。第一个表单是“训练数据”，包含20000个混合物样本，已知混合物中是否含有特定成分，其中1表示含有特定成分，0表示不含特定成分。第二个表单是“测试数据”，包含5000个混合物样本，未知混合物中是否含有特定成分。请完成以下三个任务：

1. 给出判定特定成分存在的主要指标；
2. 成分存在与否的判定是否存在模糊区域（对于7项指标），即在这些区域中很难判断成分是否存在。如果存在，试估计这一区域；
3. 建立数学模型，判断测试数据中每一个混合物是否含有特定成分。论文中给出前10个混合物的判定即可。

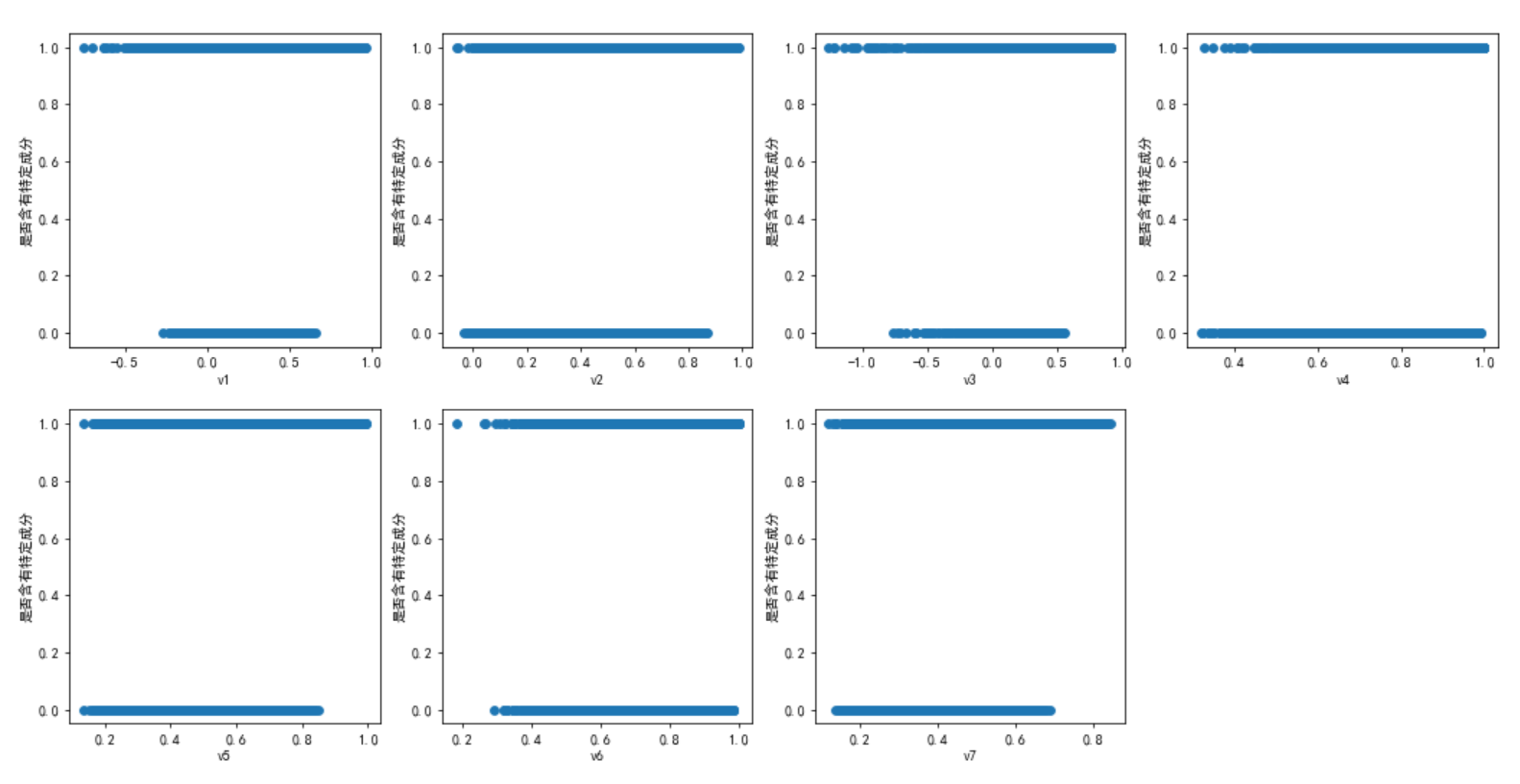
# 数据准备

### 数据来源

对某种混合物的25000个样本进行了7项指标(记为V1,V2,…,V7)的检测，判断是否含有特定成分。其中20000个样本作为训练数据，已知混合物中是否含有特定成分，1表示含有特定成分，0表示不含特定成分，5000个样本作为测试数据，未知混合物中是否含有特定成分。

### 数据清洗

在进行分析之前，需要先对数据进行清洗以及补值。因为数据集不存在缺失数据，而且通过分别绘制7项指标和是否含有特定成分的散点图可以看出，并不存在明显的异常离群值，故没有对原有数据进行任何操作。



### 数据划分

因为在模型建立之后，需要对模型进行评估，评价模型的预测准确率等性能，所以需要对包含20000个样本的训练数据进行划分。随机选取其中的80%作为模型建立的训练数据，剩余20%的样本作为模型的评估依据。

# 模型假设

1）所有数据均为原始数据，来源真实可靠。

2）因变量是二分类变量，且两个分类间互斥。

3）每个观测间相互独立。

4）假设各指标间不会发生化学反应等，其含量保持不变。

# 通用符号说明

|  |  |  |
| --- | --- | --- |
| 序号 | 符号 | 符号说明 |
| 1 | θi | 随机变量yi取1的概率 |
| 2 | β | 七项指标组成的参数向量 |
| 3 | xij | 第i个样本的第j个观测指标的数值 |
|  |  |  |

# 问题一：给出判定特定成分存在的主要指标

## 问题分析

问题一要求给出判定特定成分存在的主要指标。首先选取合适的数据分析方法来对指标进行分析，本文采用的是数据挖掘方法；然后通过建立模型分析各项指标对判定结果的影响，本文采用逻辑回归模型；最后通过比较逻辑回归结果来得出判定特定成分存在的主要指标。

## 模型建立

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中提出隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。它是一类深层次的数据分析方法。

实施数据挖掘的步骤可以分为以下步骤：确定业务对象、数据准备、数据挖掘、结果分析、知识的同化。

### 确定业务对象

清楚地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结果是不可预测的，但要探索的问题应是有预见的。为了数据挖掘而挖掘往往带有盲目性，是不会成功的。

### 数据准备

①数据选择 搜索所有与业务对象有关的内部和外部数据信息，从中选择出适用于数据挖掘应用的数据。

②数据预处理 包括对冗余、与任务无关数据的删除，缺失数据、异常数据处理，数据标准化等等，数据预处理是数据挖掘成功的关键。

### 数据挖掘

对得到的经过转换的数据进行挖掘，也就是用分析模型进行分析计算。

### 结果分析

解释并评估结果，其使用的分析方法一般应随数据挖掘操作而定，通常还会用到可视化技术。

### 知识的同化

将分析所得的知识集成到业务信息系统的组织结构中去。

## 问题求解

### 相关性检测

经过对数据的清洗后，使用相关矩阵评估变量之间关系的强度和方向（求解使用Minitab&SPSS）。较高的正相关值表明这些项目度量同一技能或特征。如果这些项目并非高度相关，则可能度量不同特征或可能未明确定义。具体结果见下表：

**相关矩阵**

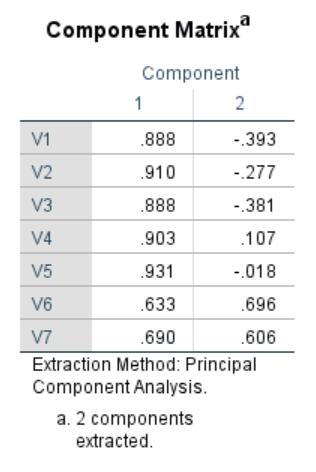
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **V1** | **V2** | **V3** | **V4** | **V5** | **V6** |
| V2 | 0.908 |  |  |  |  |  |
| V3 | 0.932 | 0.873 |  |  |  |  |
| V4 | 0.738 | 0.792 | 0.715 |  |  |  |
| V5 | 0.799 | 0.802 | 0.841 | 0.792 |  |  |
| V6 | 0.315 | 0.388 | 0.307 | 0.658 | 0.533 |  |
| V7 | 0.376 | 0.466 | 0.399 | 0.600 | 0.653 | 0.749 |

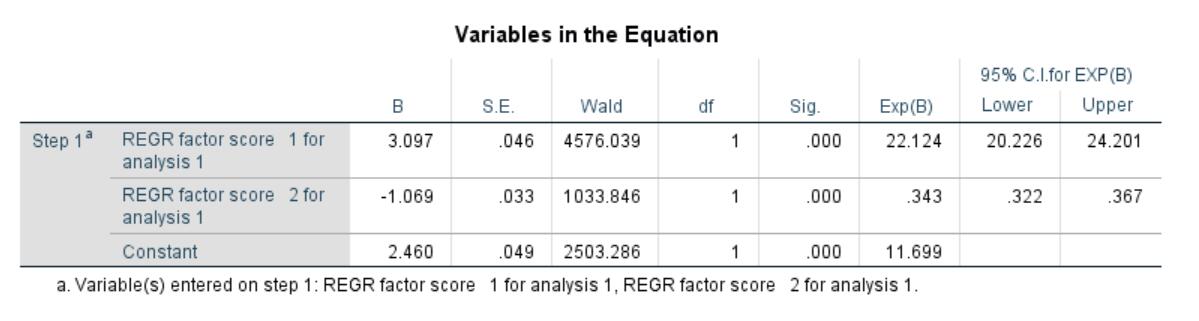
*单元格内容      Pearson 相关系数*

从上表可以看出，部分变量的相关性较强，本文考虑采用一些方法来消除变量之间的相关性。

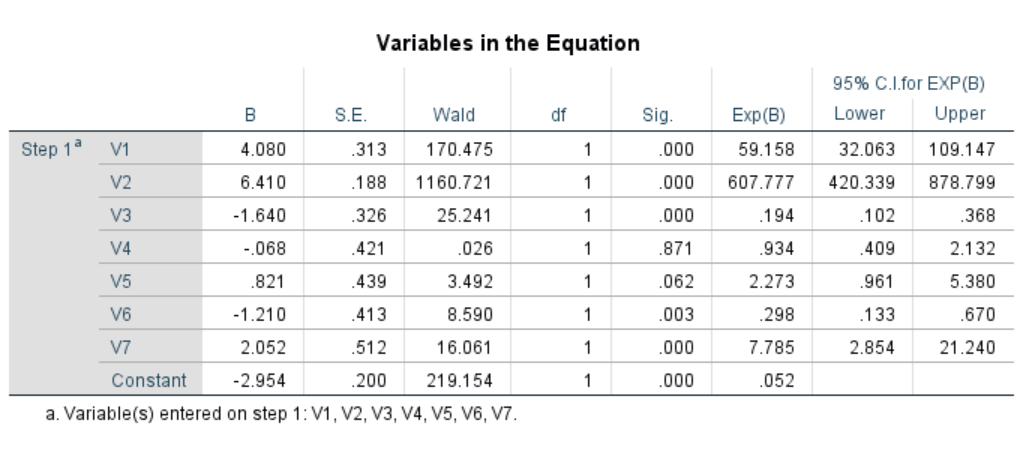
首先，我们选择使用主成分分析法。主成分分析主要思想是降维，目的是在信息损失较少的前提条件下，把多个指标转化为几个综合指标的多元统计方法。这些新的综合指标称为主成分。本文用公式对指标数据进行标准化，来消除量级的影响。将标准化后的影响判定特定成分存在的7个指标V1，V2...V7， 来表示，用Y表示标准化后的因变量。本文构建模型都是用标准化后的数据，故在本章中提到的“指标”都是指Vi，用F表示第i个主成分，则求主成分的原理如下:向量V= (V1，V2... V7)T表示7个指标构成的7维随机向量。设随机向量V的均值为μ，协方差矩阵为E。对V进行线性变换，可以形成新的主成分，由原来的变量线性表示，即满足下式:

结果如下图：





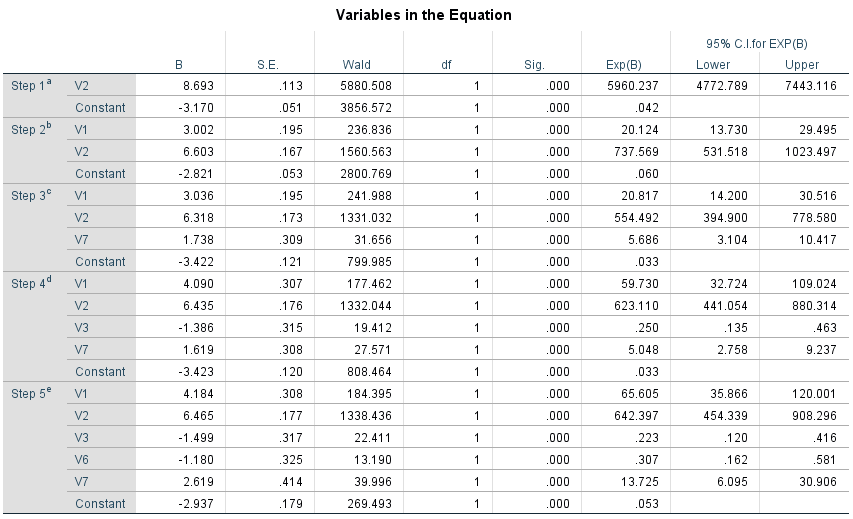
根据逻辑回归模型对上述主成分进行分析得到回归方程如下：



指标的重要性可以通过Exp(B)（优势比）的值来判定。从上表可以看出，指标V1、V2、V5、V7的优势比相较于其他指标的优势比具有显著性，所以判定特定成分存在的主要指标为V1、V2、V5、V7。

将主成分分析法的结果与Logistic回归进行比较，发现模型并没有很大的改进，而且主成分分析法有一个明显的缺点：该方法得到的新变量具有一定的模糊性，不像原始变量的含义那么清楚、确切，而且没有办法准确找到判断特定成分是否存在的主要指标，所以我们不选择这一方法。

为了消除多重共线性，我们用SPSS软件在进行Logistic回归时选用了“向前选择法：最大偏似然估计似然比检验”这一方法。向前选择法把候选的自变量逐个引入回归方程进行回归系数的显著性检验，然后决定是否把该自变量引入模型。似然比检验考虑整个模型的拟合情况，寻找能够以较高概率产生观察数据的系统发生树。因此Forward：LR能够筛选出部分变量，降低变量间的多重共线性。使用SPSS软件的分析结果如下：



Step5为最终的结果。

根据20000个训练数据中的80%的样本拟合Logistic回归模型结果如下：

β0 = -2.937 β1 = 4.184 β2 = 6.465

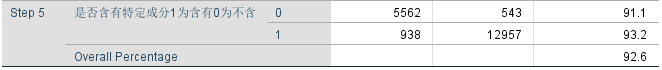
β3 = -1.499 β6 = -1.180 β7 = 2.619

由此，我们可以得到y与各个指标之间的关系式，关系如下：

y =

与最初的Logistic回归模型相比，该模型删除了V4、V5两个变量，且模型预测的准确率达到了92.6%（如图x所示）。因此通过比较，我们选用Logistic回归“Forward：LR”方法得到的模型来寻找主要指标。





指标的重要性可以通过Exp(B)（优势比）的值来判定。从（表x）可以看出，指标V1、V2、V7的优势比相较于其他指标的优势比具有显著性，所以判定特定成分存在的主要指标为V1、V2、V7。

# 问题二：成分存在与否的判定是否存在模糊区域

## 问题分析

问题二要求分析成分存在与否的判定是否存在模糊区域，即7项指标在这个区域内时，很难去判断成分是否存在。在本文采用的Logistic回归模型中，分类阈值恰好对应模糊区域。

## 模型建立

在Logistic回归模型中

当为0.5时，是Logistic回归的分类阈值，此时分到0，1这两类的可能性

是相同的，很难对数据进行分类。即当v1,v2,…,v7这7项指标满足如下条件时，很难去判断成分是否存在。

故存在模糊区域（对于7项指标），当7项指标在 这个区域内时，很难判断成分是否存在。

## 模型求解与结果分析

利用训练数据建立Logistic回归模型如下：

y =

求解7项指标的模糊区域如下：

# 问题三：建立数学模型，判断测试数据

## 问题分析

问题三要求建立数学模型对测试数据中每一个混合物是否含有特定成分进行判定，只需给出前10个混合物的判定即可。根据“训练数据”建立逻辑回归模型，然后对“测试数据”的结果进行预测。

## 模型准备

因变量Y是一个定性变量，只存在两种结果。基于因变量的特点，首先构建函数f(z)，使得经此函数变换后y的取值范围在[0,1]内。一个符合要求的函数为：

 (7.1)

其函数图像如图1所示。该函数自变量的取值范围是(-∞，+∞)，函数值的取值范围是[0,1]。



图1 f(z)函数图像

y的概率分布为：

 (7.2)

根据离散型随机变量期望值的定义，可得

 (7.3)

Y与x间的关系应满足：

E(*yi*) = ，i = 1,2,...,n (7.4)

其中， f(Xi,β) = (7.5)

在该式中，Xi’ = [1,Xi1,...,Xi7]，i = 1,2,...,n，β=[β0,β1,...,β7]’ 。β为七项指标组成的参数向量。

式7.4也可以表示为：

θi = P{yi = 1} = , i=1,2,...,n (7.6)

对模型7.6作变换可得：

y = ln =β0+β1X1+...+β7X7 (7.7)

所以最终的概率计算公式为：

P = (7.8)

## 模型建立、评估与求解

### 模型的建立

建立的Logistic回归模型如下：

y =

根据20000个训练数据中的80%的样本拟合Logistic回归模型结果如下：

β0 = -3.035 β1 = 3.601 β2 = 6.132 β3 = -1.075

β4 = 0.367 β5 = 0.986 β6 = -1.140 β7 = 1.307

由此，我们可以得到y与各个指标之间的关系式，关系如下：

y =

### 模型的评估

对剩下20%的样本是否含有特定成分进行预测，并对建立的模型进行评估。

计算混淆矩阵如下：

[[1096 100]

[ 195 2609]]

根据矩阵可以看出TP = 1096，FP = 100，FN = 195，TN = 2609，大部分观测值出现在了第二、四象限，这里的数值越多越好，反之，在第一、三象限对应位置出现的观测值越少越好，由此可以看出模型是十分优秀的。

混淆矩阵里面统计的是个数，有时候面对大量的数据，光凭算个数，很难衡量模型的优劣。因此可以再混淆矩阵的结果基础上计算如下几个指标：

准确率（Accuracy）ACC = = 0.92625

精确率（Precision）PPV = = 0.91639

灵敏度（Sensitivity）TPR = = 0.84895

特异度（Specificity）TNR = = 0.96309

F1 Score = = 0.88138

准确率表示模型所有判断正确的结果占总观测值的比重，精确率表示在模型预测混合物含有特定成分的所有结果中，模型预测对的比重，灵敏度表示在真实值是混合物含有特定成分的所有结果中，模型预测对的比重，特异度表示在真实值是混合物不含有特定成分的所有结果中，模型预测对的比重。F1-Score指标综合了Precision与Recall的产出的结果。F1-Score的取值范围从0到1的，1代表模型的输出最好，0代表模型的输出结果最差。从这几个指标都可以看出，模型是很优秀的。

### 问题求解

根据得到的模型对测试数据进行预测，这里仅给出前10个混合物的判定为例：

[1 1 1 1 0 1 1 1 1 1]

即：混合物5中不含特定成分，混合物1-4、6-10中含有特定成分。

# 模型分析与评价

## 模型评价

### 优点

1. 利用python对数据进行处理并做出散点图，简便、直接、快捷。

2. 建立的模型有成熟的理论基础，可信度较高。

3. 适当运用SPSS、Minitab软件辅助运算，提高了模型数据的处理效率。

### 缺点

1. 七个变量之间存在多重共线性，虽然最终构建的模型准确率达到了92%，但是共线性问题依旧没有解决。

2. 模型容易出现欠拟合，分类精度不够高。

3. 在模糊区域的判定划分中，区域边界界定的不够明确、科学。

## 模型改进

这里写模型改进。

**参考文献**

[1]迟畅. 基于主成分的回归和决策树对高血压数据的处理[D].吉林大学,2019.

[2]司守奎，数学建模简明教程，北京：国防工业出版社，2019。