**实验分析及实验设计**

**摘要**

C4烯烃作为重要的工业原料，在工业生产中被广泛运用。传统的制备方法通常采用化石原料，其会造成严重的环境污染案问题。随着对自然环境重视程度的提高，现采用新型的绿色环保燃料乙醇来制备丁醇。

针对问题一，对乙醇转化率、C4烯烃的选择性与温度的关系进行定量研究。首先进行数据预处理，将自变量温度标准化后，画出21组折线图，发现乙醇转化率、C4 烯烃的选择性与温度呈正相关关系。然后，运用**Shapiro-wilk检验**，判断样本数据是否符合正态分布的要求。最后，建立**相关性分析模型**，引入**皮尔逊相关系数r**，分别求得21组不同催化剂组合下乙醇转化率、C4烯烃的选择性与温度之间的r值与p值。

接着，分析附件2中的任务数据，对350度特定催化剂组合下一次实验不同时间下的结果进行分析。分析发现有些指标与时间的相关性极弱，所以需要先利用**皮尔逊相关系数法**，剔除掉p>0.05的三个指标，最后再引入**偏相关分析模型**，研究其余4个的实验指标与时间的直接关系，得到4个**偏相关系数**，划分系数范围，解释其紧密型。

针对问题二，探讨不同催化剂组合及温度对乙醇转化率以及 C4 烯烃选择性大小的影响。不同的催化剂组合中包含着4个不同的变量，其组合方式没有一定规律可循，于是，我们将21组不同的催化剂组合中的影响因素拆开，分别研究单一因素对结果的影响。除此之外还需考虑温度对结果的影响。综合所有自变量，发现其可分为两类——无序变量和有序变量。针对**有序变量**，采用**偏相关分析模型**；而针对**无序变量**，采用**控制变量法**，分别分析不同自变量对乙醇转化率以及 C4 烯烃选择性大小的影响。

针对问题三，需要选择出能提高C4烯烃收率的催化剂组合与温度。在问题2的基础上，我们将对催化剂变量作更深一步的探讨。先对无序分类变量进行**虚拟化，**然后进行回归分析，探讨无序分类变量与C4烯烃收率之间的关系。接着再对其余变量进行**多元逐步回归分析**，探讨这些变量对C4烯烃收率的影响。

针对问题四，需要在问题三的基础上，将实验数据中的21组实验样本分别两两进行分析，若判定某两个样本可以构成一组对照实验，则将其标记为“1”，反之，标记为“0”。这样，我们可以得到一个21\*21的**对称矩阵**。分析矩阵数据，挑选出在原有的实验中缺乏对照实验的催化剂组合，即含“1”最少的5种催化剂组合。进一步分析，确定5组催化剂组合中的单变量及对照实验中的数据设置，得到5组能够进一步探究5种不同实验因素对C4烯烃收率影响的**实验设计**。

**关键词：Shapiro-wilk检验 皮尔逊相关系数 偏相关分析模型 多元逐步回归分析**

# **问题重述**

## **问题的背景**

考古学家说，99%的人类历史都是由文物和化石来展现的。通过分析文物可以研究其所处的时代，它为什么出现在那里的原因，以及所具有的文化功能和社会价值。而玻璃作为丝绸之路的早期贸易往来的宝贵物证，研究其特征和化学成分，可以推动我国古代历史、文化以及生产力的研究。

但我国古代玻璃因其化学成分的不同，特别容易受环境的影响。在整个的风化过程中，玻璃内部的化学成分会和大气发生作用从而导致其成分产生变化，从而给考古工作增加了难度。因此，如何对风化过后的玻璃进行处理和分类，以及对其化学成分进行分析，便具有了重要意义。

现有一批我国古代玻璃制品的相关数据，考古工作者依据这些文物样品的化学成分和其他检测手段已将其分为高钾玻璃和铅钡玻璃两种类型。

## **问题提出**

现有一批我国古代玻璃制品的相关数据，考古学家根据其化学成分和其他检测方法，将其分为高钾和铅钡两类。给出附件1（文物的分类信息），附件2（相应的主要成分所占比例）以及附件3（未知类别玻璃文物的化学成分）。

根据题目背景和上述附件，我们需要建立数学模型解决如下问题：

1. 问题一：分析这些玻璃文物的表面风化与玻璃类型、装饰、颜色之间的关系；再从玻璃类型和玻璃有无风化分析其化学成分含量的统计规律；从风化点的数据去推测风化前的化学成分含量。
2. 问题二：根据附件1和2，推出高钾和铅钡玻璃的分类规则；选择每一类别玻璃合适的化学成分，对其进行细分，给出具体的分类方法和分类结果，评价分类的合理性和敏感性。
3. 问题三：根据附件3，根据未知玻璃的化学成分，去推算它所属的类型，并对结果进行敏感性分析。
4. 问题四：对不同类别的玻璃之间的化学成分进行分析，找出其关联关系，比较不同类别玻璃之间化学成分关联关系的差异性。

# **问题分析**

## **问题一分析**

~~题目一要求探究乙醇转化率、C4 烯烃的选择性与温度的关系。根据附件1给出的数据，可以先进行数据预处理。将自变量温度标准化后，对数据进行可视化处理，根据图像初步分析自变量（温度）与两个因变量（乙醇转化率、C4 烯烃的选择性）之间的关系。然后，运用Shapiro-wilk检验，判断样本数据是否符合正态分布的要求。最后，建立相关性分析模型，~~~~引入皮尔逊相关系数r，分别求得乙醇转化率、C4 烯烃的选择性与温度之间的r值与p值，得到自变量与因变量之间的相关性大小。~~

~~接着，题目要求分析附件2中的任务数据，对一次实验不同时间的测试结果进行分析。分析发现有些指标与时间的相关性极弱，所以需要先剔除掉这些不相关数据，最后再引入偏相关分析模型，研究各个不同的实验指标与时间的直接关系。~~

问题(1):分析这些玻璃文物纹饰、类型、颜色、表面风化的关系;这题很简单,分析关系有相关性分析和差异性分析,需要注意一点,这四个指标都是定类变量,并非连续变量,因此:对于相关性分析,不能直接使用皮尔逊相关分析,可以采用斯皮尔曼相关系数分析(Spearman相关系数)对于差异性分析,不能采用方差分析或T检验,应当采用卡方检验

问题(2) :基于这些玻璃文物的类型,分析文物样品表面有无风化化学成分含量的统计规律;对于类型只有两种,一种是高钟,一种是铅钡。在做完缺失值处理之后,我们可以对比文物样本表面有无风化的化学成分的一些统计学规律,这块主要其实也是差异性分析和相关性分析,不过在做这两个分析之前,我们可以做一些数据合或者分类汇总,观察推断出来两种玻璃文物类型的有无风化以及成分的差异或者相关情况,进而推断出来统计规律,形成公式。

问题(3) :根据风化点检测数据,预测其风化前的化学成分含量。根据问题2得到的相关关系形成公式,预测已经风化的文物在风化前的化学成分含量

## **问题二分析**

~~问题二中要求探讨不同催化剂组合及温度对乙醇转化率以及 C4 烯烃选择性大小的影响。不同的催化剂组合中包含着4个不同的变量，其组合方式没有一定规律可循，于是，我们从21组不同的催化剂组合中，分别提取出三种不同的影响因素，分别研究单一因素对结果的影响。进一步分析催化剂因素和温度后，发现其包含两种变量——无序变量和有序变量。针对有序变量，采用偏相关分析模型；而针对无序变量，采用控制变量法，分别分析变量对不同指标结果的影响~~。

问题(1):分析高钟玻璃、铅钡玻璃的分类规律;构建一个可解释的机器学习分类模型,例如决策树、逻辑回归,以类型(高钟玻璃、铅钡玻璃)为Y,尽可能构建足够多的特征X,形成可解释的分类规律,得分点关键在于模型寻优做的牛不牛以及可视化效果。

问题(2) :对于每个类别选择合适的化学成分对其进行亚类划分,给出具体的划分方法及划分结果分别对类别进行聚类模型,得分点关键在于亚类划分的可解释性以及划分效果。

## **问题三分析**

~~问题三要求选择出能得出较高C4烯烃收率的催化剂组合与温度。根据问题2中对数据的分析，我们将在此问中对催化剂变量作更深一步的探讨。我们先对无序分类变量进行虚拟化，然后进行回归分析探讨无序分类变量与C4烯烃收率之间的关系，接着，再对其余变量进行多元逐步回归分析，探讨这些变量对C4烯烃收率的影响。~~

问题(1) :对附件表单3中未知类别玻璃文物的化学成分进行分析,鉴别其所属类型;基于问题2,预测表单3未知类别玻璃文物的类型,至于化学成分分析不分析不是重点

问题(2) :对分类结果的敏感性进行分析。如果我们采用的是机器学习模型,那么这一步就无需进行分析,机器学习只需要在训练时分析敏感性,预测时模型已经固定,无需进行敏感性分析,但是如果我们是通过推断的方式来进行分析的话,那么这里需要复现一下敏感性分析

## **问题四分析**

问题(1):针对不同类别的玻璃文物样品,分析其化学成分之间的关联关系与问题1的问题(2)类似,只是减少了一个条件-有无风化

问题(2) :比较不同类别之间的化学成分关联关系的差异性。与问题1的问题(2)类似

# **模型假设**

1. 假设题目给出的数据真实可靠。
2. 绝对质量不同的Co/SiO2和HAP装料比是不同分类 (例如：50mg:50mg和200mg:200mg的Co/SiO2HAP装料比在本文中不被定义为相同的量。

# **符号说明**

# **模型的建立与求解**

## **相关性分析模型**

### **数据预分析**

附件1中给出的是不同温度区间内的对应数据，为了便于比较分析，需要先筛选出相同温度区间（250—450摄氏度符号）内的C4烯烃的选择性数值和C4烯烃收率数值，紧接着用Excel将其可视化成横坐标为温度，纵坐标为两个变量（C4 烯烃的选择性和C4 烯烃收率）的折线图。结果如下图1所示。

图 1

|  |
| --- |
|  |
|  |  |

温度分别与C4 烯烃的选择性和C4 烯烃收率的关系

### **折线图分析**

经过初步观察和分析，可以发现两个变量都与温度呈正相关，并且大致呈线性关系。少部分图像中会出现突变值，偏离图像原有趋势。

### **相关性分析模型**

基于样本容量过小的问题，本文不采用拟合回归的方法。后续经过统计分析，发现数据条件符合以下三个基本特征：

1. 实验数据是成对的来自于正态分布的总体（Shapiro-wilk检验）

皮尔逊相关系数要求样本数据满足正态分布的要求，因此本文选用来查看样本数据是否满足正态分布。运用SPSS软件，对样本数据进行Shapiro-wilk检验，得到表1：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 表1自由度与显著性 | | | | | | | | | | | | | | | | |
| **夏皮洛-威尔克正态性检验** | | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |
|  | **V1** | **5** |  | **0.967** |  |  | **V1** | **5** |  | **0.967** |  |  | **V1** | **7** |  | **0.853** |
| **A1** | **V2** | **5** |  | **0.652** |  | **A2** | **V2** | **5** |  | **0.783** |  | **A3** | **V2** | **7** |  | **0.352** |
|  | **V3** | **5** |  | **0.231** |  |  | **V3** | **5** |  | **0.157** |  |  | **V3** | **7** |  | **0.491** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |
|  | **V1** | **6** |  | **0.964** |  |  | **V1** | **6** |  | **0.964** |  |  | **V1** | **5** |  | **0.787** |
| **A4** | **V2** | **6** |  | **0.838** |  | **A5** | **V2** | **6** |  | **0.838** |  | **A6** | **V2** | **5** |  | **0.249** |
|  | **V3** | **6** |  | **0.059** |  |  | **V3** | **6** |  | **0.19** |  |  | **V3** | **5** |  | **0.0601** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |
|  | **V1** | **5** |  | **0.787** |  |  | **V1** | **5** |  | **0.787** |  |  | **V1** | **5** |  | **0.787** |
| **A7** | **V2** | **5** |  | **0.815** |  | **A8** | **V2** | **5** |  | **0.207** |  | **A9** | **V2** | **5** |  | **0.031** |
|  | **V3** | **5** |  | **0.142** |  |  | **V3** | **5** |  | **0.435** |  |  | **V3** | **5** |  | **0.579** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |
|  | **V1** | **5** |  | **0.787** |  |  | **V1** | **5** |  | **0.787** |  |  | **V1** | **5** |  | **0.787** |
| **A10** | **V2** | **5** |  | **0.06** |  | **A11** | **V2** | **5** |  | **0.015** |  | **A12** | **V2** | **5** |  | **0.145** |
|  | **V3** | **5** |  | **0.054** |  |  | **V3** | **5** |  | **0.441** |  |  | **V3** | **5** |  | **0.277** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |
|  | **V1** | **5** |  | **0.787** |  |  | **V1** | **5** |  | **0.787** |  |  | **V1** | **5** |  | **0.787** |
| **A13** | **V2** | **5** |  | **0.051** |  | **A14** | **V2** | **5** |  | **0.169** |  | **B1** | **V2** | **5** |  | **0.139** |
|  | **V3** | **5** |  | **0.471** |  |  | **V3** | **5** |  | **0.1** |  |  | **V3** | **5** |  | **0.285** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |
|  | **V1** | **5** |  | **0.787** |  |  | **V1** | **6** |  | **0.964** |  |  | **V1** | **6** |  | **0.964** |
| **B2** | **V2** | **5** |  | **0.091** |  | **B3** | **V2** | **6** |  | **0.008** |  | **B4** | **V2** | **6** |  | **0.062** |
|  | **V3** | **5** |  | **0.275** |  |  | **V3** | **6** |  | **0.383** |  |  | **V3** | **6** |  | **0.091** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 续表1 | | | | | | | | | | | | | | | | |
|  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |  |  |  | **自由度** |  | **显著性** |
|  | **V1** | **6** |  | **0.964** |  |  | **V1** | **6** |  | **0.964** |  |  | **V1** | **6** |  | **0.964** |
| **B5** | **V2** | **6** |  | **0.096** |  | **B6** | **V2** | **6** |  | **0.089** |  | **B7** | **V2** | **6** |  | **0.066** |
|  | **V3** | **6** |  | **0.333** |  |  | **V3** | **6** |  | **0.103** |  |  | **V3** | **6** |  | **0.701** |
|  | | | | | | | | | | | | | | | | |

分析发现Shapiro-wilk的显著性基本大于0.05 水平，其结果符合零假设，因此认为样本数据符合正态分布的要求。

1. 实验数据中异常点数量较小
2. 每组样本之间具有独立性

为了进一步探究乙醇转化率、C4 烯烃的选择性与温度的关系，本文引用皮尔逊相关系数，建立相关性分析模型，以此来度量温度和C4 烯烃的选择性以及温度和 C4 烯烃收率之间的相关程度，对其进行量化分析。相关系数的绝对值越大，则两变量之间的相关性越强，反之，相关性越弱。

### **相关性分析**

在此模型中，首先设立原假设和备择假设：H0：r=0，H1：r≠0。在原假设成立并且满足上述三个特征的的条件下，利用皮尔逊相关性系数r，构造出一个符合正态分布的统计量，其中的特征值为：

|  |  |
| --- | --- |
|  | () |

其中，为样本容量，将值代入，运用SPSS对样本数据进行运算，可以得到最终的检验值。接着，将的概率定义为该组数据的单侧p值，则为双侧p值，通过下列计算出最终的p值：

|  |  |
| --- | --- |
|  | () |

其中，为累计概率密度函数，为该组数据的自由度，分别将数值代入，可以得到表2结果：

表2【r值与p值】

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **催化剂组合编号** | **乙醇转化率(%)** | | |  | **C4烯烃选择性(%)** | | |
| **R皮尔逊相关系数** |  | **P值** |  | **R皮尔逊相关系数** |  | **P值** |
| A1 | 0.9655 |  | 0.0077\*\*\* |  | 0.8871 |  | 0.0448\*\* |
| A2 | 0.995 |  | 0.0004\*\*\* |  | 0.9143 |  | 0.0297\*\* |
| A3 | 0.982 |  | 0.0001\*\*\* |  | 0.9554 |  | 0.0008\*\*\* |
| A4 | 0.9975 |  | 0.0000\*\*\* |  | 0.9578 |  | 0.0026\*\*\* |
| A5 | 0.9344 |  | 0.0063\*\*\* |  | 0.9696 |  | 0.0014\*\*\* |
| A6 | 0.9836 |  | 0.0025\*\*\* |  | 0.8854 |  | 0.0458\*\* |
| A7 | 0.9994 |  | 0.0000\*\*\* |  | 0.9682 |  | 0.0068\*\*\* |
| A8 | 0.9771 |  | 0.0042\*\*\* |  | 0.9916 |  | 0.0009\*\*\* |
| A9 | 0.9206 |  | 0.0265\*\* |  | 0.9974 |  | 0.0002\*\*\* |
| A10 | 0.923 |  | 0.0254\*\* |  | 0.8615 |  | 0.0606\* |
| A11 | 0.9032 |  | 0.0356\*\* |  | 0.989 |  | 0.0014\*\*\* |
| A12 | 0.9632 |  | 0.0084\*\*\* |  | 0.9832 |  | 0.0026\*\*\* |
| A13 | 0.9366 |  | 0.0190\*\* |  | 0.9883 |  | 0.0015\*\*\* |
| A14 | 0.9639 |  | 0.0082\*\*\* |  | 0.9592 |  | 0.0082\*\*\* |
| B1 | 0.962 |  | 0.0088\*\*\* |  | 0.9858 |  | 0.0020\*\*\* |
| B2 | 0.9292 |  | 0.0224\*\* |  | 0.9848 |  | 0.0022\*\*\* |
| B3 | 0.8899 |  | 0.0175\*\* |  | 0.971 |  | 0.0013\*\*\* |
| B4 | 0.8996 |  | 0.0146\*\* |  | 0.895 |  | 0.0160\*\*\* |
| B5 | 0.9129 |  | 0.0111\*\* |  | 0.9776 |  | 0.0007\*\*\* |
| B6 | 0.9399 |  | 0.0053\*\*\* |  | 0.8818 |  | 0.0201\*\*\* |
| B7 | 0.9361 |  | 0.0060\*\*\* |  | 0.9944 |  | 0.0000\*\*\* |

分析发现，样本数据基本符合p < 0.05 ，说明在95%的置信水平上拒绝原假设，意味着皮尔逊相关系数显著异于0，两个变量(V2与V3)具有一定的相关性。最后，本文对两个变量与温度的**皮尔逊相关系数**r做进一步分析，规定：

|  |  |
| --- | --- |
| 0<r≤0.25: | 相关性极弱 |
| 0.25<r≤0.50： | 相关性弱 |
| 0.50<r≤0.75： | 相关性强 |
| 0.75<r≤1.00： | 相关性极强 |

可以得出：

1. 在催化剂组别A2、A4、A5、A7中，乙醇转化率与温度的相关性极强；在催化剂组别A6、A8—A14、B1—B7中，乙醇转化率与温度的相关性极弱。
2. 在催化剂组别A9、B7中，C4 烯烃的选择性与温度的相关性强；在催化剂组别A1、A2、A4—A7、A10、A14、B4、B6中，C4 烯烃的选择性与温度的相关性极弱。

## **偏相关分析模型：**

### **数据预处理**

由附件2中的数据可以得知在特定催化剂条件下，七种指标随时间变化的不同数值。对数据进行初步分析，发现不仅“时间”会对目标变量造成影响，其他指标也会对目标变量造成一定程度的影响，并且还发现部分数据与时间没有直接的相关性。因此，首先采用皮尔逊相关系数法，计算出的p值如表3所示，发现有3个指标的p>0.05。为了更好地分析目标指标与时间的直接关系，去除掉这些显著性大于0.05 的指标：C4烯烃选择性、甲基苯甲醛和甲基苯甲醇和其他。

表3 7个指标的显著性

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **乙醇** | **乙烯** | **C4烯烃** | **乙醛** | **碳数为4-12** | **甲基苯甲醛和甲基** | **其他** |
|  | **转化率(%)** | **选择性(%)** | **选择性(%)** | **选择性(%)** | **脂肪醇选择性(%)** | **苯甲醇选择性(%)** |
| **p值** | **0.00039141** | **0.0011** | **0.6427** | **0.00011887** | **0.0143** | **0.2819** | **0.5008** |

### **偏相关分析模型**

分别以不同的参考值（乙烯选择性、乙醛选择性和碳数为4-12脂肪醇中的任意一个）为自变量，X1（时间）为因变量，进行多元线性回归，得到X1最终的残差；再以不同的参考值为自变量，X2（乙醇转化率）为因变量，进行多元线性回归，得到X2的残差。最后，再对X1的残差和X2的残差进行简单相关计算，得到最终的偏相关系数，结果如表4所示。

表4【PR表】

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **乙醇转化率(%)** | **乙烯选择性(%)** | **乙醛选择性(%)** | **碳数为4-12脂肪醇(%)** |
| **偏相关系数** | **-0.5314** | **-0.1439** | **0.6899** | **0.0571** |

|  |  |
| --- | --- |
| 0<r≤0.25: | 相关性极弱 |
| 0.25<r≤0.50： | 相关性弱 |
| 0.50<r≤0.75： | 相关性强 |
| 0.75<r≤1.00： | 相关性极强 |

可以得出以下结论：

乙醇转化率、乙烯选择性与时间呈负相关关系，乙醇转化率与时间的相关性强，乙烯选择性与时间的相关性极弱；

乙醛选择性、碳数为4-12脂肪醇选择性与时间呈正相关关系，乙醛选择性与时间的相关性强，碳数为4-12脂肪醇选择性与时间的相关性弱。

## **结合控制变量的偏相关分析**

### **数据预处理【还需要加入zhn的数据隔离依据】**

进一步分析附件1中的数据，发现不同催化剂组合中包含不同的Co/SiO2和HAP装料比、Co负载量、乙醇浓度。其中Co/SiO2和HAP装料比为无序变量，Co负载量、乙醇浓度为有序变量。此外，我们还需考虑有序变量温度。因此，本文采用偏相关分析模型研究Co负载量、乙醇浓度和温度对2个目标指标（乙醇转化率以及 C4 烯烃选择性大小）的影响，再采用控制变量法探究不同的Co/SiO2和HAP装料比对2个目标指标（乙醇转化率以及 C4 烯烃选择性大小）的影响。

### **偏相关分析模型**

探讨不同的Co负载量(wt%)、乙醇浓度(ml/min)）及温度对乙醇转化率以及 C4 烯烃选择性大小的影响。沿用2.2的模型，在3个指标中选取一个目标值（如Co负载量），将其他2个指标定义为参考值。分别以不同的参考值（乙醇浓度及温度中的任意一个）为自变量，X1（乙醇转化率）为因变量，进行多元线性回归，得到X1最终的残差；再以不同的参考值为自变量，X2（Co负载量）为因变量，进行多元线性回归，得到X2的残差。最后，再对X1的残差和X2的残差进行简单相关计算，得到最终的偏相关系数，结果如表5所示。

表5（6个偏相关结果）

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Co负载量(%wt)** | **乙醇浓度(ml/min)** | **温度** |  |  | **Co负载量(%wt)** | **乙醇浓度(ml/min)** | **温度** |
| **乙醇转化率** | **0.1715** | **-0.4918** | **0.8083** |  | **C4烯烃选择性** | **-0.2046** | **-0.0846** | **0.7072** |

可以得出以下结论：

乙醇浓度与乙醇转化率呈负相关关系，Co负载量、温度与乙醇转化率呈正相关关系。

Co负载量、乙醇浓度与C4烯烃选择性呈负相关关系，温度与C4烯烃选择性呈正相关关系。

规定：

|  |  |
| --- | --- |
| 0<r≤0.25: | 相关性极弱 |
| 0.25<r≤0.50： | 相关性弱 |
| 0.50<r≤0.75： | 相关性强 |
| 0.75<r≤1.00： | 相关性极强 |

进一步分析得：

温度与乙醇转化率相关性极高，乙醇浓度与乙醇转化率相关性弱，Co负载量与乙醇转化率相关性极弱；

温度与C4烯烃选择性相关性高，Co负载量和乙醇浓度与C4烯烃选择性相关性极弱。

### **控制变量法**

这里筛选出相同条件的Co负载量、乙醇浓度和温度，探究不同的Co/SiO2和HAP装料比对乙醇转化率以及 C4 烯烃选择性大小的影响。例如：当Co负载量取1wt%，乙醇浓度取1.68ml/min，温度为250时，分别观察不同Co/SiO2和HAP装料比下的乙醇转化率以及 C4 烯烃选择性大小，将其可视化后得到图2和图3：

|  |  |
| --- | --- |
|  |  |
| 图二 | 图三 |

分析图像，可以发现：

Co/SiO2和HAP装料比为2和3时，乙醇转化率最高，为2.8%；

Co/SiO2和HAP装料比为1时，C4 烯烃选择性最高，为34.05%。

## **逐步回归模型**

### **数据预处理**

由于本题需要研究催化剂的组合方式，故本文在研究开始前，将附件1中催化组合变量拆分成为三个不同的变量：CO负载量、Co /SiO2 和 HAP 装料比和乙醇浓度，以此来深入讨论该如何选取高效催化剂和温度，使得在其他实验条件不变的情况下C4烯烃收率尽可能高。

接着，对本题中各个变量进行分类：

|  |  |
| --- | --- |
| 装料方式 | 二分类变量 |
| Co /SiO2 和 HAP 装料比 | 多分类无序变量 |
| CO负载量、乙醇浓度和温度 | 有序变量 |
| C4烯烃收率 | 连续型数值变量 |

在后续的讨论中，对于不同类型的变量我们会采取不同的方法去进行分析。

### **多元线性回归模型**

本文拟采用多元线性回归模型来求解问题三。

可知，若想使用多元回归分析需满足以下四个条件：

（1）需要至少2个自变量，且自变量之间互相独立；

（2）因变量为为连续变量；

（3）数据具有方差齐性、无异常值和正态分布的特点；

（4）自变量间不存在多重共线性。

由于此题中C4烯烃收率为连续型数值变量，故本文将其作为结果变量，将装料方式、CO负载量、Co /SiO2 和 HAP 装料比和乙醇浓度作为起因变量并采用多元线性回归模型去研究如何选取温度及催化剂组合，使得C4烯烃收率尽可能高。

通过观察本题中各个变量的类型，我们首先研究Co /SiO2 和 HAP 装料比（多分类无序变量）对C4烯烃收率的影响。我们构建Co /SiO2 和 HAP 装料比变量的八个变量值：

表 Co /SiO2 和 HAP 装料比变量的对应值

|  |
| --- |
| **无HAP=0** |
| **Co/SiO2 和 HAP 质量比为 200mg：200mg=1** |
| **Co/SiO2 和 HAP 质量比为 100mg：100mg=2** |
| **Co/SiO2 和 HAP 质量比为 75mg：75mg=3** |
| **Co/SiO2 和 HAP 质量比为 50mg：50mg=4** |
| **Co/SiO2 和 HAP 质量比为 25mg：25mg=5** |
| **Co/SiO2 和 HAP 质量比为 10mg：10mg=6** |
| **Co/SiO2 和 HAP 质量比为 33mg：67mg=7** |
| **Co/SiO2 和 HAP 质量比为 67mg：33mg=8** |

然后通过SPSS软件对此变量进行虚拟化操作（Dummy Coding），即将此变量设置为哑变量。通过多次调试，此多元回归符合上述四个条件，并且我们发现：在其他实验条件不变的情况下，当b被定义为参考类别的时候，在95%置信水平下，以a作为催化剂条件所得到的C4烯烃收率显著大于以b作为催化剂条件所得的C4烯烃收率(p<0.05)，且其他催化剂条件所得的C4烯烃收率与b催化剂条件所得的C4烯烃收率无显著差异（p>0.05）。

**表系数？？**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **未标准化系数** | | |  | **标准化系数** |  | **t** |  | **显著性** |
| **B** |  | **标准错误** |  | **Beta** |  |  |
| 1 | (常量) | 4.811 |  | 2.244 |  |  |  | 2.144 |  | 0.034 |
| a | 5.921 |  | 2.721 |  | 0.291 |  | 2.176 |  | 0.032 |
| c | 0.101 |  | 4.297 |  | 0.002 |  | 0.023 |  | 0.981 |
| d | -0.236 |  | 2.697 |  | -0.012 |  | -0.087 |  | 0.931 |
| e | -3.276 |  | 4.297 |  | -0.079 |  | -0.762 |  | 0.448 |
| f | -3.863 |  | 4.297 |  | -0.093 |  | -0.899 |  | 0.371 |
| g | -1.79 |  | 4.599 |  | -0.039 |  | -0.389 |  | 0.698 |
| z | -4.216 |  | 4.599 |  | -0.093 |  | -0.917 |  | 0.361 |

a. 因变量：C4烯烃收率1

### **逐步回归分析**

在确定了多分类无序变量对C4烯烃收率的影响后，我们研究二分类变量，有序变量对C4烯烃收率的影响。

通过观察C4烯烃收率与某些二分类变量或有序变量的散点图，我们发现难以判断自变量与C4烯烃收率的相关关系，故本文引入多元逐步回归分析来针对性求解此题。使用多元逐步回归的前提条件和使用多元线性回归的前提条件一样，故我们也将对其进行验证。

相较于多元线性回归分析，逐步回归具备更合理的自变量筛选机制，既可以在一定程度上避免多重共线性的影响，也可以有效地剔除不显著的自变量。在分析过程中，我们选用逐步法进行自变量筛选。逐步法结合向后法和向前法的优点，在向前引入每一个新自变量之后都要重新对已代入的的自变量进行计算，以此来检验这些自变量有无继续保留的价值，并以此为依据进行引入和剔除，直到无自变量可引入或剔除。

在多元逐步回归分析中，本文设定变量纳入方程的显著性标准为0.05，即变量需要至少在百分之九五的置信水平下显著才可以被纳入方程。通过SPSS软件我们得到了回归模型。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型摘要d？？？？？？？？？** | | | | | |
| 模型 | R | R 方 | 调整后 R 方 | 标准估算的错误 | 德宾-沃森 |
| 1 | .726 | .527 | .522 | 6.457625599220834 |  |
| 2 | 0.752 | .566 | .558 | 6.213690348328155 |  |
| 3 | .766c | .587 | .575 | 6.089264172037740 | 1.171 |
| c. 预测变量：(常量), 温度, 乙醇浓度(ml/min), 装料方式(A类催化剂为0,B类催化剂为1) | | | | | |
| d. 因变量：C4烯烃收率（删除？ | | | | | |

Durbin-Watson值位于1-2之间，说明自变量之间基本相互独立

|  |  |
| --- | --- |
|  |  |
|  |  |

从标准化残差直方图和标准化残差的正态P-P图可以看出残差基本满足正态分布曲线，所以残差的正态性基本通过检验。

**ANOVAa？？？？？？**要留下来吗？？？

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模型 | | 平方和 | 自由度 | 均方 | F | 显著性 |
| 1 | 回归 | 5195.012 | 1 | 5195.012 | 124.578 | .000b |
| 残差 | 4670.504 | 112 | 41.701 |  |  |
| 总计 | 9865.516 | 113 |  |  |  |
| 2 | 回归 | 5579.812 | 2 | 2789.906 | 72.259 | .000c |
| 残差 | 4285.704 | 111 | 38.610 |  |  |
| 总计 | 9865.516 | 113 |  |  |  |
| 3 | 回归 | 5786.811 | 3 | 1928.937 | 52.022 | .000d |
| 残差 | 4078.705 | 110 | 37.079 |  |  |
| 总计 | 9865.516 | 113 |  |  |  |
| d. 预测变量：(常量), 温度, 乙醇浓度(ml/min), 装料方式(A类催化剂为0,B类催化剂为1) | | | | | | | |
|  | | | | | | | |

由于所得模型的显著性水平小于0.05，即在95%的置信水平上此模型显著，此模型被判定为一个成功的模型。

最后我们看到调整后的R方达到了50%以上，说明了我们的变量基本可以解释C4烯烃收率变化的50%以上，所以我们得到了我们的多元逐步回归式

式子

从此式我们可以得到结果，在其他条件不变的前提下，温度越高，乙醇浓度越低，装料方式越接近A（装料方式越远离B），C4烯烃收率越高。

结合前面对多分类无序变量的讨论，我们得到最终结论：要想使C4烯烃收率尽量高，我们需要适当提高实验温度（在本题中实验温度在400度左右最佳），尽量降低乙醇浓度（在本题中降低到0.3ml/min效果最佳），尽量选择装料方式I，尽量选择催化剂Co /SiO2 和 HAP 装料比为200mg:200mg，CO负载量在合理的范围内进行选择即可**。**

**为什么空很多行？**

在去除了实验维度为350度及以上的样本后，我们仅对实验温度在350度以下的样本重复上述分析，发现在对Co /SiO2 和 HAP 装料比进行回归分析后，在95%置信水平下，以a作为催化剂条件所得到的C4烯烃收率显著大于其他配比条件下所得的C4烯烃收率。

表 系数？？？

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **模型** | | **未标准化系数** | | | **标准化系数** | **t** | **显著性** | **共线性统计** | | |
| **B** |  | **标准错误** | **Beta** | **容差** |  | **VIF** |
| 1 | (常量) | 0.911 |  | 0.994 |  | 0.917 | 0.363 |  |  |  |
| a | 2.973 |  | 1.136 | 0.473 | 2.618 | 0.011 | 0.341 |  | 2.935 |
| c | 0.123 |  | 1.649 | 0.01 | 0.075 | 0.941 | 0.673 |  | 1.485 |
| d | -0.232 |  | 1.142 | -0.036 | -0.203 | 0.84 | 0.346 |  | 2.894 |
| e | -0.72 |  | 1.649 | -0.056 | -0.437 | 0.664 | 0.673 |  | 1.485 |
| f | -0.815 |  | 1.649 | -0.064 | -0.494 | 0.623 | 0.673 |  | 1.485 |
| g | -0.728 |  | 1.815 | -0.05 | -0.401 | 0.69 | 0.73 |  | 1.37 |
| h | -0.657 |  | 1.815 | -0.045 | -0.362 | 0.719 | 0.73 |  | 1.37 |
| z | -0.9 |  | 1.815 | -0.061 | -0.496 | 0.622 | 0.73 |  | 1.37 |

因变量：C4烯烃收率

**然后，再对**二分类变量、有序变量进行讨论。我们可以得到如下模型：

式子

结果发现C4烯烃收率依旧与温度呈弱正相关，与是否使用装料方式I呈正相关。但是相比于没有去除温度大于350度的样本所作的多元逐步回归模型，去除温度大于350度的样本后所得的回归模型中，乙醇浓度并不显著影响C4烯烃收率，这说明，在温度较低的时候，乙醇浓度的变化并不会显著的增加或减少C4烯烃收率。所以当温度小于350度时，我们应该选择相对更高的实验温度，尽量选择装料方式I，尽量选择催化剂Co /SiO2 和 HAP 装料比为200mg:200mg，乙醇浓度及CO负载量在合理的范围内进行选择即可。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | 调整后 | 标准估算的错误 | 德宾-沃森 | |
| 2 | .618b | 0.382 | 0.365 | 2.34302 | | 1.372 |
| 预测变量：(常量), 温度, 装料方式(A类催化剂为0,B类催化剂为1) | | | | | |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | 平方和 | 自由度 | 均方 | F | 显著性 |
| 2 | 回归 | 237.958 | 2 | 118.979 | 21.673 | .000c |
| 残差 | 384.282 | 70 | 5.49 |  |  |
| 总计 | 622.239 | 72 |  |  |  |

预测变量：(常量), 温度, 装料方式(A类催化剂为0,B类催化剂为1)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **未标准化系数** | | | **标准化系数** | | **t** | | **显著性** | | **共线性统计** | | | |
| **B** |  | **标准错误** | | **Beta** | |  | |  | | **容差** |  | **VIF** | |
| **2** | (常量) | -15.744 |  | 3.032 | |  | | -5.192 | | 0 | |  |  |  | |
| 温度 | 0.064 |  | 0.011 | | 0.561 | | 5.95 | | 0 | | 0.994 |  | 1.007 | |
| 装料方式(A类催化剂为0,B类催化剂为1) | -1.889 |  | 0.575 | | -0.31 | | -3.288 | | 0.002 | | 0.994 |  | 1.007 | |

## **实验设计思路**

### 

依据前文的数据处理方式，将催化剂组合中的影响因素拆开分别考虑，CO负载量、装料物质的类别（第一类为Co /SiO2 和HAP，第二类为Co/SiO2和石英砂）、装料比和乙醇浓度。除此之外，还有一个变量是温度。以便后续在设计实验中可以更清晰地分类催化剂。

本题中，我们将设计5次实验，以便更好地研究催化剂及温度对各个指标的影响。我们拟采用对照实验的方法，对原实验数据中没有或少有对照组的实验样本进行对照组的补充。我们以控制变量为原则，认为若两个实验样本有且仅有一处催化剂条件的不同，则这两个实验可以构成一组对照实验，反之，若两个实验样本有超过一处催化剂条件的不同或者催化剂条件完全相同，则认为其无法构成一组对照实验。

本文将实验数据中21组实验样本分别两两进行分析，若判定某两个样本可以构成一组对照实验，则将其标记为1，反之，标记为0。这样，我们可以得到一个21\*21的对称矩阵。矩阵形式如下：



### **矩阵分析**

分析矩阵数据，可以发现除了A11，其余催化剂组合中存在可以与其形成对照的单一变量催化剂组合，这是因为A11的装料物质应用的是Co/SiO2和石英砂，所以在后续实验中，需要单独研究A11的装料物质对实验结果的影响，此为第一组实验。另外，将矩阵中出现0较多的序列拿出来单独分析，再筛选出以下4个催化剂组合：A5（1个1），A10（1个1），B5（2个1），B7（1个1）。分析这4种催化剂组合，发现其中A5 和A10缺乏关于温度这一单一变量的研究，B5和B7缺乏关于Co /SiO2和HAP 的不同装料比这一单一变量的研究。

### **实验设计**

由此，我们得出以下5次实验设计方案：

(1).将A11（50mg 1wt%Co/SiO2+ 90mg石英砂-乙醇浓度1.68ml/min，无HAP）这一催化剂组合中的投料物质设为单一变量，其余因素保持不变，增设另一组投料物质为Co /SiO2 和HAP，投料比为50:90，乙醇浓度为1.68ml/min，温度为250的对照实验。

探究在一定条件下（投料物质为Co /SiO2 和HAP，投料比为50:90，乙醇浓度为1.68ml/min，温度为250），不同投料物质对C4 烯烃收率的影响。

(2).将A5（200mg 2wt%Co/SiO2- 200mg HAP-乙醇浓度0.3ml/min）这一催化剂组合中的温度设为单一变量，其余因素保持不变，增设另一组200mg 2wt%Co/SiO2- 200mg HAP，乙醇浓度0.3ml/min的对照实验。

探究在一定条件下（200mg 2wt%Co/SiO2- 200mg HAP，乙醇浓度0.3ml/min），温度对对C4 烯烃收率的影响。

(3).将A10（50mg 5wt%Co/SiO2- 50mg HAP-乙醇浓度2.1ml/min）这一催化剂组合中的温度设为单一变量，其余因素保持不变，增设另一组50mg 5wt%Co/SiO2- 50mg HAP-乙醇浓度2.1ml/min的对照实验。

探究在一定条件下（50mg 5wt%Co/SiO2- 50mg HAP-乙醇浓度2.1ml/min），温度对C4 烯烃收率的影响。

(4)将B5（50mg 1wt%Co/SiO2- 50mg HAP-乙醇浓度2.1ml/min）这一催化剂组合中的Co /SiO2和HAP的不同装料比设为单一变量，其余因素保持不变，增设另一组投料物质为Co /SiO2和HAP，乙醇浓度为2.1ml/min，温度为400，但投料比为200:200的对照实验。

探究在一定条件下（投料物质为Co /SiO2和HAP，乙醇浓度为2.1ml/min，温度为400），Co /SiO2和HAP的不同装料比对C4 烯烃收率的影响。

(5)将B7（100mg 1wt%Co/SiO2- 100mg HAP-乙醇浓度0.9ml/min）这一催化剂组合中的Co /SiO2和HAP的不同装料比设为单一变量，其余因素保持不变，增设另一组投料物质为Co /SiO2和HAP，乙醇浓度为0.9ml/min，温度为400，但投料比为200:200的对照实验。

探究在一定条件下（投料物质为Co /SiO2和HAP，乙醇浓度为0.9ml/min，温度为400），Co /SiO2和HAP的不同装料比对C4 烯烃收率的影响。

# **模型的评价**

## **模型的优点：**

## **模型的缺点：**

# **参考文献**

1. 姜启源, 谢金星, 叶俊. 数学模型.第4版[M]. 高等教育出版社, 2011.1.
2. 薛薇. SPSS统计分析方法及应用.第3版[M]. 电子工业出版社, 2013..
3. 柳彦从, 胥月兵, 陆江银. ZSM-5催化乙醇制低碳烯烃[J]. 化学进展, 2010(4):754-759.