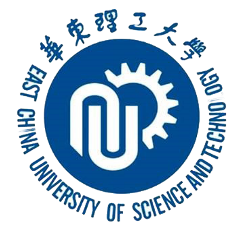
****  ** **

**中国研究生创新实践系列大赛**

**“华为杯”第十七届中国研究生**

**数学建模竞赛**

|  |  |
| --- | --- |
| **学 校** | **长春理工大学** |
| **参赛队号** | **20101860020** |
| **队员姓名** | **1.张耹铭** |
| **2.刘阳** |
| **3.谷晓雁** |

**中国研究生创新实践系列大赛**

**“华为杯”第十七届中国研究生**

**数学建模竞赛**

题 目 （方法+对象+目的）

摘 要：

本文研究了XXX问题。主要创新点在于（创新点1解决了XX问题，取得了XX有益效果），（创新点2解决了XX问题，取得了XX有益效果）以及（创新点3解决了XX问题，取得了XX有益效果）。

问题一：我们（采用了XX方法），主要包括（主要的处理步骤），（获得的有益效果，要有数据支持，关键点加粗）。

问题二：我们（采用了XX方法），主要包括（主要的处理步骤），（获得的有益效果，要有数据支持，关键点加粗）。

问题三：我们（采用了XX方法），主要包括（主要的处理步骤），（获得的有益效果，要有数据支持，关键点加粗）。

最后，对于XXX模型的优缺点进行了评价，本文提出了以下几点（未来的改进措施？）：（1）（2）（3）。

关键字：关键词1；关键词2；关键词3；关键词4；关键词5

1. 问题重述

1.1 问题背景

汽油辛烷值（RON）是车用汽油最重要的品质指标之一，辛烷值越高表示汽油的抗爆性越好，提高辛烷值对汽油动力经济性能有十分重要的意义。辛烷值与汽油中的硫、烯烃等含量有关。为了符合相关的国家标准，必须降低汽油中的硫、烯烃含量，同时也要尽量保持较高的辛烷值。

过去采用数据关联和机理分析的方法对化工过程建模，但由于工艺过程的复杂性、设备的多样性以及操作变量（控制变量）之间具有高度非线性和相互强耦联，而且辛烷值测定具有滞后性，对数据处理有较高的要求。辛烷值和操作变量之间的数学模型分析工作较少，如何控制操作变量、建立损失预测模型对于提高汽油精制的经济效益具有重要意义。

某石化企业的催化裂化汽油精制脱硫装置运行4年，积累了大量的历史数据，其汽油产品的辛烷值损失平均达到了1.37个单位，而同类装置的最小损失值只有0.6个单位。，所以在对辛烷值损失量的建模方面有较大的优化空间。

中国石油化工研究所开发的DSO选择性汽油加氢脱硫技术，不仅具有较高的加氢脱硫活性及脱硫选择性，而且具有很好的原料适应性，可以灵活地控制加氢脱硫反应深度，在达到产品硫含量要求的同时，辛烷值损失较少，能够满足装置长周期运行的要求。DSO技术在催化裂化汽油加氢装置上的应用\_陈晓华.pdf



图1.1 催化裂化汽油精制脱硫装置

1.2 问题重述

题目包含4个附件，附件一提供了某石化企业2017年4月17日至2020年5月26日共325个样本数据以及对应的367个变量，附件二是确定样本数据的方法，附件三提供了285号和313号样本的原始数据，附件四是354个操作变量的信息，要求利用上述数据解决以下问题：

问题一：参考附件一的325个数据样本的预处理结果，根据附件二的样本确定方法对附件三中285号、313号样本原始数据进行整定、筛选，并将数据求取平均值加入到附件一相应的样本号中。

问题二：在建立降低辛烷值损失的模型中，使用先降维后建模的方法，通过降维的方法从367个变量中筛选出具有代表性、独立性的主要变量，并详细说明建模主要变量的筛选过程及合理性。

问题三：采用上述样本和建模的主要变量，通过数据挖掘技术建立辛烷值损失预测模型，并进行模型验证。

问题四：要求在产品硫含量不大于5ug/g的前提下，保证优化过程中原料、待生吸附剂、再生吸附剂的性质保持不变，利用已建立辛烷值损失预测模型获得325个数据样本，求预测的辛烷值不大于附件一中原有辛烷值70%的样本对应主要变量优化后的操作条件。

问题五：针对问题四选取的已优化主要操作变量，结合其每次允许的调整幅度，对133号样本以图形展示其主要操作变量优化调整过程中对应的汽油辛烷值和硫含量的变化轨迹。



2. 模型假设

考虑到实际情况，本文做出如下假设：

1. 模型假设，如果不够可以根据题目条件进行补充。
2. 模型假设，如果不够可以根据题目条件进行补充。
3. 模型假设，如果不够可以根据题目条件进行补充。
4. 模型假设，如果不够可以根据题目条件进行补充。
5. 模型假设，如果不够可以根据题目条件进行补充。
6. 模型假设，如果不够可以根据题目条件进行补充。



3. 符号系统

|  |  |  |
| --- | --- | --- |
| 符号 | 符号说明 | 单位 |
|  | 原始数据变量 | - |
|  | 原始数据变量的公共因子 | - |
|  | 标准误差 | - |
|  | 因子载荷矩阵 | - |
|  | 特殊因子 | - |
|  | 第个原始变量对第个因子的得分 | - |
|  | 回归参数 | - |
|  | 随机扰动 | - |
|  | 残差 | - |
|  | 相对误差 | - |
|  | 辛烷值损失量 | - |
|  | 第个样本第个变量 | - |
|  | 第个变量的最小取值范围 | - |
|  | 第个变量的最大取值范围 | - |
|  | 第个变量的步进次数 | - |
|  | 第个变量的步进值 | - |
|  | 变量的个数 | - |
|  | 主要变量的个数 | - |
|  | 样本的个数 | - |
|  | 产品性质的硫含量 |  |
|  | Spearman相关系数 | - |
|  | 两组数据的等级之差 | - |
|  | 辛烷值损失量回归模型系数 | - |

4. 数据处理

4.1 问题分析

问题一要求按照附件二中的样本确定方法对附件三中285号和313号样本原始数据进行数据预处理，并将获得的数据插入到附件一中。附件一中325个样本数据和附件三的样本原始数据都需要满足附件四中规定的操作变量的取值范围，同时附件一和附件四的数据还存在数据缺失项，采用插值法补全数据缺失项，并根据拉依达准则进行异常值检测。

4.2 数据处理步骤

数据处理内容包括附件一的样本数据和附件三的原始数据，附件三为附件一中两个样本的原始数据，所以我们先对附件三中原始数据进行处理，其处理步骤如下：

1. 最大最小限幅筛选法：根据附件四提供的数据变量操作范围，对附件三的原始数据进行筛选。
2. 插值：在保留的49条原始数据中，有5处数据项存在缺失，使用前后两小时数据平均值进行了插值补全。
3. 均值处理：使用2小时内操作变量的平均值作为辛烷值操作变量数据插入附件一的样本数据中。
4. 拉依达准则（3准则）进行异常值检测。

随后我们对附件一中的样本数据进行处理：

1. 覆盖：将原始数据处理得到的两条样本数据插入到附件一中。
2. 插值：插值补全122项缺失数据，剩下的18个位点仅仅含有部分时间位点数据，无法使用插值补充，直接将此位点删除，保留了343项操作变量。
3. 最大最小限幅筛选法：根据附件四提供的操作变量范围，发现样本号为301、302、303、304的操作变量超出范围，作为无效样本数据删除，最终保留了321条有效样本数据。

4.3 数据处理与结果分析

附件三中名为“非净化风进装置流量”的部分数据在删除前后进行比较，根据附件四可知，非净化风进装置流量的取值范围为0-900，通过最大最小限幅筛选法将原始数据中超出范围的数据删除。图5.1为非净化风进装置流量删除超出范围数据前后的对比，在原始异常数据中，删除了第七次采样超出0-900范围的数据，剩余所有数据都在操作范围内。

表4.1 原始数据删除前后对比

|  |  |
| --- | --- |
| 删除前 | 删除后 |
| 543.3043 | 543.3043 |
| 595.8862 | 595.8862 |
| 648.4681 | 648.4681 |
| 511.6857 | 511.6857 |
| 446.7246 | 446.7246 |
| 463.643 | 463.6430 |
| 910.5219\* |  |
| 784.4914 | 784.4914 |
| 514.8187 | 514.8187 |
| 473.3051 | 473.3051 |
| 463.4757 | 463.4757 |
| 422.0896 | 422.0896 |
| 472.5419 | 472.5419 |
| 866.6923 | 866.6923 |
| 420.2141 | 420.2141 |

注：灰底上标\*为删除值

|  |  |
| --- | --- |
|  |  |
| 图4.1 数据删除前后 | |

（处理了多少数据，有什么有益效果）

根据附件四提供的数据变量操作范围，对附件三的原始数据进行筛选，共计删除了31条样本原始数据，保留了285号样本原始数据40 条，313号样本原始数据9条。在保留的49条原始数据中，有5处数据项存在缺失，使用前后两小时数据平均值进行了插值补全。使用2小时内操作变量的平均值作为辛烷值操作变量数据插入附件一的样本数据中。

将超出范围的异常值删除后，对缺失数据进行插值填补，以2#催化汽油进装置流量的缺失数据为例，选取SPSS数据分析软件对2#催化汽油进装置流量的4项缺失数据填补，填补效果良好。图5.2是插值前后的对比图，2#催化汽油进装置流量数据填补完成后，整体数据呈平稳趋势，保持在40-60之间，无异常值，插值效果好，可以采用。

表4.2 插值前后数据对比

|  |  |
| --- | --- |
| 插值前 | 插值后 |
| 53.86215 | 53.86215 |
| 46.35702 | 46.35702 |
| 44.51688 | 44.51688 |
| 0\* | 52.38405 |
| 0\* | 52.38405 |
| 0\* | 52.38405 |
| 0\* | 52.38405 |
| 48.84124 | 48.84124 |
| 51.11994 | 51.11994 |
| 51.83154 | 51.83154 |
| 46.70276 | 46.70276 |

注：灰底上标\*的为缺失值

|  |  |
| --- | --- |
|  |  |
| 图4.2 插值前后对比 | |

（处理了多少数据，有什么有益效果）

拉伊达准则（3准则）适用于数据量大且呈现正态或近似正态分布的样本，先假设一组检测数据只含有随机误差，对其进行计算处理得到标准偏差，按一定概率确定一个区间，认为超过这个区间的误差，就不属于随机误差而是粗大误差，含有该误差的数据应予以剔除。数学模型如下：

设对被测量变量进行等精度测量，得到，算出其算术平均值即剩余误差，按贝塞尔公式算出标准误差，若某个测量值的剩余误差，满足，则认为是含有粗大误差值的坏值，应予剔除。贝塞尔公式如下：

|  |  |
| --- | --- |
|  | （） |

5. 提取主要变量建立辛烷值损失预测模型

5.1问题分析

建立降低辛烷值损失预测模型共涉及367个变量，包括7个原料性质、2个再生吸附性质、2个产品性质以及354个操作变量。问题要求从367个变量中选取30个以下具有代表性、独立性的的主要变量作为建模的自变量，拟采用因子分析降维、多元线性回归建模的方法预测求解。

因子分析的基本目的就是用少数几个公共因子去描述许多指标或因素之间的联系，将相关比较密切的几个变量归在同一类中，每一类变量就成为一个公共因子，以较少的几个公共因子作为主要变量反映原数据的大部分信息。

多元回归适用于一个自变量多个因变量的数据处理，将选取的主要变量作为模型的自变量，辛烷值的损失量作为因变量，建立预测模型。依据选定主要变量对辛烷值损失量的贡献值的分析，拟采用多元回归的方法进行预测。

5.2模型的建立

5.2.1因子分析模型

因子分析模型用较少的综合性指标来描述和反映原有变量中所包含的指标，形成少数独立的、具有代表性的主要变量对数据进行解释，一方面减弱了变量之间的多重共线性，另一方面在一定程度上方便数据的分析和采集。

本文在建立辛烷值损失预测模型过程中采用的因子分析的数学模型如下

|  |  |
| --- | --- |
|  | （5.1） |

上式中， 是原始指标，是公共因子，即选择的主要指标，是公共因子的系数，称为因子载荷矩阵，成为因子载荷，是第个原始指标在第个因子上的载荷，是的特殊因子。

公共因子反映各原始指标的相关关系，使用公共因子代表原始指标，需要使用原始指标的观测值来计算各个公共因子的得分，其数学模型为：

|  |  |
| --- | --- |
|  | （5.2） |

式中为标准化后的数据，为第个原始指标对第个因子的得分。

5.2.2多元线性回归模型

通过因子分析可以证明所选取的主要变量对辛烷值损失有很大的影响，采用多元线性回归模型进行预测。多元线性回归描述的是一个变量受到多个不同变量的影响的模型，建立多元线性回归模型时，为了保证回归模型具有优良的解释能力和预测效果，应首先注意自变量选择，其准则是：

a) 自变量对因变量必须具有显著的影响，并呈密切的线性相关；

b) 自变量与因变量之间的线性相关必须是真实的，而不是形式上的；

c) 自变量之间应具有一定的互斥性，即自变量之间的相关程度不应高于自变量与因变量之间的相关程度；

d) 自变量应具有完整的统计数据，其预测值容易确定。

多元线性回归模型的一般表达式如下：

|  |  |
| --- | --- |
|  | （5.3） |

式中为被解释变量（因变量），为解释变量（自变量），为解释变量的数目，为回归参数矩阵，为随机扰动。

5.3模型的求解

5.3.1因子分析求解

采用因子分析法的对367个指标归类分析，提取公共因子，再以每个公共因子的方差贡献率作为权重与该公共因子的得分乘数之和构造得分函数。

因子分析法的求解步骤如下：

1. 数据标准化，将数据进行无量纲化处理，消除单位的影响。

数据标准化处理方法如下：

|  |  |
| --- | --- |
|  | （5.4） |

式中为标准化处理后的数据，为原始数据，为原始数据均值，为原始指标数据的标准差。确定待分析的原始指标是否适合进行因子分析，采用KMO检测和巴特利特 （Bartlett）球形检测

1. 因子提取，确定所需公共因子的个数。
2. 估计因子载荷矩阵，并将其旋转，便于公共因子的解释和命名。
3. 计算每一个样本因子得分。

本文采用KMO检验和巴特利特（Bartle

5）球形检验。当KMO>0.5,巴特利特球形检验的相伴概率值小于0.05时，适合因子分析。运用SPSS数据分析软件进行KMO和巴特利特检验。其结果如下：

表5.1 KMO和巴特利特检验

|  |  |  |
| --- | --- | --- |
| KMO取样适切性量数 | 0.774 | |
| 巴特利特球形度检验 | 近似卡方 | 20039.552 |
| 自由度 | 946 |
| 显著性 | 0.000 |

如表5.1所示，KMO值为0.774大于0.5，巴特利特球形检验的相伴概率值近似为0.000，在5%的显著性水平下拒绝原假设，认为原始变量间存在关联性，因此原始指标适合做因子分析。

建立343个原始指标的相关系数矩阵，求取特征值和特征向量，以碎石图的形式表示特征值，统计343个因子对应的特征值、方差百分比与累计百分比，要求选取特征值大于1的因子，且因子的方差百分比越大，表明该因子的比重越大，对公共因子的影响越大。根据特征值及方差百分比的选取原则选取因子1至因子40的部分数据，统计如下：（具体见附件1）

表5.2 解释方差总和

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 因子 | 特征值 | 方差百分比 | 累计百分比 | 因子 | 特征值 | 方差百分比 | 累计百分比 |
| 1 | 109.497 | 31.923 | 31.923 | 21 | 2.344 | 0.684 | 84.363 |
| 2 | 38.524 | 11.232 | 43.155 | 22 | 2.216 | 0.646 | 85.009 |
| 3 | 23.198 | 6.763 | 49.918 | 23 | 1.906 | 0.556 | 85.565 |
| 4 | 19.608 | 5.717 | 55.635 | 24 | 1.845 | 0.538 | 86.103 |
| 5 | 14.133 | 4.121 | 59.756 | 25 | 1.761 | 0.513 | 86.616 |
| 6 | 11.711 | 3.414 | 63.17 | 26 | 1.628 | 0.475 | 87.091 |
| 7 | 9.658 | 2.816 | 65.986 | 27 | 1.516 | 0.442 | 87.533 |
| 8 | 8.26 | 2.408 | 68.394 | 28 | 1.443 | 0.421 | 87.953 |
| 9 | 7.245 | 2.112 | 70.506 | 29 | 1.365 | 0.398 | 88.351 |
| 10 | 6.545 | 1.908 | 72.414 | 30 | 1.309 | 0.382 | 88.733 |
| 11 | 6.027 | 1.757 | 74.171 | 31 | 1.301 | 0.379 | 89.112 |
| 12 | 5.049 | 1.472 | 75.643 | 32 | 1.241 | 0.362 | 89.474 |
| 13 | 4.82 | 1.405 | 77.049 | 33 | 1.227 | 0.358 | 89.832 |
| 14 | 4.037 | 1.177 | 78.226 | 34 | 1.128 | 0.329 | 90.161 |
| 15 | 3.971 | 1.158 | 79.383 | 35 | 1.096 | 0.32 | 90.48 |
| 16 | 3.502 | 1.021 | 80.405 | 36 | 1.063 | 0.31 | 90.79 |
| 17 | 3.103 | 0.905 | 81.309 | 37 | 1.041 | 0.303 | 91.094 |
| 18 | 2.867 | 0.836 | 82.145 | 38 | 0.98 | 0.286 | 91.379 |
| 19 | 2.756 | 0.804 | 82.949 | 39 | 0.952 | 0.278 | 91.657 |
| 20 | 2.506 | 0.731 | 83.68 | 40 | 0.912 | 0.266 | 91.923 |

表5.2显示，共有37个因子的特征值大于1，基于过程中内定取特征值大于1的原则，初步筛选1至37个因子，方差积累量为91.094%，表明前37个因子的信息可以代表原始数据。鉴于问题二要求选取主要变量的个数小于30，通过对累计百分比的分析，发现第22个因子的方差累计量达到85.009%，因此还需要对第22个因子到第27个因子的特征值以及方差累计量的变化趋势进行分析，确定最终选取的因子数量。下面给出了碎石图。

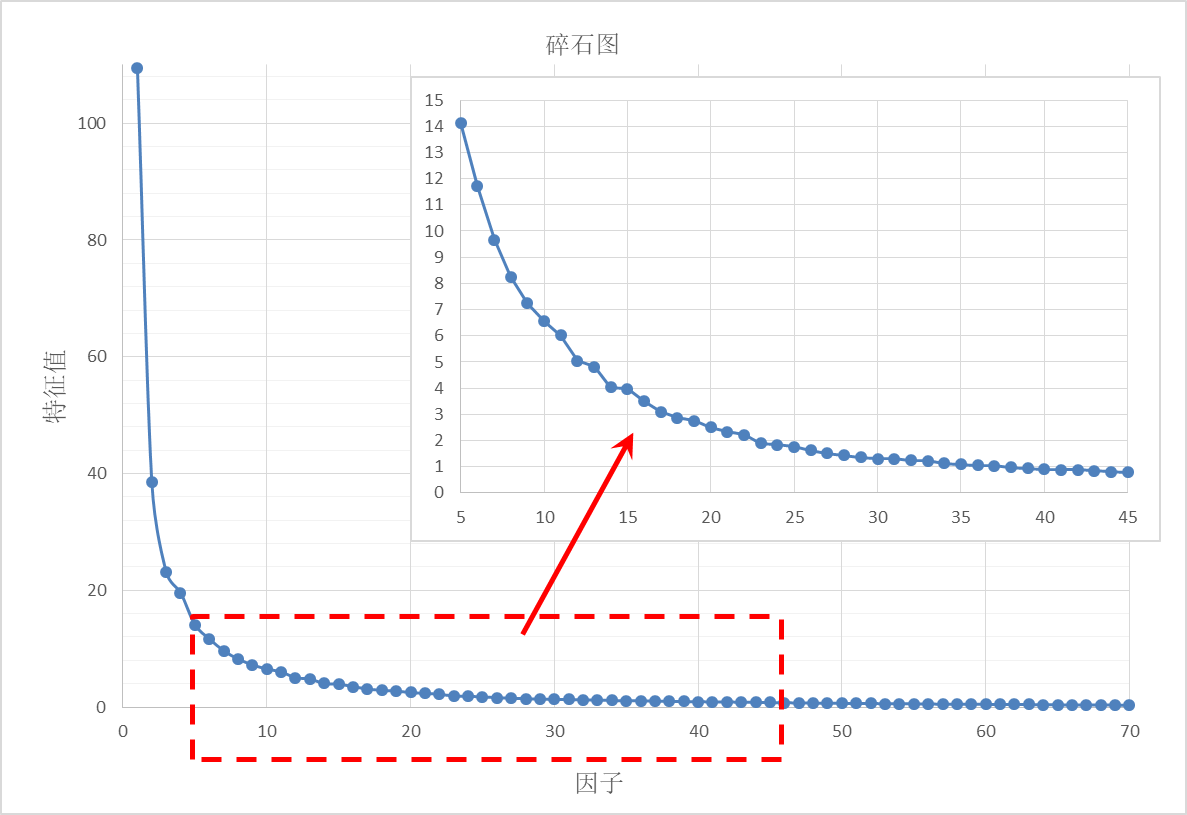


图5.1 因子碎石图

由图5.1的因子碎石图分析可知，因子1至因子20特征值变化剧烈，特征值由109.497迅速变化至2.506，因子20至因子40特征值变化趋于平缓，由2.506缓慢降至0.912，对因子5到因子40碎石图进行放大分析，观察特征值的变化趋势，结合表5.1的特征值及方差的百分比，选择1至26号因子作为主要变量，将348个原始指标降维保留至26个主要变量，即共产生了26个公共因子。

选取了合适的公共因子的个数之后，需要根据变量之间的相关系数矩阵估计因子载荷矩阵。采用主成分分析法对因子载荷矩阵进行估计。并将得到的因子载荷矩阵旋转，便于了解每个公共因子的意义，采用最大方差法将因子载荷矩阵旋转，针对旋转后的因子载荷矩阵进行分析。表5.3是部分旋转因子载荷矩阵：（具体见附件1）

表5.3 旋转因子载荷矩阵

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 因 子 | | | | | | | |
| 1 | 2 | 3 | 4 | … | 24 | 25 | 26 |
| K-103A排气压力 | 0.886 | 0.288 | 0.212 | 0.146 | … | -0.02 | 0.004 | -0.004 |
| K-103A排气温度 | -0.885 | -0.295 | -0.206 | -0.146 | … | 0.02 | -0.004 | 0.004 |
| D-123压力 | -0.48 | -0.122 | -0.33 | -0.206 | … | 0.144 | -0.096 | -0.018 |
| E-206壳程出口管温度 | -0.566 | -0.095 | -0.004 | 0.022 | … | 0.221 | 0.211 | 0.083 |
| 反吹氢气压力 | 0.984 | 0.028 | 0.029 | 0.047 | … | -0.014 | -0.004 | -0.009 |
| ME-103进出口差压 | 0.882 | 0.116 | -0.161 | -0.142 | … | -0.009 | 0.02 | -0.019 |
| 过滤器ME-101出口温度 | -0.23 | -0.345 | -0.49 | 0.014 | … | -0.12 | -0.019 | 0.023 |
| … | … | … | … | … | … | … | … | … |
| D-102温度 | 0.191 | 0.331 | 0.428 | -0.154 | … | -0.071 | 0.015 | -0.042 |
| 原料进装置流量累计 | 0.936 | -0.288 | -0.119 | -0.014 | … | 0.011 | 0.011 | 0.019 |
| 燃料气进装置压力 | -0.162 | -0.125 | -0.117 | -0.05 | … | 0.037 | -0.04 | 0.008 |
| 反吹氢气温度 | 0.366 | -0.168 | 0.183 | 0.051 | … | 0.013 | -0.025 | 0.039 |
| 再生器下部温度 | -0.135 | 0.011 | 0.051 | 0 | … | -0.004 | 0.01 | 0.012 |
| 再生风流量 | 0.481 | 0.389 | -0.093 | 0.078 | … | 0.08 | 0.032 | 0.022 |

因子分析模型共包含343个原始指标，26个公共因子，根据因子载荷矩阵写出下列数学方程：

|  |  |
| --- | --- |
|  | （） |

分析每个公共因子对应的旋转后的因子系数，除去只包含一个原始指标的公共因子，公共因子1包含的原始指标个数最多，其包含原始指标的因子系数在0.992至0.138变化，有很明显地两极分化，也存在因子系数变化不大的公共因子，比如公共因子7，其所包含的四个原始指标的因子系数最大为0.838，最小为0.447。

前9个公共因子对应的原始指标较多，其中公共因子1包含96个原始指标，占比最大，公共因子24、公共因子26各对应三个原始指标，公共因子18、公共因子23和公共因子25都只对应1个原始指标。下表是26个公共因子对应的具体原始指标及主要因素，主要因素指向每个公共因子的意义，解释公共因子对应的主要的原始指标。

表5.4 公共因子对应主要变量

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 因子 | 变量 | 主要因素 | 因子 | 变量 | 主要因素 |
| 1 | 芳烃,v% | 2与预加氢反应器有关 | 14 | D-125液位 | 与9反应温度压力有关 |
| 非净化风进装置流量 | EH101出口 |
| ... | ... |
| E-105管程出口管 | K-101B排气压力 |
| 反应器质量空速 | K-101B进气压力 |
| 2 | 闭锁料斗顶压力 | 3与预加氢反应器有关 | 15 | 再生风流量 | 与12反应有关 |
| 闭锁料斗充氢线压力控制 | 再生器顶烟气温度 |
| ... | … |
| K-101A左排气温度 | 再生器顶部/再生器接收器差压 |
| F-101循环氢出口管温度 | R-102床层吸附剂料位密度 |
| ... | ... | ... | ... | ... | ... |
| 12 | 再生器接收器顶部/再生器接收器底部差压 | 3/6与加氢脱硫反应器有关 | 25 | 密度(20℃),kg/m³  D-122入口管温度  D-103底部液位 | 与物料23/24消耗有关 |
| 再生器接收器/LH差压 |
| ... |
| 再生进料罐/LH差压 |
| 13 | 辛烷值RON | 与8/7原料有关 | 26 | 循环氢至闭锁料斗料腿流量闭锁料斗H2过滤器出口气流量  P-101A入口过滤器差压 | 与再生26器有关 |
| 催化汽油进装置总流量 |
| ... |
| 原料汽油硫含量 |

由于选取的公共因子和原始指标之间具有相关关系，可以用公共因子代表原始指标反映大部分的信息，可以用标准化后的原始指标表征26个公共因子的得分，用来评价公共因子选取的合理性。下表选取了部分原始指标以及对应的1至26号公共因子的因子得分系数矩阵：（具体见附件1）

表5.5 因子得分系数矩阵

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 因 子 | | | | | | | |
| 1 | 2 | 3 | 4 | … | 24 | 25 | 26 |
| 硫含量,μg/g | -0.001 | -0.014 | 0.008 | -0.003 | … | -0.005 | 0.028 | -0.013 |
| 辛烷值RON | 0.001 | -0.011 | -0.015 | -0.005 | … | -0.024 | -0.013 | -0.07 |
| 饱和烃,v%（烷烃+环烷烃） | -0.005 | 0.004 | 0 | 0.001 | … | -0.012 | 0.014 | 0.037 |
| 烯烃,v% | 0.008 | -0.012 | -0.002 | 0.002 | … | 0.027 | 0.008 | -0.014 |
| 芳烃,v% | -0.007 | 0.021 | 0.006 | -0.008 | … | -0.043 | -0.056 | -0.055 |
| 溴值,gBr/100g | 0.003 | -0.012 | -0.003 | -0.012 | … | -0.013 | -0.044 | 0.057 |
| 密度(20℃),kg/m³ | -0.003 | 0.017 | 0.008 | -0.001 | … | -0.023 | -0.061 | -0.138 |
| … | … | … | … | … | … | … | … | … |
| 汽油产品去气分累积流量 | 0.014 | -0.005 | 0 | 0.001 | … | -0.005 | 0.001 | -0.007 |
| 8.0MPa氢气至循环氢压缩机入口 | 0 | -0.011 | -0.006 | 0.097 | … | -0.026 | 0.029 | -0.005 |
| 8.0MPa氢气至循环氢压缩机入口 | 0.001 | -0.028 | 0 | 0.006 | … | 0.018 | 0.007 | 0.003 |
| 8.0MPa氢气至反吹氢压缩机出口 | 0 | 0.014 | 0.009 | -0.098 | … | 0.011 | -0.022 | 0.013 |
| 8.0MPa氢气至反吹氢压缩机出口 | 0 | -0.028 | 0 | 0.005 | … | 0.009 | 0.003 | -0.008 |
| D101原料缓冲罐压力 | -0.011 | 0.001 | -0.001 | -0.003 | … | 0.007 | 0.002 | -0.013 |

将表5.5的系数带入（5.3）中可以得到对应的26个公共因子的函数，同时根据各个公共因子的方差百分比计算公共因子的综合得分：

|  |  |
| --- | --- |
|  | （5.5） |

多元回归模型求解及分析

对辛烷值损失预测模型采用多元线性分析模型进行预测，预测目标是辛烷值的损失值，因子分析确定的26个公共因子作为辛烷值损失值的影响因素。通过对325条样本数据进行多元线性回归，得到的部分预测数据如下：（具体见附件2）

表5.6 多元线性回归

|  |  |  |  |
| --- | --- | --- | --- |
| 真实值 | 预测值 | 残差 | 相对误差 |
| 1.38 | 1.338296 | -0.04170 | 0.03022 |
| 1.18 | 1.271150 | 0.09115 | 0.077246 |
| 1.38 | 1.275337 | -0.10466 | 0.075842 |
| 1.38 | 1.362620 | -0.01738 | 0.012594 |
| 1.28 | 1.391571 | 0.111571 | 0.087165 |
| 1.41 | 1.351081 | -0.05892 | 0.041787 |
| 1.20 | 1.297800 | 0.09780 | 0.0815 |
| 1.30 | 1.291075 | -0.00893 | 0.006866 |
| 1.10 | 1.312487 | 0.212487 | 0.19317 |
| 1.40 | 1.340815 | -0.05919 | 0.042275 |
| … | … | … | … |
| 1.15 | 1.241704 | -0.03830 | 0.029919 |
| 1.15 | 1.131579 | -0.10842 | 0.087436 |
| 1.15 | 1.197304 | 0.047304 | 0.041134 |
| 1.05 | 1.198967 | -0.20103 | 0.143595 |
| 1.25 | 1.174686 | 0.024686 | 0.021466 |
| 1.15 | 1.166273 | 0.016273 | 0.014151 |
| 1.25 | 1.181094 | 0.031094 | 0.027038 |
| 1.15 | 1.216163 | 0.166163 | 0.15825 |
| 1.25 | 1.166173 | -0.08383 | 0.067061 |
| 1.35 | 1.240346 | 0.090346 | 0.078562 |
| 1.28 | 1.095973 | -0.154030 | 0.123221 |
| 1.25 | 1.281526 | 0.131526 | 0.114371 |



图5.2 残差图

残差是真实值与预测值的差，残差图是以自变量为横坐标，残差为纵坐标做出的散点图。通过残差所提供的信息，分析数据的可靠性、周期性或其他干扰，可用于分析关于误差项假定的合理性以及线性回归关系的假定的可能性。残差平方和反应除去自变量和因变量的线性关系外，所有其他因素影响的总和，残差平方和越小，说明预测越准确。通过对残差的分析，共有321个残差，其中大于0.6的有3个，大于0.5的有7个，大于1的只有一个，有97.8462%的残差落在[-0.5 0.5]区间内，且此区间范围内所有散点随机分布，没有固定的趋势，不存在异方差的情况，说明模型选择存在合理性。

相对误差定义为绝对误差与约定真值的比值，即

|  |  |
| --- | --- |
|  | （5.6） |

对表5.6的相对误差进行统计分析，发现相对误差大于0.3的有19个，相对误差在0.3以内的占比较重，为94.1538%，可以认为多元线性回归模型具有一定合理性。

（相关指数要不要说）

相关指数用来衡量回归模型的拟合效果，越大，残差平方和越小，回归模型的拟合效果越好，相关指数的公式如下：

|  |  |
| --- | --- |
|  | （5.7） |

式中是第个数据，是回归值，是数据均值。

得到的回归参数矩阵如下：

表5.7 回归参数矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| 回归系数 | 第项 | 回归系数 | 第项 |
| 1.264555 | 常数项 | 0.000806 | 14 |
| -0.05012 | 1 | -0.01062 | 15 |
| -0.0491 | 2 | 0.03664 | 16 |
| -0.02142 | 3 | -0.0342 | 17 |
| 0.001018 | 4 | -0.03151 | 18 |
| 0.001084 | 5 | 0.008514 | 19 |
| -0.00346 | 6 | -0.00823 | 20 |
| -0.00468 | 7 | 0.00465 | 21 |
| -0.01199 | 8 | -0.00357 | 22 |
| 0.025957 | 9 | 0.009754 | 23 |
| 0.020495 | 10 | -0.0015 | 24 |
| 0.019796 | 11 | 0.000414 | 25 |
| -0.0114 | 12 | 0.007623 | 26 |
| -0.04497 | 13 |  |  |

本模型的自变量是公共因子，因变量是辛烷值的损失值，共26个自变量，325个因变量，根据回归参数矩阵，写出下面的多元参数方程：

|  |  |
| --- | --- |
|  | （5.8） |

6. （有具体写“问题三：具体问题”，否则“问题三的建模与求解”）

6.1主要变量操作方案优化的分析

从辛烷值损失预测模型可知，该优化问题的目标找到辛烷值损失降幅大于30%样本对应主要变量优化后的操作条件，需要先求取辛烷值损失的最小值，然后找到对应样本优化后的操作条件，分析如何操作才能使辛烷值损失降幅达到目的。该优化问题共有变量343个，样本数据325个。在优化过程中原料、待生吸附剂、再生吸附剂的性质保持不变：根据附件1“325个样本数据”可知，1至7号、9至12号的变量数据保持不变。

优化的约束条件为：

1. 要求保证产品硫含量不大于5ug/g，产品硫含量与原料、待生吸附剂、再生吸附剂以及331个操作变量可能存在一定的相关关系，为了满足产品中硫含量不大于5ug/g,以产品中硫含量作为自变量，343个变量作为因变量建立逐步回归进行预测。
2. 331个操作变量数据需在各自的取值范围内优化，且优化的过程只允许以每次步进值的方式对操作变量进行调整。

6.2模型建立

6.2.1目标函数

该问题分为两步，首先是求取辛烷值损失的最小值，将因子得分系数矩阵带进多元回归模型中，得到如下模型：

|  |  |
| --- | --- |
|  | （6.1） |

其中，表示辛烷值损失量回归模型系数，表示辛烷值的损失量，表示变量的个数，表示多元回归的公共因子的个数， 为第个变量。在343个变量中，存在11个数据保持不变的变量，分别是7个原料性质、2个待生吸附性质、2个再生吸附性质，且产品硫含量与其余变量之间也可能存在一定的相关关系，所以目标函数等价于

|  |  |
| --- | --- |
|  | （6.2） |

式中，表示产品硫含量，表示331个操作变量。产品硫含量应满足约束条件，本文采用逐步回归来建立产品硫含量与其他变量之间的关系。

逐步回归的基本思想是逐个引入自变量，每次引入对因变量最显著的自变量，并对方程中的老变量逐个进行检验，把不显著的变量逐个从方程中剔除，最终的回归方程中包含对因变量具有显著影响的变量，又剔除了对因变量影响不显著的变量。本文中逐步回归的因变量是产品硫含量，逐步回归的基本步骤如下：

1. 求取全部自变量的偏回归平方大小，从大到小依次引入回归方程；
2. 对回归方程所含全部变量进行检验，剔除不显著因素，直到回归方程中所含的所有变量对因变量的影响都显著时，才考虑引入新的变量；
3. 在剩余未选因素中，选出对因变量作用最大者，检验其显著性，若存在显著性，则引入回归方程，否则不引入；
4. 最终没有显著性因素可以引入，也没有不显著因素需要剔除，得到回归方程。

求解得到回归方程如下：

|  |  |
| --- | --- |
|  | （6.3） |

式中，代表预测的产品硫含量，代表回归方程的常数项，代表回归方程的系数，为了保持与目标函数变量的序号一致，上述回归方程的自变量集采用了不连续集合表示，，即共有15个变量与产品硫含量有关，其中包含序号为的变量是不变的原料性质，可划分为常数项，将回归方程的系数带入方程：

|  |  |
| --- | --- |
|  | （6.4） |

将产品硫含量的回归方程带入到目标函数中，去掉常数项后，目标函数可等价为：

|  |  |
| --- | --- |
|  | （6.5） |

式中，集合为集合的子集，是集合去掉常数项后的集合。

6.2.2约束条件

题目给出了2个约束条件，分别是对产品硫含量和操作变量的约束。

目标函数的优化要在产品硫含量不大于5ug/g的前提下，产品硫含量会受到其他变量的影响，根据逐步回归可知，产品硫含量受到2个原料变量、13个操作变量影响。

|  |  |
| --- | --- |
|  | （6.6） |

根据附件4“354个操作变量信息”可知，331操作变量只能在各自的取值范围内变化，且为了防止避免实际操作过程中大幅度调整所带来的波动，每个操作变量都有固定的值，即每个操作变量每次允许的调整幅度，优化的过程只允许以每次步进值的方式对操作变量进行调整。

|  |  |
| --- | --- |
|  | （6.7） |

|  |  |
| --- | --- |
|  | （6.8） |

表示产品硫含量，表示第个样本、第个变量，表示第个变量的最小取值范围，表示第个变量的最大取值范围，表示第个变量的步进次数，表示第个变量的步进值，表示样本的个数。

6.2.3得出模型

经过整理，得出优化模型：

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

分析该模型可知，共有331个连续型变量，对应331个整数变量，所以该模型为整数线性模型，约束条件有两种，共有633个不等式约束，331个等式约束。

6.2.4模型简化

为进一步简化模型，可将等式约束代入目标函数：

|  |  |
| --- | --- |
|  | （6.9） |

常数项不影响优化，将上式最小化，可等价为：

|  |  |
| --- | --- |
|  | （6.10） |

代入约束条件可得：

|  |  |
| --- | --- |
|  | （6.11） |

综上所述，模型可等价为：





本模型为整数线性模型，简化后有331个整数变量，662个不等式约束。

8. 模型评价与改进

8.1 模型的优点

1. 这里写模型的优点一。

2. 这里写模型的优点二。

3. 这里写模型的优点三。

8.2 模型的缺点

1. 这里写模型的缺点一。

2. 这里写模型的缺点二。

8.3 模型的改进与推广

本文提出的模型和方法可推广应用于XXXXXXXXX：

1. 本文模型在推广过程的普遍适用性（属优点）。
2. 本文模型在对同一领域的其他研究的参考意义和价值（属优点）。
3. 本文模型在推广过程中可能遇到的困难和不足之处（属缺点）。

参考文献

[1]左苏. 基于主成分回归模型的工程项目成本预测[D].扬州大学,2014.

附录

程序1：XXXXX程序

程序2：XXXXX程序

程序3：XXXXX程序

程序4：XXXXX程序