

# 辛烷值损失预测模型的构建与优化

姜妍希, 张炳江

北京信息科技大学, 北京

收稿日期: 2021年9月21日; 录用日期: 2021年10月14日; 发布日期: 2021年10月22日

## 摘要

汽油是小型车辆的主要燃料, 汽油燃烧产生的尾气排放对大气环境有重要影响。针对我国汽油在精制过程中辛烷值损失的问题, 首先使用Lasso回归对工业样本数据进行预处理, 再运用多元线性回归建立辛烷值损失预测模型, 最后基于数据包络分析法(DEA)对汽油精制过程中的主要操作变量进行优化。该优化模型具有实际应用意义, 为汽油清洁化过程降低辛烷值损失的研究提供了理论依据。

## 关键词

辛烷值损失, Lasso回归, 多元线性回归, 数据包络分析

# Construction and Optimization of Octane Loss Prediction Model

Yanxi Jiang, Bingjiang Zhang

Beijing Information Science & Technology University, Beijing

Received: Sep. 21<sup>st</sup>, 2021; accepted: Oct. 14<sup>th</sup>, 2021; published: Oct. 22<sup>nd</sup>, 2021

## Abstract

Gasoline is the major fuel for small-sized vehicles, and the gasoline-fueled vehicle emissions generated by gasoline combustion exert an important impact on the atmosphere. Regarding the loss of octane value amid the process of gasoline refining in our country, t Lasso regression is used to process the industrial specimen data beforehand, and then a prediction model for octane loss is established through multiple linear regression. Finally, the major operating variables used in gasoline refining are optimized based on Data Envelopment Analysis (DEA) model. The optimized model is of practical significance for application and proffers theoretical ground for the reduction of the loss of octane value during the refining of gasoline.

## Keywords

Octane Loss, Lasso Regression, Multiple Linear Regression, Data Envelopment Analysis

Copyright © 2021 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

目前, 小型车辆的主要燃料是汽油, 汽油燃烧排放的产物一氧化碳、氮氧化合物、碳氢化合物和微粒等对大气环境造成了重大影响。为此, 世界主要大国均制定了日益严格的汽油质量标准。汽油清洁化的重点是降低汽油中的硫和烯烃含量, 同时尽量保持其辛烷值。辛烷值(ROK)是衡量汽油燃烧性能的最重要指标, 目前除考虑改进发动机的构造和操作条件外, 主要依靠提高汽油辛烷值, 借以提高发动机压缩比的办法来增加燃烧功率和节约汽油。然而, 现有技术在对催化裂化汽油进行脱硫和降烯烃过程中, 普遍降低了汽油辛烷值[1]。辛烷值每降低 1 个单位, 造成经济损失约 150 元/吨。以一个年产 100 万吨的催化裂化汽油精制装置为例, 若能降低 0.3 个单位 RON 损失, 其经济效益将达到 4500 万元。

为满足日益严格的清洁汽油标准而不断降低硫和烯烃含量的需求, 科研人员们在汽油清洁化领域开展了大量的研究工作。如冯连坤指出汽油辛烷值损失的影响因素为预加氢反应器温度、加氢脱硫反应器入口温度和轻汽油抽出量[2]; 刘畅采用随机森林算法去寻找降低汽油精制过程中的辛烷值损失模型中的主要变量[3], 而本文使用 Lasso 回归方法寻找主要变量。对于有不同取值的属性的数据, 取值划分较多的属性会对随机森林产生更大的影响, 所以随机森林在这种数据上产出的属性权值是不可信的。Lasso 算法是一种带有惩罚因子的线性模型估计方法, 该方法的本质是约束各个回归系数的绝对值之和小于某个常数的条件下, 最小化回归方程的残差平方和, 同时阈值的设定又可以收缩每个估计的参数值[4] [5] [6]。Lasso 方法可以有效地估计回归模型中的各个参数, 同时也可以较好地解决变量间的多重线性问题。

本文首先通过 Lasso 回归找出精制过程中的主要操作变量, 再采用多元线性回归方法建立辛烷值损失预测模型。多元线性回归不需要很复杂的计算, 在数据量大的情况下依然运行速度很快, 并且可以根据系数给出每个变量的理解。本文最后一部分运用了数据包络分析方法(DEA)对主要操作变量进行调整来达到降低辛烷值损失的目的。DEA 是一个对于多投入多产出的多个决策单元的效率评价方法, 在汽油精制过程中, 将多个主要操作变量作为输入, 辛烷值的损失作为输出, 利用输出型 CCR 模型来计算并进行调整。

## 2. 变量降维

### 2.1. 数据预处理

本文具体数据来源于某石化企业的催化裂化汽油精制脱硫装置运行 4 年所积累的历史数据。由于工厂得到的原始数据存在一定数据缺失和数据失真的情况, 需要排除对数据中的坏值或者短缺值, 对失真的数据进行修正。

数据预处理的具体步骤如下:

Step 1: 遍历每个样本, 查看样本是否大多变量为空值, 若是则剔除该样本; 查看变量是否全列为空值, 若是则删除该列, 若部分是, 用平均值补充空值。

Step 2: 根据各变量取值范围来剔除不符样本。

Step 3: 根据  $3\sigma$  准则删除不符样本。

Step 4: 对于 Matlab, `xlswrite(filename, data)` 为保存到 excel 的命令。

$3\sigma$  准则: 首先设对被测量变量进行等精度测量, 得到  $x_1, x_2, \dots, x_n$ , 算出其算术平均值  $\bar{x}$  及剩余误差  $v_i = x_i - \bar{x} (i=1, 2, \dots, n)$ , 并按贝塞尔公式算出标准误差  $\sigma$ 。若某个测量值  $x_b$  的剩余误差  $v_b (1 \leq b \leq n)$  满足  $|v_b| = |x_b - \bar{x}| > 3\sigma$ , 则认为  $x_b$  是含有粗大误差值的坏值, 应予剔除。

## 2.2. 数据降维

在解决实际问题时, 自变量过多对模型的构建弊大于利, 因此, 一般通过数据降维来简化模型。本文中原有 367 个操作变量, 因为变量数量过多且大于样本数据数量, 主成分分析法等常规方法不能很好地实行降维, 因此本文利用构造惩罚函数的 Lasso 回归对多个变量进行筛选, 选出具有代表性且独立性的 19 个建模主要变量。降维的好处有: 1) 减少数据存储所需的内存, 节约成本。2) 减少数据处理与建模的时间, 提高效率。3) 提高该算法的性能, 因为会有一些算法在此 300 维的数据上表现不佳。4) 能更直观地看出降维的结果, 有助于数据可视化。

### 2.2.1. Lasso 回归

Lasso (Least absolute shrinkage and selection operator) 方法是以缩小变量集(降阶)为思想的压缩估计方法, 在拟合广义线性模型的同时进行变量筛选和复杂调整。通过构造一个惩罚函数, 可以将变量的系数进行压缩并使某些回归系数变为 0, 进而达到变量选择的目的。普通线性模型

$$Y = X\beta + \varepsilon \quad (1)$$

其中, 响应变量  $Y = (y_1, y_2, \dots, y_n)^T$ , 协变量  $X = (X^{(1)}, X^{(2)}, \dots, X^{(d)})$ , 对于每一个  $X^{(j)}$ , 有  $X^{(j)} = (x_1^j, x_2^j, \dots, x_n^j)^T$ 。假设每个  $x_i^j (i=1, 2, \dots, n; j=1, 2, \dots, d)$  均中心化和标准化, 则随机误差项  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $(i=1, 2, \dots, n)$ ,  $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$ , 回归系数  $\beta = (\beta_1, \beta_2, \dots, \beta_d)^T$ 。

当  $X$  为列满秩设计矩阵时, 回归系数  $\beta$  可由普通最小二乘估计方法求得

$$\hat{\beta}_{OLS} = \arg \min_{\beta \in R^d} \|Y - X\beta\|^2 = (X^T X)^{-1} X^T Y$$

惩罚方法: 当设计矩阵  $X$  不满足列满秩时, 将不能采用普通最小二乘法来求解回归系数, 此时应引入惩罚方法。该方法可以同时实现变量选择和参数估计, 在参数估计时, 通过将部分参数压缩为 0 来达到变量选择的目的。惩罚方法是取惩罚似然函数最小时的值作为回归系数的估计值, 即

$$\hat{\beta} = \arg \min_{\beta \in R^d} \left( \|Y - X\beta\|^2 + P_\lambda(|\beta|) \right),$$

其中惩罚项

$$P_\lambda(|\beta|) = \lambda \sum_{j=1}^d |\beta_j|^m, m \geq 0,$$

$\lambda$  为调节参数(也可以为向量)。当  $m=2$  时, 得  $L_2$  惩罚项(Ridge 惩罚)。

Lasso 方法则是在普通线性模型中增加  $L_1$  惩罚项。具体算法如下:

1) 令  $\sigma_i = \text{sign}(\hat{\beta}^0)$ , 这里  $\hat{\beta}^0$  为通常的最小二乘估计。

2) 计算  $\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left( y_i - \sum_j \beta_j x_{ij} \right)^2 \right\}$ , s.t.  $G\beta \leq t$ , 这里  $G = \sigma_i^T$ 。

3) 验证是否满足  $\sum_j |\hat{\beta}_j| \leq t$ , 若满足则停止,  $\hat{\beta}$  即为所求, 否则继续。

4) 令  $\sigma_j = \text{sign}(\hat{\beta}_j)$  加入到  $G$  中, 即令  $G = \begin{pmatrix} \sigma_i^T \\ \sigma_j^T \end{pmatrix}$ , 然后返回(2)。

LASSO 回归复杂度调整的程度由参数  $\lambda$  来控制,  $\lambda$  越大对变量较多的线性模型的惩罚力度就越大, 从而最终获得一个变量较少的模型。LASSO 回归与 Ridge 回归同属于一个被称为 Elastic Net 的广义线性模型家族。这一家族的模型除了相同作用的参数  $\lambda$  之外, 还有另一个参数  $\alpha$  来控制应对高相关性数据时模型的性状。LASSO 回归  $\alpha = 1$ , Ridge 回归  $\alpha = 0$ , 一般 Elastic Net 模型  $0 < \alpha < 1$ 。

### 2.2.2. 主要变量筛选结果

首先, 从与辛烷损失值相关的 367 个变量中筛选出主要变量。设辛烷损失值为因变量, 原料经过一系列工艺流程得到最终产品, 此处研究的是产品的性质, 不宜作为自变量, 所以选取作为因变量使用。辛烷值损失可近似于原料辛烷值与产品辛烷值的差值, 所以若把原料性质和产品性质一起进行筛选, 那原料辛烷值和产品辛烷值对辛烷值损失值的影响是巨大的, 会让操作变量所造成的影响几乎不显现, 这与我们所求的实际目标不符。因此我们去除 2 个产品的性质变量, 只对 365 个变量使用 Lasso 回归, 根据其  $\lambda$  的变化利用 Matlab 编程实现。由于变量太多, 从 Matlab 运行结果的参数系数图中不能看出具体的系数情况。使用交叉验证选出最佳的  $\lambda$ , 其筛选的主要变量如下, 见表 1。

Table 1. Main variables

表 1. 主要变量

| 符号       | 说明                                    |
|----------|---------------------------------------|
| $x_1$    | 原料性质辛烷值                               |
| $x_2$    | S-ZORB.TE_5202.PV 精制汽油出装置温度           |
| $x_3$    | S-ZORB.FT_9001.PV 燃料气进装置流量            |
| $x_4$    | S-ZORB.FT_9302.PV 0.3MPa 凝结水出装置流量     |
| $x_5$    | S-ZORB.FT_9401.PV 净化风进装置流量            |
| $x_6$    | S-ZORB.FT_9002.DACA D203 出口燃料气流量      |
| $x_7$    | S-ZORB.SIS_TE_2802 D-102 温度           |
| $x_8$    | S-ZORB.TE_5006.DACA 稳定塔底出口温度          |
| $x_9$    | S-ZORB.TE_5201.DACA A-202A/B 出口总管温度   |
| $x_{10}$ | S-ZORB.TC_2201.OP EH101 出口            |
| $x_{11}$ | S-ZORB.AT_1001.DACA 原料汽油硫含量           |
| $x_{12}$ | S-ZORB.FC_5103.DACA 稳定塔顶回流流量          |
| $x_{13}$ | S-ZORB.FT_2803.DACA 紧急氢气去 D-102 流量    |
| $x_{14}$ | S-ZORB.PDT_1003.DACA P-101B 入口过滤器差压   |
| $x_{15}$ | S-ZORB.PT_7502.DACA K-102A 进气温度       |
| $x_{16}$ | S-ZORB.TE_7106B.DACA K-101B 左排气温度     |
| $x_{17}$ | S-ZORB.PT_6008.DACA F-101 辐射室出口压力     |
| $x_{18}$ | S-ZORB.AT-0001.DACA.PV S_ZORB AT-0001 |
| $x_{19}$ | S-ZORB.TE_1102.DACA.PV E-101 壳程入口总管温度 |

### 3. 辛烷值(RON)损失预测

在解决实际问题时, 实验结果往往受多个因素的影响, 这时可以通过多元回归分析求出试验指标(因变量)  $y$  与多个试验因素(自变量)  $x_j (j=1, 2, \dots, m-1)$  之间的近似函数关系  $y = f(x_1, x_2, \dots, x_{m-1})$ , 亦称多重回归或复回归。当影响因素与因变量之间呈线性关系时, 所进行的回归分析即为多元线性回归。多元线性回归是多元统计分析中的一种重要方法, 被广泛应用于社会、经济、技术以及众多自然科学领域的研究中[7] [8] [9] [10] [11]。其最主要的思想是多个影响因素与因变量之间是线性关系, 设  $y$  是可观测的随机变量, 它受  $m-1$  个非随机因素  $x_1, x_2, \dots, x_k$  和随机因素  $\varepsilon$  的影响。若  $y$  与  $x_1, x_2, \dots, x_{m-1}$  有如下线性关系

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{m-1} x_{m-1} + \varepsilon$$

其中  $\beta_0, \beta_1, \beta_2, \dots, \beta_{m-1}$  是未知参数;  $\varepsilon$  是均值为 0、方差为  $\sigma^2 > 0$  的不可观测的随机变量, 称为误差项, 并通常假定  $\varepsilon \sim N(0, \sigma^2)$ , 称

$$\hat{y} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_{m-1} x_{i, m-1}, i = 1, 2, \dots, n$$

为多元线性回归模型, 且称  $y$  为因变量,  $x_1, x_2, \dots, x_{m-1}$  为自变量。

本文明确定义研究的自变量与因变量, 将辛烷值的损失值设为因变量  $Y$ , 自变量为问题二中筛选出的 19 个变量, 用 SPSS 软件计算这些变量间的关联度, 计算结果中变量的 VIF 均小于 5。

VIF 是方差膨胀因子, 是容差的倒数, 说明这些变量之间不存在多重共线性, 关联度很低。因此, 先选择多元线性回归分析方法进行建模。通过 MATLAB 计算, 得:

```

beta =
    -8.8987606
     0.025846888
     0.017969824
    -0.00032382168
    -0.0000077132427
    -0.00028025878
    -0.00044747347
    -0.0018174917
     0.01428002
    -0.000011339097
    -0.0024736435
    -0.0005658502
     0.00000605529
    -0.088719864
    -0.010613126
    -0.022314031
     0.001932968
    -0.24946609
     0.025012407
     0.015629628

Pj =
     0.1087
     0.1165
     0.0430
     0.0500
     0.0763
     0.1025
     0.0973
     0.1637
     0.0005
     0.0154
     0.1783
     0.0976
     0.1162
     0.1938
     0.1031
     0.1601
     0.0772
     0.2972
     0.1774

R =
    0.5682

S =
    8.8285e-09

'方差来源'      '偏差平方和'      '自由度'      '方差'      'F值'      'F_α'
'x1'            [ 0.1174]          [ 1]          [ 0.1174]    [1.5060e+15]    [3.8721]
'x2'            [ 0.1380]          [ 1]          [ 0.1380]    [1.7708e+15]    [6.7187]
'x3'            [ 0.0160]          [ 1]          [ 0.0160]    [2.0503e+14]    []
'x4'            [ 0.0088]          [ 1]          [ 0.0088]    [1.1241e+14]    []
'x5'            [ 0.0801]          [ 1]          [ 0.0801]    [1.0278e+15]    []
'x6'            [ 0.0804]          [ 1]          [ 0.0804]    [1.0320e+15]    []
'x7'            [ 0.0707]          [ 1]          [ 0.0707]    [9.0748e+14]    []
'x8'            [ 0.2350]          [ 1]          [ 0.2350]    [3.0149e+15]    []
'x9'            [9.7884e-07]       [ 1]          [9.7884e-07] [1.2558e+10]    []
'x10'           [ 0.0019]          [ 1]          [ 0.0019]    [2.3913e+13]    []
'x11'           [ 0.3100]          [ 1]          [ 0.3100]    [3.9769e+15]    []
'x12'           [ 0.0977]          [ 1]          [ 0.0977]    [1.2537e+15]    []
'x13'           [ 0.1985]          [ 1]          [ 0.1985]    [2.5466e+15]    []
'x14'           [ 0.3255]          [ 1]          [ 0.3255]    [4.1766e+15]    []
'x15'           [ 0.0580]          [ 1]          [ 0.0580]    [7.4437e+14]    []
'x16'           [ 0.2378]          [ 1]          [ 0.2378]    [3.0510e+15]    []
'x17'           [ 0.0438]          [ 1]          [ 0.0438]    [5.6181e+14]    []
'x18'           [ 0.3630]          [ 1]          [ 0.3630]    [4.6569e+15]    []
'x19'           [ 0.3374]          [ 1]          [ 0.3374]    [4.3287e+15]    []
'回归'         [ 5.3371]          [ 19]         [ 0.2809]    [3.6039e+15]    [1.6208]
'剩余'         [2.3773e-14]       [ 305]        [7.7943e-17] []             [1.9652]
'总和'         [ 16.5325]         [ 324]        []           []             []

```

从中可以看出: 回归方程是

$$\begin{aligned} y = & -8.898656 + 0.025845x_1 + 0.017971x_2 - 0.000324x_3 \\ & - 0.000008x_4 - 0.000280x_5 - 0.000447x_6 - 0.001817x_7 \\ & + 0.014382x_8 - 0.000011x_9 - 0.002475x_{10} - 0.000566x_{11} \\ & + 0.000006x_{12} - 0.088715x_{13} - 0.010613x_{14} - 0.022320x_{15} \\ & + 0.001933x_{16} - 0.249645x_{17} + 0.025012x_{18} + 0.015629x_{19} \end{aligned}$$

由方差分析表的显著性可以看出, 变量  $x_1, x_2, \dots, x_{19}$  对  $y$  的影响都是显著的, 说明选取的变量具有一定代表性, 此方程的剩余标准差  $S$  接近 0, 但可决系数  $R = 0.5682$ , 说明回归效果一般, 有较大改进的余地。

从 325 个样本中选取最后 50 个样本作为上述模型的误差分析, 可得出该模型的预测值与 RON 损失值的真实数据平均误差为 0.0981。

## 4. 方案优化

针对主要操作变量的优化问题, 要求在保证产品硫含量不大于  $5 \mu\text{g/g}$  的前提下, 使得样本的辛烷值 (RON) 损失降幅大于 30%。本文采用数据包络分析方法 (DEA) 进行方案优化, DEA 是一种针对多投入多产出问题的效率评价以及优化模型。

### 4.1. DEA 模型

#### 4.1.1. DEA 模型概述

DEA 模型, 又称为数据包络分析方法, 是 1978 年由美国运筹学家 A. Charnes 和 W.W. Cooper 提出的一种定量分析方法, 常用的有 CCR 和 BCC 等。DEA 以相对效率概念为基础, 以凸分析和线性规划为工具, 应用数学规划模型计算比较决策单元之间的相对效率, 对评价对象做出评价 [12] [13] [14]。DEA 能充分考虑对于决策单元本身最优的投入产出方案, 因而能够更理想地反映评价对象自身的信息和特点, 同时对于评价复杂系统的多投入多产出分析具有独到之处。

#### 4.1.2. DEA 原理

设有  $n$  个决策单元, 每个决策单元都有  $m$  个输入和  $s$  个输出, 分别记第  $j$  个决策单元的输入和输出为

$$X_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T, \quad j = 1, 2, \dots, n$$

$$Y_j = (y_{1j}, y_{2j}, \dots, y_{sj})^T, \quad j = 1, 2, \dots, n$$

其中  $x_{ij} > 0$ ,  $y_{rj} > 0$  ( $i = 1, 2, \dots, m$ ,  $r = 1, 2, \dots, s$ ) 分别表示第  $j$  个决策单元  $\text{DMU}_j$  的第  $i$  种类型的输入量和第  $r$  种类型的输出量。DEA 的产出指向型  $C^2R$  模型如下所示

$$\begin{aligned} \min E_{j_0} &= w^T X_{j_0} \\ \text{s.t. } w^T X_j - \mu^T Y_j &\geq 0, \quad j = 1, 2, \dots, n \\ \mu^T Y_{j_0} &= 1 \\ w &\geq 0, \mu \geq 0 \end{aligned} \quad (2)$$

利用对偶原理得到产出指向型的对偶模型

$$\begin{aligned}
& \max z \\
& \text{s.t. } \lambda^T X_j \leq X_{j_0}, j=1, 2, \dots, n \\
& \quad \lambda^T Y_j \geq z Y_{j_0}, j=1, 2, \dots, n \\
& \quad \lambda \geq 0, z \in E^n
\end{aligned} \tag{3}$$

判断决策单元的有效性, 本质上是判断它是否位于生产可能集的生产前沿面上。如果决策单元不为 DEA 有效, 通过计算、求解, 可以对原有的投入向量和产出向量进行调整, 使其成为 DEA 有效, 我们称经过调整后的点为决策单元在生产前沿面上的“投影”, 即为优化过程。求解最优解的模型如下:

设  $\lambda_0$ ,  $S^{-0}$ ,  $S^{+0}$ ,  $z^0$  为上述模型的最优解, 令

$$\hat{X}_0 = X_0 - S^{-0}$$

$$\hat{Y}_0 = z Y_0 + S^{+0}$$

称  $(\hat{X}_0, \hat{Y}_0)$  为  $DMU_{j_0}$  在生产可能集  $T_{C^2R}$  的生产前沿面上的“投影”。

## 4.2. 优化结果

根据优化过程中原料、待生吸附剂和再生吸附剂的性质保持不变, 此时所剩变量共 19 个。根据工艺要求, 19 个变量的取值范围应作为约束条件, 再依据优化目标, 得到其他约束条件如下:

$$\begin{cases} y \leq \hat{y} * 0.7 \\ S \leq 5, \end{cases}$$

其中  $\hat{y}$  为原始数据中的辛烷(ROn)损失值。

本文将 19 个主要变量作为输入指标, 辛烷值和硫含量作为两个输出指标, 总共 325 个样本, 这满足数据包络分析法的使用要求。利用 CCR 产出模型计算后对各个样本的操作变量进行调整, 优化操作方案使其满足约束条件, 采用 Matlab 编程实现, 部分计算结果见表 2。

**Table 2.** Results of Matlab running

**表 2.** Matlab 运行结果

| 编号 | $x_2$    | $x_3$    | $x_4$    | $x_5$    | $x_6$    |
|----|----------|----------|----------|----------|----------|
| 2  | 35.00525 | 470.9836 | 5394.292 | 100.0282 | 470.7347 |
| 3  | 32.8114  | 433.0178 | 5349.392 | 98.51615 | 459.351  |

由于分析的样本数据过多, 以上为截取部分数据的计算结果。为达优化目标, 通过包络分析法分析出每组样本数据的调整情况, 图中数据即为每组样本数据所进行优化操作变量后的值。

优化方案是每个数据样本中 19 个变量所对应原数据值与此图中对应变量的调整区间。例: 用图中编号 2 样本数据为例, 为达到优化目标,  $x_2$  变量建议下调至 35.00525,  $x_3$  变量建议下调至 470.9836, 其他变量操作相同。

## 5. 结束语

本文研究了一般石化企业的催化裂化汽油精制脱硫装置的模型构建。首先对原始数据进行预处理, 基于处理后的数据运用 Lasso 回归方法通过 Matlab 软件对变量进行降维处理。然后, 通过 Matlab 语言编程分析降维后的变量与 RON 损失的线性关系, 建立 RON 损失预测模型, 并通过预测结果与实际值的对比验证了 RON 损失预测模型的可靠性。最后基于数据包络分析法(DEA)对汽油精制过程中的主要操作变

量进行优化。建立的模型对降低辛烷损失值具有重要的理论意义和实际应用价值。

## 参考文献

- [1] 王宁, 闫娜, 徐友真, 杨剑锋. 复杂多工序制造过程关键质量特性识别[J]. 统计与决策, 2021, 37(8): 177-180.
- [2] 冯连坤, 陈晓华, 李彦荣. 操作工艺对混合汽油辛烷值损失的影响[J]. 石化技术与应用, 2021, 39(1): 51-53.
- [3] 刘畅, 魏静雯, 黄甜. 寻找降低汽油精制过程中的辛烷值损失模型中的主要变量[J]. 赤峰学院学报(自然科学版), 2020, 36(12): 7-11.
- [4] 郑茗友, 王伟, 赵文杰, 王建峰, 郗英杰. 基于 PSO-Lasso 算法的电站湿法脱硫出口 SO<sub>2</sub> 浓度预测[J/OL]. 华北电力大学学报(自然科学版), 2021: 1-7.  
<http://kns.cnki.net/kcms/detail/13.1212.TM.20210412.0910.002.html>, 2021-04-12.
- [5] 李成, 谢晗进. 赣南苏区居民获得感调查及影响因素研究——基于 LASSO 筛选[J]. 商业经济, 2021(3): 136-138.
- [6] 许赞娟, 罗幼喜. 基于变量聚类的主成分 Lasso 降维算法与模拟[J]. 统计与决策, 2021, 37(4): 31-36.
- [7] 王岩, 隋思涟. 数理统计与 MATLAB 数据分析[M]. 第 2 版. 北京: 清华大学出版社, 2014: 172-216.
- [8] 梅玉虹, 王灵芝, 雷雪, 李祖兰, 周戈耀. 中药民族药企业自主创新能力影响因素分析——以贵州省为例[J]. 卫生软科学, 2021, 35(4): 93-97.
- [9] 李伟, 肖书成, 王建华, 余扬. 基于线性回归分析的陆军航空兵训练航材消耗预测[J]. 兵工自动化, 2021, 40(4): 65-68.
- [10] 邵阳, 康宇. 基于多元线性回归预测采煤工作面瓦斯涌出量[J]. 煤炭技术, 2021, 40(4): 118-121.
- [11] 王惠文, 孟洁. 多元线性回归的预测建模方法[J]. 北京航空航天大学学报, 2007, 33(4): 500-504.
- [12] 高激, 黄银倩. 广东省海洋经济发展的影响因素分析[J]. 中国集体经济, 2021(10): 21-23.
- [13] 崔海洋, 卓雯君, 虞虎, 龙娇, 刘玉芳. 基于三阶段 DEA 模型的农业生产效率及其时空特征研究——以长江经济带为例[J]. 中国生态农业学报(中英文), 2020, 29(7): 1243-1252.
- [14] 徐秋艳, 郝涵. 基于 DEA 模型的我国大蒜产业种植生产效率研究[J]. 北方园艺, 2020(17): 153-159.