基于多元线性回归模型的锅炉冷壁温度曲线研究

## 摘要

## 引言

在燃煤发电过程中，锅炉是一种重要的热能动力设备。它通过在炉膛中燃烧煤粉释放热量，将水加热成一定温度（或压力）的蒸汽，蒸汽再推动汽轮机旋转并驱动发电机发电。锅炉的主要受热部分是水冷壁，通常由数排钢管组成，分布于锅炉炉膛的四周，其内部为流动的水，用于吸收炉膛中高温燃烧产生的辐射热量，水受热蒸发产生高压蒸汽。

在实际生产过程中，希望水冷壁的温度变化尽可能平稳，同时为保证安全，水冷壁温度不宜过高，否则有烧坏的风险。按照实际经验，操作人员给出的水冷壁温度超温报警线为445℃。影响水冷壁温度的因素有很多，包括锅炉负荷、蒸汽温度、蒸汽压力、燃料量、水煤比等。除了对水冷壁温度控制以外，识别出影响温度主要的操作变量对于工业生产和应用也是非常重要的。当然这个过程可以凭借专业的先验认知来确定，但为了可复用性，同时由于影响水冷壁温度的因素众多，因此需要建立一定的数学模型对其温度变化曲线进行分析和优化研究。

本文对各个管道温度的均值进行拟合，然后将拟合模型应用到各个管道上。首先进行数据预处理，包括缺失值填补、异常值识别与处理、数据标准化、数据集划分。接下来进行降维处理后建立合适的统计模型。本文采用可接受性强、模型拟合效果较优的OLS回归与Ridge回归建立模型，并检验模型的是否有效，然后通过对比两个模型的拟合效果最终确定应用的模型。

## 数据介绍

### 数据来源

数据来源于2021年长三角高校数学建模竞赛赛题数据。给出了10个具有代表性水冷壁管道的温度值，采样频率为15秒，共5000组数据；同时给出了影响水冷壁温度的153个输入变量的5000组数据，其中包括111个操作变量和42个状态变量。

这里的操作变量，也常称作控制变量，是指在锅炉燃烧过程中，操作人员能够根据实际生产的需要而进行调节的量；而状态变量是用来描述锅炉燃烧系统运行状态的量，它的取值由相关检测设备采集得到，操作人员无法进行直接调节。

由于原数据变量名较长且比较复杂，不便于编程实现和结果展示。因此，对所有输入变量进行统一编码。

本文分析使用的软件为R语言 3.6版，主要用到的软件包为：tidyverse、glmnet、broom。

### 数据预处理

#### 缺失值填补

收集得到的数据某些字段值存在为空的情况，一般有三种处理方法：删除记录、数据填补、空值处理。填补缺失值方法有：人工填补、均数填补、中位数或众数填补、多重填补、使用最接近的样本值填补等。不同情况变量的缺失值本研究采用了不同的处理方式。当某一条观测数据缺失项大于总项的65%时，由于缺失信息较大，故选择删除记录；考虑到不偏离原数据的总体分布，故在删除记录之后用列众数对缺失数据进行填补。处理之后还剩5000条观测数据，152个输入变量。

#### 数据标准化

不同的变量有不同的量纲，比如总给水流量、锅炉烟气含氧量、锅炉给水温度等，数值间的差别比较大，归一化就是为了消除变量之间量纲的影响。归一化就是将数据映射到某一固定区间，一般为（0,1）或（-1,1）。主要是为了数据处理方便提出来的，把有量纲表达式变成无量纲表达式。归一化公式：

考虑到量纲不同问题，故本文对除结局变量外的所有变量均用归一化方法进行处理。处理之后共计还剩5000条有效观测数据，152个输入变量。

#### 异常值处理

库克距离用来判断强影响点是否为Y的异常值点：

$$D\_i(M,c)=\frac{\left(\hat{\beta}(i)-\hat{\beta}\right)'M\left(\hat{\beta}(i)-\hat{\beta}\right)}{c}$$

M是给定的正定矩阵，c是给定的常熟，当较大时，判定为异常。一般认为当D<0.5时判断为非异常值点；当D>0.5时判断为异常值点。通过计算所以观测样本的Cook距离识别到如下6个异常值。

**表**

  本文采用直接对异常值删除的方式进行处理，处理之后还剩4994条观测数据，152个输入变量。

#### 数据集划分

将上述处理之后数据按照采用时间的顺序，以7:3的比例划分为训练集和测试集。得到训练集共3491条观测，测试集共1503条观测。

## 统计描述

### 基本统计特征

通过对11个（其中新增一列均值温度）水冷壁管道的数据分析，得到各个管道温度值的相关统计特征。从平均水平来看，各个管道的温度都在338℃～466℃的范围内波动。且平均温度最高的是管道10（424.92±18.70），最低是管道1（374.29±21.45），而管道9和管道10的温度均值在400℃以上，且最高温度值均高于超温报警线445℃，温度过高，有烧坏风险。整个水冷壁中出现的最高温度是在管道10中为465.40°，最低温度出现在管道6为338.00°。同时，各个管道温度的样本均值与样本中位数都非常匹配，没有发现任何异常现象。另外，通过识别序列的波峰波谷可以发现，整体上各个管道温度的变化波动都比较频繁，在4994个采样时间点上，出现了400次左右的波峰和波谷转折。

**表**

### 可视化分析

绘制各个管道温度的时序图如下。可以发现，10个水冷壁管道的温度变化范围基本在338℃～466℃以内。其温度变化均有一定的趋势，在取样点1000附近时，各个管道的温度都在平缓地上升，并达到最高值。随后各管道的温度在小范围内波动并以较快的速度下降到较低值，在取样点3500左右时又缓慢回升，最后趋于稳定温度。

管道1～管道8的总体变化趋势较为相似，且其温度变化范围都在超温警戒线以内。而管道9和管道10均有超过警戒线445℃的取样时间段，且管道10在取样点3200后，多个时段温度超出445℃，出现明显的超温现象。

**图**

为了更好的描述各水冷壁管道温度数据的离散情况，我们使用箱型图进行进一步分析。箱型图最大的优点就是不受异常值的影响，能够以一种相对稳定的方式描述数据的离散分布情况。

由下图可以分析得出，管道1的平均温度最低，管道10的平均温度最高。管道6的最高温度相较于其他管道最低，管道10的最高温度最大，且已超过警戒值445℃。管道3、9的温度值出现了离群点，存在异常值，而其他管道温度无异常值。管道平均温度处于中间水平，能够较好代表各个管道温度。

**图**

由于管道平均温度能够较好代表各个管道温度，为了降低模型的冗余，故先在管道平均温度上进行分析，然后应用到各个单独的管道数据上。因此，绘制单独的管道平均温度进行观察。

管道平均温度的整体走势符合各个管道的温度走势，且存在同各个管道类似的时序特点，最高温度和最低温度都同上面所述的各个管道情况接近。

**图**

### 相关性分析

 通过计算各自变量与各个管道温度的相关系数，对得到的各个输入变量的相关系数绘制箱线图如下图所示。图中横轴表示不同的自变量，纵轴表示关于各个管道的相关系数的箱线图。

  从图中可以发现，不同的变量对管道温度的相关性差异明显，存在多个高度正相关的输入变量，同时也存在一些高度负相关的输入变量。除此之外，还可以注意到在相关系数（-0.5，0.5）的范围内同样存在需要变量，从相关系数上来看，这些变量与因变量温度的相关性很低甚至无线性相关性。

  所以通过相关性分析知道输入变量中存在着大量的冗余变量。为了降低模型的复杂度，有降维的必要性，进而删除一些无关变量。

## 建立模型

### 数据降维

 常见的数据降维方法有很多种，例如主成分降维、Lasso降维、相关性降维等。本文通过对比不同参数下的弹性网约束估计降维，确定一个较优的降维参数。弹性网约束模型如下：

  从理论上可知，弹性网是综合了绝对约束和平方约束之后的模型，同时具有绝对约束估计和岭估计的特点。

* 当时：弹性网约束估计就是岭回归；
* 当时：弹性网约束估计就是Lasso回归；

  本文通过设置不同的参数对比训练结果进而确定最终降维方法。设置的参数为alpha=0（Ridge约束），0.25（弹性网约束），0.75（弹性网约束），1（Lasso约束）。得到如下的20折交叉验证的收敛图，从图中可知，横轴是对数值，纵轴是均方误差。按照此参数的设置，不会进行变量降维，保留的 变量个数为99个，保留的变量个数为99个，保留的变量个数为108个，保留的变量个数为88个，且四种参数设置下均方误差都接近0。

  另外从理论上分析可以知道，岭回归约束不能对变量进行降维；而Lasso约束作为一种广泛使用的降维处理手段，在变量选择上有明显的优势且实际效果也较好。故本文采用Lasso约束进行变量降维处理。

图

 得到的变量选择结果如下表所示，共选取了xxx个变量。

表

### OLS回归

#### 建立OLS模型

在描述统计的基础上，进一步分析因变量Y和各个输入自变量之间的关系。更进一步地，希望知道解释变量是怎样影响管道温度的，其影响强度如何。因此建立一个普通最小二乘回归模型对训练集数据进行拟合。

模型的相关参数以及检验结果如下表所示。模型整体的F-检验高度限制（P-值<0.001），这说明至少有一个输入变量更管道温度显著相关。模型的调整后判决系数，说明输入变量对管道温度的拟合情况较优，但可能存在着较为严重的共线性问题，但本文主要考虑的是对管道温度的预测效果，故不考虑共线性问题。

表

检查分析每一个输入变量的t-检验结果。如下表中的结果显示，X11、X12、X14、X15、X22、X29、X55、X57、X61、X65、X68、X71、X72、X85、X90、X91、X93、X95、X106、X107、X108、X121、X123、X143、X144、X145、X152变量高度不显著（检验水平$\alpha=0.05$）。另外从表中可以发现，在高度显著的变量中，X7、X16、X111、X127在控制其他变量不变的前提下，与管道温度的变化呈现明显正相关；X18、X137在控制其他变量不变的前提下，与管道温度的变化呈现明显负相关。

表

#### OLS模型残差检验

 对于OLS模型中的随机误差向量，若高斯-马尔科夫的某些假设不成立，的分布就会发生变化，从而导致残差e的分布变化。因此有必要对所建立模型进行残差分析，残差分析的主要思想就是通过对e的表现加以研究，包括零均值、同方差、独立性，此处关于高斯-马尔科夫假设不进行详细阐述。从预测角度定义残差：

  由于量纲和度量单位的影响，直接进行比较普通残差是不适宜的，为此将他们标准化。有两种重要的标准化残差定义方式就是内学生化残差和外学生化残差。本文同时给出了这两种残差的残差图如下所示。

  由残差图可以发现，这些残差大致都落在（-2，2）的水平带域内，且不存在任何的趋势。这时数据与假设没有不一致的征兆，可以认为这个假设基本上是合理的。

### Ridge回归

### 模型结果

### 模型应用

## 结论

## 参考文献

## 附录