**形状

中度可信度描述已自动生成**

**智能制造质量预测**

组长： 张庆伟 3210103414

组员： 葛涛 3210103407

组员： 刘祥盛 3210105577

2023年11月08日

目录

[**一、** **问题介绍** 3](#_Toc150543852)

[**1.1.** **问题背景** 3](#_Toc150543853)

[**1.2.** **问题描述** 4](#_Toc150543854)

[**1.3.** **评估指标** 4](#_Toc150543855)

[**二、** **数据处理** 5](#_Toc150543856)

[**三、** **线性回归** 9](#_Toc150543857)

[**3.1** **立项阶段** 9](#_Toc150543858)

[**3.2** **外部对接1** 9](#_Toc150543859)

[**3.3** **外部对接2** 9](#_Toc150543860)

[**3.4** **筹备阶段** 9](#_Toc150543861)

[**3.5** **活动阶段1** 9](#_Toc150543862)

[**3.6** **活动阶段2** 9](#_Toc150543863)

[**3.7** **总结阶段** 9](#_Toc150543864)

[**四、** **Lasso回归** 10](#_Toc150543865)

[**4.1 Lasso回归介绍** 10](#_Toc150543866)

[**4.2 代码分析** 10](#_Toc150543867)

[**五、** **Ridge回归** 12](#_Toc150543868)

[**5.1 Ridge回归介绍** 12](#_Toc150543869)

[**5.2 代码分析** 12](#_Toc150543870)

[**5.3 运行结果** 13](#_Toc150543871)

[**5.4 分析总结** 14](#_Toc150543872)

[**六、** **小组分工** 16](#_Toc150543873)

1. **问题介绍**
   1. **问题背景**

半导体产业是一个信息化程度高的产业。高度的信息化给数据分析创造了可能性。基于数据的分析可以帮助半导体产业更好的利用生产信息，提高产品质量。

现有的解决方案是，生产机器生产完成后，对产品质量做非全面的抽测，进行产品质量检核。这往往会出现以下状况，一是不能即时的知道质量的好坏，当发现质量不佳的产品时，要修正通常都为时以晚，二是在没有办法全面抽测的状况下，存在很大漏检的风险。

在机器学习，人工智能快速发展的今天，我们希望着由机器生产参数去预测产品的质量，来达到生产结果即时性以及全面性。更进一步的，可基于预先知道的结果，去做对应的决策及应变，对客户负责，也对制造生产更加敏感。

**痛点与挑战：**

（1）TFT-LCD（薄膜晶体管液晶显示器）的生产过程较为复杂，包含几百道以上的工序。每道工序都有可能会对产品的品质产生影响，故算法模型需要考虑的过程变量较多。

（2）另外，这些变量的取值可能会存在异常（如测点仪表的波动导致、设备工况漂移等现象），模型需要足够稳定性和鲁棒性。

（3）产线每天加工的玻璃基板数以万计，模型需要在满足较高的精准度前提下尽可能实时得到预测结果，这样才能给在实际生产中进行使用。

**价值：**

（1）如果能够建立算法模型准确预测出特性值，便可以实现生产过程的实时监控和预警，提前发现当前工序的问题、避免问题流入到后道工序，减少生产资源浪费的同时也优化了产品良率。

（2）基于预测模型得到的关键参数，工艺人员能够快速地针对那些电性表现不佳的产品进行问题溯源分析，重点分析和调整那些关键的影响因子，加快不良问题的处理、提高整体工艺水平。

（3）该预测模型在部署后也可以用于减少特性检测相关的工序，能够节约检测资源并且对提升产线整体的产能有正面作用。

**问题/数据集来源：**

天池智能制造质量预测数据集：<https://tianchi.aliyun.com/dataset/140667>

具体处理见第二部分“数据处理”。

* 1. **问题描述**

本题目提供的数据集提供了生产线上的抽样数据，反应机台的温度，气体，液体流量，功率，制成时间等因子。通过这些因子，需要研究人员设计出模型，准确的预测与之相对应的特性数值。这是一个典型的回归预测问题。

* 1. **评估指标**

本任务采用常用的MSE指标作为评估指标，具体来讲：是我们建立回归模型预测的值，是真实的值。的值越小，代表预测结果和真实值越接近，效果越好。其中是：

1. **数据处理**

我们选用天池智能制造质量预测数据集， 本次大作业所用的所有数据集均在DataSet文件夹中，这里做一点说明：

训练集.xlsx：共5954列，801行；

测试集A.xlsx：共5953列，301行；

测试集B.xlsx：本次大作业中未使用；

测试集A\_答案.csv：官方提供的测试集A的答案；

train：经过数据处理后得到的训练集，用于后续的模型训练；

testa：经过数据处理后得到的测试集a，用于后续的模型训练；

训练集中共有800条样本数据，每条数据包含5954列字段。第一个字段为ID号码，最后一列为要预测的值Y。其余的5952列数据为用于预测Y的变量X。这些变量一共由多道工序组成，字段的名字可以区分不同的工序，例如 210X1, 210X2。300X1,300X2（即标题行）。字段中的TOOL\_ID或者Tool为每道工序使用的机台。

测试集A中共有300条样本数据，每条数据包含的字段数与训练集相同，但是缺少最后一列的Y值。

下面是对训练集数据处理的介绍（数据处理的所有源代码均在中）：

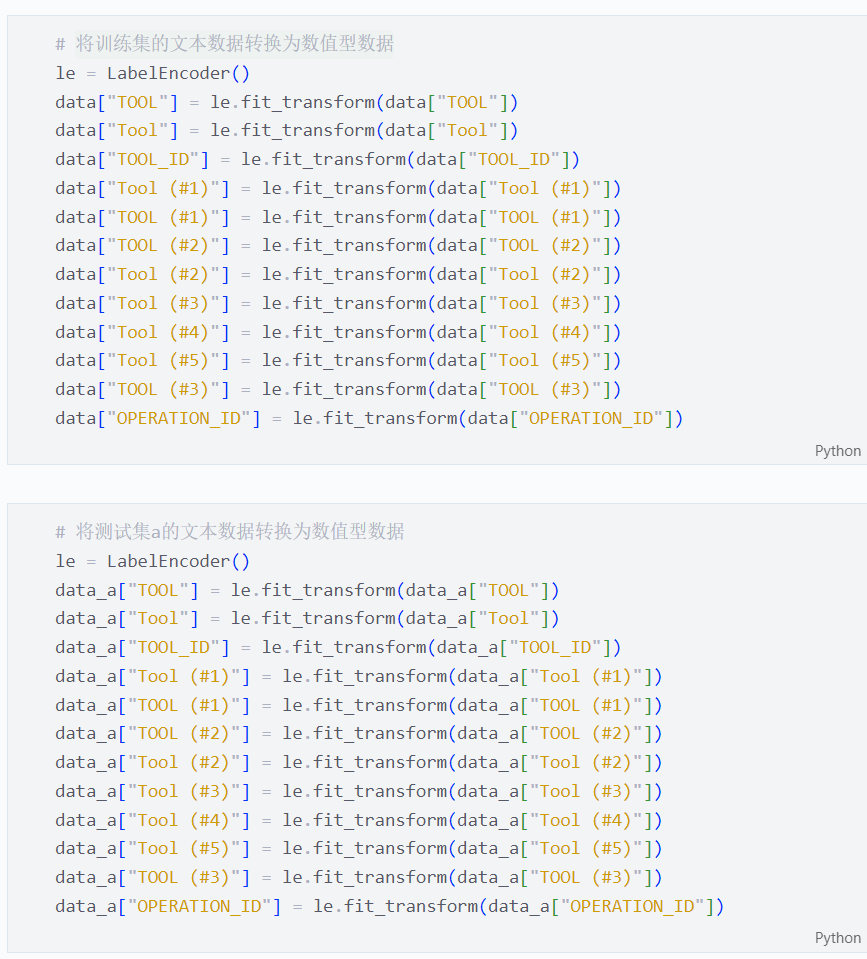
1. **预处理**

导入外部库，导入训练集.xlsx为train(data)，导入测试集A为testa(data\_a)；

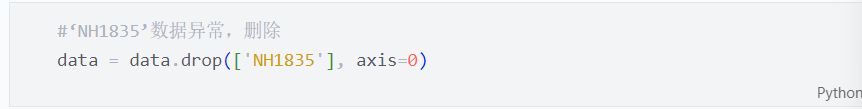


1. **类型转换**

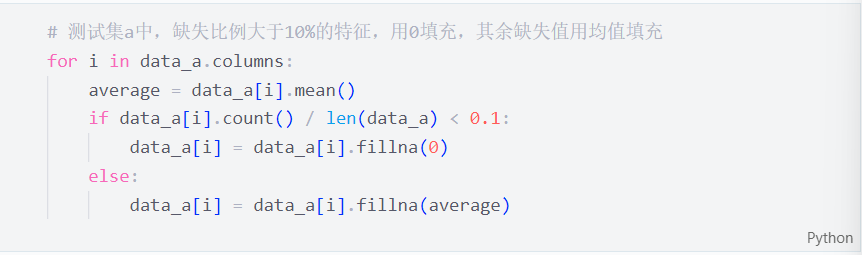
我们借助中的将训练集和测试集A的文本数据（类型，即数据中的工具列“TOOL”等），转换为数值型数据



1. **异常处理**
2. 删除重复样本，共六条；
3. 删除Y值异常的样本，共一条；

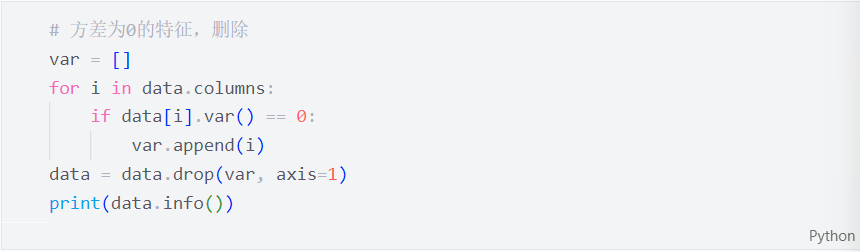


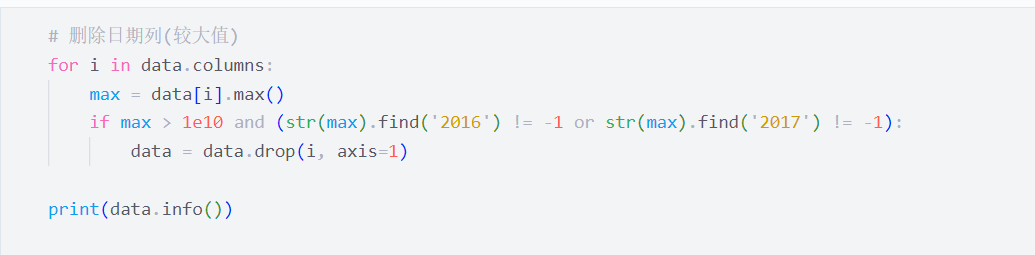
1. 测试集A和训练集中，所有特征均存在缺失值，对于缺失比例大于10%的特征，用0填充；其余缺失值用均值填充；





1. 删除方差为0的特征（即列中数据完全一样）；





1. 删除特征之间相关性（线性化）程度过高的特征，只保留一条以显示这几个特征所传递的信息；

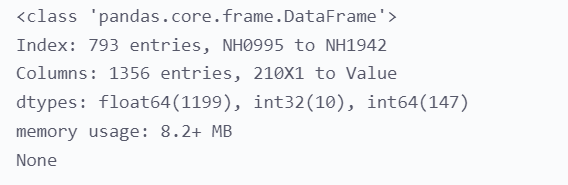


1. 后期处理中，我们注意到训练集和测试集A中的某些特征差距很大，极大地影响了回归模型的稳定性，因此我们删除训练集中与测试集A差距很大的特征。



1. 数据输出

通过数据处理，我们将训练集从原来的5954×801压缩到了1356×793。



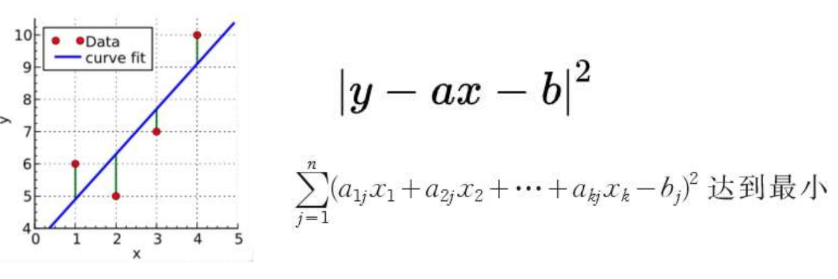
1. **线性回归**
   1. **基本含义**

在统计学中，线性回归（Linear Regression）是利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归，大于一个自变量情况的叫做多元回归。

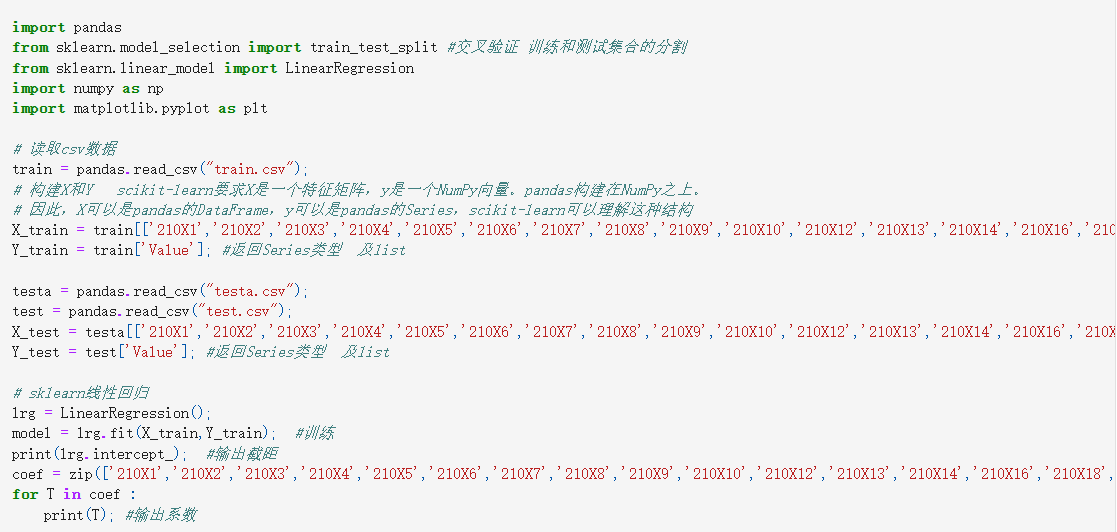
* 1. **拟合方程**

一般来说，线性回归都可以通过最小二乘法求出其方程，一般地，影响y的因素往往不止一个，假设有x1，x2，...，xk，k个因素，通常可考虑如下的线性关系式：

用“距离直线的竖直距离的平方”反应误差



* 1. **代码部分展示**

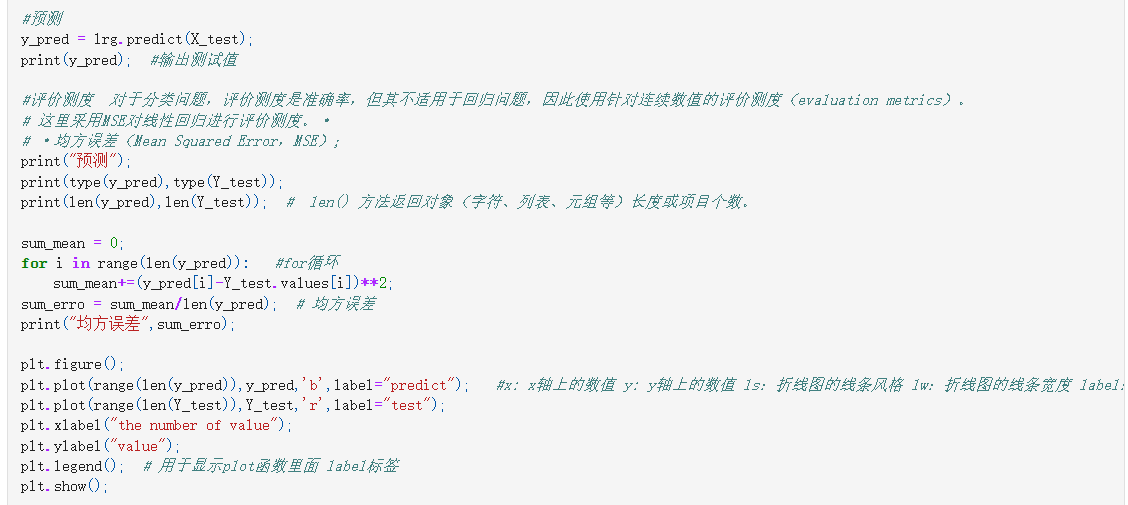


先是对训练集的数据进行线性回归的模型训练

lrg = LinearRegression();

model = lrg.fit(X\_train,Y\_train);  #训练

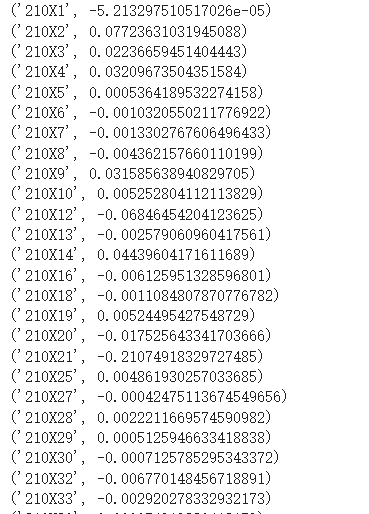
将训练集的多元数据与Value值采用最小二乘法，建立多元线性方程，得到各个变量的系数。



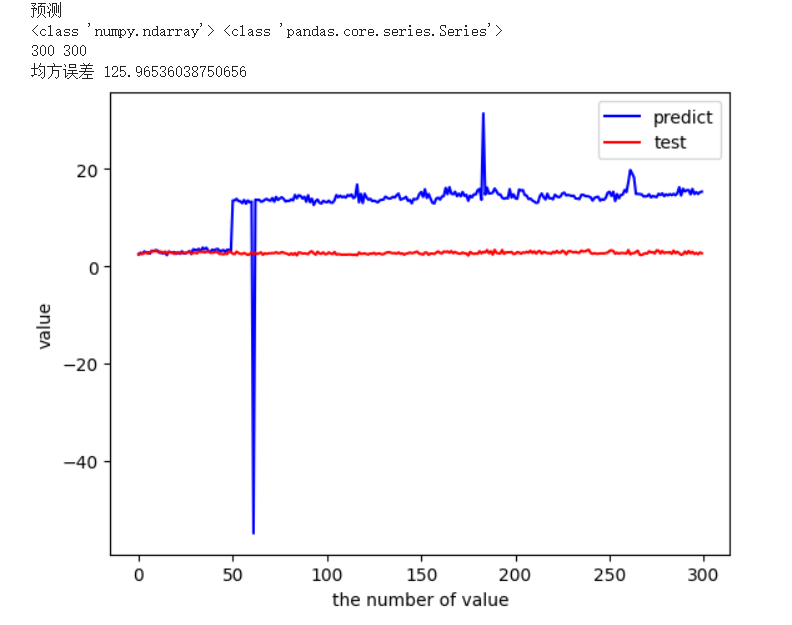
将测试集的数据代入到训练得到的多元线性方程中，并求出预测值与真实值的均方误差作为评价测度，画出预测值与实际值比较的折线图。

* 1. **运行结果**

1. 各个变量的系数：



②评价测度值以及各预测数据与实际数据比较的折线图：



* 1. **结果分析**

结果显示，线性回归算法得到结果的均方误差高达一百多，而通过折线图可以看出预测数据的前四十组数据与实际值相近，而后面的两百多组数据都与实际值有着相似较大的偏差。

显然，对训练集使用线性回归训练后的模型对测试集的预测结果不太理想，误差较大。

误差分析：

* 1. 训练集数据大小范围较为集中（即工艺相似，种类少），而测试集的数据大小跨度大（工艺种类多），导致测试结果与训练数据大小相近（工艺种类相似）的误差小（如前四五十个数据）；而后面预测的数据误差很大。
  2. 线性回归最小二乘的算法容易受“离群值”的影响，离群值影响被平方放大，拟合会“迁合”这些离群值，这是线性回归训练模型的不足。
  3. 线性回归缺乏因果推断能力，多元线性回归只能表明相关性，但无法推断出因果关系，而工艺制造的数据之间往往存在因果关系。
  4. **总结**

采用线性回归模型对该智能制造质量预测存在较大的不足，不是一个理想的预测模型。导致线性回归模型缺陷的可能有以下原因：

①对异常值和离群点敏感：

多元线性回归对异常值和离群点敏感，这些异常值可能会对回归系数产生较大影响，从而影响模型的稳健性和准确性。

②需要满足线性假设：

多元线性回归模型要求自变量与因变量之间的关系是线性的，如果实际关系存在非线性，就会导致模型拟合不佳。

③对多重共线性较为敏感：

当自变量之间存在高度相关性（多重共线性）时，多元线性回归模型的参数估计会变得不稳定，使得模型的解释能力下降。

④对数据分布要求较高：

多元线性回归模型通常要求自变量和因变量的分布符合正态分布，且误差项具有相同方差（同方差性），否则会影响参数估计的准确性。

⑤需要满足独立性假设：

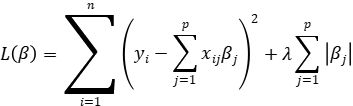
多元线性回归模型假设误差项之间是相互独立的，如果存在自相关性（误差项之间相关）就会违反该假设。

1. **Lasso回归**

**4.1 Lasso回归介绍**

Lasso回归（Least Absolute Shrinkage and Selection Operator Regression）是另一种用于处理多重共线性问题的回归方法，类似于岭回归，它也通过在最小二乘法的损失函数中加入正则化项来降低回归系数的大小，从而提高模型的稳定性和泛化能力。

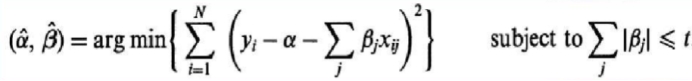
Lasso回归的损失函数可以表示为：



其中，​是第个观测值，是第个样本的特征向量,是回归系数向量，是样本数量，是特征数量，是正则化参数，也称为Lasso参数，它控制了正则化项对损失函数的影响程度。

Lasso回归使用的是，即回归系数的绝对值，作为正则化项。这使得在Lasso回归中，一些回归系数可能会被压缩到零，从而实现了特征选择的效果。因此，Lasso回归不仅可以缓解多重共线性问题，还能够在一定程度上进行特征选择，去除对预测目标贡献较小的特征。

Lasso回归的回归系数可以通过以下公式求解：



Lasso回归要正确选择合适的 值。较大的值会导致更多的回归系数被压缩到零，从而增加模型的稳定性但可能引入欠拟合，而较小的 值则可能引发过拟合。

总的来说，Lasso回归在处理多重共线性问题的同时，还具有特征选择的优势，但也需要谨慎选择正则化参数。本次通过sklearn.linear\_model库中Lasso。

**4.2 代码分析**

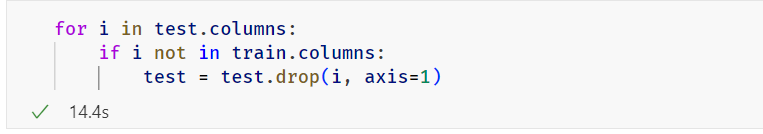
导入所需的库。



读取数据并删除测试集中训练集没有的特征

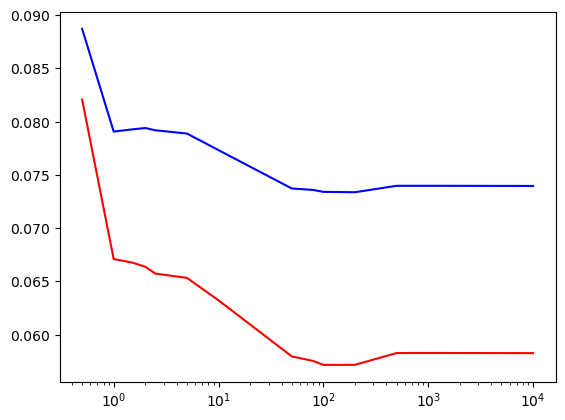


通过一系列数据转换，将导入的数据从𝑝𝑎𝑛𝑑𝑎𝑠的𝐷𝑎𝑡𝑎𝐹𝑟𝑎𝑚𝑒类型转换为𝑁𝑢𝑚𝑝𝑦的浮点类型𝑛𝑑𝑎𝑟𝑟𝑎𝑦，并将训练集中的特征数据与Y值分离。

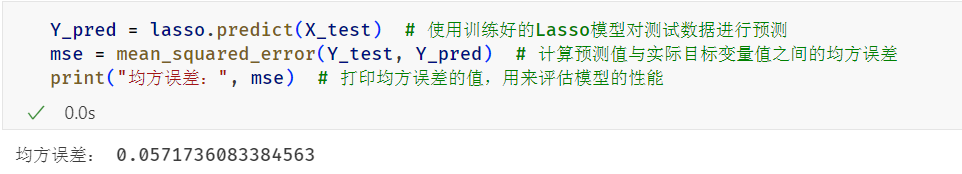


**4.3 运行结果**

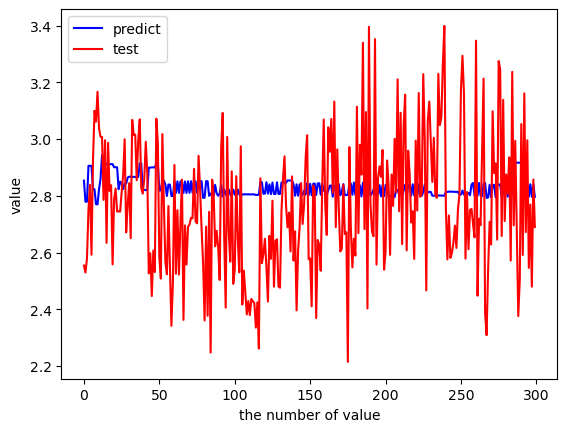
在Lasso模型的λ参数选择时，我自己设计了一个列表[0.5, 1, 1.5, 2, 2.5, 5, 10, 50, 80, 100, 200, 500, 1000, 10000]，从中找出最合适的λ值。



得到预测值后，计算预测值与真值之间的均方误差，大约为0.0572：



图像如下：



**5.4 分析总结**

数据显示， *Lasso*回归模型预测的结果与线性模型相比较好，其均方误差较小，为0.0572；但是将本模型的预测值与真实值放在一起对比，可以看出，其余预测数据与真实值相差略大，不能很好地跟踪其变化。原因可能有：

多重共线性： 当特征之间存在高度相关性时，Lasso回归可能面临困难。由于L1正则项的存在，它倾向于选择一个特征并将其他高度相关的特征的权重设为零。在存在共线性的情况下，由于多个特征与目标变量高度相关，选择其中一个而忽略其他可能导致不准确。

选择参数： Lasso回归的性能受到正则化参数（通常表示为λ）的影响。选择合适的正则化参数对于模型的性能至关重要。如果选择的参数不合适，可能会导致模型过度稀疏或不足够稀疏，从而影响模型的准确性。

非线性关系： Lasso回归是一种线性模型，数据可能包含非线性关系，Lasso可能无法很好地捕捉这些关系。

数据异常值： 数据中的异常值对Lasso回归的性能产生负面影响，因为Lasso对异常值敏感。异常值可能导致模型的参数选择不稳定，从而影响预测的准确性。比如其中某一列的数据有一个跃变，影响回归分析。

1. **Ridge回归**

**5.1 Ridge回归介绍**

回归是一种用于处理多重共线性问题的回归方法，它通过在最小二乘法的损失函数中加入一个正则化项，来减小回归系数的大小，从而提高模型的稳定性和泛化能力，这非常适合本题在线性回归模型表现不佳的情况。

岭回归的损失函数可以表示为：

其中，​是第个观测值，是第个样本的特征向量,是回归系数向量，是样本数量，是特征数量，是正则化参数，也称为岭参数，它控制了正则化项对损失函数的影响程度。当时，岭回归就退化为普通的最小二乘法；当时，岭回归就会对回归系数进行收缩，使得它们更接近于零，从而降低模型的复杂度和方差，增加模型的偏差，达到一种偏差-方差权衡的效果。

岭回归的回归系数可以通过以下公式求解：

其中，是样本的特征矩阵，是样本的观测值向量，是单位矩阵。可以看出，岭回归通过在的对角线上加上，避免了的奇异性，从而保证了逆矩阵的存在，使得回归系数可以估计出来。

岭回归的一个重要问题是如何选择合适的值，一般来说,值越大，回归系数的收缩程度越高，模型的方差越小，但是模型的偏差也越大，可能导致欠拟合的问题；反之，值越小，回归系数的收缩程度越低，模型的方差越大，但是模型的偏差也越小，可能导致过拟合的问题。因此，需要正确选择的值。

岭回归是一种简单而有效的正则化方法，它可以有效地缓解多重共线性问题，提高模型的稳定性和泛化能力，但是它也有一些局限性，比如它不能实现特征选择，因为它不会将回归系数完全压缩到零，而是保留了所有的特征。

本次回归通过中的创建。

**5.2 代码分析**

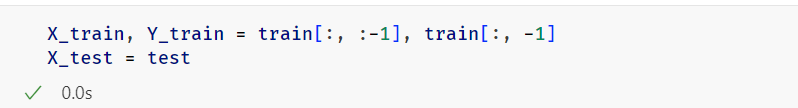
首先导入所需库：



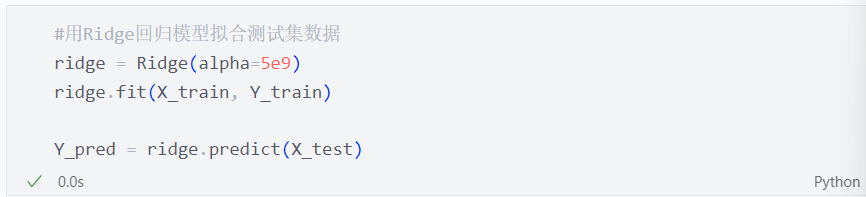
读取数据并删除测试集中训练集没有的特征



通过一系列数据转换，将导入的数据从𝑝𝑎𝑛𝑑𝑎𝑠的𝐷𝑎𝑡𝑎𝐹𝑟𝑎𝑚𝑒类型转换为𝑁𝑢𝑚𝑝𝑦的浮点类型𝑛𝑑𝑎𝑟𝑟𝑎𝑦，并将训练集中的特征数据与Y值分离。

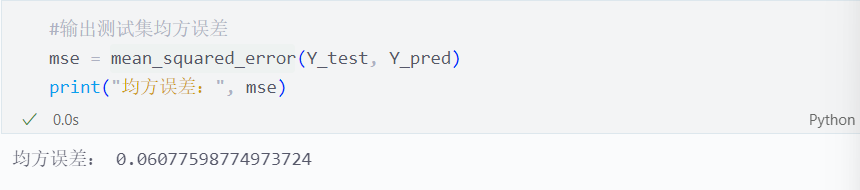


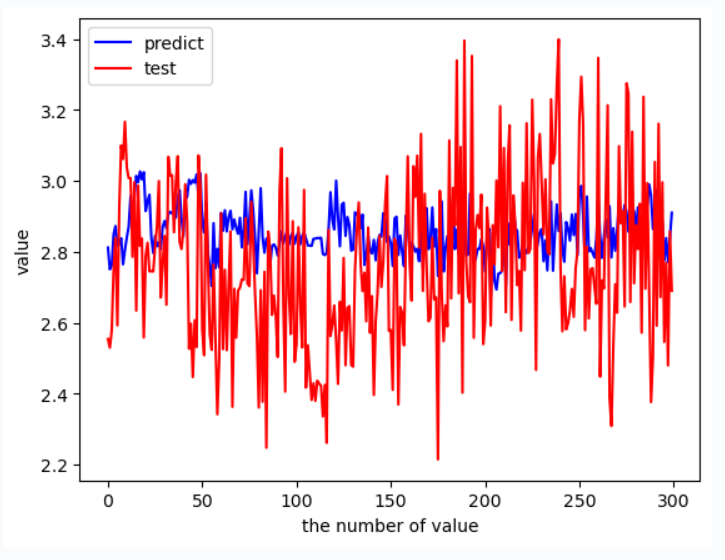
然后通过𝑠𝑘𝑙𝑒𝑎𝑟𝑛.𝑙𝑖𝑛𝑒𝑎𝑟\_通过中的𝑅𝑖𝑑𝑔𝑒创建模型，并以此模型对训练集进行训练，最后得到预测值。



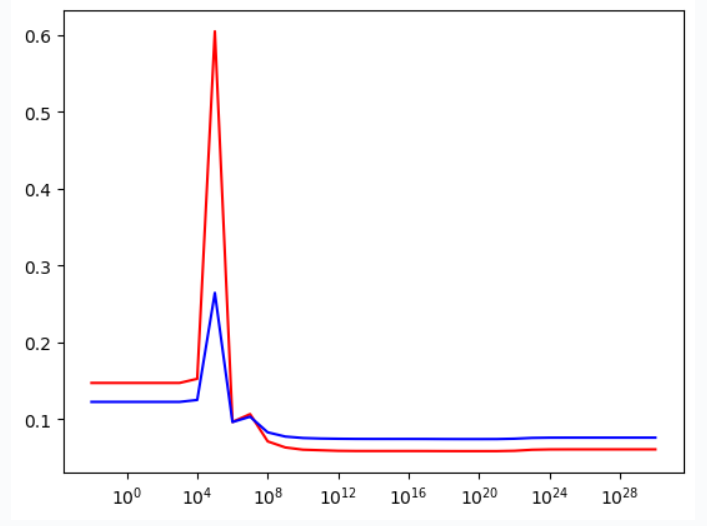
**5.3 运行结果**

得到预测值后，计算预测值与真值之间的均方误差，大约为0.0608：





在𝑅𝑖𝑑𝑔𝑒模型的𝑎𝑙𝑝ℎ𝑎(𝐿2)参数选择时，我遍历了从1𝑒−2至1𝑒30之间（十倍增长）的所有𝑎𝑙𝑝ℎ𝑎值，最终选择了5𝑒9作为我的𝑅𝑖𝑑𝑔𝑒模型的正则参数。



**5.4 分析总结**

数据显示，𝑅𝑖𝑑𝑔𝑒回归模型预测的结果与线性模型相比较好，其均方误差较小，为0.0608；但是将本模型的预测值与真实值放在一起对比，不难观察到除去第1-50组和250-300组的数据，其余预测数据与真实值相差略大，不能很好地跟踪其变化。

其好与坏的原因出自𝑅𝑖𝑑𝑔𝑒回归模型的特性。𝑅𝑖𝑑𝑔𝑒回归模型可以缓解多重共线性问题，提高模型的稳定性和泛化能力，并通过正则化参数控制模型的复杂度，防止过拟合的问题，这从数据和图像中可以看出，相对于线性回归模型，𝑅𝑖𝑑𝑔𝑒回归几乎没有过拟合的现象；且在本题提供如此高维数据的情况下（数据间的共线性性质强），𝑅𝑖𝑑𝑔𝑒回归模型仍能提供较好的预测值（𝑅𝑖𝑑𝑔𝑒回归可以处理特征数大于样本数的情况）。但是本题中我选择了较大的正则系数，这导致了预测数据的毛刺较大，没能更好地拟合真实值的大小。

1. **小组分工**

葛涛：选题，线性回归模型及其PPT、文字报告

刘祥盛：选题，Lasso回归模型及其PPT、文字报告

张庆伟：问题介绍、数据处理、Ridge回归模型，及其PPT、文字报告