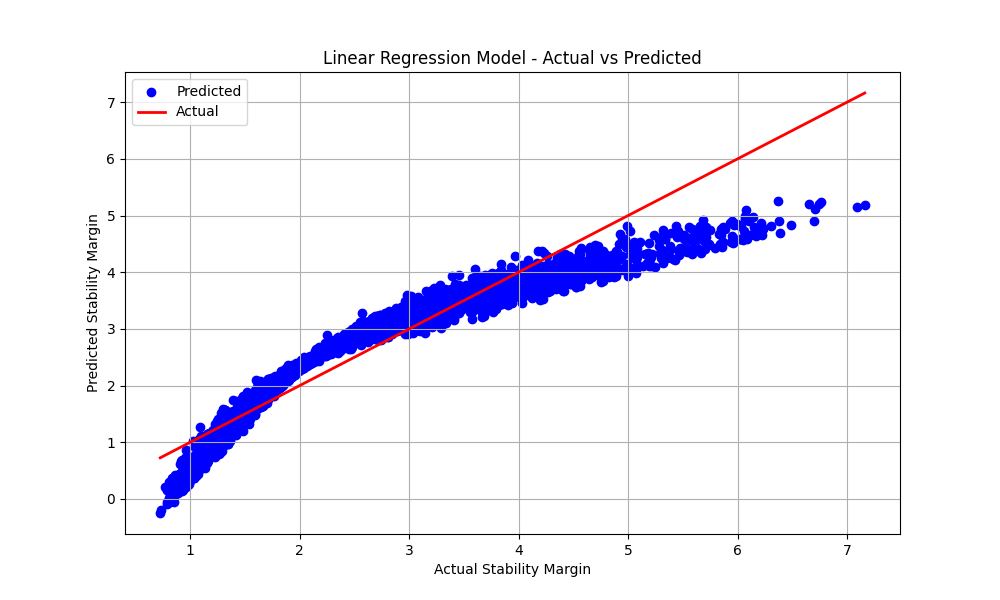
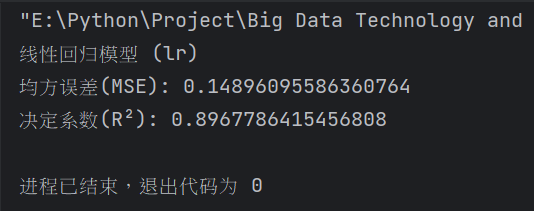
第2次作业\_回归

电 25 吴晨聪 2022010311

# （1） Given the dataset ‘hw2\_3gpf.csv’, apply the linear/nonlinear regression approaches to estimate the stability margin of the system

在回归分析的过程中应用了线性回归模型（lr）。首先对数据集进行读取，再把数据集的80%划分成训练集，20%划分为测试集，这是常见的用于大数据量的数据集比例，可以确保足够训练数据的同时也有足够数据用于测试模型的性能，然后对模型进行训练。测试得到最终数据的误差和决定系数，并以画图的形式呈现，直观比较模型的偏差。

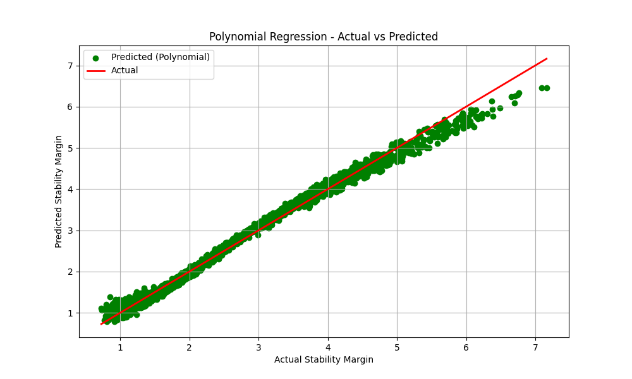
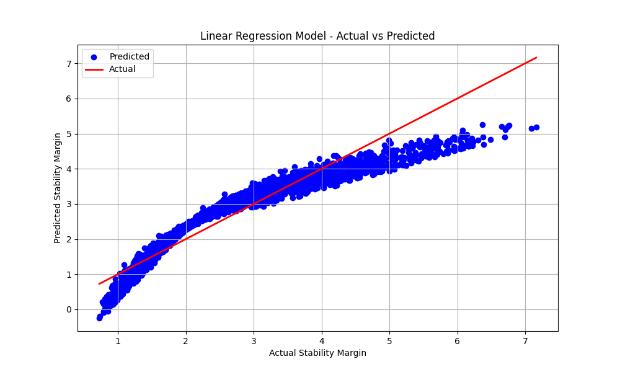




以上两张图片分别为线性回归模型对系统稳定性裕度的估计值(蓝点)和实际稳定性裕度(红线)的对比图，以及回归方程的均方误差和决定系数。可以看出，模型的线性化程度较高，线性回归的均方误差在0.1490左右，属于误差较小的范围，同时决定系数为0.8968左右也在较高的水平，说明线性回归模型能良好地应用在此问题。

# （2） Try different regression methods to find the best fitting results. Try to understand the overfitting & underfitting phenomenon when you adjust the regression models and hyperparameters.

在本题目中，对回归分析的过程中新加入了、多项式回归模型（lr2）以及Lasso回归模型（ls），前者通过引入二次项，捕捉特征与目标变量之间的非线性关系。后者则对模型进行 L1 正则化，强制模型将不重要的特征系数缩小到零，从而达到特征选择的效果。训练过程与上题相似，首先对数据集进行读取，再把数据集的80%划分成训练集，20%划分为测试集，对模型进行训练，得到最终数据的误差和决定系数，并以画图的形式呈现，直观比较三种回归模型的表现。



一張含有 繪圖, 行, 圖表, 文字 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

以上图片为各模型对系统稳定性裕度的估计值和实际稳定性裕度(红线)的对比图，其中蓝点为线性回归模型;绿点为多项式回归模型(degree=2);紫点为Lasso 回归模型，以及各模型的均方误差和决定系数。对比之下，degree=2的多项式回归模型的表现最好，误差为最低的0.0137且决定系数最高0.9905非常接近1，表明该模型比线性回归模型更适合于这个问题。

Learning curve 分析:

从learning curve 的角度判断各模型是否欠拟合或过拟合。

一張含有 文字, 圖表, 行, 繪圖 的圖片

自動產生的描述一張含有 文字, 行, 圖表, 繪圖 的圖片

自動產生的描述

一張含有 文字, 圖表, 繪圖, 行 的圖片

自動產生的描述

各回归模型的learning curve如以上各图所示。其中横坐标为训练集中数据的数量，纵坐标为方差，每个点的误差都是五次随机抽样训练后得到的均方误差的平均值，红线为训练误差，绿线为验证误差。可以看出线性回归模型和degree=2的多项式回归模型两者均无过拟合的问题，但线性回归的总体看来误差较大且误差的波动更大，存在欠拟合的问题。而对Lasso 回归模型而言, 可以看到模型的误差和误差波动都比前两个模型都要大得多，说明模型存在过拟合问题，因此对于本题目，采用degree=2的多项式模型训练能得到更为合适的结果。