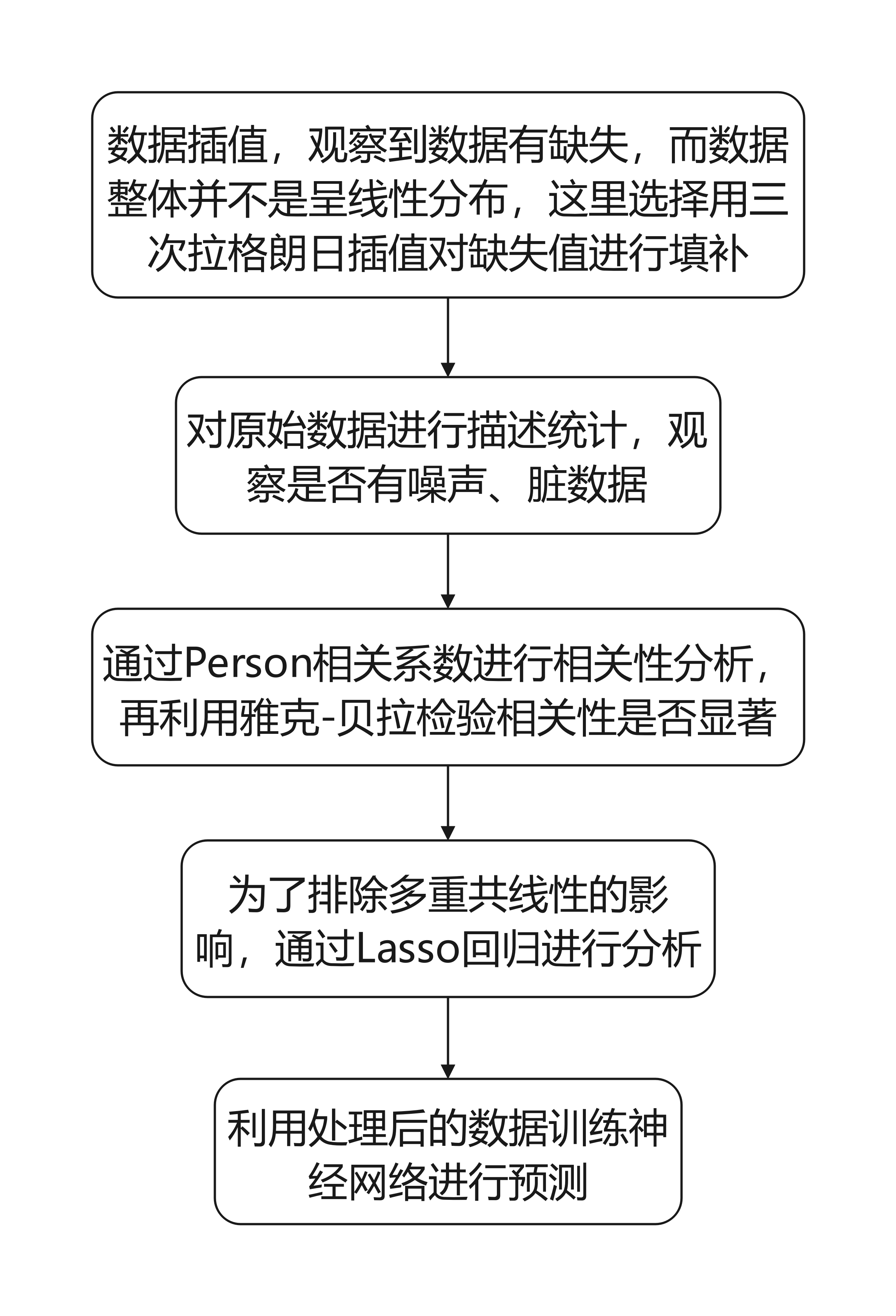
**财政预测项目报告**

* 项目背景

地方财政是国家财政收入的重要来源，地方财政收入与社会从业人数、社会消费品零售总额等密切相关，如何合理的分配地方资源，促进地方的财政的收入提高是地方政府要考虑的首要问题，对地方财政收入进行预测，能帮助政府确定当年的发展方向，提高政府的财政收入。

* 项目架构

项目架构分析图如下：

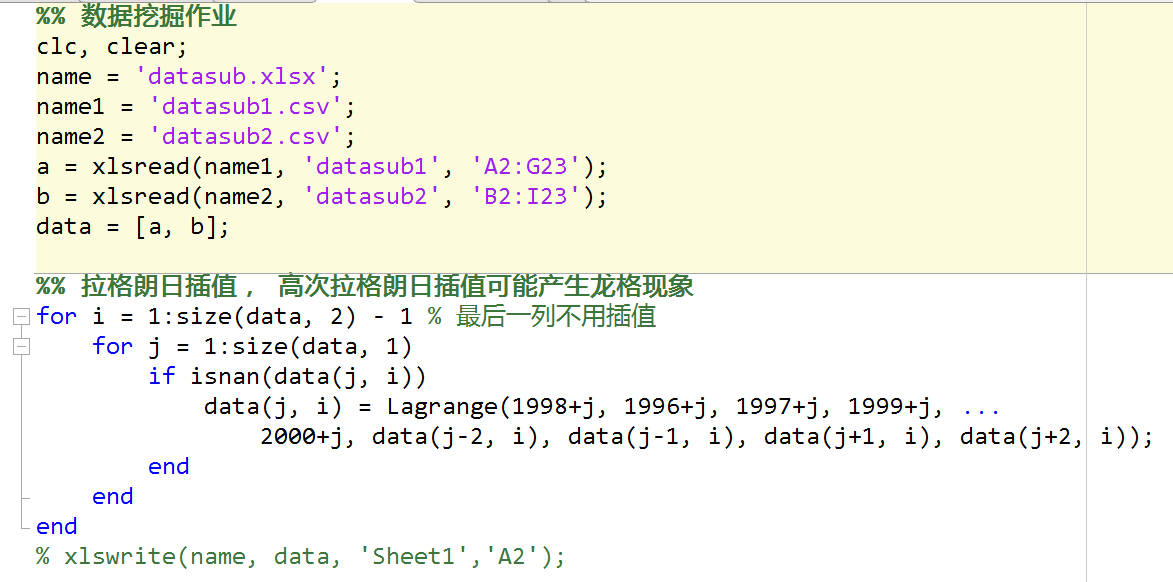


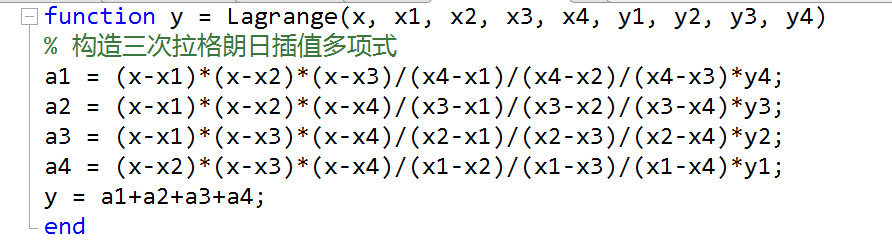
* 项目内容

1. 数据探索与预处理

首先读入数据到matlab中，对数据进行遍历，如果有缺失值则调用三次拉格朗日插值函数对数据进行插值处理。常用的插值方法有平均数插值、牛顿插值、拉格朗日插值、三次样条插值、埃尔米特插值等，注意这里使用了三次拉格朗日插值函数，因为数据整体是呈现一种加速上升的趋势，所以应该使用多次函数进行插值处理，但因为高次插值会产生龙格震荡现象，所以此处构造了三次拉格朗日插值函数。

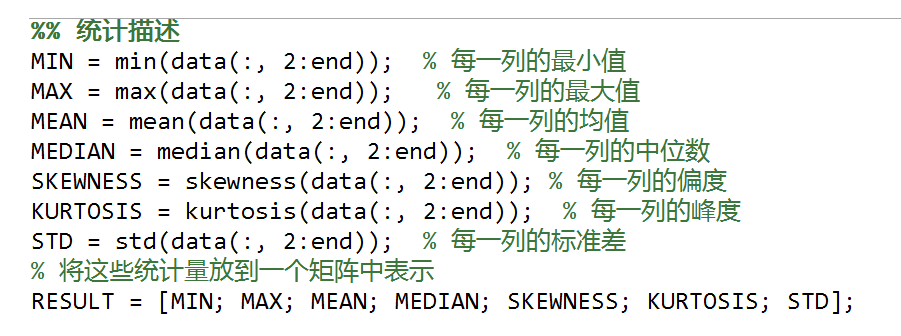
代码如下：

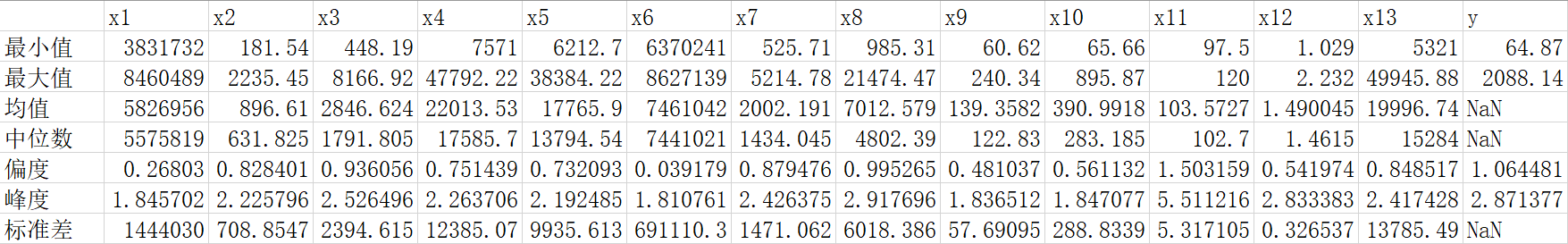




1. 数据特征可视化

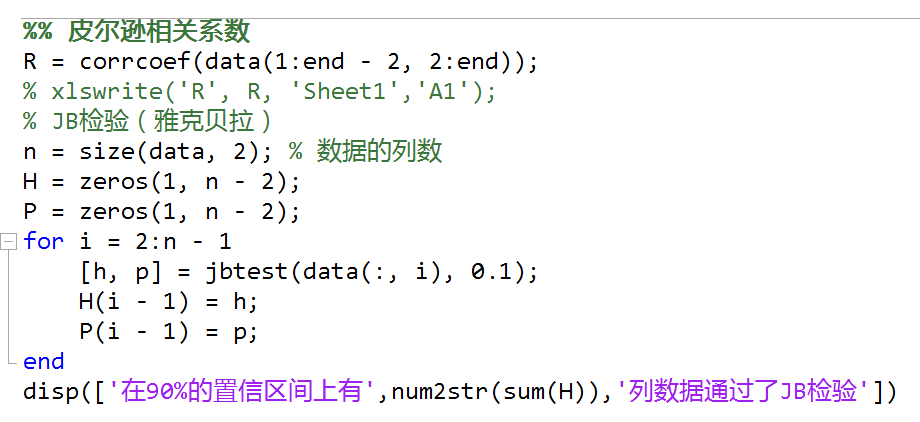
为了判断数据除了缺失值之外有无异常值，也就是脏数据，我们需要对数据特征进行检验，调用matlab库函数对数据进行数据分析后写入Excel，代码及结果如下：





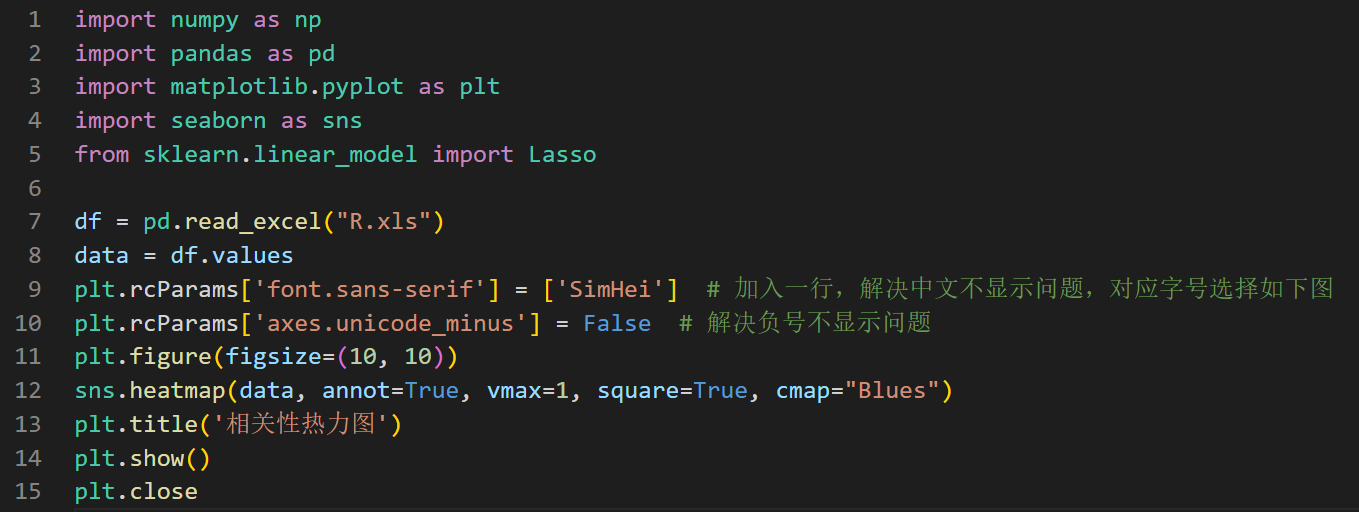
观察可以发现，数据分布较为有规律，最小值、最大值、均值、中位数的数值相对而言合乎常理，说明数据中没有脏数据。

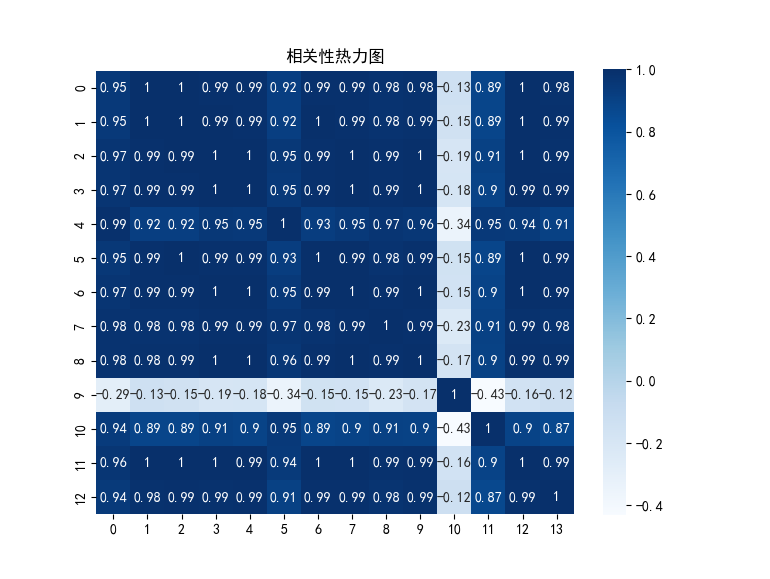
接着，我们对数据进行相关性分析，探索数据的相关性，数据都是数值型数据，计算各列之间的Pearson相关系数，将计算结果导入Excel文件利用Python绘制相关系数热力图，完成特征可视化处理。接着对数据进行雅克-贝拉（JB）检验，探索数据是否显著相关，代码及结果如下：





热力图绘制：

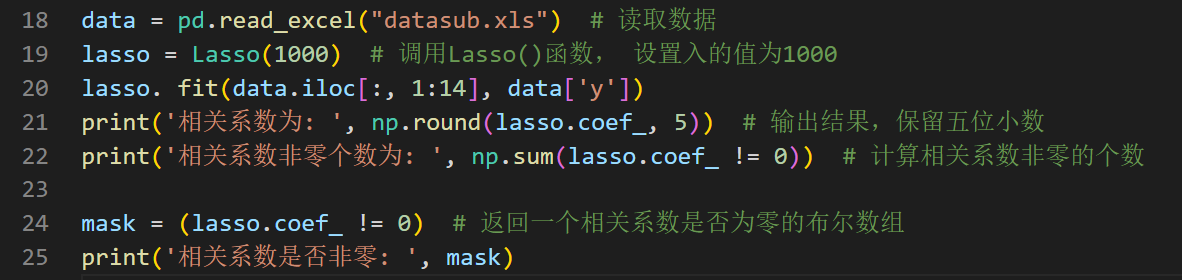


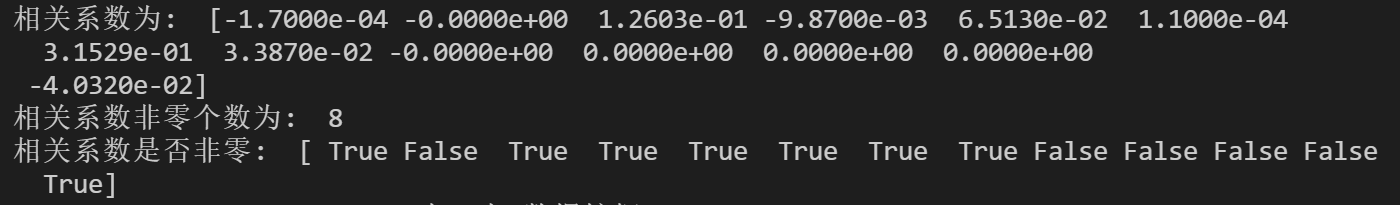


我们可以得出以下结果：除居民消费价格指数与财政收入的线性关系不显著且呈负相关外，其它属性均呈高度正相关。

1. 影响财政收入的关键因素分析

从上面的分析结果我们容易考虑到，有太多相关性较强的属性，很容易产生多重共线性现象使后面的预测产生误差，所以需要把影响最终财政收入的关键因素给筛选出来，这里不妨选用lasso回归对数据进行分析，筛去对最终预测没有帮助的列。代码及结果如下：

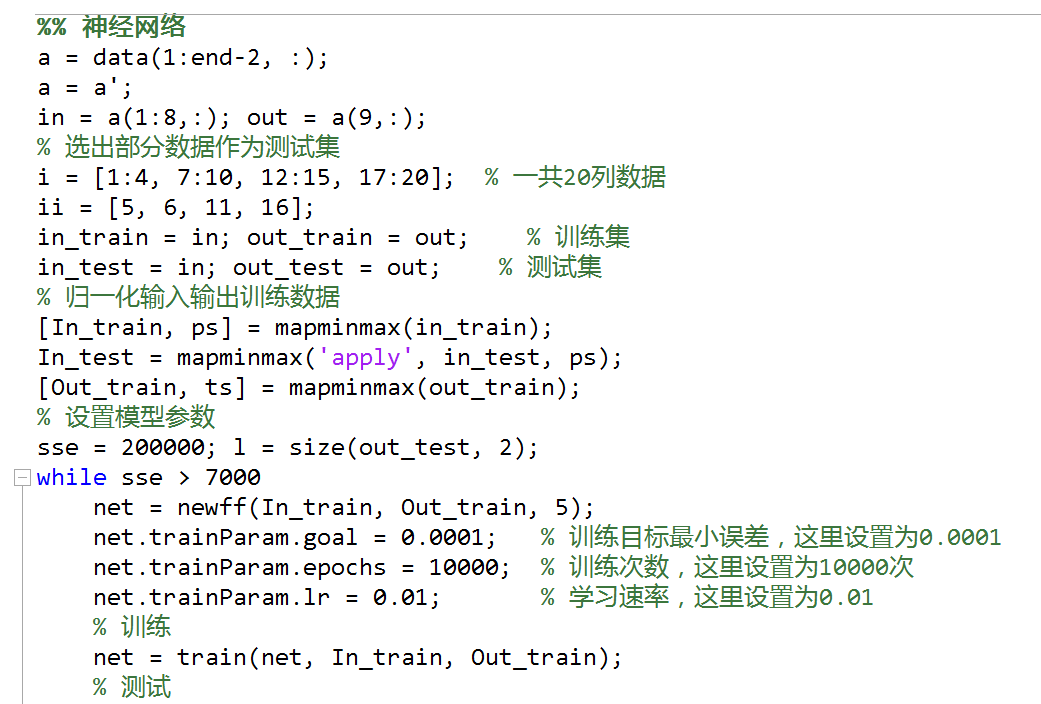




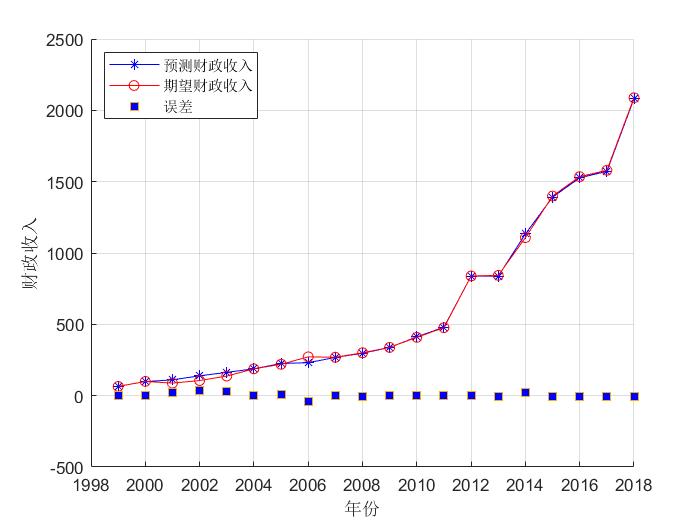
观察结果，有八列的系数值不为零，分别为：社会从业人数、社会消费品零售总额、城镇居民人均可支配收入、城镇居民人均消费性支出、年末总人口、全社会固定资产投资额、地区生产总值、居民消费水平。

1. 有监督模型的构建

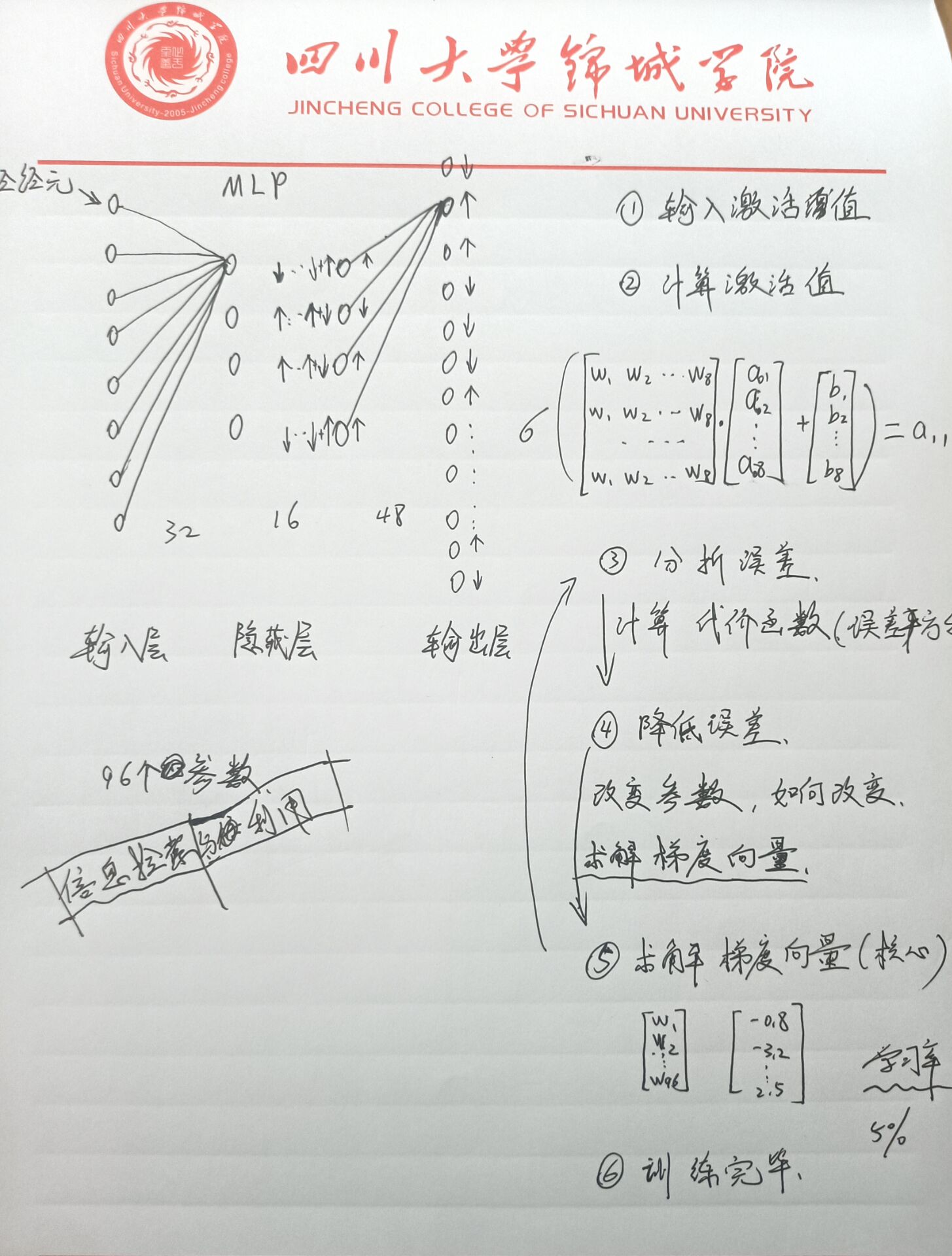
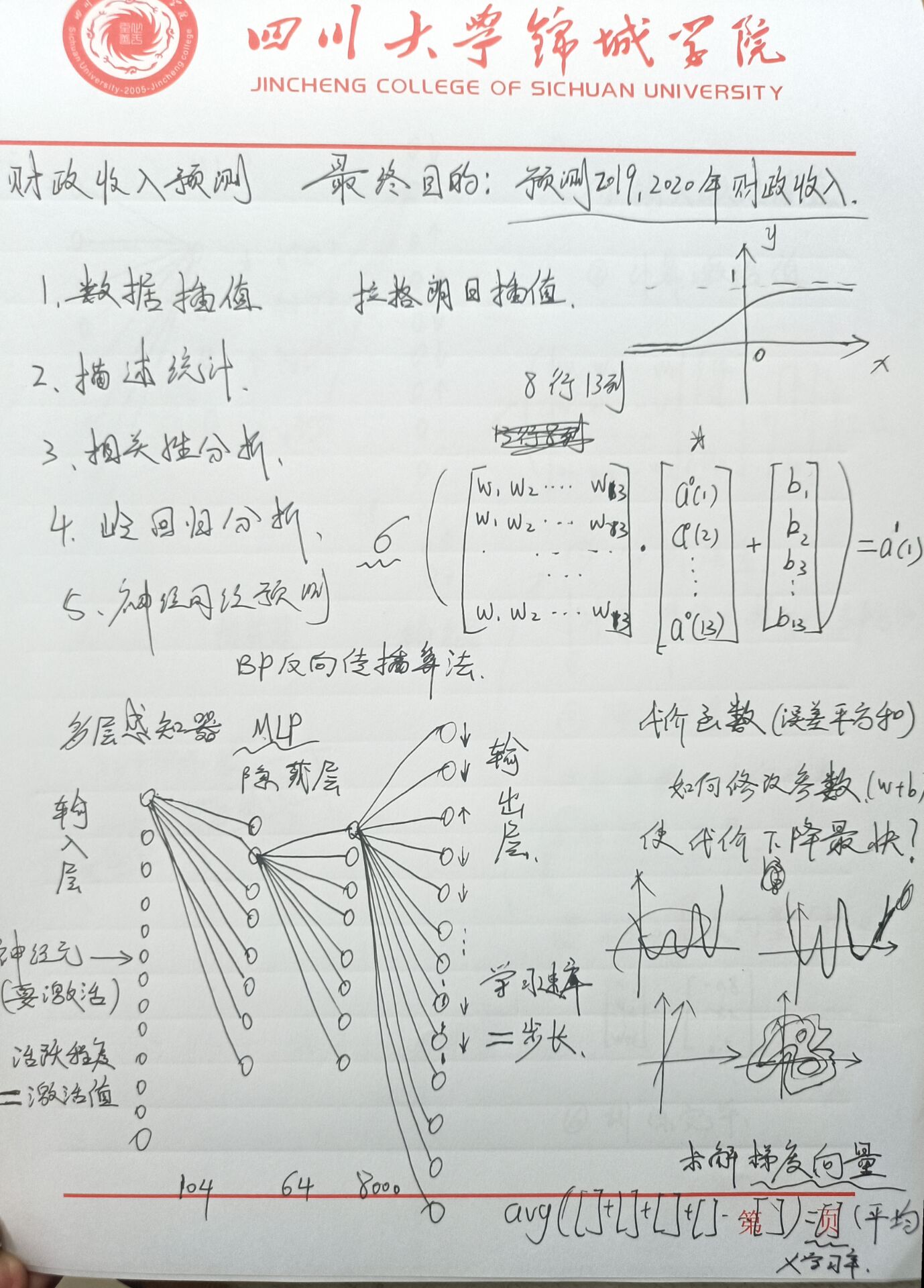
通过我们筛选出的八列数据，构建并训练神经网络模型进行预测。预测代码、结果及部分推导过程如下：



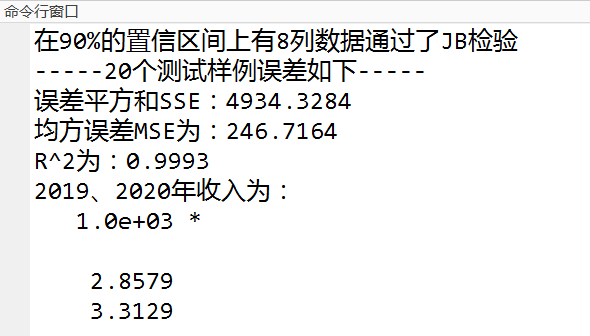




可以观察到误差全部在零刻度线左右，训练效果良好，基本0误差。



1. 模型评估



最后求出的SSE虽然很大，原因并不是因为模型不准确，而是本题目中的数据较大，又是求平方和，所以SSE达到了1e3级别，通过R2可以看出拟合优度为0.9993，拟合优度极好，最终预测结果为2857.9、3312.9，符合常理。

* 项目小结

神经网络是典型的有监督机器学习算法，其构造复杂，有着极强的非线性映射能力，在数据量较大时，神经网络往往有着较好的表象，但在数据较少时，神经网络就会产生误差，在训练神经网络时要注意学习率和学习次数的调节，学习率太低会导致预测不准，学习率太高容易过拟合，导致拟合效果极佳，但预测不准，如何配置隐层数量、学习率、学习算法是一个神经网络模型适用与否的关键。