实验 8_任务 1_多元回归

问题：

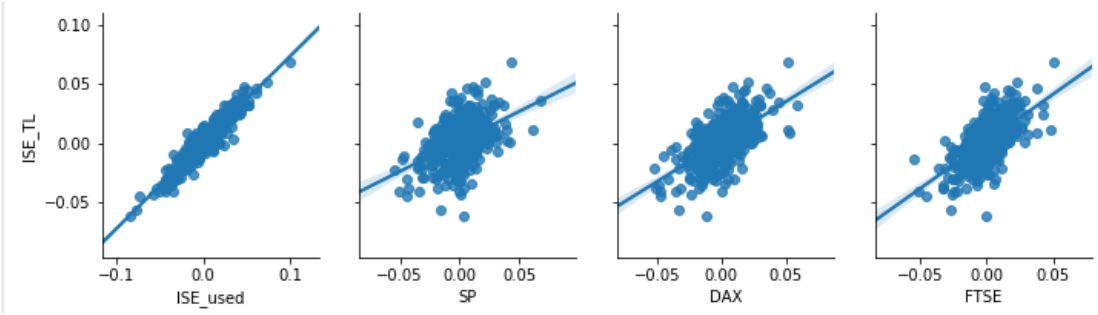
任务描述：在给定的数据文件中，每一行代表一个开盘日中的股指交易涨跌值，第一列记录具体日

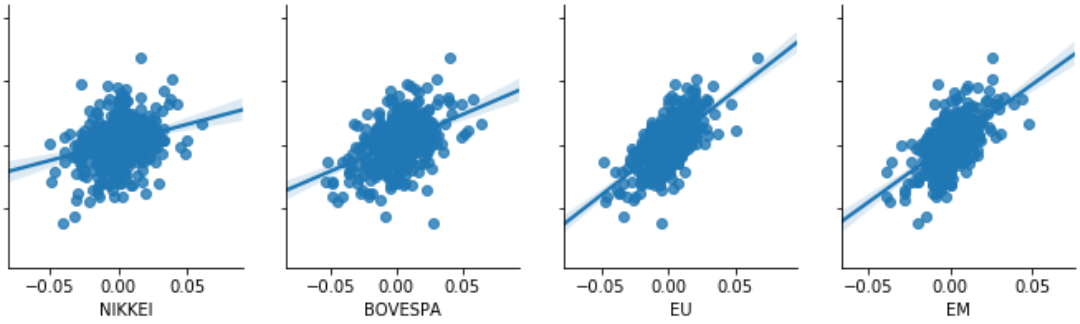
期，其后每一列代表一项股指数据，共九列，依次为：ISE(TL-based), ISE(usd), SP, DAX, FTSE, NIKKEI, BOVESPA, EU, EM。回归任务是通过后八项股指来对第一项股指（ISE(TL-based)）的数值进行预测。请完成下列工作：

1. 使用可视化的方法观察数据之间的关联，推测该数据是否适合进行回归分析/线性回归分析。
2. 使用回归分析的方法（如线性回归）进行回归分析，并与你的推测结果进行对比和思考。（实验过程中请注意评价指标、训练误差、泛化误差、测试数据划分等内容，并记录在实验报告中。）
3. （附加题）尝试使用降维前后的数据表示分别进行回归，并比较回归的结果，思考降维该回归任务的影响。

解答及实验过程记录：

1. 可视化数据分析：





皮尔相关系数：

```
  FutureWarning)
            ISE_TL  ISE_used        SP  ...      BOVESPA        EU        EM
ISE_TL    1.000000  0.942897  0.439489  ...     0.432898  0.655519  0.600295
ISE_used  0.942897  1.000000  0.449561  ...     0.446889  0.690761  0.701954
SP        0.439489  0.449561  1.000000  ...     0.722069  0.687550  0.528243
DAX       0.602081  0.629218  0.685843  ...     0.585791  0.936393  0.665162
FTSE      0.622948  0.648740  0.657673  ...     0.596287  0.948963  0.687543
NIKKEI    0.260052  0.393225  0.131250  ...     0.172752  0.283750  0.547288
BOVESPA   0.432898  0.446889  0.722069  ...     1.000000  0.621704  0.688074
EU        0.655519  0.690761  0.687550  ...     0.621704  1.000000  0.716502
EM        0.600295  0.701954  0.528243  ...     0.688074  0.716502  1.000000
```

问题 1：

通过对进行处理，获得了后 7 列数据分别与第一列数据的关系，对数据进行可视化观察和相关性分析。（利用 seaborn 库进行作图，并且绘制了拟合曲线和 95%置信区间），发现数据量较大，数据点对于除了 NIKKEI 和 BOVESPA 两列数据外的数据，其他数据都可以拟合出一条直线，并且数据在直线的投影分布离散型较大，而且 BOVESPA 和 NIKKEI 与 ISE_TL 的相关性较小，因此，我认为该数据集适合进行多元线性回归。

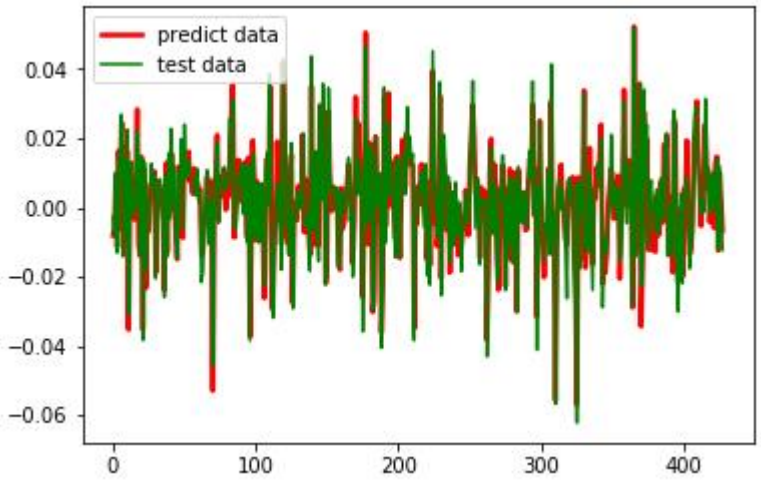
问题 2：

利用了 skilearn 工具对该数据进行了多元线性回归分析。

1. 划分数据集：80%作为训练集，20%作为测试集。
2. 评估方法：MSE 方法，将预测结果与训练集比较，利用如下公式（MSE）进行评估.

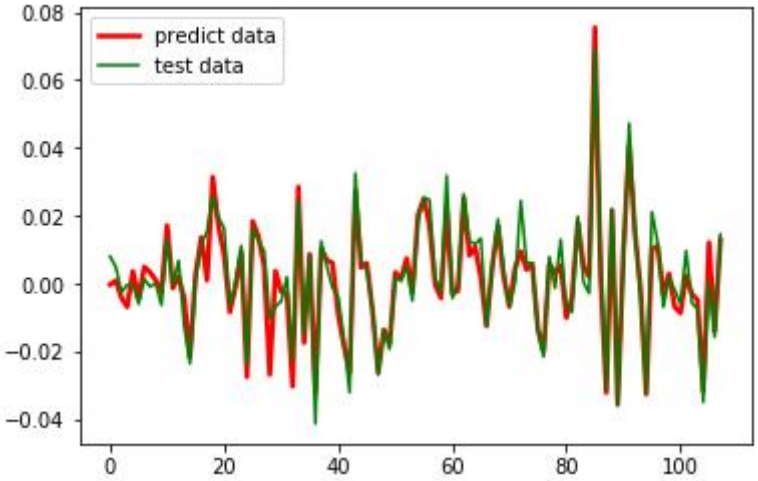
$$\frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}}[\,(f(xi)-y)(f(xi)-y)]$$

3. 训练误差：



MSE: 2.489414504745773e-05

4. 泛化误差：



MSE: 2.3850799709279053e-05