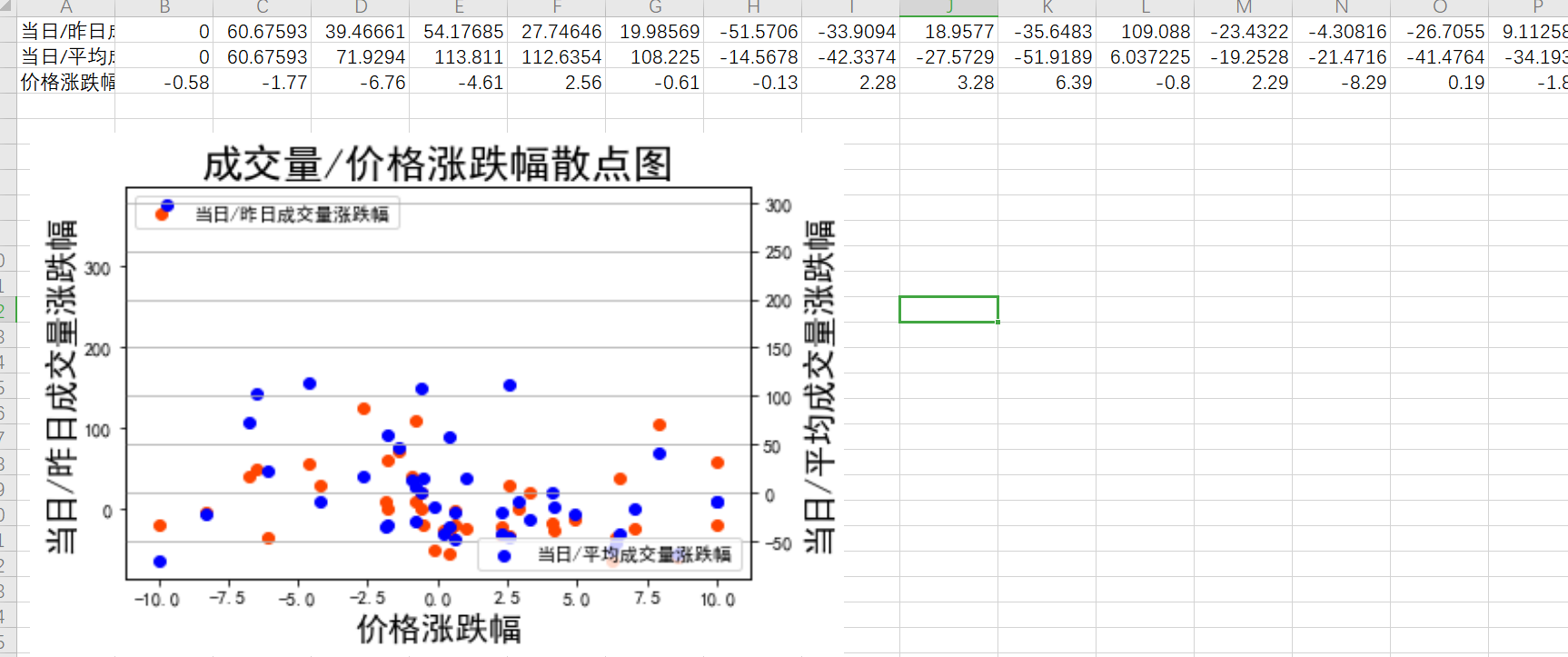
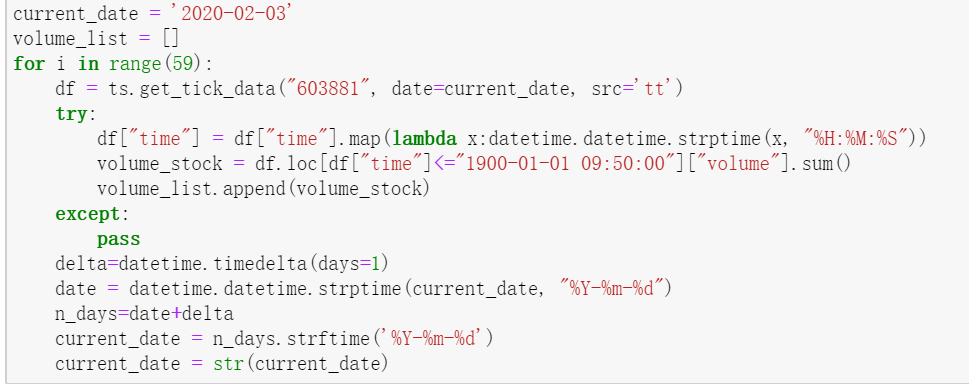
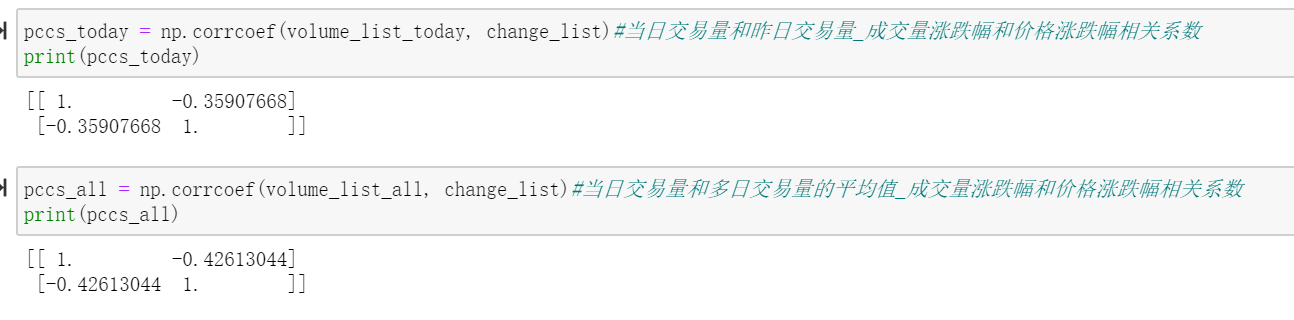
**Q1：**



题目一主要考察了Tushare和可视化库的运用，并通过xlwings存储到EXCEL中，同时考验了皮尔逊相关系数。如图，我已经将成交量（前20分钟）的两种变化率表达式求得的结果数据呈现，然后和价格涨跌幅整体形成了一个散点图，根据题目要求做了双Y轴。

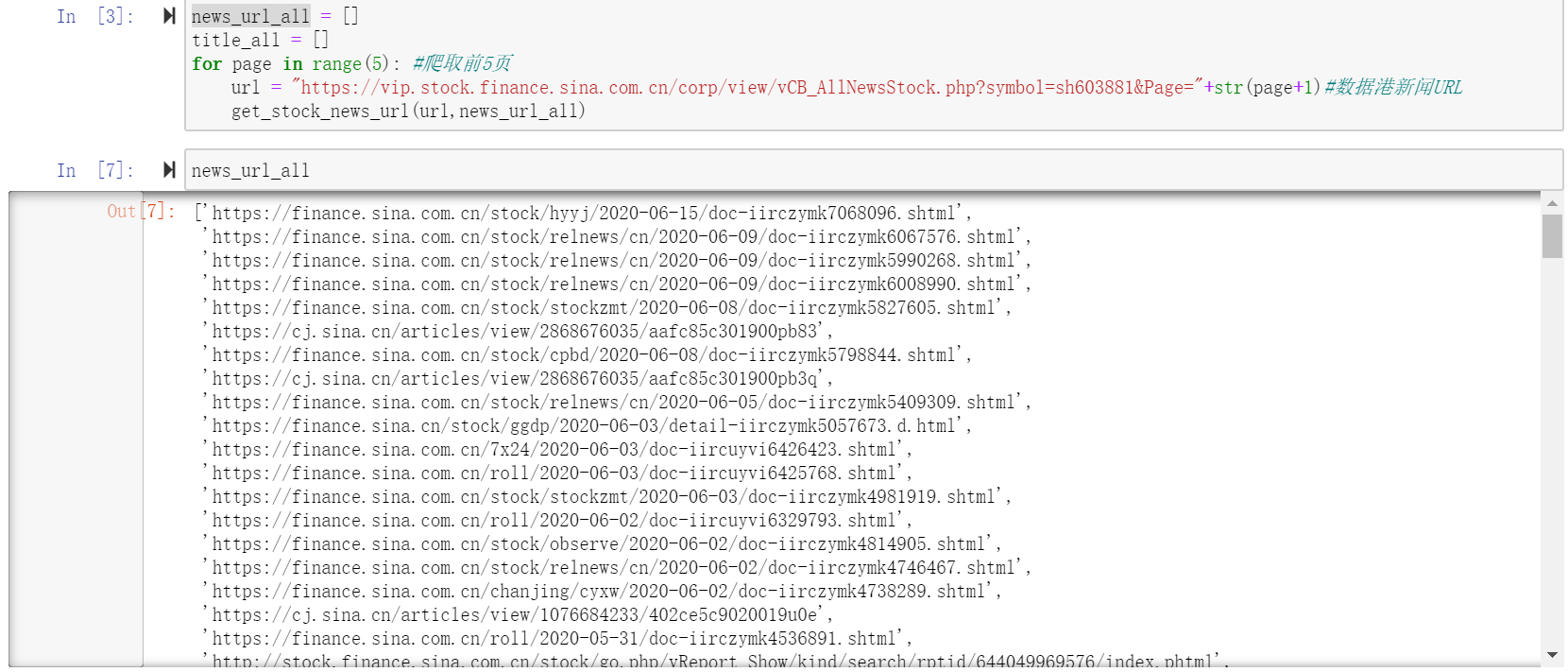
数据调取阶段，我主要采用分笔数据的加总计算方式，采用时间日期的加减等计算，同时在计算涨跌幅我采用了函数表达式计算，最终以列表的方式存储数据。

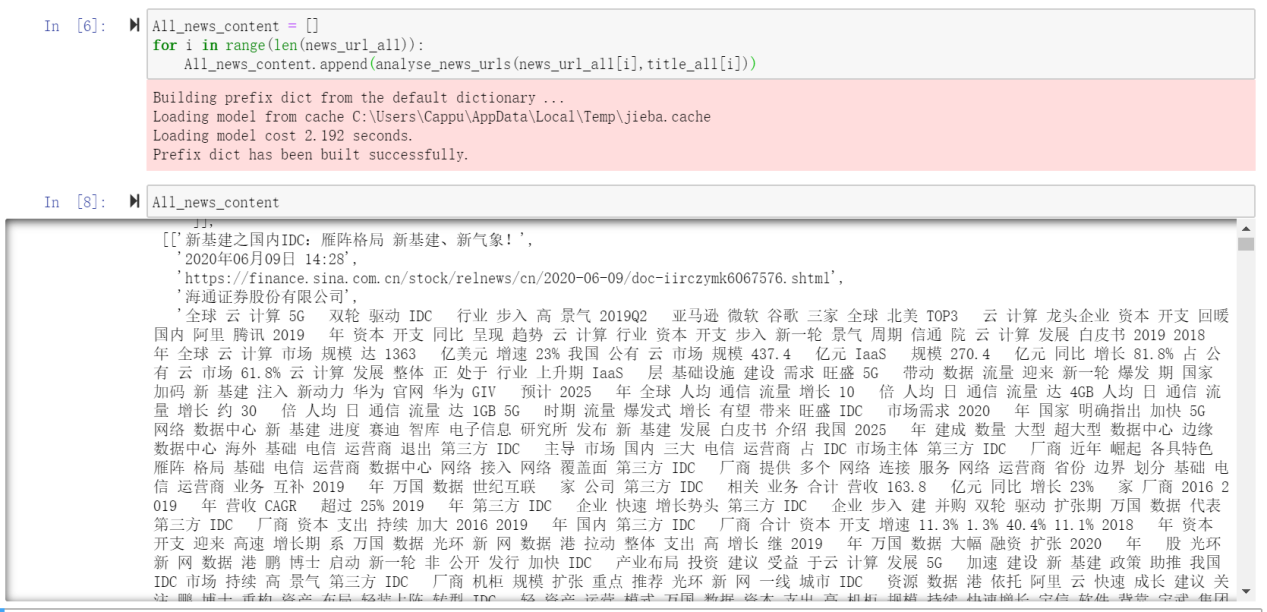
相关系数约为-50%左右，因此相关性并不高，而且是负相关的作用，我猜测是由于砸盘严重而封盘的时候交易氛围较弱的情况，即多个涨停时，成交量反而低迷，但是跌的时候多头和空头竞争激烈。

**Q2：**

问题二主要要求我们采用requests框架和etree正则表达式解析的方法爬取A（数据港）上市公司的新闻数据，并采用jieba做停用词处理（数据清洗）。

首先我们采用传入headers参数的requests.get()函数访问页面，并设置Range(5)遍历五页，采用etree整理网页源代码后，运用正则表达式和xpath定位我们需要的URL，即herf属性然后爬取。

获取所有新闻的URL之后，如法炮制，一样采用requests.get()访问这些URL，我们解析页面发现，可以获取其时间、来源和正文内容，正文内容采用jieba的停用词处理。最后，我们将内容存储到txt文件中。

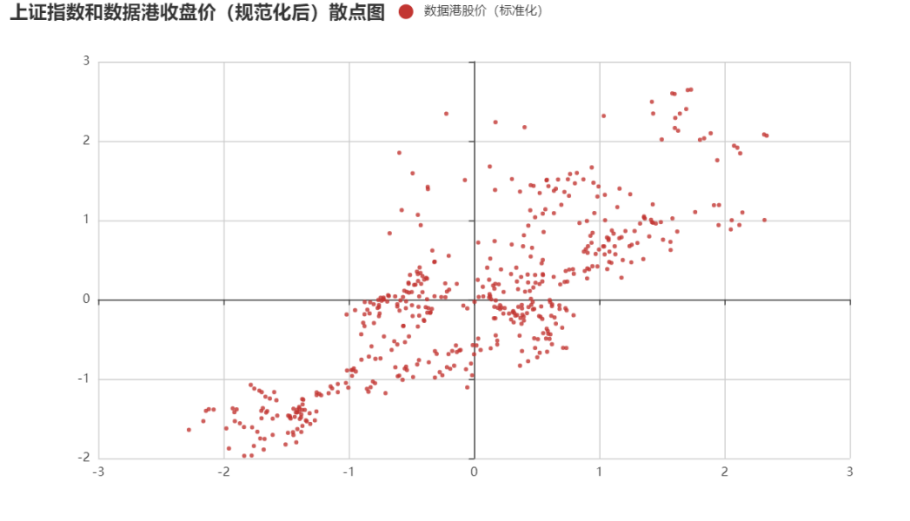
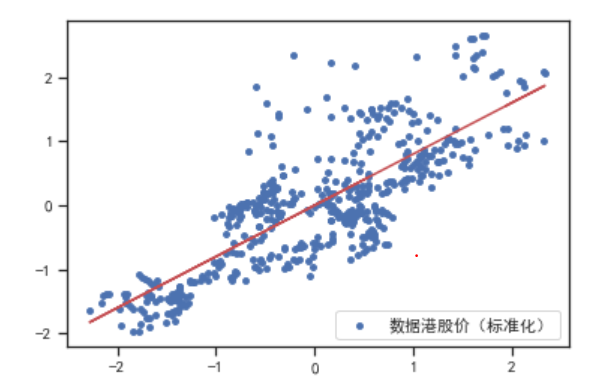
**Q3：**

问题三主要采用的是selenium自动化框架的运用，通过PyQuery库解析爬取下来的新闻网页提取有价值的信息。selenium的好处是反反爬能力较强，可以在Jupyter Notebook中持续运行，帮助爬虫代码和解析的高效编写。

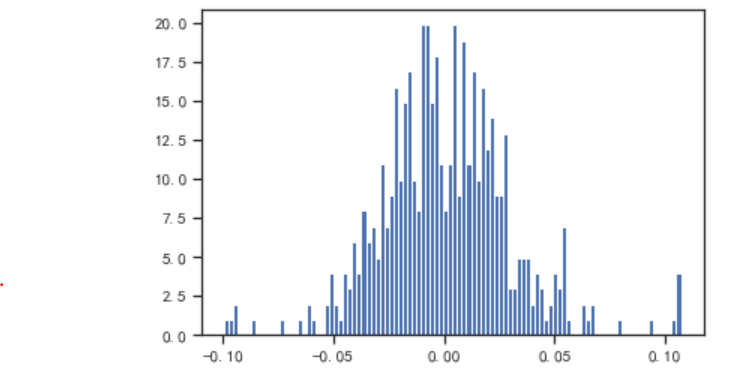
问题三主要解决思想是selenium爬取各页面的URL并循环点击进入新的Webdriver，并通过正则表达四获取新闻标题，简文，时间，来源和原本的URL。

**Q4：**

我选取了数据港的股票和上证指数做分析，采用Tushare获取了股票价格数据，随后规范化（即采用Z得分进行标准化），最终绘制了线图和散点图。



随后我计算了数据港股价和上证指数的对数收益率，首先我采用正态分布图的方式做了一些观察。如图所示：

看起来基本符合正态分布，但是还需要进行检验：

检验方法一：scipy的kstest检验，检验其P值，即显著性水平，实验结果如下：

KstestResult(statistic=0.4617910306496136,pvalue=1.4142187416242526e-95)

KstestResult(statistic=0.48333148392122344,pvalue=3.0688218572514387e-105)

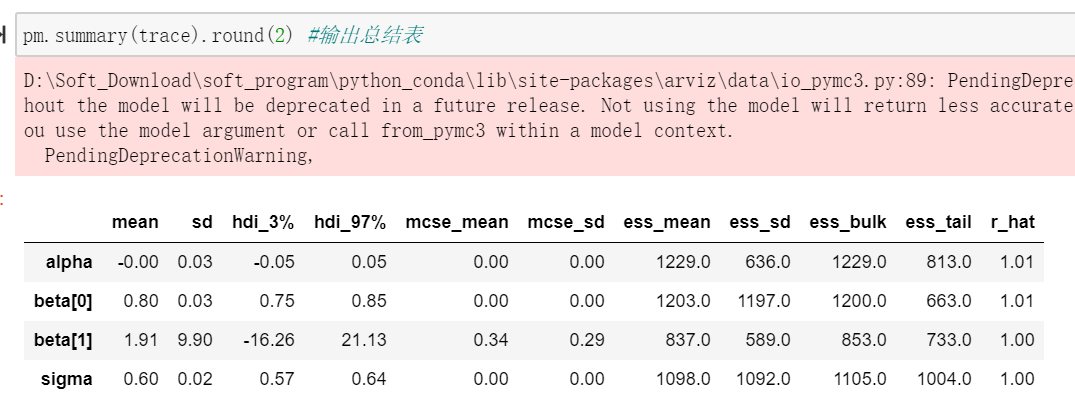
P>0.05符合正态分布，所以两个对数收益率均符合正态分布。

检验方法二：stats的shapiro检验其统计值。

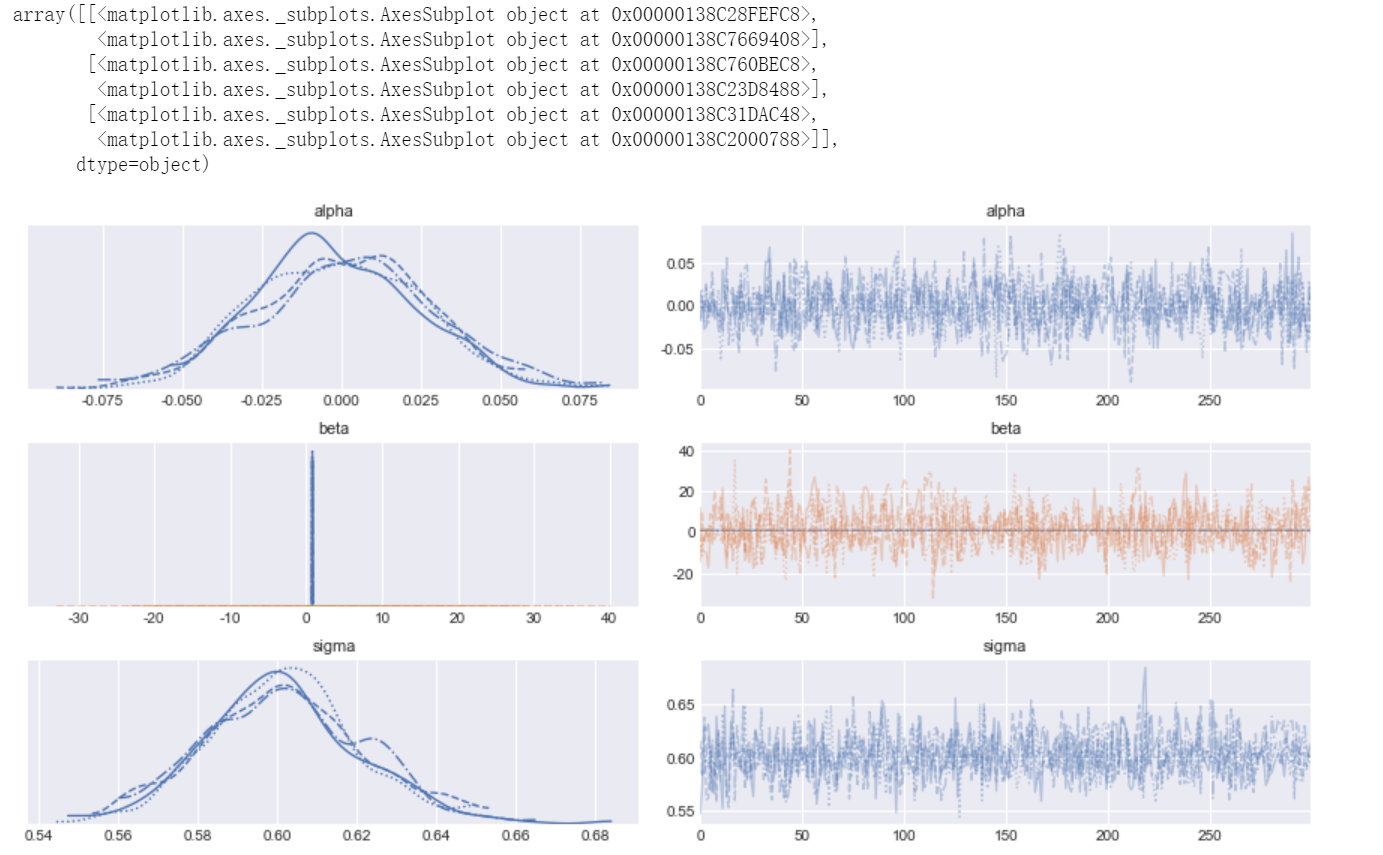
(0.9773707985877991, 7.305965823434235e-07) 统计量接近1，符合正态分布且拟合效果好(0.9576665163040161, 1.305267005591304e-10) 统计量接近1，符合正态分布且拟合效果好，所以两个对数收益率均符合正态分布。

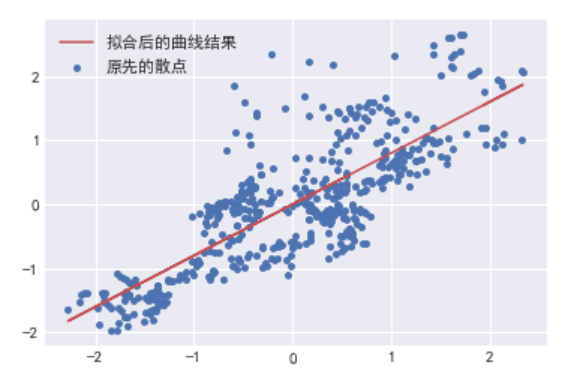
随后采用pymc3模型的贝叶斯统计：

主要采用统计概率的回归方程，以数据港规范化股价预测上证指数的规范化股价。输出的统计表如下：



后验概率指的是建立好方程后我们需要对结果做一定的检验：



后验概率检验良好，基本处于正态分布范围，因此我们认可其贝叶斯方程的统计学意义，绘制了散点-折线图：