**数学实验 exp9 实验报告**

计65 赖金霖 2016011377

**5**

取α=0.05。

（1）

选择x1和x2，计算得s=4.6484823252393985，且系数的置信区间均不包含零点；

选择x1和x3，计算得s=5.622452826419641，且x3的系数置信区间包含零点；

选择x2和x3，计算的s=5.0408331181362，且x3的系数置信区间包含零点。

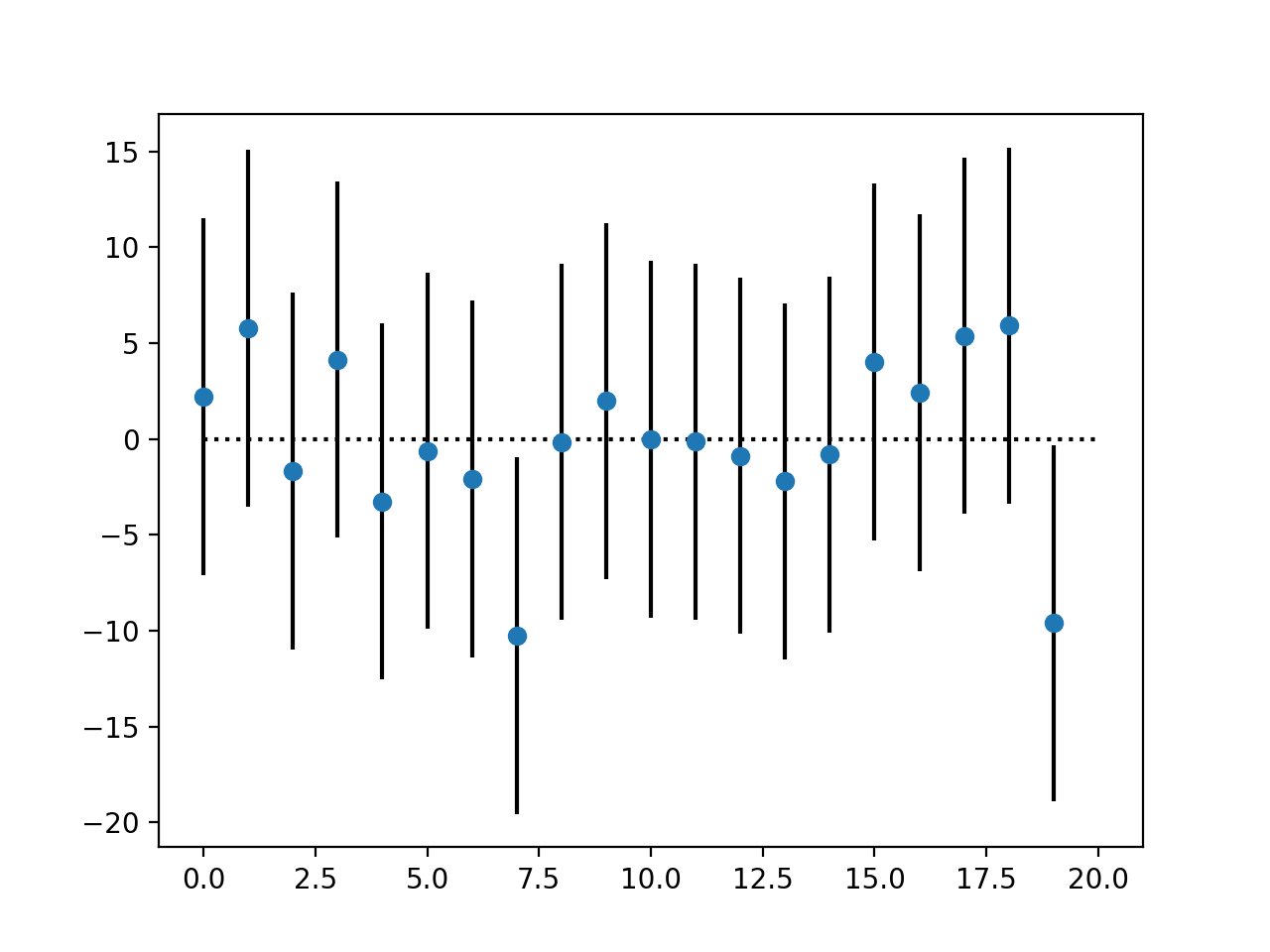
综上可知，选择x1和x2作为变量，是三者中最好的模型。

（2）

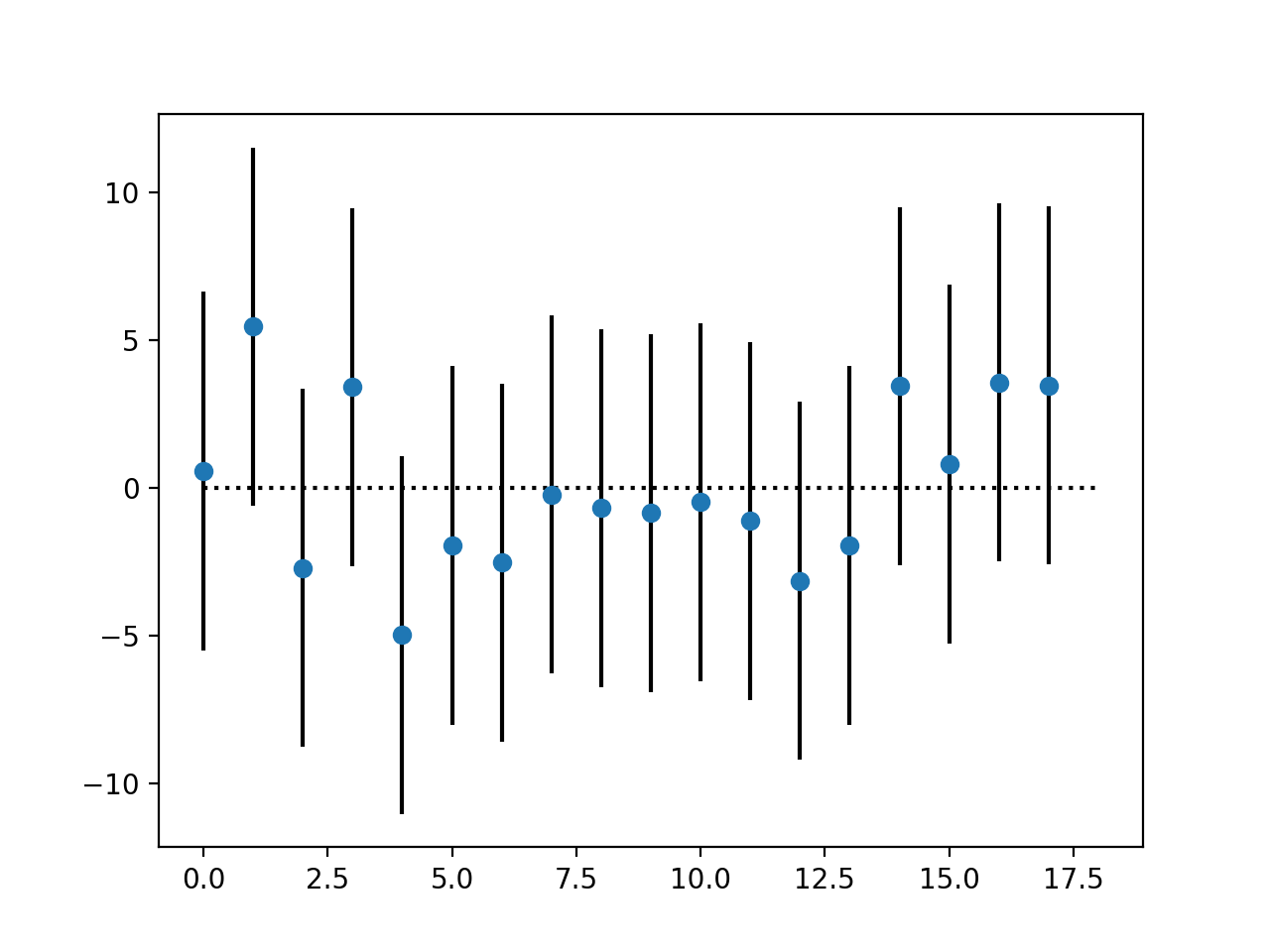
选择x1、x2、x3作为自变量，进行线性回归，计算得 s=4.5897784444326195，比单纯x1和x2要小。然而，x3的系数的置信区间为[-0.00057, 0.00210]，它的引入看上去不合理，减小的s可能是因为过拟合导致的。

所以，最好的模型还是只选择x1和x2作为自变量。

（3）

在这一模型下，残差如下（以4s为区间宽度）：

可以看出，第8个和第20个数据是异常点，取出之后，残差如下：

最终模型为y=β0+β1x1+β2x2，其中β0，β1，β2如下：

|  |  |  |
| --- | --- | --- |
| 系数 | 估计值 | 置信区间 |
| β0 | -35.709 | [-45.212, -26.207] |
| β1 | 1.602 | [0.782, 2.423] |
| β2 | 3.393 | [1.228, 5.557] |

**10**

（1）

取α=0.05。

仅以x1和x2作为自变量进行线性回归，得s= 9.251641618339645。

由于有把握认为y与x2之间有线性关系，所以可以固定x2的次数，增加x1的次数，看s的变化情况：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x1最高次 | 1 | 2 | 3 | 4 |
| s | 9.252 | **1.803** | 1.628 | 1.671 |
| 备注 | 系数均可信 | | 大于2次项系数不可信 | |

当最高次数从1变为2时，s的值剧烈减少，所以可以认为y与x1有2次关系。

（2）

以x1，x12，x2，x1\*x2为自变量进行线性回归，得s=1.803，x1\*x2的系数不可信；以x1，x12，x2，x12\*x2为自变量进行线性回归，得s=1.791，x12\*x2的系数不可信。

故可以认为x1与x2之间没有交互效应。

最终模型为y=β0+β1x1+β2x12+β3x2，其中β0，β1，β2，β3如下：

|  |  |  |
| --- | --- | --- |
| 系数 | 估计值 | 置信区间 |
| β0 | -62.349 | [-73.373, -51.324] |
| β1 | 0.840 | [0.400, 1.279] |
| β2 | 0.037 | [0.033, 0.041] |
| β3 | 5.685 | [5.265, 6.104] |

**11**

由于python没有对应工具，所以只能手动调参。

取α=0.05。

1. 测试各变量次数。

分别以x1;x1和x12；x1和x12和x13位自变量，进行线性回归。

|  |  |  |  |
| --- | --- | --- | --- |
| x1最高次 | 1 | 2 | 3 |
| s | 7.947 | **7.027** | 7.198 |
| 备注 | 系数均可信 | | 系数不可信 |

分别以x2;x2和x22；x2和x22和x23位自变量，进行线性回归。

|  |  |  |  |
| --- | --- | --- | --- |
| x1最高次 | 1 | 2 | 3 |
| s | 14.599 | - | - |
| 备注 | 系数不可信 | S0TS0矩阵奇异 | |

分别以x3;x3和x32；x3和x32和x33位自变量，进行线性回归。

|  |  |  |  |
| --- | --- | --- | --- |
| x1最高次 | 1 | 2 | 3 |
| s | 14.892 | 15.209 | 15.585 |
| 备注 | 系数不可信 | | |

故暂时以x1和x12为基础变量（s值已加粗）。

1. 测试相关性（策略为在当前变量和一次变量上乘其他变量）

加入x1\*x2，得s=6.298，系数均可信，保留。

加入x1\*x3，得s=5.953，系数不可信，舍弃。

加入x2\*x3，得s=6.400，系数不可信，舍弃。

加入x1\*x1\*x2，得s=6.416，系数不可信，舍弃。

加入x1\*x1\*x3，得s=5.532，系数可信，保留。

1. 迭代尝试舍弃的变量（由于是手动做的，不一定符合stepwise策略）

加入x2，得s=5.683，系数不可信，舍弃。

加入x3，得s=**4.313**，系数可信，保留。

加入x3\*x3，得s=4.294，系数不可信，舍弃。

加入x1\*x3，得s=4.263，系数不可信，舍弃。

加入x2\*x3，得s=4.348，系数不可信，舍弃。

加入x1\*x1\*x2，得s=4.364，系数不可信，舍弃。

加入x2，得s=4.437，系数不可信，舍弃。

结束（最终s=4.313）。

最终模型为y=β0+β1x1+β2x12+β3x3+β4x1\*x2+β5x12x3，其中β0，β1，β2，β3，β4，β5如下：

|  |  |  |
| --- | --- | --- |
| 系数 | 估计值 | 置信区间 |
| β0 | 52.680 | [42.041, 63.320] |
| β1 | -10.748 | [-13.748, -7.748] |
| β2 | 0.810 | [0.536, 1.083] |
| β3 | 25.064 | [10.791, 39.338] |
| β4 | 0.955 | [0.408, 1.502] |
| β5 | -0.597 | [42.041, 63.320] |

**阅读报告**

本文的主要内容是应用多种统计模型对1994~1998中国股票数据进行拟合和预测。方法主要参考了Ou and Penman和Abarbanell and Bushee两篇工作，简记为OP和AB。

 论文首先采用OP方法，使用37个财务变量对EPS是否增加进行logistic回归。模型如下：

而由于每年的资本环境不同，所以θ会随年份改变。这五年一共得到4个模型，在α=0.05时，这些模型都是有效的(p值可信)。

 采用OP方法仅能定性分析，论文进而通过AB方法尝试定量分析。AB方法通过11个“基本信号”对EPS的增量进行分析：

实验显示，这个模型在七年间的三组数据上均有效，准确率在85%、90%、95%的置信区间内都达到85%以上。这11个“基本信号”里，除3和5外，都至少被选择一次，验证了这些变量的预测能力。

论文将针对美国股市的分析方法引入国内，验证了国内公司财务报表对股市表现的预测性。值得思考的是，部分公司是否会学来这套方法，调整报表使得有更好的预测值？此外，我认为文章创新点不够多，OP方法的37个变量和AB方法的11个变量都是从其他论文中挑出来的。如果能针对中国股市选择一些更适合的新变量，可能会得到更好的结果。

代码可在<https://github.com/lll6924/math_exp/tree/master/exp9>下找到。

吐槽：本次作业对python选手过于不友好，已做好挂科准备。