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。

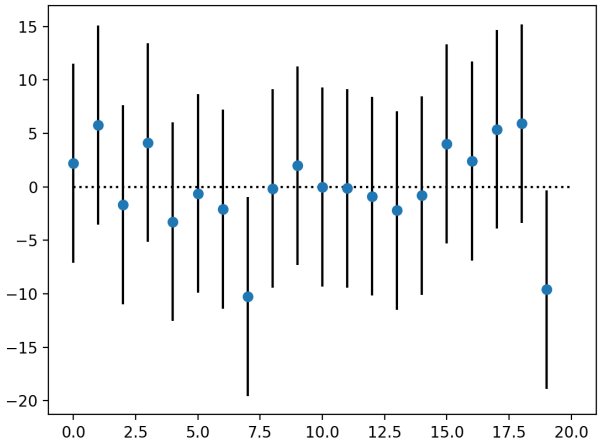
（2）

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。

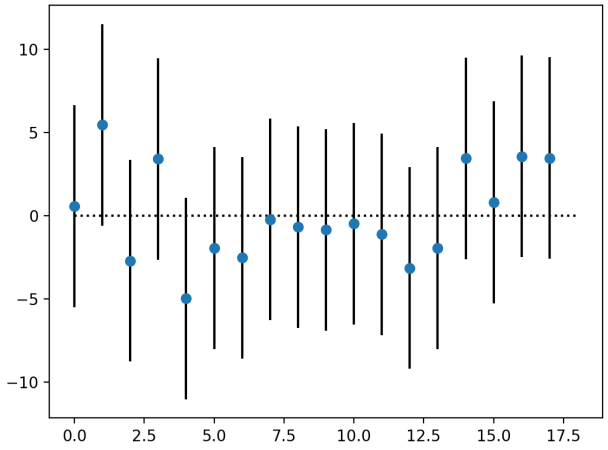
所以，最好的模型还是只选择 x1 和 x2 作为自变量。

（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 $y=\beta_0+\beta_1x_1+\beta_2x_2$，其中 β_0，β_1，β_2 如下：

系数	估计值	置信区间
β_0	-35.709	[-45.212, -26.207]
β_1	1.602	[0.782, 2.423]
β_2	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，$x1^2$，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，$x1^2$，x2，$x1^2$*x2 为自变量进行线性回归，得 s=1.791，$x1^2$*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 $y=\beta_0+\beta_1 x_1+\beta_2 x_1^2+\beta_3 x_2$，其中 β_0，β_1，β_2，β_3 如下：

系数	估计值	置信区间
β_0	-62.349	[-73.373, -51.324]
β_1	0.840	[0.400, 1.279]
β_2	0.037	[0.033, 0.041]
β_3	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 $\alpha=0.05$。

（1） 测试各变量次数。

分别以 x1;x1 和 $x1^2$；x1 和 $x1^2$ 和 $x1^3$ 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 $x2^2$；x2 和 $x2^2$ 和 $x2^3$ 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	$S_0^T S_0$ 矩阵奇异	

分别以 x3;x3 和 $x3^2$；x3 和 $x3^2$ 和 $x3^3$ 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 $x1^2$ 为基础变量（s 值已加粗）。

（2） 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。

（3） 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 $y=\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_3+\beta_4x_1*x_2+\beta_5x_1^2x_3$，其中 β_0， β_1， β_2， β_3， β_4， β_5如下：

系数	估计值	置信区间
β_0	52.680	[42.041, 63.320]
β_1	-10.748	[-13.748, -7.748]
β_2	0.810	[0.536, 1.083]
β_3	25.064	[10.791, 39.338]
β_4	0.955	[0.408, 1.502]
β_5	-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

$$y = log\frac{P}{1-P} = \theta^T \mathbf{X}$$

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

$$y = \triangle EPS = b_0 + \sum_{j=1}^{11} b_i S_i$$

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。