

上市公司股价异动的数据建模诊断

刘天

(东北财经大学金融学院, 辽宁大连, 116025)

摘要: 上市公司股价异动为证券市场监管层和投资者所重点关注, 因其与证券市场的秩序和投资者的切身利益密切相关。以中国证监会公布的三起违规案例为样本, 以股价为响应变量, 以市盈率和涨跌幅绝对值为自变量建立回归模型。利用数据建模诊断方法, 根据学生化残差、杠杆值、Cook 距离、马氏距离等诊断统计量, 对股价是否存在异动进行检测, 并进行综合交叉印证, 确定重点怀疑数据, 然后, 删除这些可疑数据, 再将删除前后表征模型优劣的若干个指标的变化情况进行比对。实证研究表明, 三起查处案例中交易异常行为都较好地得到定位, 与实际结果相符较好, 这对于规范证券市场健康发展及保护投资者合法权益有积极意义。

关键词: 证券市场; 上市公司; 股价异动; 数据建模诊断

中图分类号: F832.5

文献标识码: A

文章编号: 1672-3104(2013)06-0071-08

上市公司股价异动是证券市场监管层与投资者都很关心的问题, 在这方面已有学者做出了一些探索, 如, 肖淑芳、李阳^[1]运用事件分析法, 对重大信息披露与股价异动的相关性进行了研究。史永东、蒋贤峰^[2]以 Logistic 模型为分析工具, 建立了违法违规行为的判别模型。Thierry Ane, Loredana Ureche Rangau 等^[3]采用稳健统计方法对亚太股市指数收益中异常点做了检测分析。曾伟^[4]运用资产定价回归模型的拟合系数来捕捉股价波动的同步性, 研究了上市公司质量与市场波动性的关系。瞿宝忠、徐启帆^[5]利用残差系数法来研究重大并购事件首次公告之前股价的异常波动。Aurea Grane, Helena Veiga^[6]基于小波变换技术对金融时间序列中的异常点进行了检测研究。

本文试图从数据建模诊断的角度, 对这个问题进行探讨。

我们知道, 通过数据建立模型来对经济现象进行分析时, 我们对数据本身是做了很多严格的假设条件的, 只有这些条件真正满足时, 由此得到的模型及其以后基于此所做的推断和结论才是可靠的, 否则就值得怀疑。对于数据本身, 我们经常假定数据是均匀同质的, 即, 假定数据集中每一个点对建模的影响是基本相同的, 每个点对建模都有影响, 但都很微小, 单独一个或若干个点对模型的总体变化趋势产生决定性的影响。而实际中, 这个条件往往不能得到满足。一个数据集中, 经常会有那么一个或几个“不安

分”的点, 它们经常基于现有建模手段的“漏洞”来“兴风作浪”, 它们就是数据集中的异常点, 杠杆点及强影响点。

本文就是从这个角度, 来寻找对建模有“不同寻常”影响的点, 从而在数据集中发现这些“异动点”。那么, 什么是异常点、杠杆点、强影响点呢? 一般来讲, 异常点是指那些与既定模型有较大偏离的数据点, 杠杆点是指那些远离数据主体的点, 强影响点是指对统计推断影响特别大的点。为了能检测出这些点, 我们需要了解几个重要的诊断统计量。

我们知道, 线性回归模型可表示为

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i, i=1, 2, \cdots, n$$

其中: y_i 为因变量; $x_{i1}, \cdots, x_{i(p-1)}$ 为自变量; ε_i 为随机误差; 其第 i 组观察值为 $(y_i, x_{i1}, \cdots, x_{i(p-1)})$ 。通常可表示为矩阵形式如下:

$$Y = X\beta + \varepsilon \quad (1)$$

其中: $Y = (y_1, \cdots, y_n)^T$, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$, $\beta = (\beta_0, \beta_1, \cdots, \beta_{p-1})^T$, X 为 $n \times p$ 阶列满秩矩阵, 其第 i 行为 $(1, x_{i1}, \cdots, x_{i(p-1)})$, 对于随机误差项 ε , 通常假定其分量 $\varepsilon_1, \cdots, \varepsilon_n$ 相互独立, 数学期望为零, 方差具有齐性, 即 $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma^2 I$, 其中 σ^2 为未知常数, I 为 n 阶单位矩阵, 可记为

$$\varepsilon \sim (0, \sigma^2 I) \quad (2)$$

在多数情况下还假定 ε 服从标准正态分布, 即

$$\varepsilon \sim N(0, \sigma^2 I) \quad (3)$$

通常的线性回归, 大多采用了这些假设。这里有一个值得注意的重要问题, 即给定的数据集

$$(y_i, x_{i1}, \dots, x_{i(p-1)}), i=1, 2, \dots, n,$$

是否符合关于模型的假定(1)(2)或(3)式?

现考虑回归分析中常用的投影阵, 在模型(1)式中, X 的投影阵常记为 P , 并记为 $Q=I-P$, Q 为 X 的正交补空间的投影阵, I 为单位阵。由于 P 作用到 Y 上可以得到拟合值 $\hat{Y}=PY$, 因此有些统计学家也称这种特定的投影阵为帽子矩阵(hat matrix)。

在(1)式中, 把 X 的列向量记为 $I=(1, \dots, 1)^T$, 矩阵 X 可写成分块形式如下:

$$X = (I, \tilde{X})$$

$$\tilde{X} = (X_1, \dots, X_{p-1})$$

由于 $P_1=11^T/n$, $Q_1=(I-11^T/n)$, 由二次投影公式可知, 帽子矩阵 P 可表示为

$$P = \frac{J}{n} + X_c(X_c^T X_c)^{-1} X_c^T$$

其中, $J=11^T$, $X_c = (I - \frac{1}{n}J)\tilde{X}$ 。

X_c 称为矩阵 $\tilde{X}$ 的中心化, 它在 (i, j) 处的元素 x_{cij} 为

$$x_{cij} = x_{ij} - \bar{x}_{.j} \quad \bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (4)$$

现记帽子矩阵的元素为 p_{ij} , 则 $P=(p_{ij})$ 具有以下性质:

$$p_{ij} = \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_j;$$

$$p_{ij} = \frac{1}{n} + (\mathbf{x}_{ci})^T (X_c^T X_c)^{-1} (\mathbf{x}_{cj}) \quad (5)$$

帽子矩阵 P 的对角元素 p_{ii} 在回归诊断中起着十分重要的作用。矩阵 $\tilde{X}$ 就是观测矩阵 X 去掉已知的第一个向量 1 而得到。 $\tilde{X}$ 的每一行就是自变量的一组数据, 现记

$$\tilde{\mathbf{x}}_i = (x_{ij}, \dots, x_{i(p-1)})^T, i=1, \dots, n$$

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_i$$

则由(4)可知, (5)式可写为

$$p_{ii} = \frac{1}{n} + (\tilde{\mathbf{x}}_i - \bar{\mathbf{x}})^T (X_c^T X_c)^{-1} (\tilde{\mathbf{x}}_i - \bar{\mathbf{x}}) \quad (6)$$

其中第二项表示点 $\tilde{\mathbf{x}}_i$ 到 $\bar{\mathbf{x}}$ 的一种距离(关于矩阵 $(X_c^T X_c)^{-1}$ 的距离), 通常称为马氏距离(Mahalanobis distance), 平均值 $\bar{\mathbf{x}}$ 可看作数据的中心点, 因此(6)式表明, p_{ii} 越大, 则第 i 组数据点 $\mathbf{x}_i$ 离数据中心越远, 反之, 若 $\mathbf{x}_i$ 离数据中心较远, 则 p_{ii} 比较大。 $p_{ii}=1$ 或 $p_{ii} \approx 1$ 的数据点, 通常称为高杠杆点(high leverage point), 这种点对建模有很大影响, 在模型诊断分析中应引起特别的注意, p_{ii} 也称为杠杆值(leverage value)。模型(1)式的学生化残差定义为

$$r_i = \frac{\hat{e}_i}{\hat{\sigma} \sqrt{1-p_{ii}}} \quad (i=1, \dots, n) \quad (7)$$

其中: $\hat{\sigma}^2 = RSS/(n-p)$, $\hat{e}_i$ 为普通残差, 学生化残差也称学生化内残差(internal studentized residuals), 若取 $\hat{\sigma}^2(i)$ 作 $\hat{\sigma}^2$ 的估计量, 则得到学生化外残差(external studentized residuals), 即,

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}(i) \sqrt{1-p_{ii}}}, i=1, \dots, n \quad (8)$$

为了考察数据集 $(y_i, \mathbf{x}_i^T)$ 与模型的符合情况 ($i=1, \dots, n$), 一个重要的方法就是逐个考察每组数据点的作用, 如数据删除模型。

对于下面模型,

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i=1, \dots, n \quad (9)$$

其中:

$$\mathbf{x}_i = (1, x_{i1}, \dots, x_{i(p-1)})^T$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$$

那么, 删除第 i 组数据点 $(y_i, \mathbf{x}_i^T)$ 后的模型就是数据删除模型, 即,

$$y_j = \mathbf{x}_j^T \boldsymbol{\beta} + \varepsilon_j, j \neq i$$

或

$$Y(i)=X(i)\boldsymbol{\beta}+\varepsilon(i) \quad (10)$$

由线性模型的理论可知, 模型(1)式中参数 $\boldsymbol{\beta}$ 的置信域可表示为以下形式:

$$\frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T X^T X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{p \hat{\sigma}^2} \leq F(p, n-p, 1-\alpha) \quad (11)$$

在参数空间 R^p 中, 它表示一个以 $\hat{\boldsymbol{\beta}}$ 为中心的椭

球, 易见, 落在椭圆以外的 β 点可能性很小, 其概率只有 α 。现考虑 $\hat{\beta}(i)$, 如果落在椭圆之外, 则说明 $\hat{\beta}(i)$ 与 $\hat{\beta}$ 的差异非常大, $(y_i, \mathbf{x}_i^T)$ 作为模型(1)式的点是不可接受的。同理, 若 $\hat{\beta}(i)$ 的值代入(11)式左端所得的值较大, 则说明 $\hat{\beta}(i)$ 离置信域中心 $\hat{\beta}$ 较远, 因而 $\hat{\beta}(i)$ 与 $\hat{\beta}$ 有较大的差异, 从而可以认为 $(y_i, \mathbf{x}_i^T)$ 对模型(1)式的影响也较大, 基于这种考虑 Cook 提出在(11)式左端以 $\hat{\beta}(i)$ 代替 β , 作为度量第 i 个数据点 $(y_i, \mathbf{x}_i^T)$ 影响大小的数量指标, 即给定模型(1)式和(10)式, 第 i 个数据点 $(y_i, \mathbf{x}_i^T)$ 的 Cook 距离定义为

$$D_i = \frac{(\hat{\beta}(i) - \hat{\beta})^T X^T X (\hat{\beta}(i) - \hat{\beta})}{p \hat{\sigma}^2} \quad (12)$$

D_i 也称为 Cook 统计量, Cook 距离表示 $\hat{\beta}(i)$ 与 $\hat{\beta}$ 之间的一种加权距离, 其权重为 $X^T X / p \hat{\sigma}^2$ 。

现考虑数据点 $(y_i, \mathbf{x}_i^T)$ 删除前后对 x_i 处拟合值的影响。其拟合值分别为 $\hat{y}_i = \mathbf{x}_i^T \hat{\beta}$ 和 $\hat{y}_i(i) = \mathbf{x}_i^T \hat{\beta}(i)$, 二者的差值 $\mathbf{x}_i^T (\hat{\beta} - \hat{\beta}(i))$ 可用来度量第 i 个数据点对于拟合带来的影响。另外, 为了消除尺度的影响, 还要除以拟合值的均方误差 $\text{var}(\hat{y})$ 。易见,

$$\text{var}(\hat{y}_i) = \text{va}(\mathbf{x}_i^T \hat{\beta}) = \sigma^2 \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i = \sigma^2 p_{ii}$$

为了研究去掉第 i 个点以后对于拟合的影响, 应由 $\hat{\sigma}^2(i)$ 来代替 σ^2 , 因此, 可得 Welsch-Kuh 距离的定义如下:

给定模型(1)式和(10)式, 第 i 个数据点 $(y_i, \mathbf{x}_i^T)$ 对于拟合值的影响定义为

$$WK_i = \frac{\hat{y}_i - \hat{y}_i(i)}{\hat{\sigma}(i) \sqrt{p_{ii}}} = \frac{\mathbf{x}_i^T (\hat{\beta} - \hat{\beta}(i))}{\hat{\sigma}(i) \sqrt{p_{ii}}} \quad (13)$$

WK_i 称为 Welsch-Kuh 距离, 简称为 W-K 统计量, 也有称之为 DFFITS, DF 表示差异(Difference), FIT 即拟合。

由上述可知, D_i 主要度量了位置参数 β 的估计量 $\hat{\beta}$ 和 $\hat{\beta}(i)$ 之间的差异, 而 WK_i 综合考虑了位置参数和尺度参数之间的差异, 即 WK_i 度量了 $(\hat{\beta}, \hat{\sigma}^2(i))$ 与 $(\hat{\beta}, \hat{\sigma}^2)$ 之间的差异。

我们知道, $\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ 是度量 $\hat{\beta}$ 优良性的统计量, 广义方差 $|\text{var}(\hat{\beta})| = |\sigma^2 (X^T X)^{-1}|$ 也有这个作用, $|A|$ 表示矩阵 A 的行列式。由于 $|\text{var}(\hat{\beta})| \leq |\text{var}(\hat{\beta}(j))| = |\sigma^2 (X^T(j) X(j))^{-1}|$, 可知, 如果 $|\text{var}(\hat{\beta}(j))|$ 的值增加越大, 即, $|X^T(j) X(j)|$ 越小, 则说明 (Y_j, X_j) 对 $\hat{\beta}$ 的影响越大, 所以可定义协方差比作为度量影响的统

计量:

$$\overline{CR_j} = \frac{|\text{var}(\hat{\beta}(j))|}{|\text{var}(\hat{\beta})|} = \frac{|X^T X|}{|X^T(j) X(j)|} \quad (14)$$

$\overline{CR_j} \geq 1$, $\overline{CR_j}$ 越大, 则说明 (Y_j, X_j) 对于 $\hat{\beta}$ 的影响越大。

下面根据中国证监会公布的处罚案例, 选取了中捷股份(002021)、科冕木业(002354)、富临运业(002357)三支股票作为样本, 对其进行建模诊断分析。

首先, 对中捷股份进行研究(见表 1)。

表 1 中捷股份数据

序号	交易日期	收盘价	涨跌幅绝对值(%)	市盈率
1	2009.10.12	5.180	0.380	304.710
2	2009.10.13	5.330	2.900	313.530
3	2009.10.14	5.550	4.130	326.470
4	2009.10.15	5.470	1.440	321.760
5	2009.10.16	5.540	1.280	325.880
6	2009.10.19	5.710	3.070	335.880
7	2009.10.20	5.840	2.280	343.530
8	2009.10.21	5.790	0.860	340.590
9	2009.10.22	5.820	0.520	342.350
10	2009.10.23	5.790	0.520	340.590
11	2009.10.26	5.650	2.420	332.350
12	2009.10.27	6.220	10.090	365.880
13	2009.10.28	6.260	0.640	368.240
14	2009.10.29	6.460	3.190	380.000
15	2009.10.30	6.410	0.770	377.060
16	2009.11.2	6.900	7.640	405.880
17	2009.11.3	6.870	0.430	404.120
18	2009.11.4	6.900	0.440	405.880
19	2009.11.5	7.180	4.060	422.350
20	2009.11.6	7.020	2.230	412.940
21	2009.11.9	6.880	1.990	404.710
22	2009.11.10	6.970	1.310	410.000
23	2009.11.11	6.930	0.570	407.650
24	2009.11.12	6.940	0.140	408.240
25	2009.11.13	6.780	2.310	398.820
26	2009.11.16	7.120	5.010	418.820
27	2009.11.17	7.270	2.110	427.650
28	2009.11.18	7.190	1.100	422.940
29	2009.11.19	7.170	0.280	421.760
30	2009.11.20	7.220	0.700	424.710

根据表 1, 我们以股价(price)为响应变量, 市盈率(ratio)和涨跌幅绝对值(variance)为自变量的作回归, 得到表 2 和表 3。

表 2 完全样本数据下回归结果

项目	贝塔值	标准误差	回归系数	标准误差	t(27)值	P 水平
截距			0.000 013	0.000 084	0.16	0.873 615
Ratio	1.000 000	0.000 013	0.017 000	0.000 000	77 024.50	0.000 000
Variance	0.000 028	0.000 013	0.000 009	0.000 004	2.15	0.040 401
R ²	1.000 000	R [*]	1.000 000	F(2, 27)	2967E6	P<0.000 0
DW	1.962 422	SC	-0.040 803			

表 3 完全数据下主要诊断统计量值

序号	学生化外残差	Cook 距离	杠杆值	马氏距离	WK 统计量	协方差比
1	-1.468 2	0.136 934	0.165 5	3.831 81	-0.653 8	1.056 7
2	-0.519 6	0.013 304	0.123 6	2.616 99	-0.195 1	1.238 8
3	-0.284 9	0.003 550	0.115 2	2.375 28	-0.102 8	1.254 0
4	1.825 5	0.115 043	0.101 8	1.986 06	0.614 6	0.868 3
5	0.899 7	0.028 078	0.094 2	1.764 26	0.290 1	1.127 7
6	0.572 8	0.009 169	0.075 5	1.223 93	0.163 7	1.166 7
7	-0.338 3	0.002 333	0.057 5	0.701 78	-0.083 6	1.172 7
8	-0.516 6	0.007 297	0.072 9	1.147 40	-0.144 9	1.171 5
9	1.296 5	0.045 936	0.077 0	1.264 92	0.374 3	1.005 4
10	-0.455 9	0.006 294	0.079 6	1.341 70	-0.134 1	1.188 0
11	0.909 1	0.023 016	0.076 6	1.255 03	0.261 8	1.104 1
12	-0.871 0	0.226 712	0.464 7	12.511 01	-0.811 6	1.919 3
13	-1.540 9	0.040 169	0.050 6	0.502 12	-0.355 9	0.907 6
14	-0.227 7	0.000 819	0.040 7	0.213 46	-0.046 9	1.160 4
15	-0.221 1	0.002 957	0.046 5	0.381 49	-0.048 8	1.167 9
16	-0.158 5	0.007 747	0.253 9	6.395 95	-0.092 5	1.496 6
17	-0.545 3	0.036 227	0.069 5	1.049 89	-0.149 1	1.163 2
18	1.191 6	0.021 911	0.071 4	1.104 14	0.330 4	1.028 2
19	0.766 2	0.004 002	0.100 6	1.949 48	0.256 2	1.164 4
20	0.421 8	0.037 739	0.060 5	0.788 93	0.107 1	1.167 9
21	-1.507 2	0.000 519	0.049 7	0.474 27	-0.344 6	0.916 6
22	0.158 1	0.016 055	0.061 4	0.813 78	0.040 4	1.189 7
23	-0.783 7	0.055 359	0.070 7	1.083 55	-0.216 2	1.123 5
24	-1.391 2	0.023 361	0.082 2	1.415 85	-0.416 2	0.983 7
25	1.253 9	0.023 361	0.043 4	0.292 45	0.267 1	0.981 7
26	0.814 2	0.031 846	0.124 5	2.643 02	0.307 0	1.186 0
27	-1.050 8	0.035 476	0.087 5	1.571 82	-0.325 5	1.083 3
28	0.654 3	0.013 932	0.085 9	1.524 97	0.200 6	1.166 6
29	2.304 1	0.170 700	0.100 4	1.943 78	0.769 6	0.712 9
30	-1.247 7	0.054 426	0.096 5	1.830 88	-0.407 7	1.041 0

根据表 3，可得表 4。

由表 4 可知，10 月 12 日、10 月 27 日在六个诊断

统计量中皆被检测出 5 次，10 月 15 日、11 月 2 日、1 月 19 日皆被检测出 3 次，因此，将这些点作为重点怀

表 4 各主要诊断统计量处于前五位的观测列表

		I	II	III	IV	V
学生化外残差绝对值	观测序号	第 29 号	第 4 号	第 13 号	第 21 号	第 1 号
	交易日期	11 月 19 日	10 月 15 日	10 月 28 日	11 月 9 日	10 月 12 日
	统计量值	2.304 1	1.825 5	1.540 3	1.507 2	1.468 2
Cook 距离	观测序号	第 12 号	第 29 号	第 1 号	第 4 号	第 24 号
	交易日期	10 月 27 日	11 月 19 日	10 月 12 日	10 月 15 日	11 月 12 日
	统计量值	0.222	0.170	0.137	0.116	0.056
杠杆值	观测序号	第 12 号	第 16 号	第 1 号	第 26 号	第 2 号
	交易日期	10 月 27 日	11 月 2 日	10 月 12 日	11 月 16 日	10 月 13 日
	统计量值	0.464 7	0.253 9	0.165 5	0.124 5	0.123 6
马氏距离	观测序号	第 12 号	第 16 号	第 1 号	第 26 号	第 2 号
	交易日期	10 月 27 日	11 月 2 日	10 月 2 日	11 月 16 日	10 月 13 日
	统计量值	12.511 01	6.395 95	3.831 81	2.643 02	2.616 99
WK 统计量绝对值	观测序号	第 12 号	第 29 号	第 1 号	第 4 号	第 24 号
	交易日期	10 月 27 日	11 月 19 日	10 月 12 日	10 月 15 日	11 月 12 日
	统计量值	0.811 6	0.769 6	0.653 8	0.614 6	0.416 2
协方差比	观测序号	第 12 号	第 16 号	第 3 号	第 2 号	第 22 号
	交易日期	10 月 27 日	11 月 2 日	10 月 14 日	10 月 13 日	11 月 10 日
	统计量值	1.919 3	1.496 6	1.254 0	1.238 8	1.189 7

表 5 删除 2009.10.12 等数据后回归结果

项目	贝塔值	标准误差	回归系数	标准误差	t(22)值	P 水平
截距			0.000 033	0.000 085	0.39	0.698 340
Ratio	1.000 003	0.000 013	0.017 000	0.000 000	77 684.40	0.000 000
Variance	0.000 035	0.000 013	0.000 017	0.000 006	2.72	0.012 615
R ²	1.000 000	R ^{*2}	1.000 000	F(2,22)	3037E6	P<0.000 0
DW	2.416 597	SC	-0.258 155			

疑对象，在原数据集中，将这些点删除，再作回归。

对比表 2 与表 5 可知，复相关系数平方 R^2 及修正的复相关系数平方 R^{*2} ，删除前后均相同，没有变化，均为 1。 F 统计量，删除前后均显著，均可认为响应变量 price 与自变量 ratio, variance 之间存在多元线性关系。对于各回归系数显著性，删除前，ratio 的 $P=0$ ，variance 的 $P=0.04$ ，前者很显著，后者在 0.05 水平下，也显著，删除后，ratio 的 $P=0$ ，很显著，variance 的 $P=0.01$ ，也比较显著，即，删除后，variance 的显著性有所提高。关于自变量之间复共线性，删除前，条件指标 $\eta_1=1$ ， $\eta_2=2.57$ ， $\eta_3=21.832\ 84$ ，删除后， $\eta_1=1$ ， $\eta_2=3.279\ 10$ ， $\eta_3=23.891\ 61$ ，删除前后没有明显变化，根据判定标准，若条件指标处于(0, 100]，则可以认为不存在复共线性关系，因此删除前后自变量之间均符合无复共线性要求，同时，方差膨胀因子 VIF，删除前， $c_{11}=1.000\ 12$ ， $c_{22}=1.000\ 12$ ，删除后， $c_{11}=1.006\ 48$ ，

$c_{22}=1.006\ 48$ ，删除前后没有明显变化，根据一般的规则，若 VIF 处于(0, 10]，则可判定为没有复共线性，因此删除前后，方差膨胀因子也表明自变量之间不存在复共线性。随机误差项之间序列相关问题，删除前，DW=1.96，半偏相关系数 SC=-0.04，根据判定标准，在 DW 接近与 2，SC 接近与 0 时，可以认为不存在序列相关，显然，均符合无序列相关要求，删除后，DW=2.42，SC=-0.25，两者已不符合无序列相关要求，因此，可以认为存在序列相关，总体上，删除前后序列相关性变化较大。

为了判定随机误差项之间是否为方差齐性，我们作残差之递减趋势概率图，如图 1、图 2，可知，删除前存在明显的异方差，删除后基本不存在异方差。分析是这样，尽管两图基本上均符和趋势线应与横坐标轴应尽可能重合的要求，但是，对于散点图应具有随机性，不应具有明显的规律性的要求，图 1 这些点不

符合要求,其先是具有明显的上升趋势,然后有下降,然后再上升,再下降特点,已有规律性,图2则符合随机性要求,另外,关于散点图应相对于横坐标轴尽可能对称的要求,图1也完全不能满足,而图2则基本满足,因此,删除前不满足方差齐性要求,删除后则满足,删除前后变化较大。

对于随机误差项正态性的判断,由正态概率图3图4可知,删除前后均较好符合正态性,没有变化。

综上所述,可知,对于中捷股份,在对重点怀疑的数据删除前后,其随机误差项的序列相关性和方差齐性均有较大变化,说明这些数据点是强影响点,已对建模产生较大影响。其中特别值得指出的是2009年10月27日的数据,在六个诊断统计量中有五个将其检测出来,并且这五项指标值均处于最大值,可见这个日期股价异动应该很明显,这一点有证监会的处罚文件为证。根据中国证监会行政处罚决定书(陈国生)(2011)10号内容的陈述,可知,2009年10月23日,陈国生使用“陈晓彤”证券账户买入“中捷股份”股票761827股,2009年10月27日9时33分41秒,

“陈晓彤”证券账户买入“中捷股份”股票297699股。2009年10月27日10时14分12秒起至11时04分17秒,陈国生利用其资金优势,大量申报买入“中捷股份”股票,并反复多次使用了“拉抬股价,虚假申报”相结合的操作手法,即先以比市场最后一笔成交价高几个价位的少量申报买入并成交,导致股价上涨,随即挂出大量低于同期市场价几个价位的买入申报,委托主要集中在第2档和第3档的位置,造成大单在低档位买入的假象,并迅速撤单。该期间,陈国生使用“陈晓彤”“陈国生”账户以5.75元至6.10元连续28笔申报买入“中捷股份”股票共27920000股,占该时段市场申买量的62%;11时06分16秒前,上述28笔申报撤单26笔(含部分撤单),撤单数量24165426股,撤单量占其申买量的87%,上述26笔撤单平均驻留时间73秒,最短驻留时间19秒;上述申报实际只成交4笔(含部分成交),成交数量3754574股,占该时段其申买量的13%,占该时段市场全部成交量的27%,同期股价上涨7%。11时04分42秒04,“中捷股份”涨停。在“中捷股份”股票处于涨停、

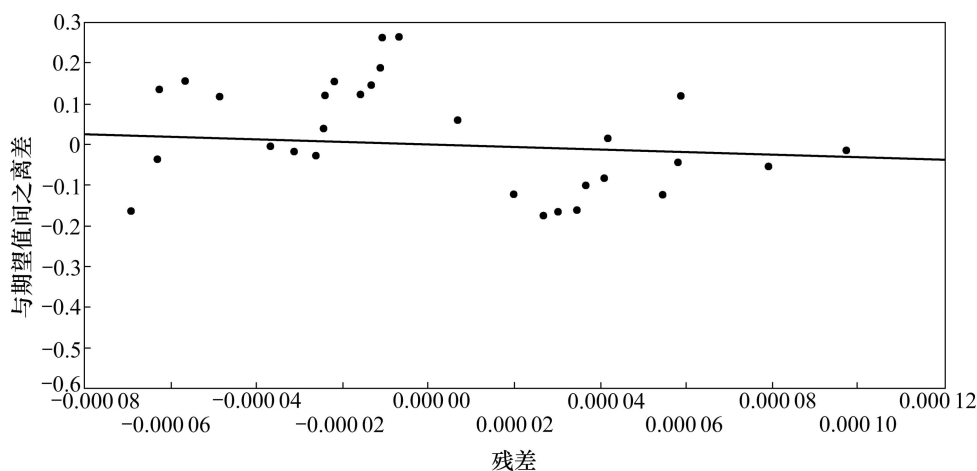


图1 完全数据下回归残差序列趋势概率图

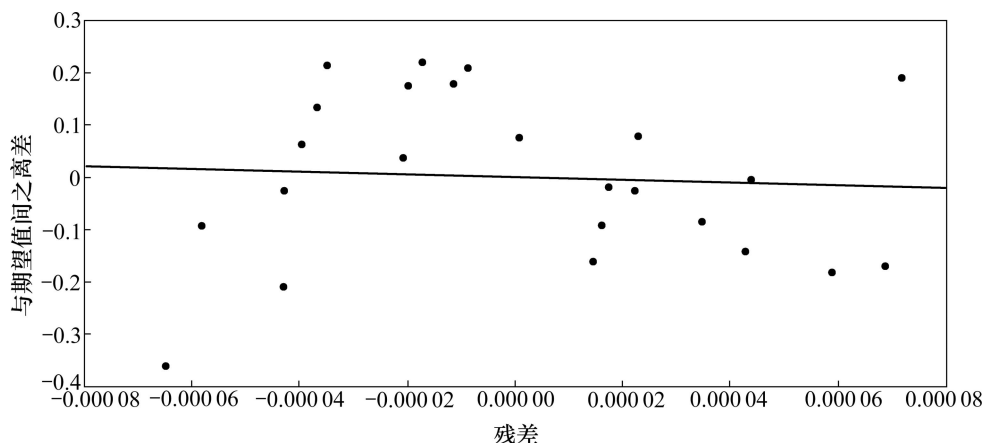


图2 删除数据2009.10.12等数据回归残差序列趋势概率图

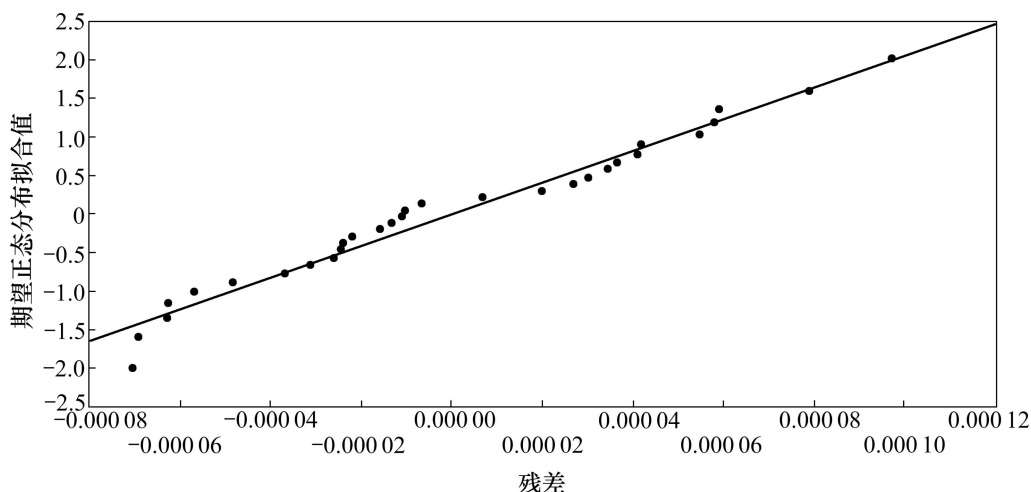


图3 完全数据下回归残差正态概率图

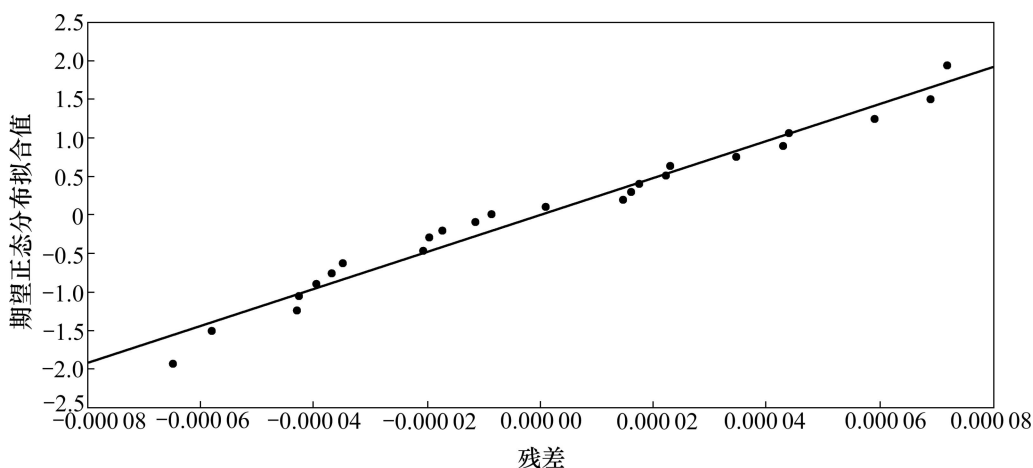


图4 删除2009.10.12等数据回归残差正态概率图

市场本身已有大量未成交买单、该阶段买入申报明显无法成交的情况下，陈国生仍然以涨停价大量申报买入，以相对较高的价格卖出已建仓的股票，同时影响其他投资者对相应股票供求和价格走势的判断，为以后交易日出货牟利提供便利。11时04分42秒04，市场以涨停价申报的待成交买单为8 084 155股。11时05分11秒11至11时26分55秒03期间，“陈晓彤”“陈国生”证券账户以涨停价6.22元的价格连续44笔申报买入42 000 000股，占该时段市场涨停价申买量的47%，占当日市场涨停价申买量的26%；上述44笔申报全部撤单，其中28笔为营业部强制撤单，撤单数量25 656 429股，占该时段其中买量的61%(剔除营业部强制撤单的影响为23%)，占该时段市场撤单量的47%(剔除营业部强制撤单的影响为17%)，平均驻留时间315秒，最短驻留时间33秒；申买成交2笔(含部分成交)，成交数量343 571股，占该时段其中买量的1%，占该时段市场成交量的2%；申卖成交1笔，

共761 827股，占该时段市场成交量的5%。2009年10月27日，“陈晓彤”“陈国生”证券账户全天申买量占市场申买量比例为31%，全天成交量仅占其中买量的6%，占全天市场成交量的6%，全天撤单量占其中买量的94%，占全天市场撤单量的45%。当日，深成指下跌3.14%，“中捷股份”涨幅为10.09%，偏离13.23%。“中捷股份”股票当日市场成交量比前一交易日市场成交量增加319%。

另外，2009年11月2日的数据也被检测为异常，被检测出三次，根据此处罚书的描述，在11月3日、11月4日，“陈晓彤”“陈国生”证券账户将所持的5 157 671股“中捷股份”股票全部卖出，获利3 355 767.86元。可见，此方法对股价异动的检测，还是比较准确的。

根据相同的方法，可将科冕木业、富临运业的结果给出。

对于科冕木业，在选定的2010年3月10日至2010

年4月21日的30个样本数据中,将3月24日、3月25日、4月1日、4月2日、4月7日、4月8日等列为重点怀疑数据。根据中国证监会行政处罚决定书(袁郑健)(2011)42号的描述,袁郑健于2010年3月22日至4月8日使用4个证券账户,连续交易科冕木业股票,在自己实际控制的证券账户之间交易科冕木业股票,在持有科冕木业股票的情况下发表博客文章推荐科冕木业股票。由于袁郑健连续交易数量较大,在自己实际控制的证券账户之间交易的数量较大,博客文章阅读次数较多,致使科冕木业股票价格从2010年3月22日的17.92元上升至4月2日的31元,之后科冕木业股票价格调整至4月8日的27.85元。根据以上事实,证监会认定,自2010年3月22日至4月8日,袁郑健操纵科冕木业股票价格。可见,本方法对于检测股价异动还是可以的。

对于富临运业,选定了2010年4月1日至2010年5月14日共30个样本数据,作回归诊断分析,其中,将4月12日、4月15日、5月10日、5月11日、5月14日等列为可疑数据。根据中国证监会行政处罚决定书(袁郑健)(2011)42号的描述,袁郑健于2010年4月9日至5月13日使用18个证券账户,连续交易富临运业股票,在自己实际控制的证券账户之间交易富临运业股票,虚假申报买入富临运业股票,在持有富临运业股票的情况下发表博客文章推荐富临运业股票。由于袁郑健连续交易数量较大,在自己实际控制

的证券账户之间交易的数量较大,虚假申报买入数量较大,博客文章阅读次数较多,致使富临运业股票价格从2010年4月9日的27.09元上升至5月4日的35.30元,之后富临运业股票价格调整至5月13日的23.95元。根据以上事实,证监会认定,自2010年4月9日至5月13日,袁郑健操纵富临运业股票价格。可见,本方法也较好地检测出股价异动的范围。

参考文献:

- [1] 肖淑芳,李阳.上市公司重大信息披露与股价异动的相关性研究[J].北京理工大学学报,2004,6(6):53-56.
- [2] 史永东,蒋贤峰.中国证券市场违法违规行为的判别——基于内部交易与市场操纵的案例分析[J].预测,2005(3):76-80.
- [3] Thierry Ane, Loredana Ureche Rangau, Jean-Benoit Gambet, Julien Bouverot. Robust Outlier detection for Asia-Pacific stock index returns [J]. Journal of International Financial Markets, Institutions & Money, 2008(18): 326-343.
- [4] 曾伟.中国A股市场异常波动机理及抑制波动研究[D].重庆:重庆大学,2009.
- [5] 瞿宝忠,徐启帆.股价异动:基于并购信息的残差系数法研究[J].审计与经济研究,2009,24(3):87-91.
- [6] Aurea Grane, Helena Veiga. Wavelet-based detection of outliers in financial time series [J]. Computational Statistics and Data Analysis, 2010(54): 2580-2593.

Data Modeling Diagnostic of Listed Companies in the Stock Transaction

LIU Tian

(College of Finance, Dongbei University of Finance and Economics Financial, Dalian116025, China)

Abstract: Securities market regulators and investors are very concerned about the abnormal fluctuation issues in stock price, because it is important to keep the securities order and investors. The author selected three stocks which had been punished as samples, simultaneously, a regression model was established with the price as response variable and the earning ratio and the absolute value of the change as independent variable. The author used data modeling diagnostic methods, according to the the diagnostic statistics of the studentized residuals, leverage values, cook distance, mahalanobis distance, then, this paper detected whether the price was abnormal and determined the suspect data by integrated crossing-confirms. After deleting the suspicious data, the author compared the characteristics about the qualities of model. Empirical studies have shown that the abnormal behaviors of the three cases had been located and conformed actual results, which was of positive significance for the keeping the healthy securities market and protecting the legitimate rights and benefits of investors.

Key Words: the securities market; listed companies; stock price abnormal fluctuations; data modeling diagnostic

[编辑:汪晓]