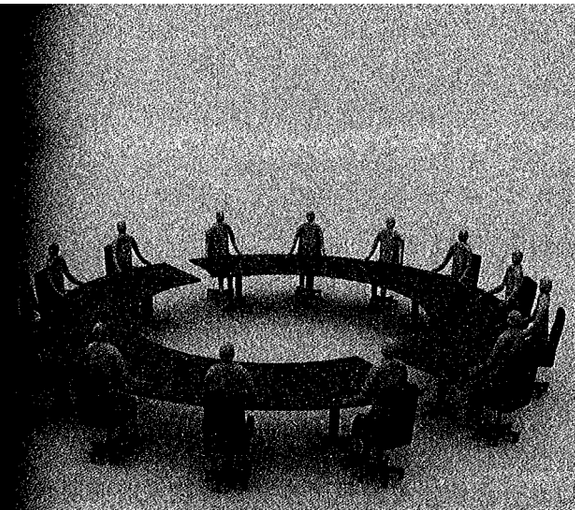
```
> lm1=lm(Y~X1*X2+X3+X4,data=a)
> anova(lm1)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X1          4 13.4161  3.3540 59.9170 < 2.2e-16 ***
X2          1  2.2944  2.2944 40.9884 1.439e-09 ***
X3          3  1.3516  0.4505  8.0485 4.792e-05 ***
X4          1  0.0133  0.0133  0.2375    0.6266
X1:X2       4  0.3153  0.0788  1.4082    0.2333
Residuals 170  9.5162  0.0560
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

去除交互作用 $X_1 \times X_2$ 以及 X_4 后的分析结果：

```
> lm2=lm(Y~X1+X2+X3,data=a)
> anova(lm2)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X1          4 13.4161  3.3540 59.6205 < 2.2e-16 ***
X2          1  2.2944  2.2944 40.7856 1.481e-09 ***
X3          3  1.3516  0.4505  8.0087 4.946e-05 ***
Residuals 175  9.8448  0.0563
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lm2)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = a)

Residuals:
      Min        1Q    Median        3Q       Max
-0.723791 -0.126765  0.006326  0.124980  0.734824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.47950    0.04767  31.037  < 2e-16 ***
X13至4环     -0.06256    0.05984  -1.046   0.2972
X14至5环     -0.17128    0.06794  -2.521   0.0126 *
X15至6环     -0.40837    0.07241  -5.640 6.73e-08 ***
X16环以外    -1.02414    0.11725  -8.734 1.92e-15 ***
X2毛坯       -0.19259    0.04160  -4.630 7.10e-06 ***
X3丰台       -0.25072    0.05427  -4.620 7.41e-06 ***
X3海淀       -0.02894    0.04940  -0.586   0.5588
X3通州       -0.11002    0.07118  -1.546   0.1240
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2372 on 175 degrees of freedom
Multiple R-squared: 0.6341,     Adjusted R-squared: 0.6174
F-statistic: 37.91 on 8 and 175 DF,  p-value: < 2.2e-16
```

第3章

逻辑回归

——以上市企业特别处理 ST 为例

3.1 背景介绍

本案例主要得益于作者以及合作伙伴关于我国特别处理政策的前期研究，相关的主要研究成果以及观点发表如下：

[1] 姜国华，王汉生. 财务报表分析与上市公司 ST 预测研究. 审计研究，2004 (6)：60～63

[2] 姜国华，王汉生. 上市公司两年亏损就应该被 ST 吗？经济研究，2005 (3)：100～101

[3] 姜国华，王汉生. ST：不怕烂货就怕假货. 新财经，2007 (6)：90

[4] 岳衡，王汉生，姜国华. 大股东资金占用与上市公司 ST 关系的研究. 金融学季刊，2008，4 (2)：1～19

[5] 姜国华，王汉生. 取消 ST 制度，完善退市制度，促进股市健康发展. 证券市场导报，2010 (6) 增刊：42～48

[6] Jiang，G. and Wang，H. Should earnings thresholds be used as a delisting criterion in stock market? Journal of Accounting and Public Policy，2008，27：409～419

本章案例从统计分析的角度主要借鉴了上述第一篇文章。因此难免发现本章的文字内容同这些文章有相似之处，特此声明！如果读者感到本章的学习对解决类似问题有所帮助，那么这是作者和合作伙伴共同努力的结果。

特别处理（special treatment，ST）政策是我国股市特有的一项旨在保护投资者利益的政策。具体地说，当上市公司出现财务状况或其他状况异常，导致投资者难以判断公司前景，投资者利益可能受到损害时，交易所要对该公司股票交易实行特别处理。被特别处理的股票每日涨跌幅度是受到限制的。正常情况下，证监会规定一只股票的每日最高涨跌幅为 10%，而被特别处理的股票其日涨跌幅被限制在 5%以内，这样就通过政策性的限制约束了该股票的日内波动程度。如果把一只股票收益率的波动程度看作其风险的一个重要含义，限制股票的每日涨跌幅度似乎可以在一定程度上控制风险。不过对一只被特别处理的股票而言，虽然其每日涨跌幅度不能超过 5%，但是它可以通过连续的涨停板或者跌停板使得

（例如）周度收益率变化幅度极大。因此，限制日度收益率的波动幅度能否减小周度（甚至月度）收益率是一个非常有争议的话题。很多学者对此有不同的看法。本章不对该问题发表过多看法，只是希望通过这样一个背景介绍告诉大家：特别处理是我国资本市场的一项重要政策，值得关注。除了涨跌幅度限制以外，对被特别处理的股票，证监会要求在原股票名称之前加上提醒性注释“ST”。同时，该上市公司的中期报告必须审计。如果一个 ST 企业仍然持续亏损，那么它将有被退市的风险。

那么什么样的企业会被 ST 呢？在上海证券交易所公布的《上海证券交易所股票上市规则（2008 年修订）》第十三章特别处理中有详细规定。具体摘录如下：

13.2.1　上市公司出现以下情形之一的，本所对其股票交易实行退市风险警示：

（一）最近两年连续亏损（以最近两年年度报告披露的当年经审计净利润为依据）；

（二）因财务会计报告存在重大会计差错或者虚假记载，公司主动改正或者被中国证监会责令改正后，对以前年度财务会计报告进行追溯调整，导致最近两年连续亏损；

（三）因财务会计报告存在重大会计差错或者虚假记载，被中国证监会责令改正但未在规定期限内改正，且公司股票已停牌两个月；

（四）未在法定期限内披露年度报告或者中期报告，且公司股票已停牌两个月；

（五）公司可能被解散；

（六）法院受理关于公司破产的案件，公司可能被依法宣告破产；

（七）本所认定的其他情形。

由此可见，判定一只股票是否应该被特别处理是一个重大而又复杂的过程。被特别处理可能有很多原因，其中最主要的原因是第一条：“最近两年连续亏损（以最近两年年度报告披露的当年经审计净利润为依据）。”当然，这是否是一个合理的规定，历来学术界、业界对此都颇有争议，作者和合作伙伴对此历来持保留态度。

我们认为一个更加合理的规则应该惩罚的是披露虚假信息的企业（不管其是否盈利），而不应该惩罚那些亏损但是诚实的企业。因为一个企业是否有投资价值，能否为投资者带来合理回报，同其最近两年是否亏损没有必然联系。实际工作中，如何理解会计报表上的“亏损”也是一个很困难的事情。一个企业的账面亏损有可能因为该企业真的运作有问题，也有可能是因为会计准则的保守性造成的。还有一个可能的原因：企业正处在成长扩张阶

段，没有立刻的盈利能力，但其未来的盈利能力很好，因此，这类企业仍然有很好的投资价值。这样的企业在国外的资本市场上比比皆是。如果按照我国的特别处理政策，那么诸如苹果电脑、福特汽车、通用汽车这样的知名企业统统都被 ST 过；而诸如雅虎、朗讯这样的公司会被暂停上市；亚马逊则在 1998 年就被退市了！

图 3—1 是中美上市公司股东权益回报率的比较。

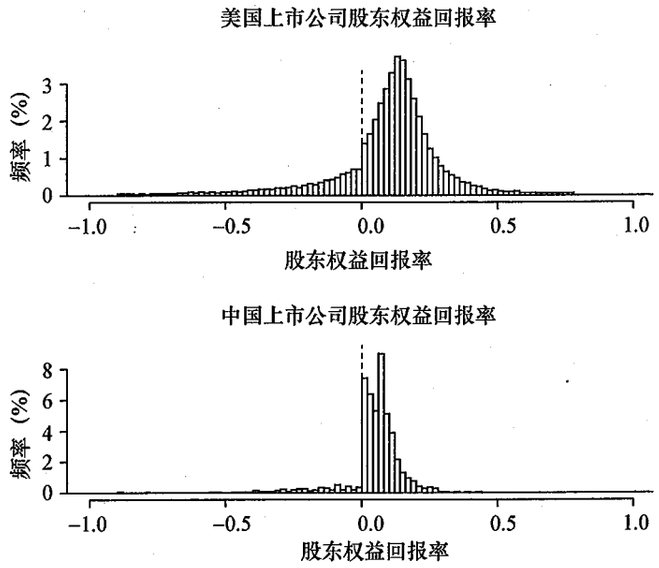


图 3—1　中美上市公司股东权益回报率比较

现行 ST 政策的后果是什么呢？如果将我国上市企业的股东权益回报率做成直方图（histogram），如图 3—1 所示，可以看到在 0 点有一个巨大的不规则的跳跃。这说明很多企业为了避免账面亏损，避免被 ST，千方百计通过各种手段将企业做成微利。这样做的后果是什么？第一，误导投资者对企业价值的判断；第二，更严重地影响企业的正常经营管理活动，伤害了企业的长期盈利能力。

限于篇幅，我们不再对 ST 政策本身做过多探讨。但是，由于被特别处理的企业面临着退市的风险，因此投资者需要对这类企业多加小心。所以，投资者有必要关心什么样的企业更有可能被 ST、它们有什么共同特征、通过正常的财务报表分析能否察觉，这就是本章接下来要解决的问题。

3.2　数据介绍

在详细介绍本案例的数据之前，首先对我国股市的 ST 状况做一个简单描述，这里的主要统计数字来自论文《大股东资金占用与上市公司 ST 关系的研究》（岳衡、王汉生、姜国华，2008）。在这篇论文中，作者对我国 2001—2007 年的 ST 企业做了系统研究，得到统计数字如表 3—1 所示。

表 3—1　　每年 ST 企业的数量

被 ST 年度	样本数	ST 样本数	ST 样本数/样本数
2001	624	21	3.37%
2002	738	41	5.56%
2003	819	52	6.35%
2004	882	37	4.20%
2005	922	31	3.36%
2006	1 010	59	5.84%
2007	1 044	46	4.41%
总计	6 039	287	4.75%

从表 3—1 可以看到，在 2001—2007 年间，随着时间推移，上市企业越来越多，被 ST 的企业的数量也呈现整体上升趋势。但是，相对百分比保持在 5%左右。

除此以外，岳衡、王汉生、姜国华（2008）还特别关心大股东占款行为同企业 ST 的关系。为此，构造了一个衡量大股东占款程度的指标（详细解释在 3.3 节中给出），根据该指标的大小将数据等分十组，其中组号越高，说明大股东占款情况越严重。然后对每一组中 ST 企业的比例做了简单计算，具体结果请见图 3—2。从中可以很清晰地看到，大股东占款比例越高，ST 企业的比例也会越高，最高的一组，该比例超过 14%。这从一个侧面暗示：企业的 ST 状况有可能同大股东的占款行为紧密相关。当然，该相关关系是否是一个因果关系，很难回答，这需要更多的经验以及常识

判断。

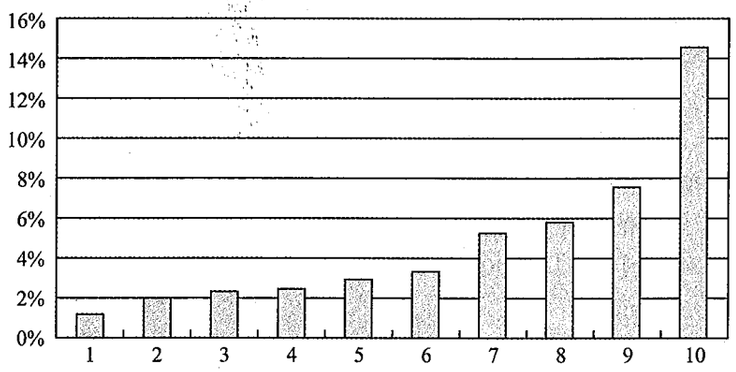


图 3—2　大股东占款程度与企业 ST 状况关系

本案例所用的数据都来自于某商业数据库。值得一提的是，我们的 ST 样本是在当年（记作第 t 年）被 ST 的深沪两市公司，而相应的财务指标（即解释变量）取自于第 $t-3$ 年，即被 ST 之前第三年的数据。我们没有考虑用第 $t-2$ 年（被 ST 之前第二年）的财务指标预测第 t 年的 ST 状况，这主要同 ST 的决定机制有关。如前所述，在第 t 年被 ST 的公司，往往是由于在第 $t-2$ 年和第 $t-1$ 年里连续亏损。因此，对于某一给定公司，在 $t-1$ 年的时候我们就已经知道该公司会不会在下一年（即第 t 年）被 ST。所以，使用第 $t-1$ 年的信息预测第 t 年公司的 ST 状态没有意义。与之类似，我们也没有采用第 $t-2$ 年的数据。因为如果一个公司在第 $t-2$ 年有利润，那么该公司即使在第 $t-1$ 年亏损，它在第 t 年也肯定不会被 ST；而如果一个公司在第 $t-2$ 年亏损，基于这一年数据对第 t 年 ST 的预测将变成简单的对第 $t-1$ 年亏损还是盈利的预测。使用第 $t-3$ 年，即被 ST 之前第三年的数据，则不存在这些问题。这样我们的样本能够覆盖大部分上市公司。

由于我们关心的问题是利用第 $t-3$ 年的财务指标预测第 t 年 ST 状况，因此本问题的因变量就是该企业的 ST 状态。一共有多少种可能的状态呢？答：两种，即 ST 以及非 ST。因此，数学上我们可以用一个 0—1 变量表示。例如，可以定义 $Y=0$ 表示非 ST，而 $Y=1$ 表示 ST。这样一个因变量同第 1 章的因变量（利润变化）以及第 2 章的因变量（房屋均价）有什么区别呢？显然，第 1、2 章的因变量是定量变量，可以支持至少一种代数运算，但本章的因变量表达的是一个属性特征（即是否被 ST），没有任何数值特征，因此是一个定性变量。更进一步，该因变量不能支持任何代数运算，如

我们不能说：ST 企业（$Y=1$）+非 ST 企业（$Y=0$）=ST 企业（$Y=1$），这显然不对。此外，作为一个定性变量，ST 状态也很特殊，其特殊性在于它只有两个（不是三个、四个，或者更多）的取值，这就难怪它可以通过一个 0—1 变量表达。0—1 变量有什么特点呢？它的所有随机特征都由概率 $P(Y=1)=p$ 决定。当然，因为不同企业的财务状况不同，所有不同样本的概率 p 也不同，它是一个企业财务状况的函数，如果该企业的财务状况已经被某解释变量 X 充分表达，那么 p 应该是 X 的一个函数，记作 $p(X)$。

如果因变量是取值多个的定性变量，那又该怎么办？例如，消费者对品牌的选择。以手机为例，常见的大品牌有诺基亚、摩托罗拉、三星等。要研究什么因素影响了消费者的品牌选择，本章将要讲述的逻辑回归显然不再适用。因为逻辑回归要求因变量只能有两个可能的取值，多个取值定性因变量如何分析将在后面的章节讨论。

3.3　指标设计

本案例应该考虑哪些解释变量 X 呢？换句话说，哪些公开的财务指标会和公司是否被特殊处理相关？我们考虑了以下财务指标：

1. ARA（X_1）

该指标是应收账款与总资产的比例，它反映的是盈利质量。对于绝大多数企业来说，没有应收账款是不可能的。一定的回款期限已经是行业内的惯例。如果对方确实是诚实可信的，能够在约定的时间内完成现金支付，应收账款就不是一个大问题。但是，天有不测风云，应收账款只要还没兑现，就一定存在违约的风险。因此，从这个角度看，对于一个企业来说，其资产中应收账款所占的比重应该是越少越好。比重越少说明该企业的盈利质量越好；相反，比重越高说明该企业的盈利质量越差。有学者的研究表明，对我国上市企业而言，应收账款在资产中所占的比率同大股东对小股东的资金侵占挪用也有紧密关系，这也是我们对此变量异常关注的另外一个原因。

2. ASSET（X_2）

该指标是对数变换后的资产规模，用于反映公司规模。这不是反映公司规模的唯一指标，甚至不是最好指标。例如，我们还可以考虑净资产收益率。但是很多经营不善的企业，资不抵债，因此净资产收益率为负数。如何合理解读负的净资产收益率同公司规模之间的抽象关系，没有一个显而易见的答案。此外，可以注意到，对于金融类企业（如基金公司）来说，资产规模其实是一个相当不错的指标。但是对于制造类企业，也许员工数量也是不错的选择。因此，依赖于具体问题，什么叫做“公司规模”是一个很复杂的问题，几乎不大可能有一致的答案。我们这里采用了资产规模，为了提高该指标的实际解读能力，做了对数变换。对于对数变换的实际意义，请参见第2章2.3节中的详细讨论。

3. ATO（X_3）

该指标是资产周转率。按照定义，它是一个企业在一定时期内（如一年以内）的销售收入净额除以资产平均总额而得。假设两个不同的企业（A和B）都有1个单位的平均资产。在一年以内，A企业总共做了10单生意，而B企业只做了1单。因此，A企业的营业额会高于B企业，A企业的资产周转率也会高于B企业。这说明，A企业对资产的利用率要高于B企业。所以，ATO量化的是一个企业对资产的利用效率。

4. ROA（X_4）

该指标是资产收益率。按照定义，它是一个企业在一定时期内（如一年以内）的利润总额除以总资产而得，它反映的是每单位资产能够给企业带来的利润如何。因此，该指标可以看作对企业盈利能力的反映。但它不是反映企业盈利的唯一指标，甚至不是最好的指标。例如，我们还可以考虑净资产收益率，如果净资产不是负数。此外，对于某些特定行业，人们还会关注销售收益率等。总而言之，资产收益率不是最好的盈利指标，但它是一个常用指标。

5. GROWTH（X_5）

该指标是销售收入增长率。按照定义，它是一个企业在一定时期内（如

一年以内）的销售总额除以前一个时期的销售总额而得，它反映的是企业的增长速度。对于很多新兴的高科技企业，在其成立之初，很难实现盈利，但这并不妨碍企业高速增长。企业的高速增长会反映在什么指标上呢？首先是销售收入，也可能是资产、净资产，还可能是市场占有率等其他非财务指标。企业的高速成长会如何影响其盈利，进而影响其被特别处理的概率呢？这不是一个简单的问题。简而言之，如果企业的销售是盈利的，那么高速成长的销售带来的应该是更好的盈利和更小的特别处理概率；但是，如果企业的销售是亏损的，那么高速成长的销售带来的应该是更大的亏损和更大的特别处理概率。

6. LEV（X_6）

该指标是债务资产比率，也叫做杠杆比率。按照定义，它是一个企业债务在其总资产中所占的比率，反映的是企业的总资产中来自于债权人的比率。企业的债务资产比率是如何影响企业盈利，进而影响特别处理的概率呢？这是一个颇有争议的问题。首先，过高的债务资产比显然不好，这会使企业背上沉重的债务负担，企业每年盈利中的一大部分将用于偿还利息，因此损伤盈利。但是，另外一方面，几乎没有企业不举债。适当举债能给企业带来很多好处，如很多创业初期的企业，发展势头很好，但是缺乏资金，那么合理举债能够帮助企业迅速成长，占领市场，确立优势。因此，从平均水平上来讲，债务资产比率到底如何影响特别处理概率不是一个显而易见的问题。

7. SHARE（X_7）

该指标是企业第一大股东的持股比率，反映的是该企业的股权结构。如果企业的第一大股东持股比例很高（如大于70%），说明该企业一股独大，其持有者对企业的方方面面具有绝对权威；如果企业的第一大股东持股比例很低（如小于10%），说明该企业股权分散。企业的股权结构如何影响盈利呢？过度分散的股权结构是不好的，因为这使得所有人都不会真正地关心企业，承担责任；过度集中的股权结构也不好，因为这使得第一大股东有能力侵害小股东的利益。怎样才是一个合理的比例，使得企业的利润最大化，特别处理概率最小化，是一个值得研究的问题。

同前面几章一样，我们需要特别强调，以上设计的指标体系有一定的实

际意义，但也可以肯定地说是不完备的。例如，如果我们认为企业的 ST 状态是一个随着时间变化的动态过程，就应该在解释变量里面加入该企业再前一期的指标。如果我们怀疑企业 ST 同行业有关系，行业特征也应该作为解释变量考虑进来。总而言之，不同的学者结合自己的研究经历和目的，完全有可能提出自己的指标体系。

3.4 描述分析

同前面几章一样，假设本案例所涉及的数据存放在目录“D:\商务数据分析与应用\案例数据”下的 CSV 文件“第 3 章. csv”中。简单展示如下：

	A	B	C	D	E	F	G	H
1	ARA	ASSET	ATO	ROA	GROWTH	LEV	SHARE	ST
2	0.19231	19.85605	0.0052	0.08771	-0.95073	0.44588	26.89	0
3	0.22012	20.91086	0.0056	0.01682	-0.94266	0.398686	39.62	0
4	0.325292	19.35262	0.0166	0.042468	-0.93744	0.303348	26.46	0
5	0.025729	21.43893	0.0028	0.018152	-0.853	0.75825	60.16	0
6	0.533591	21.61334	0.2552	0.004147	-0.8167	0.726875	54.24	1
7	0.061275	21.04117	0.1248	0.051081	-0.81029	0.40175	57.14	0
8	0.441472	20.51676	0.0785	0.060003	-0.80149	0.422824	29	0
9	0.213081	20.61706	0.0606	0.029295	-0.75594	0.559263	29.82	0
10	0.416293	20.51604	0.4747	0.090226	-0.72622	0.569279	36.15	1
11	0.010404	21.3777	0.0643	0.061089	-0.67999	0.375268	49.87	0
12	0.634684	20.2088	0.112	0.032006	-0.67366	0.490132	27.27	0
13	0.501477	20.57185	0.0674	0.051879	-0.64986	0.242655	30.58	0
14	0.06831	20.5835	0.1187	0.111056	-0.63264	0.380344	52.5	0
15	0.042897	20.03443	0.1039	0.024356	-0.59398	0.638925	52.29	1

其中值得注意的是最后一列 ST，这是因变量，取值为 0 或者 1。然后 SAS 读入：

```
data A0;
    infile "D:\商务数据分析与应用\案例数据\第3章.csv"
        firstobs=2 delimiter=",";
    input ARA ASSET ATO ROA GROWTH LEV SHARE ST;
run;
```


原始数据“第 3 章. csv”被读入 SAS 环境，并存放在数据集 A0 中。在 SAS 的资源管理器下可以找到该数据，展示如下：

	ARA	ASSET	ATO	ROA	GROWTH	LEV	SHARE	ST
1	0.192309634	19.85604835	0.0052	0.087709802	-0.950727316	0.445880057	26.89	0
2	0.220119957	20.91086312	0.0056	0.016820383	-0.942656324	0.398686374	39.62	0
3	0.325291689	19.35262341	0.0166	0.042468332	-0.937440417	0.303348107	26.46	0
4	0.025728678	21.43892774	0.0028	0.01815183	-0.852995319	0.758250185	60.16	0
5	0.533590893	21.61333919	0.2552	0.004146607	-0.816703929	0.726875299	54.24	1
6	0.061275206	21.04117011	0.1248	0.051080619	-0.810268401	0.401750312	57.14	0
7	0.441471658	20.51675708	0.0785	0.060003237	-0.801490036	0.422824133	29	0
8	0.213081404	20.61706383	0.0606	0.029295059	-0.755937539	0.559262963	29.82	0
9	0.416292804	20.51603545	0.4747	0.090225874	-0.726220637	0.569279399	36.15	1
10	0.010403584	21.37770103	0.0643	0.061089344	-0.679994765	0.375268384	49.87	0
11	0.634684249	20.2088267	0.112	0.032005969	-0.673661076	0.490131711	27.27	0
12	0.501476997	20.5718533	0.0674	0.05167865	-0.64986085	0.242655012	30.58	0
13	0.068310417	20.58350214	0.1187	0.111056226	-0.632636764	0.380343512	52.5	0
14	0.042896615	20.03442823	0.1039	0.024355732	-0.593975962	0.638924541	52.29	1
15	0.05284031	21.86492839	0.1337	0.030484036	-0.592011661	0.448451148	37.56	0

请注意，本案例涉及的所有解释变量都有数值意义。因此，可以模仿第 1 章的办法，对其做描述分析如下：

```
data A0; set A0; id=_n_; run;
proc transpose data=A0 out=A1; by id; var ARA ASSET ATO ROA GROWTH LEV SHARE ST; run;
proc sort data=A1; by _name_; run;
proc univariate data=A1 noprint;
    by _name_; var col1;
    output out=A2 n=n mean=mean std=std min=min median=median max=max;
run;
data A2; set A2; format mean 5.3 std 5.3 min 5.3 median 5.3 max 5.3; run;
```

具体的分析结果如下表数据集 A2 所示：

	以前的变量名	COL1 的非缺失值数	COL1 的均值	COL1 的标准差	COL1 的最大值	COL1 的中位数	COL1 的最小值
1	ARA	684	0.095	0.092	0.635	0.068	0.000
2	ASSET	684	20.78	0.834	24.02	20.70	18.66
3	ATO	684	0.520	0.363	3.151	0.433	0.003
4	GROWTH	684	0.115	0.307	0.999	0.102	-.951
5	LEV	684	0.406	0.166	0.980	0.407	0.018
6	ROA	684	0.056	0.039	0.311	0.051	0.000
7	SHARE	684	46.03	17.68	88.58	44.96	4.160
8	ST	684	0.053	0.223	1.000	0.000	0.000

首先讨论上表的最后一行 ST。由于 ST 是一个 0—1 变量，因此讨论其最大最小值没有特别大的意义。但是，讨论 ST 的均值还是很有意义的，因为这反映了样本中 ST 企业所占的比例。在 $n=684$ 个样本中 ST 企业占了 5.3%，这同我们在前一节中谈到的常识基本一致。然后我们讨论 ARA。从上表得知，一般企业平均的应收账款比例为 9.5%（以均值计），或者 6.8%（以中位数计）。这似乎是一个不大的水平。但是，值得注意的是其最大值高达 63.5%，这是一个很夸张的比例。从 ASSET

的结果看，样本中的平均资产规模（以中位数计）为 exp（20.70）＝9.77 亿元。此外还可以看到，企业的平均资产周转率为 ATO＝52.0%，一般的销售成长速度为 GROWTH＝11.5%，债务资产比平均保持在 LEV＝40.6%，平均盈利水平为 ROA＝5.6%。最后值得注意的是，第一大股东一般持股比例都很高，平均水平为 SHARE＝46.0%。

前面的描述分析固然有用，但是有一个缺点，那就是都是单变量的，缺乏对比。例如，如果我们能够对每一个解释变量对比其在 ST 组（即 $Y=1$）与非 ST 组（即 $Y=0$）的差异，会获得很多有益信息。要达此目的，一种做法是模仿第 2 章的描述分析手法，分组计算各种统计量，然后再做成对比较。毋庸置疑，这是一个不错的办法。但本章将和大家一起分享一种统计图形（statistical graphics）方法，该方法（或者图）叫做盒状图或箱形图（box plot）。以第一个解释变量 ARA 为例，在 SAS 环境下，可以简单实现如下：

```
proc sort data=A0; by ST; run;
proc boxplot data=A0; plot ARA*ST; run;
```

具体结果如图 3—3 所示：

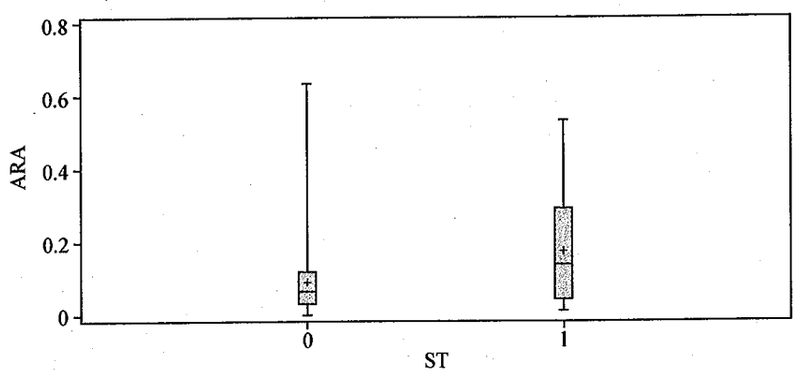


图 3—3　盒状图

接下来，我们仔细解释如何理解该盒状图。根据横坐标的指示，我们知道第一个盒子对应的是 ST＝0（即非 ST 组），第二个对应的是 ST＝1（即 ST 组）。以 ST＝1 组为例，我们看到盒子的中间有一根横线和一个加号“＋”。其中横线的纵坐标代表的是 ST＝1 组的 ARA 中位数（请见纵轴标

示），而加号对应的纵轴位置代表了 ARA 均值。因此，对本组而言，ARA 均值略高于 ARA 中位数。均值和中位数两个指标在一定程度上反映了该组数据 ARA 的中心位置。那么它的变异性（variability）如何判断呢？请注意盒子的上沿，它所对应的纵坐标是 ARA 的第 1 个四分位数，而该盒子的下沿对应的是第 3 个四分位数。因此，盒子的厚度（上沿到下沿的距离）就是 ARA 的第 1、3 个四分位数的间距。按照四分位数的定义，我们知道有 50%的数据被覆盖在此范围以内。如果数据的变异性小，那么其第 1、3 个四分位数的间距应该偏小，相应地，盒子厚度也会小；相反，如果数据的变异性大，那么其第 1、3 个四分位数的间距应该偏大，相应地，盒子厚度就大。因此，在一定程度上，盒子的厚度就反映了 ARA 的变异性的大小。最后值得注意的是在盒子的上下沿以外还延伸出去两条垂直直线。在直线的顶端，各有一个小小的横杠。该横杠的纵坐标对应的分别是 ST＝1 组的 ARA 的最大最小值。对比 ST＝0 组和 ST＝1 组，不难发现 ST＝1 组的均值要比 ST＝0 组的高（请注意“＋”的位置）。类似地，不难发现 ST＝1 组的中位数要比 ST＝0 组的高（请注意横线的位置）。最后还可以注意到，ST＝1 组的变异性要比 ST＝0 组的高（请注意盒子的厚度）。所有这一切都表明，ARA 这个指标在 ST＝0 组和 ST＝1 组之间的规律是不一样的。因此，我们可以合理地预期该指标对于判断预测企业 ST 有重要作用。由此可见，盒状图是一种非常有用的描述分析工具。它不仅能够展示数据的中心位置（均值、中位数），还能够同时展示数据的变异性（四分位间距，inter-quartile distance）。因此，有必要对其他各个解释变量也做类似分析如下：

```
proc boxplot data=A0; plot (ASSET ATO ROA GROWTH LEV SHARE)*ST; run;
```

SAS 程序会自动对其他所有的解释变量按照 ST 分组，然后生成盒状图。分析方法与上文类似，不再赘述。

3.5　统计模型

在描述分析的基础上，讨论如何做回归分析。我们关心的问题是：哪些财务指标能够影响企业 ST，如何影响？因此，可以首先考虑使用第 1 章的线性回归模型来研究此问题，即

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\varepsilon$$

其中 Y 是因变量 ST。乍一看，这似乎是一个不错的模型。如果该模型可以接受，那么第 1 章讲的最小二乘估计、假设检验等方法都适用。但是，我们很快就发现这是一个无法接受的模型，该模型等号的左右两边是矛盾的。等号的右边是一个取值任意的量，尤其是在随机噪音 ε 存在的情况下；左边却是一个取值 0—1 的量 Y。因此，牛头不对马嘴。如何纠正该问题？

Y 是一个取值 0—1 的变量，因此，在给定 $X=(1, X_1, X_2, \cdots, X_7)'$ 的情况下，其随机规律完全由条件概率 $P(Y \mid X)=P(X'\beta)$ 决定，其中 $\beta=(\beta_0, \beta_1, \beta_2, \cdots, \beta_7)'$，而 $X'\beta=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7$。因此，只要能够对 $p(X'\beta)$ 的函数形式做出一个合理的、数学上不矛盾的假设，就能够获得一个合理的模型。关于 $p(X'\beta)$，什么样的函数形式才是合理的呢？$p(X'\beta)$ 是概率，取值 0—1 之间，而 $X'\beta$ 作为一个一般的线性组合取值任意。因此，我们需要这样一个单调函数：它能够把一个取值任意的线性组合 $X'\beta$ 变换到实数 0—1 之间。这样的变换有很多，哪一种最好没有定论，但逻辑变换是最常用的，具体如下：

$$p(X'\beta)=\frac{\exp(X'\beta)}{1+\exp(X'\beta)}$$

或者等价地有

$$\text{logit}\{p(X'\beta)\}=\log\left\{\frac{p(X'\beta)}{1-p(X'\beta)}\right\}=X'\beta$$

这就是我们常说的逻辑回归模型。

同普通线性回归类似，对于逻辑回归而言，人们关心回归系数 β。对于一个给定的解释变量 X_j，$\beta_j=0$ 意味着在给定其他解释变量的前提下，该指标对于解释条件概率 $p(X'\beta)$ 没有任何帮助。因此，对于解释因变量 Y 的随机行为也没有任何帮助。但是，如果 $\beta_j>0$，那么在给定其他解释变量不变的前提下，指标 X_j 的上升会带来条件概率 $p(X'\beta)$ 的上升。也就是说，因变量 Y 取值为 1 的可能性会变大。从某个角度看来，这似乎是一种“正”相关。当然，如果 $\beta_j<0$，那么在给定其他解释变量不变的前提下，指标 X_j 的上升会带来条件概率 $p(X'\beta)$ 的下降。也就是说，因变量 Y 取值为 0 的可能性会变大。从某个角度看来，这似乎是一种“负”相关。

逻辑回归模型应该如何估计呢？从上式看来，逻辑回归似乎也是一个线

性回归，只不过“因变量”不是 Y，而是 $\text{logit}\{p(X'\beta)\}$。因此如果能够观测到 $p(X'\beta)$ 的取值；那就能以 $\text{logit}\{p(X'\beta)\}$ 作为因变量做一个普通的最小二乘估计，所有的问题都能够迎刃而解。问题的难处在于 $p(X'\beta)$ 本身就是一个未知参数，不知道大小，因此，这种天真的最小二乘想法无法实施。为了解决该问题，我们考虑极大似然准则。具体地说，我们用 (Y_i, X_i) 代表来自第 i 个个体的数据。其中 Y_i 是因变量，而 $X_i=(X_{i1}, X_{i2}, \cdots, X_{ip})'$ 是相应的解释变量。给定 X_i 后，Y_i 取值为 0 或者 1 的概率分别为：

$$P(Y_i \mid X_i)=\begin{cases}\dfrac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}, Y_i=1\\ \dfrac{1}{1+\exp(X_i'\beta)}, Y_i=0\end{cases}$$

把这两个表达式整合在一起，可得到：

$$P(Y_i \mid X_i)=\left\{\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right\}^{Y_i}\left\{\frac{1}{1+\exp(X_i'\beta)}\right\}^{1-Y_i}$$

假设不同的样本之间是互相独立的，它们的联合似然函数（likelihood function）为：

$$\prod_{i=1}^{n}P(Y_i \mid X_i)=\prod_{i=1}^{n}\left\{\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right\}^{Y_i}\left\{\frac{1}{1+\exp(X_i'\beta)}\right\}^{1-Y_i}$$

对它做对数变换，得到对数似然函数（log-likelihood function）为：

$$\begin{aligned}\mathcal{L}(\beta)&=\sum_{i=1}^{n}\log\{p(Y_i \mid X_i)\}\\&=\sum_{i=1}^{n}\left[Y_i\log\left\{\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right\}+(1-Y_i)\log\left\{\frac{1}{1+\exp(X_i'\beta)}\right\}\right]\end{aligned}$$

然后通过极大化该对数似然函数获得极大似然估计（maximum likelihood estimator），即：$\hat{\beta}=\text{argmax}_\beta\,\mathcal{L}(\beta)$。标准的统计学理论告诉我们，该估计量是渐进无偏的（asymptotically unbiased）、相合一致的（consistent），而且是极限正态的（asymptotically normal）。因此，可以对每个系数的估计误差有所判断，进而计算相应的 p-值，并做统计学推断，即假设检验 H_0：$\beta_j=0$，H_1：$\beta_j\neq0$。

这里提到的是一个局部检验。如果关心的是全局检验 H_0：$\tilde{\beta}=0$，H_1：

$\tilde{\beta}\neq 0$呢？其中$\tilde{\beta}=(\tilde{\beta}_1, \tilde{\beta}_2, \cdots, \tilde{\beta}_p)'$。或者同方差分析一样，解释变量是定性指标，有很多不同的水平，应该如何检验该因素的显著性呢？这里我们回忆一下第1章和第2章是怎么检验模型的整体显著性和因素显著性的，细心的读者一定会发现它们都借助了残差平方和RSS的作用。更具体地说，都是通过对比不同模型的残差平方和来做推断。因此，对于逻辑回归，可以考虑类似的手段。但是，难点是逻辑回归没有“残差”这个概念。前辈学者在极大似然理论的指引下，找到了一个很有用的、比残差平方和更加广泛的概念，即离差（deviance，DEV）。具体定义如下：

$$DEV=-2\mathcal{L}(\hat{\beta})$$

可以看到，离差其实很简单，它就是负2倍的极大对数似然函数。对普通线性回归而言，残差平方和其实就是一种离差，如果随机扰动项是正态分布的。对于不是线性回归的模型而言，残差平方和不一定存在，但是离差却永远存在，除非连似然函数都写不出来。

基于离差应该如何检验模型整体显著性呢？同第1章的想法类似，考虑两个不同的、互为竞争关系的模型，其中模型A允许$\tilde{\beta}\neq 0$，模型B假设$\tilde{\beta}=0$，把它们各自相应的离差记为：DEV_A和DEV_B。由于模型A比模型B更加灵活，其相应的似然函数一定更大，一定有$DEV_B>DEV_A$。但是，如果原假设$\tilde{\beta}=0$是正确的，那么该差异DEV_B-DEV_A应该不会特别大。到底会多大呢？标准的似然比检验（likelihood ratio test）理论告诉我们，该差异在样本量足够大的情况下，应该服从一个自由度为p的卡方分布（chi-square distribution）。该结论似乎和第1章的结论有点不一样。第1章涉及的分布是一个自由度为$(p, n-p-1)$的F-分布。但是，如果样本量n足够大，该分布其实就会变成一个自由度为p的卡方分布。因此，本章的结论同第1章的结论是高度一致的。在这样一个极限理论的指导下，可以近似地计算出模型全局检验的p-值，并以此为依据，做相应的统计推断。

类似地，如果解释变量中涉及定性的、多水平的因素，我们可以模仿第2章的做法解决所有感兴趣的问题。其中，唯一要做的是把残差平方和用离差代替，然后把F-分布用渐进的卡方分布代替。为节省篇幅，这里不再赘述。

3.6　预测评估

在实际工作中，逻辑回归一个很重要的应用就是预测。例如，对本案例而言，我们能否利用已经建立的逻辑回归模型对未来企业的ST状态予以预测？如果能，那么投资者可以受到警示，然后根据自己对风险收益的偏好，决定最优投资策略。要达到此目的，我们必须具备两种能力：第一，需要具备预测的能力；第二，需要具备对预测精度评估的能力。这就是本节讲述的重点。

首先考虑如何预测。请注意$(Y_i, X_i)$$(i=1, \cdots, n)$代表历史数据，再假设$(Y_i^*, X_i^*)$$(i=1, \cdots, m)$是未来数据。对于未来数据，解释变量$X_i^*$是已知的，但因变量$Y_i^*$是未知的。就本案例而言，$Y_i$是某企业当年的ST状态，$X_i$是它两年前的财务指标。那么，$X_i^*$可以是另外一个企业当年的财务指标，而$Y_i^*$是它两年后的ST状态。如何预测呢？首先通过分析历史数据建立逻辑回归模型，获得极大似然估计$\tilde{\beta}$。然后，将此估计应用于未来数据X_i^*，对其因变量Y_i^*取值为1的概率估计如下：

$$P(Y_i^*=1|X_i^*)\approx P(X_i^{*\prime}\hat{\beta})=\frac{\exp(X_i^{*\prime}\hat{\beta})}{1+\exp(X_i^{*\prime}\hat{\beta})}$$

此概率量化了该企业未来被特别处理的可能性。显然，如果该可能性很大，我们更趋向于将Y_i^*预测为$\hat{Y}_i^*=1$；否则，我们更乐于将Y_i^*预测为$\hat{Y}_i^*=0$。但是，这里的实际难题是到底多大的概率才叫大。显然，我们需要一个阈值α，然后定义一个预测规则如下：

$$\hat{Y}_i^*=\begin{cases}1, p(X_i^{*\prime}\hat{\beta})>\alpha\\ 0, p(X_i^{*\prime}\hat{\beta})\leqslant\alpha\end{cases}$$

要决定α的取值，首先需要了解预测的目的是什么。显然，我们的目的是希望预测得准。那么，如何评判一个预测结果的准确度呢？不同的评判方法、量化手段会产生不同的α选取方法，最常见的莫过于错判率（mis-classification rate，MCR），它的严格定义如下：

$$MCR=\frac{1}{m}\sum_{i=1}^{m}I(Y_i^*\neq\hat{Y}_i^*)$$

简单地说，*MCR* 刻画的就是错误判断的比率。*MCR*=0，意味着所有预测都正确；*MCR*=1，意味着所有预测都错误。如果我们的目标是极小化 *MCR*，那么最优的 α=50%。这是一个非常简单而优美的结论，非常有用。但值得注意的是，*MCR* 隐含着一个假设，即不管真实的因变量 Y_i^* 是 0 还是 1，只要判断错误所带来的损失都是一样的。这常常是一个合理的假设，如果在整个样本中$Y_i^*=0$ 的样本和 $Y_i^*=1$ 的样本分布比较均匀可比。如果该分布非常不均匀，情况就不一样了。

以本章 ST 案例为例，前面的描述分析提到，在我们的数据中，ST 样本只占了整个样本的 5.3%。这说明样本分布极其不均。如果以 α=50%为界去预测，就会把所有的样本都预测为 0（即非 ST）。这样的预测结果显然不好。因此，对此类数据 *MCR* 不再是一个合理的判断标准。我们应该考虑一种加权的 *MCR*，对那些稀有的样本（即 ST=1）赋予更多的权重，而对另外的丰富的样本（即 ST=0）赋予较小的权重。因此，我们考虑加权错判率（weighted mis-classification rate，WMCR）如下：

$$WMCR=\frac{1}{m}\sum_{i=1}^{m}\left\{\frac{I(Y_i^*\neq\hat{Y}_i^*\mid Y_i^*=0)}{\pi_0}+\frac{I(Y_i^*\neq\hat{Y}_i^*\mid Y_i^*=1)}{\pi_1}\right\}$$

其中 $\pi_0=1-\pi_1=P(Y_i^*=0)$ 刻画的是总体中 $Y_i^*=0$ 的比率。对本案例而言，可以大概估计 $\pi_0\approx94.7\%$而 $\pi_1\approx5.3\%$。如果以 WMCR 为标准，重新讨论阈值的选取，我们可以在一定假设限制下获得另外一个结论，那就是最优的 $\alpha=\pi_1$。这个结论细想一下，很符合人们的常识。如果没有任何解释变量，我们知道 $Y_i^*=1$ 的先验概率为 π_1。但在解释变量 X_i^* 的帮助下，如果发现新的概率 $P(X_i^{*\prime}\hat{\beta})$ 大于先验 π_1，那么可以认为该样本似乎更有可能取值 $Y_i^*=1$，否则为 0。

值得注意的是，在 WMCR 的定义中，我们涉及两种不同的错误预测行为。第一种是把真实的 $Y_i^*=0$ 预测为 $\hat{Y}_i^*=1$；而另外一种是把真实的 $Y_i^*=1$预测为 $\hat{Y}_i^*=0$。这与假设检验中的两类错误有些类似，区别在于，对逻辑回归的预测来讲，到底哪一种预测错误带来的损失更大，要具体问题具体分析。为了讨论方便，人们定义了两个不同的概念：

$$\text{TPR(True Positive Rate)}=P(\hat{Y}_i^*=1|Y_i^*=1)$$

$$\text{FPR(False Positive Rate)}=P(\hat{Y}_i^*=1|Y_i^*=0)$$

如果把 $Y_i^*=0$ 比喻成一个“好人”，把 $Y_i^*=1$ 比喻成一个“坏人”，TPR 定义了一个方法抓住坏人的概率（假如真的是“坏人”），因此应该越大越好。相反，FPR 定义了一个方法冤枉“好人”的概率（假如真的是“好人”），因此应该越小越好。但是，对一个实际问题而言，人们不可能无限制地增加 TPR，还同时降低 FPR，它们之间是互相制约的。例如，如果把所有的样本都预测为 $Y_i^*=1$，那么 TPR 是 100%，但同时 FPR 也是 100%；如果把所有的样本都预测为 $Y_i^*=0$，那么 FPR=0，同时 TPR 也是 0。实际工作中，如何平衡 TPR 和 FPR 不是一个容易的问题，这牵涉到两种错误所带来的损失差别有多大。如果实际工作能够提供该信息，那么可以适当选取阈值 α，使得总的损失最小化。但是，对大多数实际工作而言，两种错误所带来的损失很难量化，这时 WMCR 不失为一种最简单而有效的选择。

3.7 SAS 编程

到此为止，逻辑回归的基本理论就介绍完了。我们回到本章的 ST 案例，通过 SAS 程序具体分析如下：

```
proc genmod data=A0 descending;
    model ST=ARA ASSET ATO ROA GROWTH LEV SHARE
      /dist=binomial link=logit;
    output out=A3 pred=pred;
run;
```

SAS 所产生的主要输出如下：

Response Profile

Ordered Value	ST	Total Frequency
1	1	36
2	0	648

PROC GENMOD is modeling the probability that ST='1'.

可以看到，样本中有 648 个非 ST 企业（即 ST=0），同时有 36 个 ST 企业（即 ST=1）。从最后一句话还可以看出，SAS 模拟的是 ST=1 的概率，而不是 ST=0 的概率，这对我们正确解读后面的参数估计结果非常重要。如果不巧 SAS 模拟的是 ST=0 的概率，那么你会发现所有参数估计的

正负号都和预期相反。

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.8692	4.6359	-17.9554	0.2169	3.66	0.0557
ARA	1	4.8797	1.4925	1.9546	7.8049	10.69	0.0011
ASSET	1	0.2468	0.2241	-0.1926	0.6858	1.21	0.2711
ATO	1	-0.5074	0.6575	-1.7960	0.7812	0.60	0.4403
ROA	1	-0.6366	6.2236	-12.8347	11.5615	0.01	0.9186
GROWTH	1	-0.8333	0.5671	-1.9448	0.2781	2.16	0.1417
LEV	1	2.3542	1.2014	-0.0005	4.7088	3.84	0.0501
SHARE	1	-0.0111	0.0111	-0.0330	0.0107	0.99	0.3189
Scale	0	1.0000	0.0000	1.0000	1.0000		

在10%的水平下，可以看到只有两个解释变量是显著的。一个是ARA，它的极大似然估计为4.88，是正值，这说明ARA的取值越高（即应收账款比率越高），该企业被特别处理的可能性越大。类似地，LEV的估计量也是显著的（p-值为0.05），其极大似然估计量为2.35，是正值，这说明LEV的取值越高（即债务水平越高），该企业被特别处理的可能性越大。对于其他所有变量，基于现有数据，无法下确定性结论。

值得注意的是，在上面这个SAS程序中，有一个选项是pred＝pred。该选项决定了在数据A3中有一列叫做pred。该列记录的就是根据模型估计所测算出来的ST概率，即$P(X_i^{*\prime}\hat{\beta})$，但存在一点区别。理论上讲，对历史数据测算该概率没有什么意义，因为历史数据的真实ST状态都是已知的。因此，只有对未来数据测算该概率才有实际意义。但是，真正的未来数据的ST状态又彻底未知，不利于我们讲解诸如MCR，TPR，FPR之类的概念。这里方便起见，对历史数据也测算了该概率，这就是变量pred的实际含义。根据pred的取值，可以对ST状态重新预测如下。

假设我们的目的是极小化MCR，那么50%是一个最优的取值。在SAS中，我们可以简单实现如下：

```
data A3; set A3;
    if pred< 0.5 then SThat=0;
    if pred>=0.5 then SThat=1;
run;
```

预测的ST状态被记录在变量SThat中。然后结合真实的ST状态，对预测精度分析如下：

```
proc freq data=A3; table ST*SThat; run;
```


相应的输出为：

FREQ 过程

ST * SThat 表

ST 频数 百分比 行百分比 列百分比	SThat 0	1	合计
0	647 94.59 99.85 94.87	1 0.15 0.15 50.00	648 94.74
1	35 5.12 97.22 5.13	1 0.15 2.78 50.00	36 5.26
合计	682 99.71	2 0.29	684 100.00

从上表可以看到，在总共684个样本中，SThat预测错误的个数为1＋35＝36个。其中，1个是把ST＝0预测为SThat＝1，另外35个是把ST＝1预测为SThat＝0。整体的预测错误概率为MCR＝36/684＝5.26%，非常小。因此，如果以MCR为标准，该结果确实不错。但是，对本案例而言，更重要的是要把那些真正的ST＝1企业成功抓住。因此，MCR不是一个很好的标准。再仔细分析一下，在整个684个样本中，真实ST＝1的企业有36个，我们抓住多少个呢？1个。相应的TPR＝1/36＝2.8%，非常差，无法接受！因此，对本案例而言，α＝50% 是一个无法接受的结果。

如果以WMCR为标准考虑，结果又如何？从上表可以看到，市场ST＝1的企业比重为36/684＝5.26%。因此，可以考虑α＝5.26%，重新预测分析如下：

```
data A3; set A3;
    if pred< 0.0526 then SThat=0;
    if pred>=0.0526 then SThat=1;
run;

proc freq data=A3; table ST*SThat; run;
```

相应的输出结果为：

FREQ 过程

ST * SThat 表

ST	SThat		
频数 百分比 行百分比 列百分比	0	1	合计
0	463 67.69 71.45 97.68	185 27.05 28.55 88.10	648 94.74
1	11 1.61 30.56 2.32	25 3.65 69.44 11.90	36 5.26
合计	474 69.30	210 30.70	684 100.00

从整体错判概率 MCR 的角度看，该结果并不令人满意。因为在整个 684 个样本中，预测错误的样本总数为 185＋11＝196，相应的比例为 MCR＝196/684＝28.7%。这可比前面 $\alpha=50\%$ 时的 MCR＝36/684＝5.26%大太多了！但是，好处是在整个 36 个 ST＝1 企业中，我们抓住了 25 个，因此 TPR＝25/36＝69.4%。该结果比前面 $\alpha=50\%$ 时的 TPR＝1/36＝2.8%又好了太多。

对比分析一下这两个结果。一个以 $\alpha=50\%$ 为阈值，一个以 $\alpha=5.26\%$ 为阈值，它们的核心区别就在 TPR 还有 FPR 上。前面已经提到，如果以 TPR 为标准，$\alpha=5.26\%$ 所产生的 TPR＝69.4%，远远优于 $\alpha=50\%$ 所产生的 TPR＝2.8%。但是，如果以 FPR 为标准，以 $\alpha=50\%$ 为阈值所产生的 FPR＝1/648＝0.2%，非常小，而以 $\alpha=5.26\%$ 为阈值所产生的 FPR＝185/648＝28.5%，非常大。一句话概括，$\alpha=50\%$ 保护“好人”ST＝0，而 $\alpha=5.26\%$ 能够抓住“坏蛋”ST＝1。对本案例而言，哪一方面更加重要呢？如果以投资为例，并假设投资 ST 股有可能带来重大损失（实际中可能不见得这样），那么抓住“坏蛋”，并避免对其投资是关键。另一方面，冤枉“好人”的后果就是不给“好人”投资。但是，在资本市场上超过 95%的企业都是“好人”，没有任何一个投资者能够全部覆盖，冤枉几个“好人”不是什么了不起的问题。因此，对本案例而言，提高 TPR 是重点。

3.8　总结讨论

到此为止，我们通过 ST 案例，对逻辑回归模型的核心理论做了简要论述，对相应的 SAS 编程做了详细展示。从理论上讲，逻辑回归同线性回归极其类似，技术上的区别就是极大似然估计取代了最小二乘估计，离差取代了残差平方和。最后值得一提的是，逻辑回归是一种常见的处理 0—1 因变量的回归模型。但如前所述，这不是唯一的。例如，PROBIT 回归就是另外一个很好的替代。限于篇幅和精力，不再赘述。有兴趣的读者可以通过查阅相关统计学教材以及 SAS 文档学习。对于一个具体数据而言，到底哪一种模型更好，这很难回答，得具体问题具体分析。幸运的是，作者有限的经验表明，在绝大多数情况下，差别不大。

另外，为了讨论简便，本案例对预测精度的讨论局限于历史数据。实际工作中用历史数据建立模型，再用它评估精度是不好的。因为这样做等同于历史数据既扮演了“运动员”（建立模型）的角色，又扮演了“裁判员”（预测评估）的角色。更科学的做法是将数据随机分成两组，用其中一组做“运动员”建立模型，用另外一组扮演“裁判员”评估预测精度。

附录 3A　分析报告

财务报表分析与 ST 预测

1. 研究目的

通过分析上市公司的公开财务报表信息，达到预测其未来两年内被 ST 的可能性，并以此警示投资风险。

2. 背景介绍

特别处理（special treatment，ST）是我国股市特有的一项旨在保护投资者利益的政策。具体地说，当上市公司出现财务状况或其他状况异常，导致投资者难以判断公司前景，投资者利益可能受到损害时，交易所要对该公司股票交易实行特别处理。被特别处理的股票每日涨跌幅度是受到限制的。正常情况下，证监会规定一只股票的每日最高涨跌幅为 10%，而被特别处理的股票其日涨跌幅被限制在 5%以内，这样就通过政策性的限制约束了该股票的日内波动程度。如果把一只股票收益率的波动程度看作其风险的一个重要含义，限制股票的每日涨跌幅度似乎可以在一定程度上控制风险。除了涨跌幅度限制以外，对被特别处理的股票，证监会要求在原股票名称之前加上提醒性注释“ST”。同时，该上市公司的中期报告必须审计。如果一个 ST 企业仍然持续亏损，那么它将有被退市的风险。

上海证券交易所公布的《上海证券交易所股票上市规则（2008 年修订）》第十三章特别处理中对什么样的企业会被 ST 有详细规定，具体摘录如下：

（一）最近两年连续亏损（以最近两年年度报告披露的当年经审计净利润为依据）；

（二）因财务会计报告存在重大会计差错或者虚假记载，公司主动改正或者被中国证监会责令改正后，对以前年度财务会计报告进行追溯调整，导致最近两年连续亏损；

（三）因财务会计报告存在重大会计差错或者虚假记载，被中国证监会责令改正但未在规定期限内改正，且公司股票已停牌两个月；

（四）未在法定期限内披露年度报告或者中期报告，且公司股票已停牌两个月；

（五）公司可能被解散；

（六）法院受理关于公司破产的案件，公司可能被依法宣告破产；

（七）本所认定的其他情形。

由此可见，判定一只股票是否应该被特别处理是一个重大而又复杂的过程。被特别处理可能有很多原因，其中最主要的原因是第一条：“最近两年连续亏损（以最近两年年度报告披露的当年经审计净利润为依据）。”

根据表 3—2 中汇总的数据可以看到，在 2001—2007 年间，随着时间推移，上市企业越来越多，被 ST 的企业的数量也呈现整体上升趋势。但是，相对百分比保持在 5%左右。

表 3—2　　每年 ST 企业的数量

被 ST 年度	样本数	ST 样本数	ST 样本数/样本数
2001	624	21	3.37%
2002	738	41	5.56%
2003	819	52	6.35%
2004	882	37	4.20%
2005	922	31	3.36%
2006	1 010	59	5.84%
2007	1 044	46	4.41%
总计	6 039	287	4.75%

由于被特别处理的企业面临退市的风险，因此投资者需要对这类企业多加小心。所以，投资者有必要关心什么样的企业更有可能被 ST、它们有什么共同特征、通过正常的财务报表分析能否察觉，这就是本研究要解决的问题。

3. 指标设计

为了能够准确预测企业未来的 ST 可能性，我们考虑了下面的常见财务

指标，并对考虑它们的原因做了详细讨论。

- ARA（X_1）

该指标是应收账款与总资产的比例，它反映的是盈利质量。简单地说，两笔不同的生意，同样为企业收获了1个单位的盈利。其中一个带回来的是现金收益，而另外一个带回来的是一个“承诺”，即对方承诺在一定期限内，以一定的方式支付现金。对于企业来讲，这部分收入是应得的，但是还没有立刻兑现，因此称为应收账款。对于绝大多数企业来说，没有应收账款是不可能的。一定的回款期限已经是行业内的惯例。如果对方确实是诚实可信的，能够在约定的时间内完成现金支付，那应收账款不是一个大问题。但是，天有不测风云，应收账款只要还没兑现，就一定存在违约的风险。因此，从这个角度看，对于一个企业来说，其资产中应收账款所占的比重应该是越少越好。比重越少说明该企业的盈利质量越好；相反，比重越高说明该企业的盈利质量越差。有学者的研究表明，对我国上市企业而言，应收账款在资产中所占的比率同大股东对小股东的资金侵占挪用也有紧密关系，这也是我们对此变量异常关注的另外一个原因。

- ASSET（X_2）

该指标是对数变换后的资产规模，用于反映公司规模。这不是反映公司规模的唯一指标，甚至不是最好的指标。例如，我们还可以考虑净资产收益率。但是很多经营不善的企业，资不抵债，因此净资产收益率为负数，如何合理解读负的净资产收益率同公司规模之间的抽象关系，没有一个显而易见的答案。此外，可以注意到，对于金融类企业（如基金公司）来说，资产规模其实是一个相当不错的指标。但是对于制造类企业，也许员工数量也是一个不错的选择。因此，依赖于具体问题，什么叫做“公司规模”是一个很复杂的问题，几乎不大可能有一致的答案。我们这里采用了资产规模，为了提高该指标的实际解读能力，做了对数变换。

- ATO（X_3）

该指标是资产周转率。按照定义，它是一个企业在一定时期内（如一年以内）的销售收入净额除以资产平均总额而得。假设两个不同的企业（A和B）都有1个单位的平均资产。在一年以内，A企业总共做了10单生意，而B企业只做了1单。因此，A企业的营业额会高于B企业，A企业的资产周转率也会高于B企业。这说明，A企业对资产的利用率要高于B企业。所以，ATO量化的是一个企业对资产的利用效率。

- ROA（X_4）

该指标是资产收益率。按照定义，它是一个企业在一定时期内（如一年以内）的利润总额除以总资产而得，它反映的是每单位资产能够给企业带来的利润如何。因此，该指标可以看作对企业盈利能力的反映。但它不是反映企业盈利的唯一指标，甚至不是最好的指标。例如，我们还可以考虑净资产收益率，如果净资产不是负数。此外，对于某些特定行业，人们还会关注销售收益率等。总而言之，资产收益率不是最好的盈利指标，但它是一个常用指标。

- GROWTH（X_5）

该指标是销售收入增长率。按照定义，它是一个企业在一定时期内（如一年以内）的销售总额除以前一个时期的销售总额而得，它反映的是企业的增长速度。对于很多新兴的高科技企业，在其成立之初，很难实现盈利，但这并不妨碍企业高速增长。企业的高速增长会反映在什么指标上呢？首先是销售收入，也可能是资产、净资产，还可能是市场占有率等其他非财务指标。企业的高速成长会如何影响其盈利，进而影响其被特别处理的概率呢？这不是一个简单的问题。简而言之，如果企业的销售是盈利的，那么高速成长的销售带来的应该是更好的盈利和更小的特别处理概率。但是，如果企业的销售是亏损的，那么高速成长的销售带来的应该是更大的亏损和更大的特别处理概率。

- LEV（X_6）

该指标是债务资产比率，也叫做杠杆比率。按照定义，它是一个企业债务在其总资产中所占的比率，它反映的是企业的总资产中来自债权人的比率。企业的债务资产比率是如何影响企业盈利，进而影响特别处理的概率呢？这是一个颇有争议的问题。首先，过高的债务资产比显然不好，这会使企业背上沉重的债务负担，企业每年盈利中的一大部分将用于偿还利息，因此损伤盈利。但是，另一方面，几乎没有企业不举债。适当举债能给企业带来很多好处，如很多创业初期的企业，发展势头很好，但是缺乏资金，那么合理举债能够帮助企业迅速成长，占领市场，确立优势。因此，从平均水平上来讲，债务资产比率到底如何影响特别处理概率不是一个显而易见的问题。

- SHARE（X_7）

该指标是企业第一大股东的持股比率，反映的是该企业的股权结

构。如果企业的第一大股东持股比例很高（如大于70%），说明该企业一股独大，其持有者对企业的方方面面具有绝对权威；如果企业的第一大股东持股比例很低（如小于10%），说明该企业股权分散。企业的股权结构如何影响盈利呢？过度分散的股权结构是不好的，因为这使得所有人都不会真正地关心企业，承担责任；过度集中的股权结构也不好，因为这使得第一大股东有能力侵害小股东的利益。怎样才是一个合理的比例，使得企业的利润最大化、特别处理概率最小化，是一个值得研究的问题。

我们需要特别强调，以上设计的指标体系是有一定的实际意义的，但也可以肯定地说是不完备的。例如，如果我们认为企业的ST状态是一个随着时间变化的动态过程，就应该在解释变量里面加入该企业再前一期的指标。如果我们怀疑企业ST同行业有关系，行业特征也应该作为解释变量考虑进来。所有这些，都可能是我们未来的研究课题。

4. 描述分析

在正式的模型分析之前，首先对因变量以及自变量做必要的描述分析。主要结果如表3—3所示。首先讨论表3—3的最后一行ST，从中可以知道，在n=684个样本中ST企业占了5.3%。然后讨论ARA，从表3—3中得知，一般企业平均的应收账款比例为9.5%（以均值计），或者6.8%（以中位数计）。这似乎是一个不大的水平，但是值得注意的是其最大值高达63.5%，这是一个很夸张的比例。从ASSET的结果看，样本中的平均资产规模（以中位数计）为exp（20.70）=9.77亿元。此外还可以看到，企业的平均资产周转率为ATO=52.0%，一般的销售成长速度为GROWTH=11.5%，债务资产比平均保持在LEV=40.6%，平均盈利水平为ROA=5.6%。最后值得注意的是，第一大股东一般持股比例都很高，平均水平为SHARE=46.0%。

表3—3　　因变量的描述分析

变量名称	样本量	均值	标准差	最大值	中位数	最小值
ARA	684	0.095	0.092	0.635	0.068	0.000
ASSET	684	20.780	0.834	24.020	20.700	18.660
ATO	684	0.520	0.363	3.151	0.433	0.003
GROWTH	684	0.115	0.307	0.999	0.102	−0.951
LEV	684	0.406	0.166	0.980	0.407	0.018
ROA	684	0.056	0.039	0.311	0.051	0.000
SHARE	684	46.030	17.680	88.580	44.960	4.160
ST	684	0.053	0.223	1.000	1.000	0.000

我们还将各个解释变量按照ST状态分组做盒状图对比。其中发现变量ARA的组间差异最大，如图3—4所示。

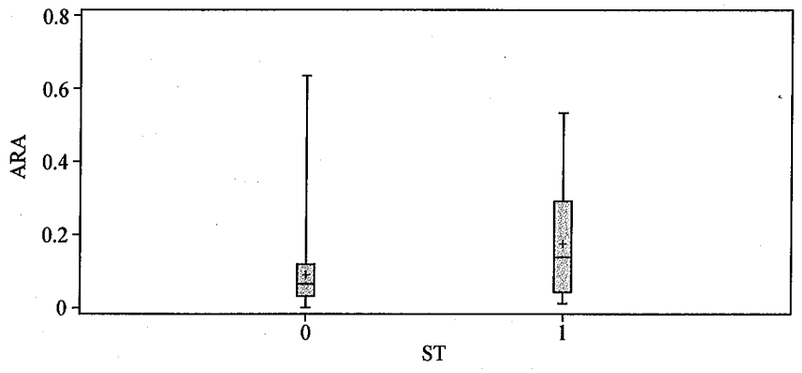


图3—4　盒状图

从上图可以看到，非ST组（即ST=0）的ARA中位数明显低于ST组（即ST=1），或者说明，ST企业常常伴随着较高的ARA水平（以中位数计，请注意横线的位置），即较差的盈利质量。同样的规律也表现在均值上（请注意“+”的位置）。最后还可以注意到，ST=1组的变异性要比ST=0组的高（请注意盒子的厚度）。所有这一切都表明，ARA这个指标在ST=0组和ST=1组之间的规律是不一样的。因此，我们可以合理地预期该指标对于判断预测企业ST有重要作用。

5. 模型分析

在描述分析的基础上，通过方差分析对各个因素同ST状态之间的关系做了逻辑回归模型。参数估计如表3—4所示。从中可以看到只有两个解释变量是显著的。一个是ARA，它的极大似然估计为4.88，是正值。这说明ARA的取值越高（即应收账款比率越高），该企业被特别处理的可能性越大。类似地，LEV的估计量也是显著的（p-值为0.05），其极大似然估计量为2.35，是正值。这说明LEV的取值越高（即债务水平越高），该企业被特别处理的可能性越大。对于其他所有变量，基于现有数据，无法下确定性结论。

表 3—4　　各参数估计以及检验结果

因素名称	参数估计	标准误差	卡方统计量	p-值
截距项	−8.869	4.636	3.66	0.055 7
ARA	4.880	1.493	10.69	0.001 1
ASSET	0.247	0.224	1.21	0.271 1
ATO	−0.507	0.658	0.60	0.440 3
ROA	−0.637	6.224	0.01	0.918 5
GROWTH	−0.833	0.567	2.16	0.141 7
LEV	2.354	1.201	3.84	0.050 1
SHARE	−0.011	0.011	0.99	0.318 9

6. 预测评估

基于上述模型，可以对每个企业的 ST 概率予以测算。然后以某阈值为界，将其预测为 ST 企业（即 ST=1）或者非 ST 企业（即 ST=0）。如果以 50%为界，我们发现除了一个样本之外，所有样本都会被预测为 ST=0。因此总体预测精度优良，错判概率 MCR 为 5.26%；但 TPR 很差，只有 2.8%，对于实际工作没有任何价值。因此，我们重新考虑了加权后的错判概率 WMCR，相应地将阈值设为 5.26%。因此而产生的总体错判概率为 MCR=28.7%，但是，TPR 大大提高为 TPR=69.4%，同时 FPR 仍然可以得到一定的控制（FPR=28.5%）。因此，5.26%是一个比较好的阈值，可以推荐。

7. 总结讨论

本研究分析上市企业的公开财务报表信息，建立了对企业未来 ST 状态具有一定预测能力的逻辑回归模型。我们的分析表明企业的盈利质量（以 ARA 计）是影响企业未来 ST 可能性的最重要的因素，值得关注。

附录 3B　课后习题

移动通信客户流失规律分析

1. 研究目的

通过对某移动通信公司客户的流失数据分析，了解客户流失规律，建立流失预警系统，为客户关系管理服务。

2. 数据介绍

某年度随机抽取的 1 000 个移动通信客户。因变量是他们来年的流失行为（0=未流失，1=流失）。为了能够预测客户的未来行为，我们采集了下面这些来自当年的指标：客户等级（区分 VIP 客户等级）：1，2，3，4；主叫次数（%）：7 日内日均主叫次数/90 日内日均主叫次数；被叫次数（%）：7 日内日均被叫次数/90 日内日均被叫次数；通话时长（%）：7 日内日均通话时长/90 日内日均通话时长；费用（%）：7 日内日均通话费用/90 日内日均通话费用。该数据存放在目录“D：\商务数据分析与应用\课后练习”下 CSV 文件“课后练习 3.csv”中。

3. 作业要求

- 问题理解：请参阅相关媒体报道，理解客户流失对移动通信类企业的重大意义。
- 做完整的逻辑回归分析，包括参数估计、假设检验，以及预测评估。
- 将分析结果汇总成为如附录 3A 所示的简短研究报告。

附录 3C　R 程序演示

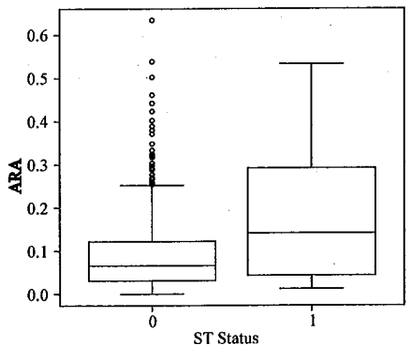
通过下面程序读入数据：

```
> a=read.csv("D:/商务数据分析与应用/案例数据/第3章.csv",header=T)
> a[c(1:5),]
        ARA    ASSET    ATO         ROA      GROWTH       LEV SHARE ST
1 0.19230963 19.85605 0.0052 0.087709802 -0.9507273 0.4458801 26.89  0
2 0.22011996 20.91086 0.0056 0.016820383 -0.9426563 0.3986864 39.62  0
3 0.32529169 19.35262 0.0166 0.042468332 -0.9374404 0.3033481 26.46  0
4 0.02572868 21.43893 0.0028 0.018151630 -0.8529953 0.7582502 60.16  0
5 0.53359089 21.61334 0.2552 0.004146607 -0.8167039 0.7268753 54.24  1
```

描述分析：

```
> N=sapply(a,length)
> MU=sapply(a,mean)
> SD=sapply(a,sd)
> MIN=sapply(a,min)
> MED=sapply(a,median)
> MAX=sapply(a,max)
> result=cbind(N,MU,SD,MIN,MED,MAX)
> result
           N          MU          SD          MIN          MED         MAX
ARA      684  0.09504945  0.09228931  0.00000000  0.06832718  0.6346842
ASSET    684 20.77785347  0.83352322 18.66070036 20.70050279 24.0176107
ATO      684  0.51977383  0.36282648  0.00280000  0.43340000  3.1513000
ROA      684  0.05567011  0.03859391  0.00008170  0.05125798  0.3111300
GROWTH   684  0.11525745  0.30702005 -0.95072732  0.10228264  0.9985565
LEV      684  0.40606356  0.16576397  0.01843107  0.40673974  0.9803218
SHARE    684 46.03451754 17.68437717  4.16000000 44.95500000 88.5800000
ST       684  0.05263158  0.22346029  0.00000000  0.00000000  1.0000000
```

盒状图：



正式逻辑回归：

```
> glm1=glm(ST~ARA+ASSET+ATO+GROWTH+LEV+ROA+SHARE,family=binomial(link=logit),data=
> summary(glm1)

Call:
glm(formula = ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE,
    family = binomial(link = logit), data = a)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4165  -0.3354  -0.2536  -0.1959   3.0778

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.86924    4.63586  -1.913  0.05573 .
ARA          4.87974    1.49245   3.270  0.00108 **
ASSET        0.24660    0.22409   1.100  0.27115
ATO         -0.50738    0.65744  -0.772  0.44026
GROWTH      -0.83335    0.56706  -1.470  0.14167
LEV          2.35415    1.20138   1.960  0.05005 .
ROA         -0.63661    6.22354  -0.102  0.91853
SHARE       -0.01111    0.01115  -0.997  0.31891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 282.07  on 683  degrees of freedom
Residual deviance: 251.51  on 676  degrees of freedom
AIC: 267.51

Number of Fisher Scoring iterations: 6
```

按照 WMCR 做模型预测以及评估：

```
> pred=predict(glm1,a)
> prob=exp(pred)/(1+exp(pred))
> yhat=1*(prob>0.0526)
> table(a$ST,yhat)
   yhat
      0   1
  0 463 185
  1  11  25
```