第3章
分类模型

线性回归模型假设因变量 Y 是定量的（quantitative）。但在很多情况下，因变量反而是定性的（qualitative）。定性变量也称为分类（categorical）变量，两者的统计含义是一样的。例如，资产价格是否上涨是定性变量，取值：上涨或不上涨。这一章我们将学习预测定性因变量的方法及分类（classification）的过程。[1, 2] 预测一个观测的定性变量需要对其分类（classifying），涉及将观测匹配到一个类别中。另一方面，大部分的分类方法先从预测定性变量的不同类别的概率开始，将分类问题作为概率估计的一个结果。从这个角度上看，分类与回归方法有许多类似之处。

目前有许多分类技术或分类模型（classifier）已被开发出来用于预测定性因变量值。[3, 4] 这一章，我们将讨论应用最广泛的分类方法逻辑回归（logistic regression）和支持向量机（support vector machine, SVM）。我们将在第 4 章树模型中讨论更多的分类方法。

3.1 分类问题概述

现实中分类问题是比较常见的，甚至比回归问题还要多，下面举几个例子。在金融市场中我们关心的问题有：（1）IPO 上市首日收益率为正数还是负数？[5, 6]（2）上市公司 i 在 t 年是否违规？[7]（3）金融市场中股票、期货的价格在某个时间段内是否上涨？与回归一样，在分类中假定有一系列训练观测 $(x_1\ y_1)$，$\cdots$，$(x_n\ y_n)$，可以根据训练数据建立一个分类模型，使模型不仅较好地拟合训练数据，而且在测试集上也能有较好的效果。[8]

在这一章，我们将通过我国上市公司的财务违规（il）数据集来阐述分类模型的概念。[9] 在该问题中，我们感兴趣的是基于上市公司递延所得税资产异动和总资产收益率预测其财务违规的状态，数据集如图 3-1 所示。图 3-1 是根据 24 817 个样本子集绘出的 ROA（总资产收益率）和 abDTA（递延所得税资产异动）的关系图。在图中红色表示某个月份出现违约的上市公司，蓝色表示未违约的上市公司。图中表明违规的上市公司比未违规的上市公司倾向于拥有更高的递延所得税资产异动额。图 3-2 显示了两对箱线图。第一对根据 il 变量的两个状态显示 ROA 的分布，第二对是类似的做法，表示的是 abDTA 的分布图形。这一章，我们将学习如何通过建立模型，使用任意给定

的变量递延所得税资产异动 abDTA（X_1）与资产收益率 ROA（X_2）来预测违规状态 il（Y）。由于变量 Y 不是定量的，所以第 2 章介绍的线性回归模型就不适用了。

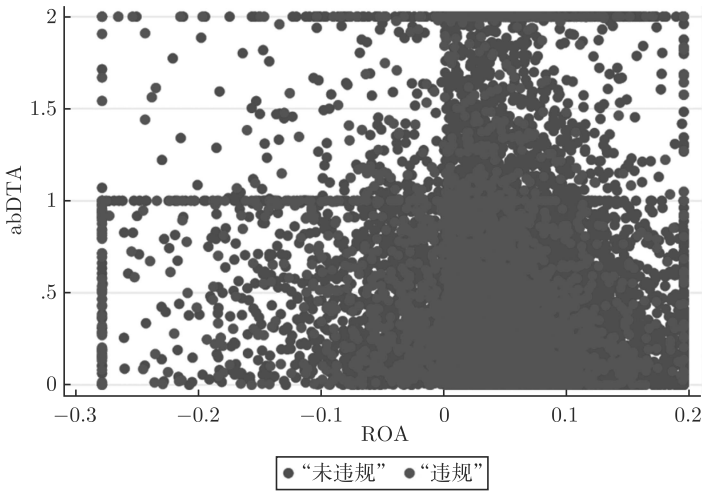
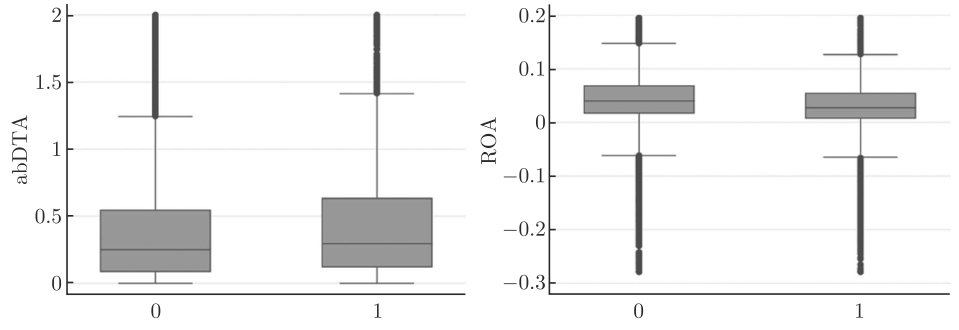


图 3-1（彩色）

图 3-1　财务违规（il）数据集[①]



（a）财务违规(il)关于递延所得税资产异动(abDTA)函数的箱线图

（b）财务违规(il)关于总资产收益率(ROA)函数的箱线图

图 3-2　箱线图

值得注意的是，从图 3-1 中并没有非常直观地显示预测变量 abDTA 和因变量 il 关系显著。在大多数应用中，如果仅从线性规律的角度看，预测变

① 上市公司的递延所得税资产异动与总资产收益率关系。财务违规标示为红色，未违规标示为蓝色。

量和因变量之间的关系并没有很强，那么就需要根据具体问题从不同的维度深入分析。

3.2 为什么线性回归不可用

上一节已经说明线性回归在定性因变量的情况下是不适用的，但为什么会这样呢？我们进行深入分析。假设我们现在要通过一个上市公司的财务状况和交易数据来预测其股价在未来（一周）是否上涨。在这个简化的例子中，有三种可能的情况：上涨、平盘和下跌。考虑用一个定量的因变量 Y 对这些值编码，如式（3.1）所示：

$$Y=\begin{cases}2, & \text{如果一周后股价上涨；}\\ 1, & \text{如果一周后股价不变；}\\ 0, & \text{如果一周后股价下跌。}\end{cases} \tag{3.1}$$

根据这些编码，结合一系列预测变量 $X_1,\cdots,X_p$，我们可以通过最小二乘法建立线性回归模型。这样做的一个问题是，编码实际默认了一个有序的输出，“如果一周后股价不变”放在“如果一周后股价上涨”和“如果一周后股价下跌”之间，表明不变和上涨的差距与下跌和不变的差距是一样的。实际上并没有特别的原因表明必须这样。比如，可以另选一个同样合理的编码表示如下：

$$Y=\begin{cases}1, & \text{如果一周后股价上涨；}\\ 0, & \text{如果一周后股价不变；}\\ -1, & \text{如果一周后股价下跌。}\end{cases} \tag{3.2}$$

如上编码给出了一个完全不同的三值关系。每种编码都会产生完全不同的线性模型，导致测试观测产生不同的预测结果。如果因变量值确实存在一种自然的程度排序，比如温和、中等和剧烈，其中温和和中等之间的程度差距与中等和剧烈之间的程度差距是相近的，那么 0、1、2 编码是合理的。需要注意的是，对一个二元（binary）的定性因变量而言，这样做不会对结果有影响。例如，我们考虑股价一周后是否上涨：上涨和不上涨。那么，就可

以利用哑变量（dummy variable）方法将相应变量编码，具体如下：

$$Y=\begin{cases}1, & \text{如果一周后股价上涨；}\\ 0, & \text{如果一周后股价未上涨。}\end{cases} \tag{3.3}$$

然后对二元因变量建立合适的线性回归模型，如果 $Y>0.5$，那么就预测 1，反之则为 0。在二元的情况下，不难证明，即使调换编码的顺序，线性回归最后依然会产生相同的预测。对一个如上 0/1 编码的二元因变量，最小二乘法对应的线性回归是有意义的。在本例中，由线性回归得到的 $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 实际上是 $P(\text{上涨}|X)$ 的估计。如果使用线性回归，这个估计值可能会在 $[0,1]$ 范围外（见图 3-3），这个数值很难被当作概率来解释数据。尽管此估计值可视为一个预测概率大小顺序的粗略估计，其仍然有一定的解释力。但是对两个以上分类的定性因变量，哑变量的方法不能任意推广。因此，找到一种真正适合分析定性因变量的方法才是合理的。

3.3　逻 辑 回 归

这一部分，我们分析财务违规数据集来讲解逻辑回归。[10] 数据集中因变量 il（上市公司在某年是否被查证违规）只取两个值 Yes（违规）或 No（不违规）。逻辑回归对 il 属于某一类的概率建模而不直接对因变量 il 建模。对财务违规数据而言，我们可以用逻辑回归建立违规概率模型。例如，给定 abDTA 时，可以记为：

$$P\left(\text{il}=\text{Yes}|\text{abDTA}\right) \tag{3.4}$$

$P(\text{il}=\text{Yes}|\text{abDTA})$ 的值，简记为 $p(\text{abDTA})$，取值范围在 0 到 1 之间。那么任给一个 abDTA 值，就可以根据这个概率对 il 进行预测。例如，如果某个企业的 $p(\text{abDTA})>0.5$，可以预测这个企业的 $\text{il}=\text{Yes}$。另一方面，如果希望对预测一家企业是否发生违规持谨慎态度，那么预测模型应该选择一个更低的阈值，比如 $p(\text{abDTA})>0.1$。

3.3.1　逻辑回归模型

那么该怎样建立 $p(X)=P(Y=1|X)$ 之间的关系呢？为方便起见，本节的因变量按常规 0/1 编码取值。之前的章节里面，我们已经讨论过使用线性

回归模型表示这些概率。

根据该方法用变量 abDTA 预测 il = Yes 的概率，结果如图 3-3 所示：在图 3-3 左侧图中，当递延所得税异动指标足够大时，会产生一个负的财务违规概率；同样，观察图 3-3 右侧图可知，总资产收益率超过一定值的时候，模型会得出上市公司违规概率为负数。这类预测值是没有意义的，因为无论递延所得税异动指标取何值，正确的财务违规概率值一定是落在 0 到 1 之间。这个问题不只是出现在上市企业财务违规数据上，其他的分类问题也类似。用一条直线拟合一个编码为 0,1 的二元因变量，原则上总可以找到 X 的一些值，使预测的 $p(X)<0$，而对 X 的另一些值 $p(X)>1$（除非 X 的范围是限定的）。

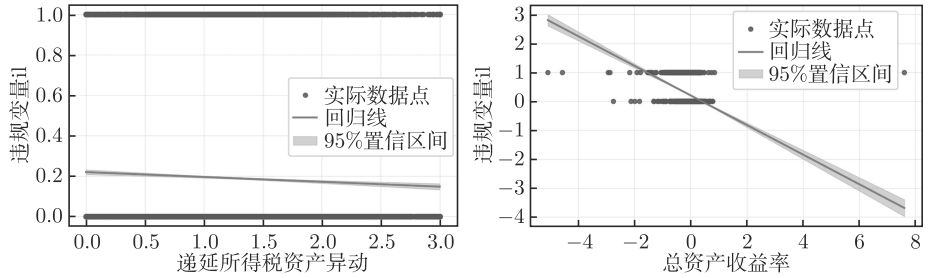


图 3-3　用 OLS 拟合违规变量 il[①]

为避免这类问题，我们必须找到一个函数建立针对 $p(X)$ 的模型，使对任意 X 值该函数的输出结果都在 0 和 1 之间。有许多函数满足这项要求。在逻辑回归中，使用逻辑函数（logistic function）[11]，

$$p(X)=\frac{1}{1+\exp\left(-\left(\beta_0+\beta_1 X\right)\right)} \tag{3.5}$$

通过整理，可得

$$\frac{p(X)}{1-p(X)}=\exp\left(\beta_0+\beta_1 X\right) \tag{3.6}$$

$p(X)/[1-p(X)]$ 的值称为发生比（优势比，odds ratio），取值范围为 0 到 ∞，其值接近于 0 表示违规概率非常低，接近于 ∞ 则表示违规概率非常高。

对式（3.6）两边同时取对数，得到

① 用线性回归估计财务违规（il）概率，在其中一些数据上预测的概率为负值。

$$\log\left(\frac{p(X)}{1-p}\right)=\beta_0+\beta_1 X \tag{3.7}$$

等式的左边称为对数发生比（优势比）（log-odds）或分对数（logit），于是，逻辑回归模型（3.5）可以视为分对数变换下关于 X 的一个线性模型。

在一个线性模型中，β_1 表示 X 值每增加一个单位时因变量的变化量。相比之下，在一个逻辑回归模型中，X 每增加一个单位，对数发生比的变化为 β_1（式 3.7）或者说发生比要乘以 e^{β_1}，见式（3.6）。但是，在式（3.5）中 $p(X)$ 和 X 的关系并不是线性的，β_1 不是当 X 增加一个单位时 $p(X)$ 的变化量，$p(X)$ 随 X 增加一个单位的改变量取决于 X 现在的取值。但是如果不考虑 X 的取值，若 β_1 值是正的，$p(X)$ 则随 X 的增加而增加，若 β_1 值是负的，$p(X)$ 则随 X 的增加而减少。

3.3.2　估计回归系数

式（3.5）中的系数 β_0、β_1 是未知的，必须通过有效的训练数据估计这些参数。虽然也可以用（非线性）最小二乘拟合模型（3.7），但由于极大似然有更好的统计性质，所以一般采用极大似然方法估计系数。[12] 极大似然法拟合逻辑回归模型的基本思想是：寻找 β_0,β_1 的一个估计，使得由式（3.6）得到的每个企业的财务违规概率 $\hat{p}(x_i)$ 最大可能地与财务违规的观测情况接近。换句话说，求出的 $\hat{\beta}_0,\hat{\beta}_1$ 估计，代入式（3.5）给出的模型中，使所有违规企业的值接近于 1，而未违规企业的值接近于 0。这个思想可以表达为数学方程的似然函数（likelihood function），其形式如下：

$$l(\beta_0,\beta_1)=\prod_{\{i/y_i=1\}}p(x_i)\prod_{\{j/y_i=0\}}(1-p(x_j)) \tag{3.8}$$

所估计的系数 $\hat{\beta}_0,\hat{\beta}_1$ 应使得似然函数值最大。在线性回归下，最小二乘法实际是极大似然方法的特例。极大似然的数学细节在本书不做详解。

表 3-1 展示了在控制行业与年份后，向 Logistic 模型中加入了财务、公司治理和市场变量等控制变量的回归结果。由于篇幅限制，此处仅列出递延所得税异动指标（abDTA）对应的回归系数。回归结果中，abDTA 的系数为 0.225，在 1%水平下显著，这表明，递延所得税资产的异常变动值越大，相应公司当期违规概率越高。回归结果中 abDTA 对应的优势比为 1.253，表

明当 abDTA 增加 1 时，当期存在违规事件的优势比提升 25.3%。表 3-1 中逻辑回归模型的输出结果与第二章线性回归输出的结果是类似的。例如，系数估计的准确性可通过计算标准误来衡量。表 3-1 中的统计量和线性回归模型输出的统计量的作用是一样的，如 β_1 统计量等于 $\hat{\beta}_1/\mathrm{SE}(\hat{\beta}_1)$，当 z 统计量的绝对值很大时说零假设 $H_0 : \beta_1 = 0$ 不成立, 表示 il 财务违规概率不依赖于 abDTA，由于 abDTA 的 p 值很小，因此拒绝 H_0。零假设也就是 $p(X) = \dfrac{\mathrm{e}^{\beta_0}}{1+\mathrm{e}^{\beta_0}}$, 表示 il 财务违规概率不依赖于 abDTA，由于 abDTA 的 p 值很小，因此拒绝 H_0，表明 il 财务违规概率与 abDTA 之间确实存在关系。

表 3-1　财务违规数据示例[①]

解释变量	abDTA	稳健标准差	财务变量	公司治理变量	市场变量	行业和年份	观察数
il	0.225	0.091	控制	控制	控制	控制	8651

3.3.3　分类预测

模型中的系数估计结束后，对任意给定的递延所得税异动指标（abDTA）及其他变量计算 il 违规概率就比较容易了。例如，用表 3-1 中的系数估计，当某个企业的递延所得税异动指标（abDTA）为 0.1819 时，预测该企业违规的概率为

$$\hat{p}(X) = \frac{1}{1+\exp\left(-\left(\hat{\beta}_0+\hat{\beta}_1 X\right)\right)} = 0.5734 \tag{3.9}$$

可见企业违规概率大于 50%。在预测是否违规时我们设置阈值，如设置为 0.5，则 0.57 的违规概率可以分类为违规；若阈值设置为 0.6，则 0.57 的违规概率可以分类为不违规。

3.3.4　多元逻辑回归

现在我们考虑预测一个二元因变量受多个自变量（因素）影响的情况。类似于第 2 章推广至多元线性回归的情况，可以对式（3.7）做如下推广：

① 建立用 abDTA 预测 il 概率的逻辑回归模型的系数估计。abDTA 每增加一个单位，il 对数发生比增加 0.253 个单位。

$$\log\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p \tag{3.10}$$

这里 $\boldsymbol{X}=(X_1,\cdots,X_p)$ 是 p 个预测变量。方程（3.10）可以重新写成：

$$p(\boldsymbol{X})=\frac{1}{1+\exp\left(-\left(\beta_0+\beta_1X_1+\cdots+\beta_pX_p\right)\right)} \tag{3.11}$$

和 3.3.2 节一样，我们可以根据极大似然方法估计出 $\beta_0,\beta_1,\cdots,\beta_p$。

使用一个预测变量做逻辑回归时，如果其他预测变量与之有关系，那么预测模型会存在风险。与线性回归一样，只用一个预测变量得到的结果可能与多个预测变量得到的结果完全不一样，尤其是当这些因素之间存在相关性时更是如此。

有时因变量取值多于两类。比如，我们可以将股市指数明天的涨幅分为三类：1，对应上涨 1%以上；0，对应上涨或下跌幅度小于 1%；−1，对应下跌 1%以上。逻辑回归模型可以推广到多分类问题，但实际应用中并不常用。[13] 后续我们将在树模型中介绍更多算法。

3.3.5　代码示例

不同编程基础的同学可以根据自己的需求快速阅读此部分，或者使用代码进行直接分析。[14]

1. 导入数据

在 IPO 的例子中，我们准备了美国 1980 年至 2019 年的数据，存放在程序相应的文件夹的数据中，路径为 path='./data/book_ipo.csv'。为了读取数据，我们需要导入 python 程序包 pandas，对应的代码为：

```
import pandas as pd
```

读取数据的代码为：

```
data=pd.read_csv('./data/book_ipo.csv',index_col=0)
```

我们可以查看 ir、nasdaq15_pos、lrassets、vc 和 odate 这几列数据，其中 ir 为美股上市公司上市首日的涨幅（%），nasdaq15_pos 为纳斯达克指数过去 15 日的收益率是否为正数的虚拟变量，lrassets 为上市公司招股书中

的资产的自然对数，vc 为上市公司 IPO 时是否有风投机构持股的虚拟变量，odate 为上市当年的年份。对应的代码为：

```
data[['ir','nasdaq15_pos', 'lrassets','vc','odate']]
```

其他所有的变量列名可以通过命令查看：

```
print(data.columns)
```

2. 变量转换

因为 IPO 首日涨幅本身是定量变量，我们为了展示逻辑回归的分类效果，需要将其转换为二元变量，如涨幅大于 0 为 1，其他为 0，对应代码如下：

```
data['ir_g2']=1*(data['ir']>0)
```

同样，我们可以生成三元变量，如涨幅大于 10%为 1，小于 −10%为 −1，中间的幅度为 0，对应的代码如下：

```
data['ir_g3']=1*(data['ir']>10)-1*(data['ir']<-10)
```

3. logistic 回归——二元

接下来，我们生成自变量 X 和因变量 Y，对应的代码如下：

```
X = data[['nasdaq15_pos', 'lrassets','vc']]
Y = data['ir_g2']
```

这里，我们只是使用三个示例自变量做展示，其他变量读者可以自行测试。我们使用 statsmodels 中的逻辑回归程序包，导入代码如下：

```
import statsmodels.api as sm
from statsmodels.discrete.discrete_model import Logit,
    Probit, MNLogit
```

之后就可以进行逻辑回归，代码如下：

```
logist_model = Logit(Y, sm.add_constant(X))
result = logist_model.fit()
```

此处的 fit() 代表数据拟合过程，result 为回归结果，可以通过以下代码查看：