第10章：含定性变量的回归模型

马学俊(主讲) 杜悦(助教)

苏州大学

数学科学学院

https://xuejunma.github.io/

Outline

自变量含定性变量的回归模型

- 简单情况

首先讨论定性变量只取两类可能值的情况，例如研究粮食产量问题，y为粮食产量，x为施肥量，另外再考虑气候问题，分为正常年份和干旱年份两种情况，对这个问题的数量化方法是引入一个$0-1$型变量D，令：

$$D_i = 1\text{表示正常年份}$$

$$D_i = 0\text{表示干旱年份}$$

粮食产量的回归模型为

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \epsilon_i$$

干旱年份的粮食平均产量为

$$E(y_i|D_i = 0) = \beta_0 + \beta_1 x_1$$

正常年份的粮食平均产量为：

$$E(y_i|D_i = 1) = (\beta_0 + \beta_2) + \beta_1 x_1$$

假设条件：干旱年份和正常年份回归直线的斜率β_1相等

自变量含定性变量的回归模型

- 复杂情况

某些场合定性自变量可能取多类值，例如某商厦策划营销方案，需要考虑销售额的季节性影响，季节因素分为春、夏、秋、冬4种情况。为了用定性自变量反应春、夏、秋、冬四季，我们初步设想引入如下4个$0-1$自变量：

$$\begin{cases} x_1 = 1, \text{春季} \\ x_1 = 0, \text{其他} \end{cases} \quad \begin{cases} x_2 = 1, \text{夏季} \\ x_2 = 0, \text{其他} \end{cases}$$

$$\begin{cases} x_3 = 1, \text{秋季} \\ x_3 = 0, \text{其他} \end{cases} \quad \begin{cases} x_4 = 1, \text{冬季} \\ x_4 = 0, \text{其他} \end{cases}$$

可是这样做却产生了一个新的问题，即$x_1 + x_2 + x_3 + x_4 = 1$，构成完全多重共线性。解决这个问题的方法很简单，我们只需去掉一个$0-1$型变量，只保留3个$0-1$型自变量即可。例如去掉x_4，只保留x_1、x_2、x_3。

对一般情况，一个定性变量有k类可能的取值时，需要引入$k-1$个$0-1$型自变量。当$k=2$时，只需要引入一个$0-1$型自变量即可。

分段回归

例10.2 表10.2给出某工厂生产批量x与单位成本y(美元)的数据。试用分段回归建立回归模型。

- 单位成本y对生产批量的回归
  - 在x_p点内服从一种线性回归
  - 在x_p点外服从另一种线性回归

例10.2

```
rm(list=ls())
x <- c(650,340,400,800,300,570,720,480)
y <- c(2.57,4.40,4.52,1.39,4.75,3.55,2.49,3.77)
plot(x, y)
```

- 由图可看出数据在生产批量$x_p = 500$时发生较大变化，即批量大于500时成本明显下降。

分段回归

我们考虑由两段构成的分段线性回归,这可以通过引入一个$0-1$型虚拟自变量实现。假定回归直线的斜率在$x_p = 500$处改变，建立回归模型

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - 500) D_i + \epsilon_i$$

来拟合，其中

$$\begin{cases} D_i = 1, \text{当} x_i > 500 \\ D_i = 0, \text{当} x_i \leq 500 \end{cases}$$

引入两个新的自变量

$$x_{i1} = x_i, \qquad x_{i2} = (x_i - 500) D_i$$

这样回归模型转化为标准形式的二元线性回归模型：

$$y_i = \beta_0 + \beta_1 x_{i1} +]beta_2 x_{i2} + \epsilon_i \tag{10.3}$$

（10.3）式可以分解为两个线性回归方程：

分段回归

当$x_1 \leq 500$时，

$$E(y) = \beta_0 + \beta_1 x_1 \tag{10.4}$$

当$x_1 > 500$时，

$$E(y) = (\beta_0 - 500\beta_2) + (\beta_1 + \beta_2)x_1 \tag{10.5}$$

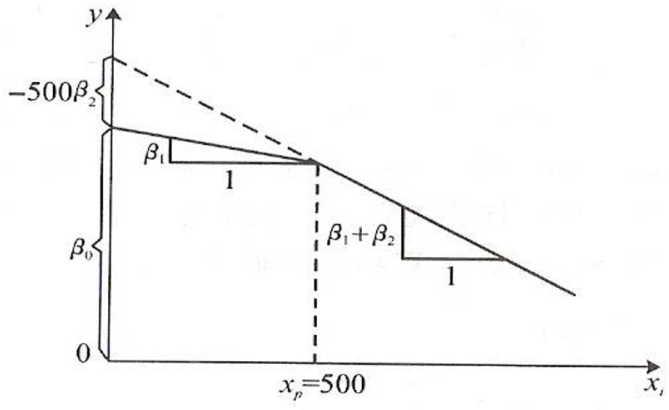


图10.2

用普通最小二乘法拟合模型(10.3)式得回归方程为：

$$\hat{y} = 5.895 - 0.00395x_1 - 0.00389x_2$$

利用此模型可说明生产批量小于500时，每增加1个单位批量，单位成本降低0.00395美元；当生产批量大于500时，每增加1个单位批量，估计单位成本降低到$0.00395 + 0.00389 = 0.00784$(美元)。

回归系数相等的检验

例10.3 回到例10.1的问题，例10.1引入$0-1$型自变量的方法是假定储蓄增加额y对家庭收入的回归斜率β_1与家庭文化程度无关，家庭文化程度只影响回归常数项β_0，这个假设是否合理，还需要做统计检验。检验方法是引入如下含有交互效应的回归模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \tag{10.8}$$

其中y为上一年家庭储蓄增加额，x_1为上一年家庭总收入，x_2表示家庭学历，高学历家庭$x_2 = 1$,低学历家庭$x_2 = 0$。

回归模型（10.8）式可以分解为对高学历和对低学历家庭的两个线性回归模型，分别为：

高学历家庭$x_2 = 1$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1} + \epsilon_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} + \epsilon_i \tag{10.9}$$

低学历家庭$x_2 = 0$

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \tag{10.10}$$

回归系数相等的检验

可见，高学历回归家庭的回归常数为$\beta_0 + \beta_2$，回归系数为$\beta_1 + \beta_3$。高学历回归家庭的回归常数为β_0，回归系数为$\beta_1 0$。要检验两个回归方程的回归系数是否相等，等价于对回归模型式(10.8)做参数假设检验

$$H_0 : \quad \beta_3 = 0$$

当拒绝H_0时，认为$\beta_3 \neq 0$，这时高学历与低学历家庭的储蓄回归模型实际上被拆分为两个不同的回归模型(10.9)和(10.10)式。

当不拒绝H_0时，认为$\beta_3 = 0$，这时高学历与低学历家庭的储蓄回归模型是如下形式的联合回归模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

定性因变量的回归方程的意义

在许多社会经济问题中，所研究的因变量往往只有两个可能结果，这样的因变量也可用虚拟变量来表示，虚拟变量的取值可取0或1。

定性因变量的回归方程的意义

一、定性因变量的回归方程的意义

设因变量y是只取$0,1$两个值的定性变量，考虑简单线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \tag{10.12}$$

在这种y只取$0,1$两个值的情况下，因变量均值$E(y_i) = \beta_0 + \beta_1 x_i$有着特殊的意义。

由于y_i是$0-1$型贝努利随机变量，得如下随机分布

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

根据离散型随机变量期望的定义，可得

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$

进而得到

$$E(y_i) = \pi_i = \beta_0 + \beta_1 x_1$$

所以，作为由回归函数给定的因变量均值，$E(y_i) = \beta_0 + \beta_1 x_i$是自变量水平为$x_i$时$y_i = 1$的概率。对因变量均值的这种解释既适用于这里的简单线性回归函数，也适用于复杂的多元回归函数，当因变量为$0-1$变量时，因变量均值总是代表给定自变量时$y = 1$的概率。

定性因变量回归的特殊问题

(1) 离散非正态误差项。

对一个取值为0和1的因变量，误差项$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$只能取两个值：

当$y_i = 1$时，$\epsilon_i = 1 - (\beta_0 + \beta_1 x_i) = 1 - \pi_i$

当$y_i = 0$时，$\epsilon_i = -(\beta_0 + \beta_1 x_i) = -\pi_i$

显然，误差项ϵ_i是两点型离散分布，当然正态误差回归模型的假定就不适用了。

(2) 零均值异方差性。

当因变量是定性变量时，误差项ϵ_i仍然保持零均值，这时出现的另一个问题是误差项ϵ_i的方差不相等。$0-1$型随机变量ϵ_i的方差为

$$\begin{aligned} D(\epsilon_i) &= D(y_i) \\ &= \pi_i(1-\pi_i) \\ &= (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i) \end{aligned}$$

ϵ_i的方差依赖于x_i，是异方差，不满足线性回归方程的基本假定。

定性因变量回归的特殊问题

(3) 回归方程的限制

当因变量为$0-1$虚拟变量时，回归方程代表概率分布，所以因变量均值受到如下限制：

$$0 \leq E(y_i) = \pi \leq 1$$

对一般的回归方程本身并不具有这种限制，线性回归方程$y_i = \beta_0 + \beta_1 x_i$将会超出这个限制范围。

对于普通的线性回归所具有的上述3个问题，虽然可以找到一些相应的解决办法。例如，对于误差项不是正态的情形，最小二乘法求得的无偏估计量在绝大多数情况下是渐近正态的。因此，当样本容量较大时，未知参数的估计与误差项假设为正态分布时的方式相同；对于异方差情况，可以用加权最小二乘法来处理；对受回归方程限制的情况，对模型范围内的x来说，可以通过确保拟合模型的因变量均值不小于0和不大于1来处理。但是这些并不是从根本上解决问题的办法，为了从根本上解决问题，我们需要构造一个自动满足以上限制的模型来处理。

分组数据的Logistic回归模型

针对$0-1$型因变量产生的问题，我们对回归模型应该做两个方面的改进。

第一，回归函数应该改用限制在$[0,1]$区间内的连续曲线，而不能再沿用直线回归方程。限制在$[0,1]$区间内的连续曲线有很多，例如所有连续型随机变量的分布函数都符合要求，我们常用的是Logistic函数与正态分布函数。Logistic函数的形式为

$$f(x)=\frac{e^x}{1+e^x}=\frac{1}{1+e^{-x}}$$

Figure: $\frac{1}{1+e^{-x}}$

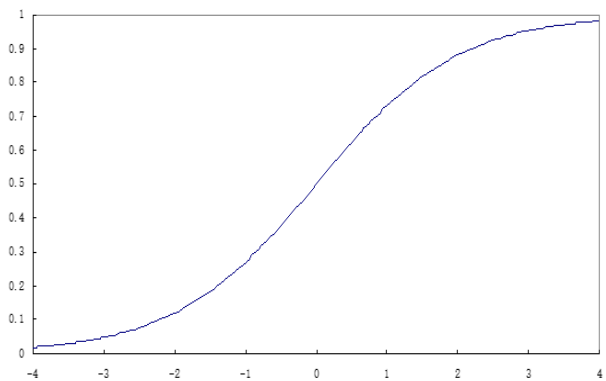
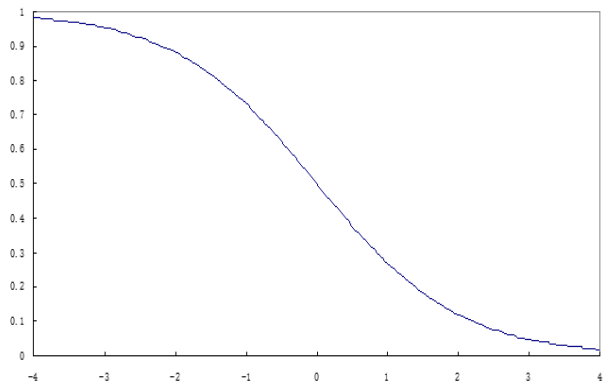


Figure: $\frac{1}{1+e^{x}}$



分组数据的Logistic回归模型

假设Y仅有两个状态0和1
假定概率$p = P(y = 1)$与p个因素$x_1, \cdots, x_p$有关则称

$$\ln(\frac{p}{1-p}) = g(x_1, \cdots, x_p)$$

为二值Logistic回归模型。最重要的Logistic回归模型是Logistic线性回归模型

$$\ln(\frac{p}{1-0}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

可以得到优势比$\frac{P(y=1)}{P(y=0)} = \frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$ 从而得到概率p的计算公示

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

分组数据的Logistic回归模型

第二，因变量y_i本身只取0、1两个离散值，不适于直接作为回归模型中的因变量。

由于回归函数$E(y_i) = \pi_i = \beta_0 + \beta_1 x_i$表示在自变量为$x_i$的条件下$y_i$的平均值，而$y_i$是$0-1$型随机变量，因而$E(y_i) = \pi_i$就是在自变量为$x_i$的条件下$y_i$ 等于1的比例。这提示我们可以用y_i 等于1的比例代替y_i本身作为因变量。

未分组数据的Logistic回归模型

设y是$0-1$型变量，$x_1, x_2, \cdots, x_p$是与y相关的确定性变量，n组观测数据为$(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i), i = 1, 2, \cdots, n$，$y_i$与$x_{i1}, x_{i2}, \cdots, x_{ip}$的关系为：

$$E(y_i) = \pi_i = f(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

其中函数$f(x)$是值域在$[0, 1]$区间内的单调增函数。对于Logistic回归

$$f(x) = \frac{1}{1 + e^x}$$

y_i服从均值为$\pi_i = f(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$的$0-1$型分布，概率函数为：

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

未分组数据的Logistic回归模型

可以把y_i的概率函数合写成：

$$P(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \quad y_i = 0, 1; \quad i = 1, 2, \cdots, n$$

于是，$y_1, y_2, \cdots, y_n$的似然函数为

$$L = \prod_{i=1}^{n} P(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

对数似然取自然对数，得

$$\ln L = \sum_{i=1}^{n}[y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)] = \sum_{i=1}^{n}[y_i \ln \frac{\pi_i}{1-\pi_i} + \ln(1-\pi_i)]$$

未分组数据的Logistic回归模型

将$\pi_i = \frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$代入得

$$\ln L = \sum_{i=1}^{n}\{y_i(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p) - \ln[1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)]\}$$

极大似然估计就是选取$\beta_0, \beta_1, \cdots, \beta_p$的估计值使上式达极大。

Probit回归模型

Probit回归称为单位概率回归，与Logistic回归相似，也是拟合$0-1$型因变量回归的方法，其回归函数是

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

用样本比例p_i代替概率π_i，表示样本回归模型

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

多类别Logistic回归

当定性因变量y取k个类别时，记为$1,2,\cdots,k$。这里的数字$1,2,\cdots,k$只是名义代号，并没有大小顺序的含义。因变量y取值于每个类别的概率与一组自变量$x_1,x_2,\cdots,x_p$ 有关，对于样本数据$(x_{i1},x_{i2},\cdots,x_{ip};y_i)$，多类别Logistic回归模型第$i$组样本的因变量$y_i$取第$j$个类别的概率为：

$$\pi_{ij}=\frac{\exp(\beta_{0j}+\beta_{1j}x_{i1}+\cdots+\beta_{pj}x_{ip})}{\exp(\beta_{01}+\beta_{11}x_{i1}+\cdots+\beta_{p1}x_{ip})+\cdots+\exp(\beta_{0k}+\cdots+\beta_{pk}x_{ip})}$$
$$i=1,2,\cdots,n;\quad j=1,2,\cdots,k$$

上式中各回归系数不是惟一确定的，每个回归系数同时加减一个常数后的数值保持不变。为此，把分母的第一项中的系数都设为0，得到回归函数的表达式

$$\pi_{ij}=\frac{\exp(\beta_{0j}+\beta_{1j}x_{i1}+\cdots+\beta_{pj}x_{ip})}{1+\exp(\beta_{02}+\beta_{12}x_{i1}+\cdots+\beta_{p2}x_{ip})+\cdots+\exp(\beta_{0k}+\cdots+\beta_{pk}x_{ip})}$$
$$i=1,2,\cdots,n;\quad j=1,2,\cdots,k$$

这个表达式中每个回归系数都是唯一确定的，第一个类别的回归系数都取做0，其他类别回归系数数值的大小都以第一个类别为参照。

多类别Logistic回归

当定性因变量y取k个类别时，记为$1,2,\cdots,k$。这里的数字$1,2,\cdots,k$仅表示顺序的大小。因变量y取值于每个类别的概率仍与一组自变量$x_1,x_2,\cdots,x_p$有关，对于样本数据$(x_{i1},x_{i2},\cdots,x_{ip};y_i)$，$i=1,2,\cdots,n$，顺序类别回归模型有两种主要类型，一种是位置结构(分量)（Location component）模型，另一种是规模结构(分量)（Scale component）模型。

(1) 位置结构模型

$$link(\gamma_{ij})=\theta_j-(\beta_1x_{i1}+\beta_2x_{i2}+\beta_px_{ip}) \qquad (10.36)$$

其中，$link(\cdot)$是联系函数；$\gamma_{ij}=\pi_{i1}+\cdots+\pi_{ij}$是第$i$个样品小于等于$j$个的累积概率，由于$\gamma_{ik}=1$，所以式(10.36)只针对$i=1,2,\cdots,n;$ $j=1,2,\cdots,k-1$。θ_j是类别界限值(threshold)

多类别Logistic回归

(2) 规模结构模型

$$link(\gamma_{ij}) = \frac{\theta_j - (\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_p x_{ip})}{\exp(\tau_1 z_{i1} + \tau_2 z_{i2} + \cdots + \tau_m z_{im})} \qquad (10.36)$$

其中，$z_1, z_2, \cdots, z_m$是$x_1, x_2, \cdots, x_p$的一个子集，作为规模结构解释变量。

联系函数的几种主要类型

联系函数类型	形式	应用场合
Logit	log(γ / (1-γ))	各类别均匀分布
Complementary log-log	log(-log(1-γ))	高层类别出现几率大
Negative log-log	-log(-log(γ))	低层类别出现几率大
Probit	$\Phi^{-1}(\gamma)$	正态分布
Cauchit (inverse Cauchy)	tan(π(γ-0.5))	两端的类别出现几率大