广义线性模型与收缩估计量

司继春

上海对外经贸大学

2023年10月

二元选择模型

二元选择模型用于解决被解释变量y只有两个取值的情况：

- 是否选择去上大学（上大学=1，不上大学=0）
- 某个国家是否爆发内战（爆发=1，不爆发=0）
- 企业是否选择出口（出口=1，不出口=0）
- ……

针对这类问题，由于被解释变量只有两种取值（二元变量），因而统称为『二元选择模型（**binary choice model**）』

二元选择模型

如果y的取值范围$support\left(y\right)=\{0,1\}$，此时其条件期望函数：

$$\begin{aligned}\mathbb{E}\left(y|x\right)&=1\cdot P\left(y=1|x\right)+0\cdot P\left(y=0|x\right)\\&=P\left(y=1|x\right)\end{aligned}$$

即给定x，$y=1$的概率。由于概率在$[0,1]$区间范围以内，因而我们不能设定：

$$\mathbb{E}\left(y|x\right)=P\left(y=1|x\right)=x'\beta$$

因为$x'\beta$可能小于0或者大于1。

二元选择模型

- 我们可以使用一个分布函数$F(\cdot)$将$x'\beta$压缩到$[0,1]$范围以内：

$$\mathbb{E}(y|x) = P(y=1|x) = F\left(x'\beta\right)$$

- 如果$F(\cdot)$取标准的Logistic分布的分布函数，则称为Logistic回归或者Logit回归：

$$\mathbb{E}(y|x) = P(y=1|x) = \frac{e^{x'\beta}}{1+e^{x'\beta}}$$

Logistic回归系数的解释

我们将$y_i = 1$的概率与$y_i = 0$的概率的比值成为几率（**odds**）：

$$odds = \frac{P\left(y = 1|x, \beta\right)}{P\left(y = 0|x, \beta\right)}$$

- 几率度量了$y = 1$的概率相对于$y = 0$的概率的大小
  - 如果几率等于1，意味着$y = 1$的概率等于$y = 1$的概率
  - 如果几率大于1，意味着$y = 1$的概率大于$y = 0$的概率（比如$odds = 2$代表概率为2倍）
  - 如果几率小于1，意味着$y = 1$的概率小于$y = 0$的概率（比如$odds = 0.5$代表概率为一半）

Logistic回归

对于Logistic回归：

$$P\left(y=1|x\right)=\frac{e^{x'\beta}}{1+e^{x'\beta}},P\left(y=0|x\right)=\frac{1}{1+e^{x'\beta}}$$

从而几率（odds）为：

$$odds=\frac{P\left(y=1|x,\beta\right)}{P\left(y=0|x,\beta\right)}=\frac{\frac{e^{x'\beta}}{1+e^{x'\beta}}}{1-\frac{e^{x'\beta}}{1+e^{x'\beta}}}=e^{x'\beta}$$

从而对数几率（log odds，也称为logit）为：

$$logit=\ln\left(odds\right)=x'\beta$$

因而以上模型被称为对数几率回归（Logit regression）。

估计

- 估计：以上模型的条件密度函数为：

$$f\left(y_i|x_i,\beta\right)=\left[\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right]^{1\{y_i=1\}}\left[\frac{1}{1+e^{x_i'\beta}}\right]^{1\{y_i=0\}}$$

因而对数似然函数为：

$$L\left(\beta|y,x\right)=\sum_{i=1}^{N}\left[y_i\ln\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)+\left(1-y_i\right)\ln\left(\frac{1}{1+e^{x_i'\beta}}\right)\right]$$

最大化以上对数似然函数即可。

- 预测：

$$\widehat{\mathbb{E}\left(y|x\right)}=\hat{p}\left(x\right)=F\left(x'\hat{\beta}\right)$$

Probit模型

- 如果$F(\cdot)$取标准正态分布的分布函数，则称为Probit回归：

$$\mathbb{E}(y|x) = P(y=1|x) = \Phi(x'\beta) = \int_{-\infty}^{x'\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

其对数似然函数为：

$$L(\beta|y,x) = \sum_{i=1}^{N} \left[y_i \ln\left(\Phi\left(x_i'\beta\right)\right) + (1-y_i) \ln\left(1-\Phi\left(x_i'\beta\right)\right) \right]$$

Probit、Logit回归

论是Logit回顾还是Probit回归，其对数似然函数都可以写为：

$$L(\beta|y,x)=\sum_{i=1}^{N}\left[y_i\ln\left(F\left(x_i'\beta\right)\right)+(1-y_i)\ln\left(1-F\left(x_i'\beta\right)\right)\right]$$

其一阶条件为：

$$\begin{aligned}\frac{\partial L(\beta|y,x)}{\partial\beta}&=\sum_{i=1}^{N}\left[y_i\frac{1}{F(x_i'\beta)}f(x_i'\beta)x_i-(1-y_i)\frac{1}{1-F(x_i'\beta)}f(x_i'\beta)x_i\right]\\&=\sum_{i=1}^{N}\left[\frac{y_i}{F(x_i'\beta)}-\frac{(1-y_i)}{1-F(x_i'\beta)}f(x_i'\beta)x_i\right]\\&=\sum_{i=1}^{N}\left[\frac{y_i-F(x_i'\beta)}{F(x_i'\beta)\left[1-F(x_i'\beta)\right]}f(x_i'\beta)x_i\right]=0\end{aligned}$$

预测

- 无论是Logistic回归还是Probit回归，在得到β的估计$\hat{\beta}$后，我们可以将$\hat{\beta}$带入到分布函数中，得到概率的估计：

$$\hat{p}_i \triangleq P\left(\widehat{y_i = 1}|x_i\right) = F\left(x_i'\hat{\beta}\right)$$

- 从而使得对数似然函数最大时的对数似然函数值为：

$$L\left(\hat{\beta}|y,x\right) = \sum_{i=1}^{N}\left[y_i \ln\left(\hat{p}_i\right) + \left(1-y_i\right)\ln\left(1-\hat{p}_i\right)\right]$$

- 而如果假设没有其他解释变量的信息，只有常数项时，对于$P\left(y_i = 1\right)$的估计就是$\bar{y}$，而此时的似然函数值为：

$$L_0 = \sum_{i=1}^{N}\left[y_i \ln\left(\bar{y}\right) + \left(1-y_i\right)\ln\left(1-\bar{y}\right)\right]$$

Pseudo-R^2

类比于线性回归中的R^2，McFadden(1974)建议使用“伪R^2”（pseudo-R^2）：

$$R^2 = 1 - \frac{L\left(\hat{\beta}|y,x\right)}{L_0}$$
$$= 1 - \frac{\sum_{i=1}^{N}\left[1\left\{y_i = 1\right\}\ln\left(\hat{p}_i\right) + 1\left\{y_i = 0\right\}\ln\left(1 - \hat{p}_i\right)\right]}{N\left[\bar{y}\ln\bar{y} + \left(1 - \bar{y}\right)\ln\left(1 - \bar{y}\right)\right]}$$

作为拟合优度的度量。

评价分类的指标

- 对于所有分类任务：选定一个临界值c，并令预测值$\hat{y}_i = 1\{\hat{p}_i > c\}$，之后将样本分为四类：
  - 真正（True Positive， TP）：$y_i = 1, \hat{y}_i = 1$;
  - 假正（False Positive， FP）：$y_i = 0, \hat{y}_i = 1$;
  - 真反（True Negative， TN）：$y_i = 0, \hat{y}_i = 0$;
  - 假反（False Negative， FN）：$y_i = 1, \hat{y}_i = 0$。

评价分类的指标

使用以上四个分类分别定义：

- 查准率（precision），即所有预测为正的样本中，正确的比例：$Precision = \frac{TP}{TP+FP}$；
- 查全率（或者召回率，recall），即所有正的样本中，正确的比例：$Recall = \frac{TP}{TP+FN}$；
- 精度（accuracy），即所有样本中预测正确的比例：$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$。
- F1度量，即查准率和查全率的调和平均：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{1}{\frac{1}{2}\left(\frac{1}{Precision} + \frac{1}{Recall}\right)}$$

查准率和查全率之间通常存在着权衡：

- 比如，如果我们希望提高查准率，需减少预测为正的比例，需要较大的c，从而降低查全率。
- 使得查准率等于查全率的点称为平衡点（break-event point, BEP）。

ROC曲线

对于任意的临界值c，我们都可以定义：

- **敏感性（sensitivity）**：观察到的正的样本中，预测正确的比例，即

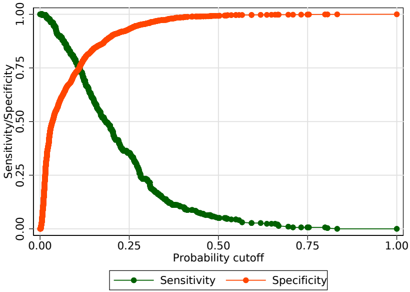
$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

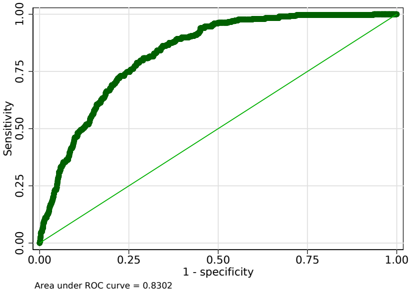
- **特异性（specificity）**：观察到的反的样本中，预测正确的比例，即

$$Specificity = \frac{TN}{TN + FP}$$

受试者工作特征曲线（receiver operating characteristic curve, ROC curve）：即当$c \in [0, 1]$时，以$1 - Specificity$作为横坐标，以$Sensitivity$作为纵坐标所画出来的图。

ROC曲线





(a)敏感性曲线、特异性曲线 （b） ROC曲线

（logit_and_roc.do）

广义线性模型

- 为了拟合条件期望函数，以上针对不同的因变量的数据形式，我们分别使用了：线性回归、Probit回归以及Logit回归
- 以上回归都是广义线性模型（generalized linear model）的特例
- 广义线性模型是估计条件期望的一个比较一般的方法，特别是针对不同形式的因变量：
  - 分类变量
  - 计数变量
  - 受限的support
  - ……

广义线性模型

为了构建广义线性模型，需要以下几个组成部分：

1. 首先将自变量x进行线性组合：$\eta_i = x_i'\beta$
2. 使用一个链接函数（link function）对y_i的条件期望进行建模：$\mu_i = \mathbb{E}\left(y_i|x_i\right) = \mu\left(x_i'\beta\right)$
3. 使用一个期望为μ_i的分布$\Psi\left(\cdot\right)$（属于指数分布族）将μ_i与y_i联系起来，即$y_i|x_i \sim \Psi\left(\mu_i\right)$。

广义线性模型

线性回归

如果我们假设链接函数为$\mu_i = \mu\left(x_i'\beta\right) = x_i'\beta$，而选取

$$y_i|x_i \sim N\left(\mu_i, \sigma^2\right)$$

（假设σ^2已知），我们知道正态分布属于指数分布族，因而以上就建立了一个广义线性模型。注意到，对于以上模型，我们可以定义：

$$u_i = y_i - \mu_i = y_i - \mathbb{E}\left(y_i|x_i\right) \sim N\left(0, \sigma^2\right)$$

或者：

$$y = x_i'\beta + u_i, u_i \sim N\left(0, \sigma^2\right)$$

从而我们得到了普通线性回归模型。

广义线性模型

Probit/Logit回归

如果我们假设$y_i = 0/1$为二元变量，取链接函数为

$$\mu_i = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}$$

并假设$y_i|x_i \sim Ber\left(\mu_i\right)$，由于Bernoulli分布也属于指数分布族，因而以上也建立了一个广义线性模型。注意到，其条件密度函数为：

$$f\left(y_i|x_i\right) = \left[\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right]^{y_i}\left[1-\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right]^{1-y_i}$$

也就是Logit回归。如果我们取链接函数为$\mu_i = \Phi\left(x_i'\beta\right)$即标准正态分布的分布函数，那么同理，我们就得到了Probit回归。

广义线性模型：常用分布和链接函数组合

分布	支撑集	链接函数名	链接函数
正态分布	$\mathbb{R}$	identity	$\mu_i = x_i'\beta$
指数分布	$(0, +\infty)$	negative inverse	$\mu_i = -\left(x_i'\beta\right)^{-1}$
Gamma分布			
泊松分布	$\mathbb{Z}$	log	$\mu_i = \exp\left(x_i'\beta\right)$
负二项分布			
伯努利分布	$\{0, 1\}$	logit/probit	$\mu_i = N \cdot \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}$
二项分布	$\{0, 1, ..., N\}$		$\mu_i = N \cdot \Phi\left(x_i'\beta\right)$

广义线性模型：Stata

Logistic回归也可以使用如下广义线性模型的命令：

```
glm exit_labor `x', family(binomial) link(logit)
```

其中：

- family 为指数分布族
- link为链接函数

广义线性模型：计数回归

计数回归

如果被解释变量y是计数数据，即取值范围为自然数集$\mathbb{Z}$，通常我们可以使用泊松分布对其进行建模：

$$y_i|\mu_i \sim P\left(\mu_i\right)$$

其中μ_i为$\mathbb{E}\left(y_i|x_i\right)$。考虑到$y_i$的取值范围为自然数，所以其期望不可能为负，为此在建模μ_i时，可以使用log链接函数，即：$\mu_i = \exp\left(x_i'\beta\right)$以上回归被称为泊松回归（**Possion regression**）

广义线性模型：计数回归

OHIE数据

在数据集“OHIE_QJE.dta”中，记录了一些人参与某项健康保险计划之后的身体健康数据，其中treatment变量为是否参加该计划，而“baddays_phys_12m”为参与该计划12个月以后一个月内身体不舒服的天数，所以是一个计数数据。我们可以使用glm命令或者possion命令对该回归进行计算：

```
use datasets/OHIE_QJE.dta
local x "treatment english_list female_list
    birthyear_list"
// 泊松回归
glm baddays_phys_12m `x', family(poisson) link(log
    )
// 与以下回归等价：
poisson baddays_phys_12m `x'
```

以上两条命令的结果是完全等价的。