

二元选择与广义线性模型

司继春

上海对外经贸大学

2024年10月

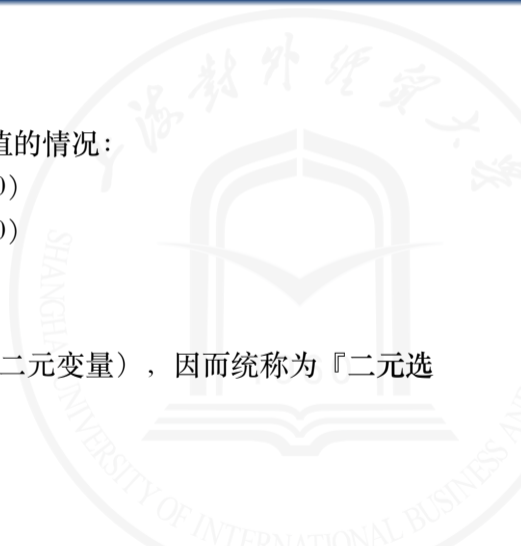


二元选择模型

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 $support(y) = \{0, 1\}$ ，此时其条件期望函数：

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

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：

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

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：

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

- 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) :

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

- 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 $y = 0$ 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)

几率比

几率比 (odds ratio)：两个不同组的几率的比率，如果等于1则代表两个组的 $y = 1$ 的概率相同。

几率比

我们使用CFPS数据计算了是否退出劳动力市场与性别的频数，并计算男性和女性退出劳动力市场的odds：

	女性	男性
未退出	1289	1722
退出	255	42
Odds	0.1978	0.0244

从而，女性退出劳动力市场的odds为男性的 $0.1978/0.0244 = 8.1$ 倍，女性退出劳动力市场的概率更高。

Logistic回归

对于Logistic回归：

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

从而几率 (odds) 为：

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

从而对数几率 (log odds, 也称为logit) 为：

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

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：

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

因而对数似然函数为：

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

最大化以上对数似然函数即可。

- 预测：

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

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：

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

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：

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

其一阶条件为：

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

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：

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

- 从而使得对数似然函数最大时的对数似然函数值为：

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

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

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

Pseudo- R^2

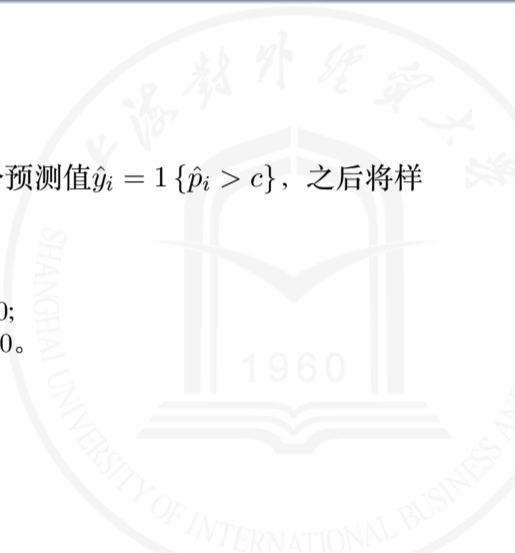
类比于线性回归中的 R^2 ，McFadden(1974)建议使用“伪 R^2 ” (pseudo- R^2)：

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

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：
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：
例： $Recall = \frac{TP}{TP+FN}$ ；
- 精度 (accuracy)，即所有样本中预测正确的比例：
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

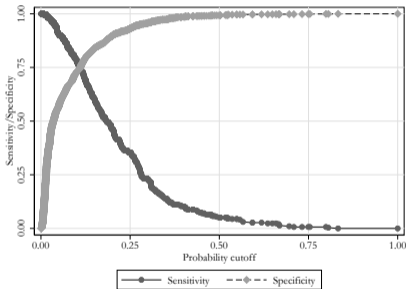
$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

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

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

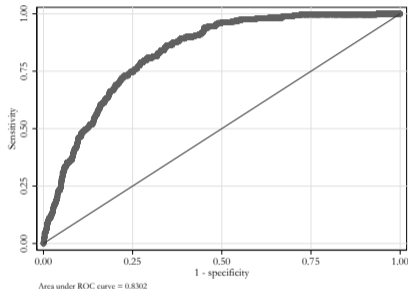
受试者工作特征曲线 (receiver operating characteristic curve, **ROC curve**)：即当 $c \in [0, 1]$ 时，以 $1 - Specificity$ 作为横坐标，以 $Sensitivity$ 作为纵坐标所画出来的图。

ROC曲线



(a) 敏感性曲线、特异性曲线

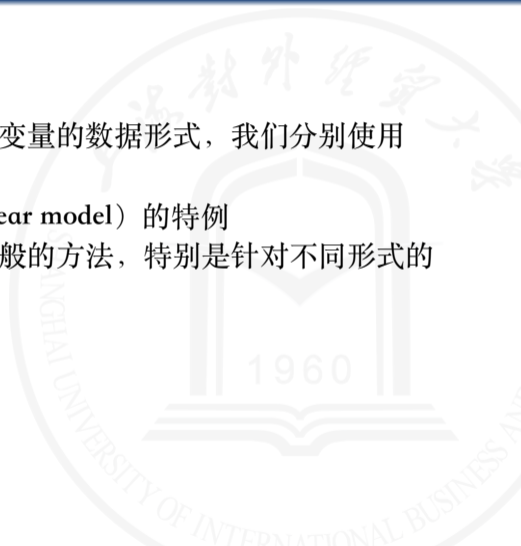
(logit_and_roc.do)



(b) ROC曲线

广义线性模型

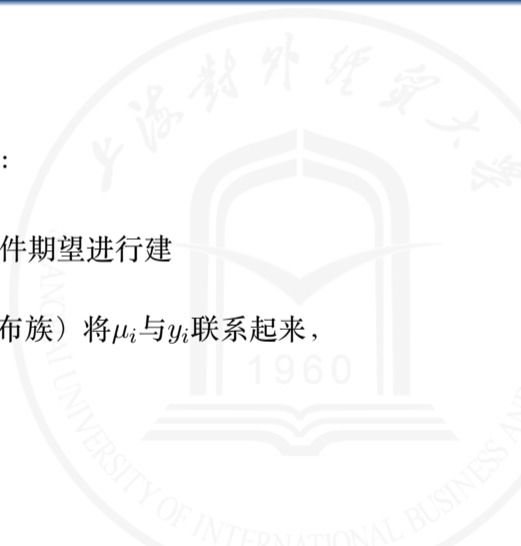
- 为了拟合条件期望函数，以上针对不同的因变量的数据形式，我们分别使用了：线性回归、Probit回归以及Logit回归
- 以上回归都是广义线性模型（**generalized linear model**）的特例
- 广义线性模型是估计条件期望的一个比较一般的方法，特别是针对不同形式的因变量：
 - 分类变量
 - 计数变量
 - 受限的support
 -



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： $\mu_i = \mathbb{E}(y_i | x_i) = \mu(x_i' \beta)$
- ③ 使用一个期望为 μ_i 的分布 $\Psi(\cdot)$ （属于指数分布族）将 μ_i 与 y_i 联系起来，即 $y_i | x_i \sim \Psi(\mu_i)$ 。



广义线性模型

线性回归

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

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

(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：

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

或者：

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

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

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

并假设 $y_i|x_i \sim Ber(\mu_i)$ ，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：

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

也就是Logit回归。如果我们取链接函数为 $\mu_i = \Phi(x_i'\beta)$ 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	$\mu_i = -(x_i' \beta)^{-1}$
Gamma分布			
泊松分布	\mathbb{Z}	log	$\mu_i = \exp(x_i' \beta)$
负二项分布			
伯努利分布	$\{0, 1\}$	logit/probit	$\mu_i = N \cdot \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)}$
二项分布	$\{0, 1, \dots, N\}$		$\mu_i = N \cdot \Phi(x_i' \beta)$

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：

```
1 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：

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

其中 μ_i 为 $\mathbb{E}(y_i | x_i)$ 。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： $\mu_i = \exp(x_i' \beta)$ 以上回归被称为泊松回归 (Poisson 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 `poisson` 命令对该回归进行计算：

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

以上两条命令的结果是完全等价的。