Logit模型的原理及应用

王裕 15320171151908

1.问题的提出

在实际经济问题中，被解释变量也可能是定性变量。因变量取值是离散的，这类回归模型称为离散选择模型或“定性反应模型”。例如通过一系列解释变量的观测值观察人们对某项提议的态度，某件事情的成功和失败等。这类模型被称为“离散选择模型”：二值选择模型、多值选择模型、计数模型。

2.线性概率模型（Tobit）

线性概率模型的形式如下：

$$y_i = \alpha + \beta x_i + u_i \tag{1}$$

其中u_i为随机误差项，x_i为定量解释变量，y_i为二元选择变量。设若是第一种选择$y_i = 1$，若是第二种选择$y_i = 0$。

对$y_i = \alpha + \beta x_i + u_i$取期望，

$$E(y_i) = \alpha + \beta x_i \tag{2}$$

下面研究y_i的分布。因为y_i只能取两个值，0和1，所以y_i服从两点分布。把y_i的分布记为，

$$P(y_i = 1) = p_i$$

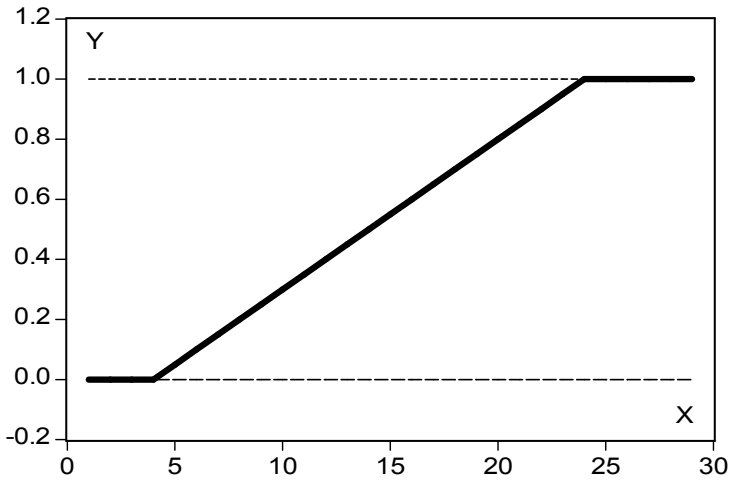
$$P(y_i = 0) = 1 - p_i$$

则$E(y_i) = p_i$ (3)

由(2)和(3)式有

$p_i = \alpha + \beta x_i$（y_i的样本值是0或1，而预测值是概率) (4)

以$p_i = -0.2 + 0.05x_i$为例，说明x_i每增加一个单位，则采用第1种选择的率增加0.05。



假设用模型(4)，$p_i = -0.2 + 0.05x_i$，进行预测，当预测值落在[0,1]区间之内（即x取值在[4,24]之内）时，则没有什么问题。但当预测值落在[0,1]区间之外时，则会暴露出该模型的严重缺点。因为概率的取值范围是[0,1]，所以此时必须强令预测值（概率值）相应等于0或1。线性概率模型常写成如下形式，

$p_i = 1, if \alpha + \beta \mathrm{x}_i \geq 1$

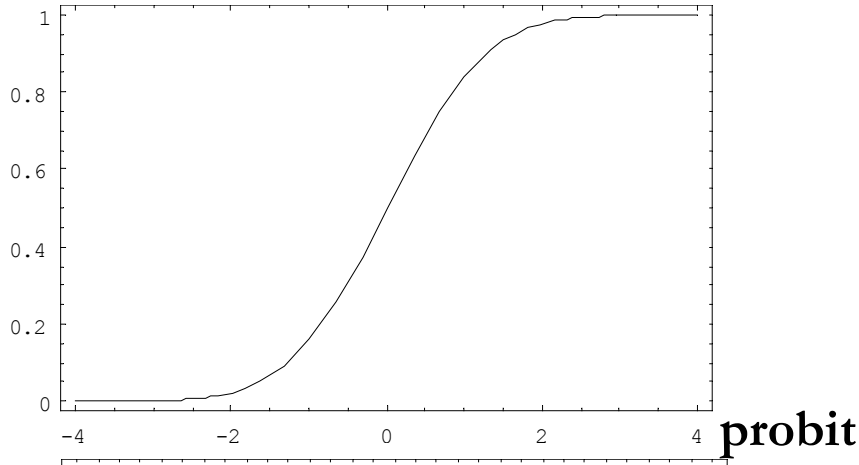
$p_i = \alpha + \beta \mathrm{x}_i, if\, 0 < \alpha + \beta x_i < 1$

$p_i = 0, if \alpha + \beta \mathrm{x}_i \leq 0$

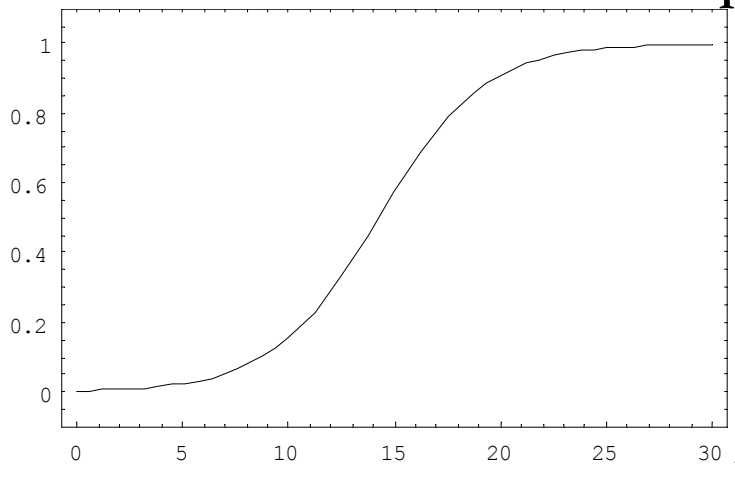
此模型由 James Tobin 1958年提出，因此称作Tobit模型（James Tobin1981年获诺贝尔经济学奖）。

然而这样做是有问题的。假设预测某个事件发生的概率等于1，但是实际中该事件可能根本不会发生。反之，预测某个事件发生的概率等于0，但是实际中该事件却可能发生了。虽然估计过程是无偏的，但是由估计过程得出的预测结果却是有偏的。

由于线性概率模型的上述缺点，希望能找到一种变换方法，使解释变量x_i所对应的所有预测值（概率值）都落在(0,1)之间。同时对于所有的x_i，当x_i增加时，希望y_i也单调增加或单调减少。显然累积概率分布函数$F(z_i)$能满足这样的要求。采用累积正态概率分布函数的模型称作Probit模型。用正态分布的累积概率作为Probit模型的预测概率。另外logistic函数也能满足这样的要求。采用logistic函数的模型称作logit模型。



probit



logit

3.Logit 模型

（1）提出

该模型是 Mcfadden于1973年首次提出。其采用的是 logistic概率分布函数其形式是，$\mathrm{p}_i = F(\mathrm{y}_i) = F(\alpha + \beta x_i) = \dfrac{1}{1+e^{-y_i}} = \dfrac{1}{1+e^{-(\alpha+\beta x_i)}}$，

其中 p_i 表示概率，$F(y_i)$ 表示 logistic累积概率密度函数。对于给定的 x_i，p_i 表示相应个体做出某科选择的概率。y_i 称作隐(潜)变量，y_i 的取值范围是 $(-\infty,+\infty)$，y_i 通过logistic函数被转换为概率。

Probit 曲线和logit曲线很相似。两条曲线都是在 p_i=0.5处有拐点，但 logit曲线在两个属部要比 Probit曲线厚。

对log曲线作如下变换，

$$p_i(1+e^{-y_i})=1$$

对上式除以 p_i，并减1得，

$$e^{-y_i}=\frac{1}{p_i}-1=\frac{1-p_i}{p_i}$$

取倒数后，再取对数，

$$y_i=\ln\left(\frac{p_i}{1-p_i}\right)$$

所以

$$\ln\left(\frac{p_i}{1-p_i}\right)=y_i=\alpha+\beta x_i$$

由上式知，回归方程的因变量是对数的某个具体选择的机会比。logit模型的一个重要优点是把在[0,1]区间上预测概率的问题转化为在实数轴上预测一个事件发生的机会比问题。 logit累积概率分布函数的斜率在 p_i=0.05时最大，在累积分布两个尾端的斜率逐渐减小。说明相对于 p_i=0.05附近的解释变量 x_i 的变化对概率的变化影响较大，而相对于 p_i 接近0和1附近的 x_i 值的变化对概率的变化影响较小。

类型	分类(因变量)		例	方法	分布	备注
定量	连续/计量		利润	普通回归	正态	可运算
	离散/计数		人口	普通或 Log 回归	Poiison 分布	可运算
定性(名义)	二分类		性别	Logit 回归	二项分布	不可运算
	多分类	无序	职业	基准一类别 Logit 回归	多项分布	不可运算
		有序	学历	累积 Logit 回归	Poiison 分布	不可运算

因变量 y	自变量 $x_1,\cdots,x_k$	方法	分布	
定量(连续,离散)	定量(连续,离散),定性	普通回归模型		
二分类	连续，定性(二分类,多分类)	Logit 模型	二项分布	SAS 中可非线性
多分类	多分类(有序)	Logit 模型	Poiison 分布	SAS 中可非线性
	多分类(无序)	Logit 模型	多项分布	
定量，定性	定量，定性	?		

（2）二分类

如果影响$\ln\frac{p}{1-p}$的因素有$x_1,x_2,...,x_p$则多元logit线性归方程为，

$$\ln\frac{p}{1-p}=\alpha_0+\alpha_1 x_1+\alpha_2 x_2+...+\alpha_p x_p$$

多元logit线性回归方程还有以下等价形式

$$p=\frac{e^{\alpha_0+\alpha_1 x_1+\alpha_2 x_2+...+\alpha_p x_p}}{1+e^{\alpha_0+\alpha_1 x_1+\alpha_2 x_2+...+\alpha_p x_p}}$$

$$p=\frac{1}{1+e^{-(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+...+\alpha_p x_p)}}$$

若将$\ln\frac{p}{1-p}$看成是因变量，则logit线性回归模型与多元线性回归模型的形式是一致的，且有很多共性。不同的是：

1、logistic回归模型中因变量是二分类的，而且非连续，其误差的分布不再是正态分布，而是二项分布，且所有的分析均建立在二项分布的基础上。

2、由于上述原因，logit回归系数的估计不能再用最小二乘法，而要用极大似然估计法。回归模型和回归系数的检验也不是F检验和t检验，而要用Wald检验、似然比检验等。

（3）多分类

前面讨论的logit模型为二分数据的情况,有时候响应变量有可能取三个或更多值，即多类别的属性变量。

根据响应变量类型的不同，分两种情况：响应变量为定性名义变量；响应变量为定性有序变量；

当名义响应变量有多个类别（即名义、无序）时，多项logit模型应采取把每个类别与一个基线类别配成对，通常取最后一类为参照，称为基线-类别logit。

有些协变量为定量数据，logistic回归模型的协变量可以是定性名义数据。这就需要对名义数据进行赋值。

通常某个名义数据有k个状态，则定义变量$M_1,...,M_{K-1}$，代表前面的k-1状态，最后令k-1变量均为0或-1来代表第k个状态。

如婚姻状况有四种状态：未婚、有配偶、丧偶和离婚，则可以定义三个指示变

量M_1、M_2、M_3，用(1,0,0)、 (0,1,0) 、(0,0,1) 、(0,0,0)或(-1,-1,-1)来对以上四种状态赋值。

参考文献：

https://wenku.baidu.com/view/9e3cc86a11a6f524ccbff121dd36a32d7275c718.html

https://chrisyeh96.github.io/2018/06/11/logistic-regression.html

https://baike.baidu.com/item/Logit%E6%A8%A1%E5%9E%8B/7286579