计量经济学

限值因变量模型

张晨峰

2016年6月1日

华东理工大学商学院

7. 限值因变量模型

主要内容

- 二值因变量：线性概率模型
- 二值响应的对数单位和概率单位模型
- 用于角点解响应的托宾模型
- 泊松回归模型
- 删截和断尾回归模型
- 样本选择纠正

7.1 二值因变量：线性概率模型

线性概率模型(LPM)

$$E(y|x) = P(y = 1|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

其中$P(y = 1|x)$被称为响应概率，在保持其他因素不变的情况下，β_j度量了因x_j的变化导致成功概率的变化。

$$\Delta P(y = 1|x) = \beta_j \Delta x_j$$

7.1 二值因变量：线性概率模型

event $y = 1$. As an example, let *inlf* ("in the labor force") be a binary variable indicating labor force participation by a married woman during 1975: *inlf* = 1 if the woman reports working for a wage outside the home at some point during the year, and zero otherwise. We assume that labor force participation depends on other sources of income, including husband's earnings (*nwifeinc*, measured in thousands of dollars), years of education (*educ*), past years of labor market experience (*exper*), *age*, number of children less than six years old (*kidslt6*), and number of kids between 6 and 18 years of age (*kidsge6*). Using the data in MROZ.RAW from Mroz (1987), we estimate the following linear probability model, where 428 of the 753 women in the sample report being in the labor force at some point during 1975:

$$\widehat{inlf} = .586 - .0034\,nwifeinc + .038\,educ + .039\,exper$$
$$\quad\quad (.154)\ \ (.0014)\quad\quad\quad (.007)\quad\quad (.006)$$
$$-.00060\,exper^2 - .016\,age - .262\,kidslt6 + .013\,kidsge6 \qquad [7.29]$$
$$\quad\quad (.00018)\quad\quad (.002)\quad\quad (.034)\quad\quad\quad (.013)$$
$$n = 753,\ R^2 = .264.$$

7.1 二值因变量：线性概率模型

线性概率模型的缺点

- 如果代入自变量的某些特定组合数值，就能得到小于0或大于1的预测值。
- 概率不可能与自变量所有的可能值线性相关。

线性概率模型的特点

- 它通常对自变量取值在均值样本附件特别奏效
- 当y是一个二值变量时，其以x为条件的方差为$Var(y|x) = p(x)[1 - p(x)]$。这意味着，线性概率模型中一定存在着异方差。

7.2 二值响应的对数单位和概率单位模型

设定对数单位和概率单位模型

$$P(y=1|x) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) = G(\beta_0 + x\beta)$$

在**对数单位模型(logit model)**中，G是对数函数

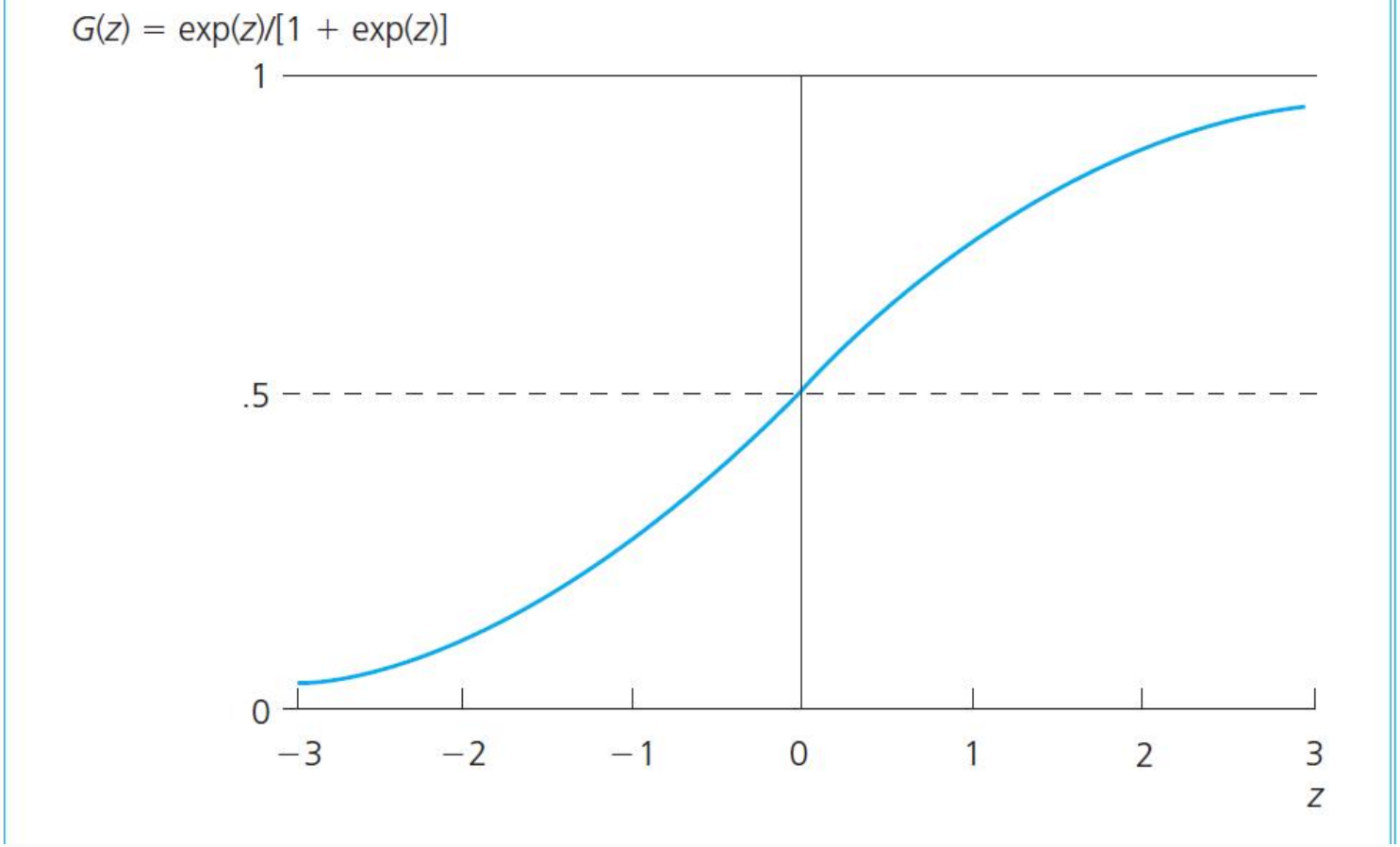
$$G(z) = exp(z)/[1 + exp(z)] = \Lambda(z)$$

它是一个标准逻辑斯蒂随机变量的累积分布函数。在**概率单位模型(probit model)**中，G是标准正态的累积分布函数，可表示为积分

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(\nu) d\nu$$

7.2 二值响应的对数单位和概率单位模型

FIGURE 17.1 Graph of the logistic function $G(z) = \exp(z)/[1 + \exp(z)]$.



7.2 二值响应的对数单位和概率单位模型

潜变量模型(latent variable model)

$$y^* = \beta_0 + x\beta + e, y = 1[y^* > 0]$$

假定e独立于x，并服从标准的逻辑斯蒂分布或标准正态分布。因此，

$$P(y = 1|x) = P(y^* > 0|x) = P[e > -(\beta_0 + x\beta)|x] = G(\beta_0 + x\beta)$$

7.2 二值响应的对数单位和概率单位模型

变量对响应概率的偏效应

连续变量对响应概率的偏效应

$$\frac{\partial p(x)}{\partial x_j} = g(\beta_0 + x\beta)\beta_j, g(z) \equiv \frac{dG(z)}{dz}$$

偏效应总是具有与β_j一样的符号。任何两个连续解释变量的相对影响都与x无关：x_j和x_h的偏效应之比为β_j/β_h。在g关于0对称分布的典型情形中，在$\beta_0 + x\beta = 0$时出现了最大的影响。例如，在概率单位情形中，$g(0) = 0.40$；在对数单位情形中，$g(0) = 0.25$。 若x_1是一个二值解释变量，那么在保持其他变量不变的情况下，x_1从0到1的偏效应无非是

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + ... + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + ... + \beta_k x_k)$$

7.2 二值响应的对数单位和概率单位模型

对数单位和概率单位模型的极大似然估计

由于$E(y|x)$的非线性性质，所以OLS和WLS都不适用，可以使用极大似然估计。

多重假设的检验

- LM(拉格朗日乘数)检验——约束模型
- Wald(瓦尔德)检验——无约束模型
- LR(似然比)检验——约束模型和无约束模型

7.2 二值响应的对数单位和概率单位模型

正确预测百分比

定义一个二值预测元在预测概率至少为0.5时取值1，否则取值0。

解释对数单位和概率单位模型的估计值

若x_j是(大致)连续的，估计成功概率的变化大致为

$$\frac{\partial p(x)}{\partial x_j} = g(\hat{\beta}_0 + x\hat{\beta})\hat{\beta}_j$$

如果将每个解释变量都代之以样本平均值，则得到平均个人偏效应(PEA)

$$g(\hat{\beta}_0 + x\hat{\beta}) = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + ... + \hat{\beta}_k\bar{x}_k)$$

通过在样本中对个体偏效应的平均得到的结果，被称为平均偏效应(AME)，即乘以$\hat{\beta}_j$的项是比例因子

$$n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + x_i\hat{\beta})$$

7.2 二值响应的对数单位和概率单位模型

TABLE 17.1 LPM, Logit, and Probit Estimates of Labor Force Participation

Dependent Variable: *inlf*			
Independent Variables	**LPM (OLS)**	**Logit (MLE)**	**Probit (MLE)**
nwifeinc	−.0034 (.0015)	−.021 (.008)	−.012 (.005)
educ	.038 (.007)	.221 (.043)	.131 (.025)
exper	.039 (.006)	.206 (.032)	.123 (.019)
$exper^2$	−.00060 (.00018)	−.0032 (.0010)	−.0019 (.0006)
age	−.016 (.002)	−.088 (.015)	−.053 (.008)
kidslt6	−.262 (.032)	−1.443 (.204)	−.868 (.119)
kidsge6	.013 (.013)	.060 (.075)	.036 (.043)
constant	.586 (.151)	.425 (.860)	.270 (.509)
Percentage correctly predicted	73.4	73.6	73.4
Log-likelihood value	—	−401.77	−401.30
Pseudo *R*-squared	.264	.220	.221

7.3 用于角点解响应的托宾模型

托宾模型(Tobit model)

托宾模型用一个基本的潜变量来表示所观测到的响应y

$$y^* = \beta_0 + x\beta + \mu, \mu|x \sim Normal(0, \sigma^2), y = max(0, y^*)$$

由于μ/σ服从标准正态分布且独立于x，则

$$p(y = 0|x) = 1 - \Phi(x\beta/\sigma)$$

对托宾估计值的解释

$$E(y|x) = \Phi(x\beta/\sigma)[x\beta + \sigma\lambda(x\beta/\sigma)] = \Phi(x\beta/\sigma)x\beta + \sigma\phi(x\beta/\sigma)$$

7.3 用于角点解响应的托宾模型

TABLE 17.2 OLS and Tobit Estimation of Annual Hours Worked

Dependent Variable: *hours*		
Independent Variables	**Linear (OLS)**	**Tobit (MLE)**
nwifeinc	−3.45 (2.54)	−8.81 (4.46)
educ	28.76 (12.95)	80.65 (21.58)
exper	65.67 (9.96)	131.56 (17.28)
$exper^2$	−.700 (.325)	−1.86 (0.54)
age	−30.51 (4.36)	−54.41 (7.42)
kidslt6	−442.09 (58.85)	−894.02 (111.88)
kidsge6	−32.78 (23.18)	−16.22 (38.64)
constant	1,330.48 (270.78)	965.31 (446.44)
Log-likelihood value	—	−3,819.09
R-squared	.266	.274
$\hat{\sigma}$	750.18	1,122.02

7.4 泊松回归模型

泊松回归模型

对于非负因变量时计数变量，可以将期望值模型转化为一个指数函数

$$E(y|x_1, x_2, ...x_k) = exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)$$

将方程取对数，得到

$$log[E(y|x_1, x_2, ...x_k)] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

泊松回归模型

对计数数据来说，令人满意的分布则是泊松分布。于是，以x为条件，y等于h的概率是

$$p(y = h|x) = exp[-exp(x\beta)][exp(x\beta)]^h/h!, h = 0, 1, ...$$

7.4 泊松回归模型

TABLE 17.3 Determinants of Number of Arrests for Young Men

Dependent Variable: *narr86*		
Independent Variables	**Linear (OLS)**	**Exponential (Poisson QMLE)**
pcnv	−.132 (.040)	−.402 (.085)
avgsen	−.011 (.012)	−.024 (.020)
tottime	.012 (.009)	.024 (.015)
ptime86	−.041 (.009)	−.099 (.021)
qemp86	−.051 (.014)	−.038 (.029)
inc86	−.0015 (.0003)	−.0081 (.0010)
black	.327 (.045)	.661 (.074)
hispan	.194 (.040)	.500 (.074)
born60	−.022 (.033)	−.051 (.064)
constant	.577 (.038)	−.600 (.067)
Log-likelihood value	—	−2,248.76
R-squared	.073	.077
$\hat{\sigma}$	.829	1.232

7.5 删截和断尾回归模型

删截回归模型

典型的截取是因为调查设计，有时候也可能是因为制度上的约束。我们将数据截取问题与角点解结果分开处理，并用一个删截回归模型(censored regression model)来解决数据截取的问题。实质上，用一个删截回归模型解决的问题是响应变量y的数据缺失问题。

断尾回归模型

当我们在抽样方案中以y为依据排除了总体的一个子集时，就出现了断尾回归模型(truncated regression model)。

7.5 删截和断尾回归模型

删截回归模型

$$y_i = \beta_0 + x_i\beta + u_i, u_i|x_i, c_i \sim Normal(0, \sigma^2)$$

$$\omega_i = min(y_i, c_i)$$

必须知道，在随机抽样的情况下，我们可以像在线性回归模型中那样解释β_j。删截回归模型的一个重要应用是持续期间分析(duration analysis)。

7.5 删截和断尾回归模型

TABLE 17.4 Censored Regression Estimation of Criminal Recidivism

Dependent Variable: log(*durat*)	
Independent Variables	**Coefficient (Standard Error)**
workprg	−.063 (.120)
priors	−.137 (.021)
tserved	−.019 (.003)
felon	.444 (.145)
alcohol	−.635 (.144)
drugs	−.298 (.133)
black	−.543 (.117)
married	.341 (.140)
educ	.023 (.025)
age	.0039 (.0006)
constant	4.099 (.348)
Log-likelihood value $\hat{\sigma}$	−1,597.06 1.810

7.5 删截和断尾回归模型

断尾回归模型

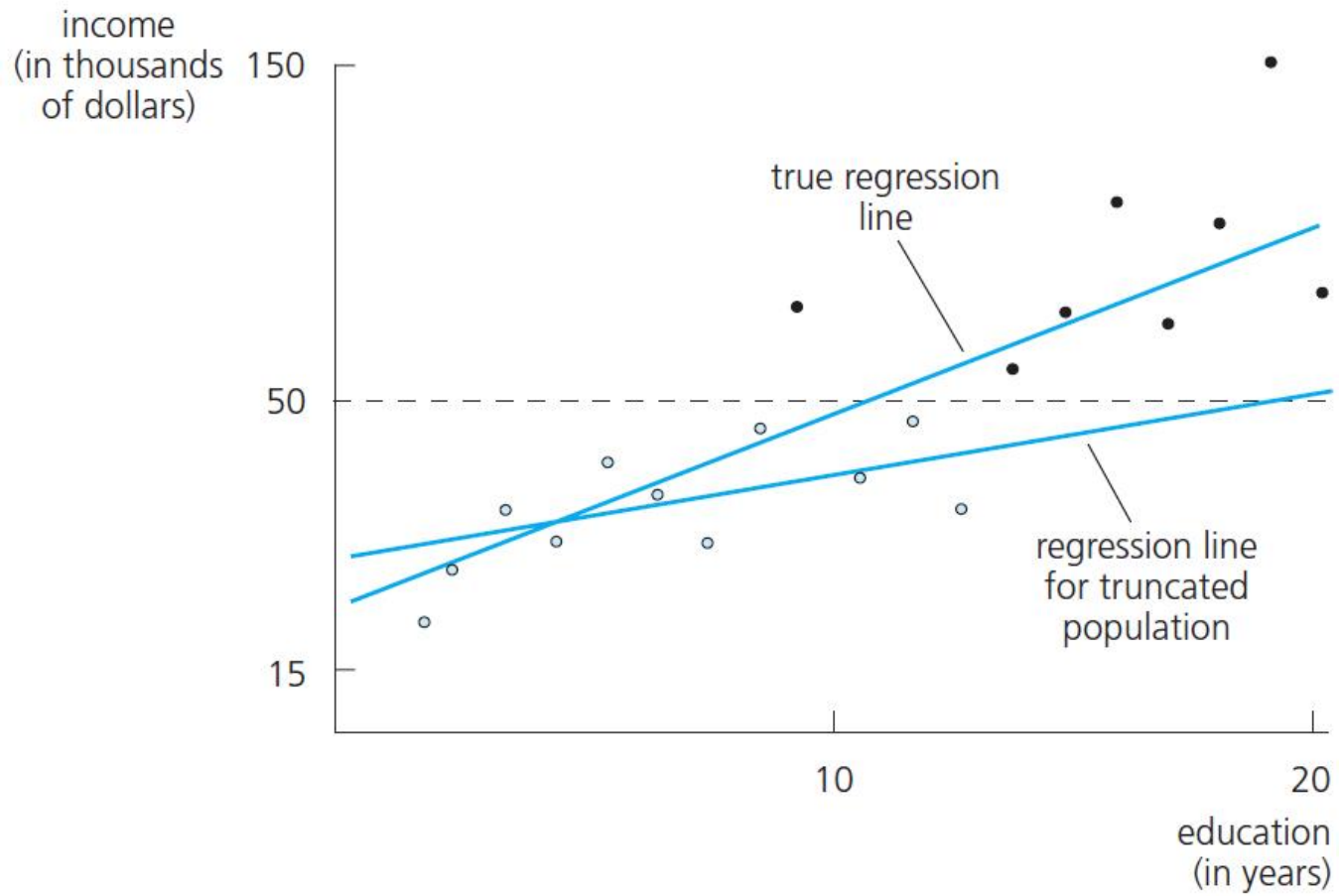
在断尾回归中，我们会首先选取总体的一部分，然后再在其中选取样本。

豪斯曼和怀斯的负收入税实验

他们研究收入的各种决定因素，数据中一个家庭的收入必须低于1967年贫困线(贫困线取决于家庭规模)的1.5倍才会包括在研究中。

7.5 删截和断尾回归模型

FIGURE 17.4 A true, or population, regression line and the incorrect regression line for the truncated population with observed incomes below $50,000.



7.6 样本选择纠正

OLS什么时候对选择样本是一致的？

一个总体模型

$$y_i = x_i\beta + \mu_i$$

如果处于某种原因，某个观测i的y_i或某些自变量不能观测到。为每个i定义一个选择指标s_i，若我们观测到(y_i, x_i)的全部，则$s_i = 1$；否则$s_i = 0$。显然，我们是估计方程

$$s_i y_i = s_i x_i \beta + s_i u_u$$

若s_i与u_i相关，那么选择样本的OLS不能一致地估计β_j。

7.6 样本选择纠正

赫克曼方法

从属断尾问题

$$y = x\beta + \mu, E(\mu|x) = 0, s = 1[z\gamma + \nu \geq 0]$$

可以得到

$$E(y|z, s = 1) = x\beta + \rho\lambda(z\gamma)$$

样本选择纠正：

1. 利用所有n个观测，估计一个s_i对z_i的概率单位模型，并得到估计值$\hat{\gamma}_h$。对每个i计算反米尔斯比$\hat{\lambda}_i = \lambda(z_i\hat{\gamma})$
2. 利用选择样本，做如下回归：y_i对x_i和$\hat{\lambda}_i$回归，则$\hat{\beta}_j$就是一致的。

7.6 样本选择纠正

TABLE 17.5 Wage Offer Equation for Married Women

Dependent Variable: log(*wage*)		
Independent Variables	**OLS**	**Heckit**
educ	.108 (.014)	.109 (.016)
exper	.042 (.012)	.044 (.016)
$exper^2$	−.00081 (.00039)	−.00086 (.00044)
constant	−.522 (.199)	−.578 (.307)
$\hat{\lambda}$	—	.032 (.134)
Sample size *R*-squared	428 .157	428 .157