微观计量经济建模的数据暴露问题及对策

王忠玉[1,2]　赵振权[1]

1 吉林大学数量经济研究中心　　2 哈尔滨工业大学

目前，微观计量经济模型在经济学的各个分支领域中得到广泛应用，为不同研究者提供了多种多样的建模工具，但是由于建立模型中用到大量的可能涉及到微观个体的“隐私”数据，因此出现了数据暴露问题。针对微观计量经济建模的数据暴露问题及解决方法，本文考察无个性化处理的二值变量作为二值 probit 模型的因变量，而解释变量仍保持最初形式，提出一种‘变动’策略克服数据暴露问题——后随机化因变量，研究因变量后随机化对统计推断的影响。

1、引言

经济学实证研究很长时间遭受到无法利用个体‘微观数据’，而且迫使经济计量学研究者利用加总时间序列进行研究，例如消费函数。与之相比，心理学、生物测量学等对微观数据分析的探索已经持续几十年了。随着经济社会的发展，日益增长的微观个体数据不断出现，并被广泛利用。

2000 年度诺贝尔奖授予赫克曼和麦克法登，以表彰他们在微观数据分析方面的创新性贡献，为现代微观计量经济学打下了坚实的基础。随着他们研究成果及工作的广泛传播，目前建立和应用微观计量经济模型已经渗透到许多应用经济学领域，并创立了微观计量经济模型软件。然而，由于统计局或者管理机构考虑到微观数据的保密性，有时候这类数据会受到一定程度的限制。实际上，数据管理机构面临着双重的服务对象，即数据的提供者（受访者）与数据需求者（政府、研究者等），这就要求其能兼顾好，在满足用户对微观数据需求的同时，又能有效保护数据的机密性。目前，数据管理机构一般采用三种方法进行数据的保密工作：完全禁止使用，在一定限制条件下使用以及公布处理后的数据[1]。

2、二值 probit 模型与随机化

通常，二值probit模型考察某个解释变量x对潜连续变量Y^*的影响，即可以假定模型

$$Y^* = \alpha + \beta x + \varepsilon \tag{1}$$

成立，其中表示正态分布随机误差，$E(\varepsilon) = 0$ 且$Var(\varepsilon) = 1$。然而，Y^*是可观测的，即在二值变量Y 满足下述门限模型时才是可观测的，Y 是由

$$Y = \begin{cases} 1, & Y^* > 0 \\ 0, & Y^* \le 0 \end{cases} \tag{2}$$

定义的。样本信息是由n个数对(x_i, y_i)给出的，其中$y_i \in \{0, 1\}$而x_i是任意实数。对未知参数α 与 β 可直接运用极大似然法加以估计。为了讨论方便，这里仅仅考察只有一个回归元且假定它是连续的情形。

现在，考察二值变量 Y 的随机化，即它的值以某个规定转移概率来进行变换。实际上，

本文得到中国博士后科学研究基金项目的资助（20070410316）

最初随机化是为了避免调查出现无响应而提出的，比如调查中的敏感问题像AIDS疾病或药物消费等。目前，关于连续变量的无个性化方法及其对微观计量经济模型的估计及影响得到了研究，取得一些可以利用的方法[2]。然而，基于随机化的对离散变量进行无个性化处理方法是后随机化，即以某个规定概率对类型加以变换或者转移。关于随机化响应和后随机化两者之间的关系如下：在随机化背景下，随机模型在数据收集之前就已被定义，然而在后随机化中，这一方法将被用于已获得的数据上。

针对微观计量经济建模的数据暴露问题及解决方法，本文考察无个性化处理的二值变量作为二值 probit 模型的因变量，而解释变量仍保持最初形式，研究在因变量后随机化对统计推断的影响。

对二值变量的随机化可描述如下：设表示源自后随机化的‘修饰’变量，于是，通过 $P_{jk} \equiv P(\mathrm{Y}^{\mathrm{m}} = j \mid \mathrm{Y} = k, j, k \in \{0, 1\}$ 且 $p_{\mathrm{j0}} + p_{\mathrm{j1}} = 1$ 对于 $j = 0, 1)$ 定义转移概率。倘若令 $p_{00} \equiv \pi_0$ 且 $p_{11} \equiv \pi_1$ 定义两个不变概率，而概率矩阵写成

$$P_y = \begin{pmatrix} \pi_0 & 1-\pi_0 \\ 1-\pi_1 & \pi_1 \end{pmatrix}$$

由于在后随机化方法中，已知两个概率，从而对两种状态可进行对称地讨论，因此，下面只考虑特殊情况：

$$\pi_0 = \pi_1 \tag{3}$$

当因变量样本接受随机化，将拥有 n 个观测值 y_i^m，其中 y_i^m 表示通过随机化方法从 y_i 中获得的二值变量。随机化具有下述优点：的最初分布可以从修饰的观测值中估计出来[3]。

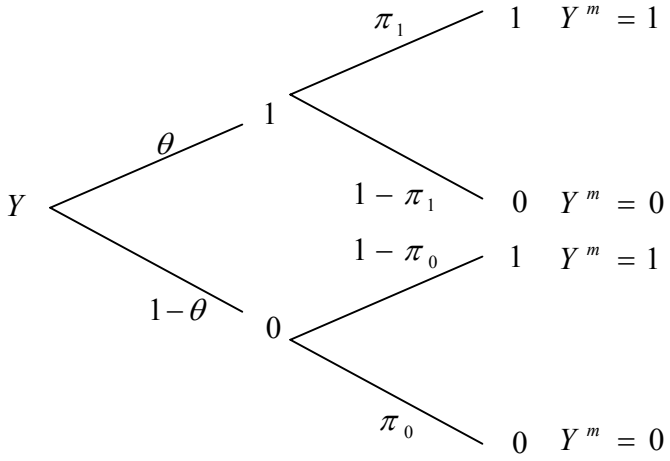

图 1. 在随机化下数据生成过程

现在研究由（1）和（2）给出的 probit 模型估计。由图 1 中以及限制（3），很明显，在随机化下拥有下面数据生成过程

$$Y = \begin{cases} 1, & \text{以概率} \Phi_i \pi + (1-\Phi_i)(1-\pi) \\ 0, & \text{以概率} \Phi_i (1-\pi) + (1-\Phi_i)\pi \end{cases} \tag{4}$$

这里Φ_i表示在正态分布下的条件概率，即未修饰因变量Y_i取值 1，对于给定x_i，也就是$\Phi_i \equiv (\alpha + \beta x_i) = P(Y_i^* > 0 \mid x_i)$。由（4），可以获得似然函数

$$\mathcal{L}(\alpha, \beta \mid (y_i^m, x_i)) = \prod_{i=1}^{n} (\Phi_i \pi + (1-\Phi_i)(1-\Phi\pi))^{y_i^m} (\Phi_i (1-\Phi\pi) + (1-\Phi_i)\pi)^{(1-y_i^m)}$$

其中 i =1,…, n.

通过对此似然函数求一阶与二阶偏导数，可知这个函数关于 α 与 β 具有全局凹性，

$$\mathrm{L} \equiv \log(\mathcal{L}) = \sum_{i=1}^{n} y_i^m \log[\Phi_i \pi + (1-\Phi_i)(1-\Phi\pi)] + (1-y_i^m) \log[\Phi_i (1-\Phi\pi) + (1-\Phi_i)\pi]$$

关于 α 与 β 的一阶偏导数是由

$$\frac{\partial \mathcal{L}}{\partial \alpha} = (2\pi - 1) \sum_i \left[y_i^m \frac{\phi_i}{W_i} - (1-y_i^m) \frac{\phi}{1-W_i} \right] = (2\pi - 1) \sum_i \left[\frac{(y_i^m - W_i)\phi_i}{W_i (1-W_i)} \right]$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = (2\pi - 1) \sum_i \left[y_i^m \frac{\phi_i}{W_i} - (1-y_i^m) \frac{\phi}{1-W_i} \right] x_i = (2\pi - 1) \sum_i \left[\frac{(y_i^m - W_i)\phi_i}{W_i (1-W_i)} \right] x_i \qquad (5)$$

给出的，其中

$$\phi_i \equiv \phi(\alpha + \beta x_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (\alpha + \beta x_i)^2 \right\} \text{(标准正态密度)}$$

$$W_i \equiv \pi \Phi_i + (1-\pi)(1-\Phi_i) \quad \text{(观测到 } y_i^m \text{ =1 的概率)}$$

下面，假定

$$\pi \neq \frac{1}{2} \qquad (6)$$

可以证明，源自二阶偏导数的海赛矩阵不再是负定的，这点和标准的 probit 情况相反。首先注意到

$$\frac{\partial W_i}{\partial \alpha} = (2\pi - 1)\, \phi_i \quad , \quad \frac{\partial W_i}{\partial \beta} = (2\pi - 1)\, \phi_i$$

$$\frac{\partial (1-W)_i}{\partial \alpha} = -(2\pi - 1)\, \phi_i \quad , \quad \frac{\partial (1-W)_i}{\partial \beta} = -(2\pi - 1)\, \phi_i x_i$$

$$\frac{\partial W_i (1-W)_i}{\partial \alpha} = (2\pi - 1)\, \phi_i (1 - 2W_i) \quad , \quad \frac{\partial W_i (1-W)_i}{\partial \beta} = (2\pi - 1)\, \phi_i x_i (1 - 2W_i),$$

而且，一旦利用 $z_i \equiv \alpha + \beta x_i$，则有

$$\frac{\partial \phi_i}{\partial \alpha} = (\alpha + \beta x_i) == -\, \phi_i z_i$$

$$\frac{\partial \phi_i}{\partial \beta} = (\alpha + \beta x_i) = -\, \phi_i z_i x_i$$

关于 α 的二阶偏导数为

$$\frac{\partial^2 L}{(\partial\alpha)^2}=(2\pi-1)\left\{\sum_i\left[\left(\frac{\partial}{\partial\alpha}\frac{y_i^m-W_i}{W_i(1-W_i)}\right)\phi_i+\frac{y_i^m-W_i}{W_i(1-W_i)}\left(\frac{\partial}{\partial\alpha}\phi_i\right)\right]\right\}$$

$$=-(2\pi-1)\left\{\sum_i\left[\frac{\phi_i(2-1)(-W_i^2-y_i^m(1-2W_i))-W_i(1-W_i)(y_i^m-W_i)z_i}{(W_i(1-W_i))^2}\right]\phi_i\right\}$$

$$=-(2\pi-1)\left\{\sum_i\left[\frac{(2\pi-1)(y_i^m-2y_i^mW_i+W_i^2)\phi_i+(y_i^m-W_i)W_i(y_i^m-W_i)z_i}{(W_i(1-W_i))^2}\right]\phi_i\right\}$$

$$=-(2\pi-1)\left\{\sum_i\frac{g_i\phi_i}{(W_i(1-W_i))^2}\right\} \tag{7}$$

其中，$g_i=(2\pi-1)(y_i^m-2y_i^mW_i+W_i^2)\phi_i+(y_i^m-W_i)W_i(y_i^m-W_i)z_i$ (8)

关于其他两个二阶偏导数的相应结果会得出下面的海赛矩阵：

$$H^{\mathrm{m}}=-(2\pi-1)\sum_i\frac{g_i\phi_i}{W_i(1-W_i)}\mathrm{u}_i\mathrm{u}_i' \tag{9}$$

其中 $\mathrm{u}_i=(1\quad x_i)$。由于(2π−1)要么为正的，要么为负的，而式(8)中的函数g_i 比标准情况要复杂一些，因此用于标准probit情况的有关负定性的证明这里行不通。(参看Amemiya, 1985, p274)。

然而，在随机化下的估计蕴含着效率的损失，从这点上我们可获得信息矩阵的简单公式。注意，$\mathrm{E}[Y_i^m]=W_I$，因而作为 Y_i^m 函数的g_i的预期值是由

$$E[g(Y_i^m)]=E[(2\pi-1)(Y_i^m-2Y_i^mW_i+W_i^2)+(Y_i^m-W_i)W_i(1-W_i)z_i]$$

$$=(2\pi-1)(W_i-2W_i^2+W_i^2)\phi_i+(W_i-W_i)W_i(1-W_i)z_i=(2\pi-1)W_i(1-W_i)\phi_i$$

给出的。因此，在修饰数据下，信息矩阵是由

$$\mathcal{T}=(2\pi-1)\sum_i\frac{\phi_i^2}{W_i(1-W_i)}\mathrm{u}_i\mathrm{u}_i' \tag{10}$$

给出的，而未修饰数据情况下的矩阵为

$$\mathcal{T}=\sum_i\frac{\phi_i^2}{\Phi_i(1-\Phi_i)}\mathrm{u}_i\mathrm{u}_i' \tag{11}$$

现在，想要证明$\mathcal{T}-\mathcal{T}^m$是非负定的。注意到，由前面定义$W_i\equiv\pi\,\Phi_i+(1\text{-}\pi)(\,1\text{-}\Phi_i)$，由此可得，

$$W_i(1-W_i)=(2\pi-1)^2\Phi_i(1-\Phi_i)+\pi(1-\pi)$$

因此对于任意的 i, 可以证明

$$\frac{1}{\Phi_i(1-\Phi_i)} > \frac{(2\pi-1)}{W_i(1-W_i)} \tag{12}$$

或者$W_i(1-W_i) > (2\pi-1)^2\Phi_i(1-\Phi_i)$，所以$\mathcal{I}-\mathcal{I}^m$非负定的。

π的什么值会蕴含着最大效率的损失呢？对于（10）式中的‘权’，由于

$$h(\pi) = \frac{\pi(1-\pi)}{(2\pi-1)^2}$$

是对称的，也就是$h(\pi) = h(1-\pi)$，单调递增的，在π=1/2 趋于无穷大，因此我们能写成

$$\frac{(2\pi-1)^2}{W_i(1-W_i)} = \frac{(2\pi-1)}{(2\pi-1)\Phi_i(1-\Phi_i)+\pi(1-\pi)} = \frac{1}{\Phi_i(1-\Phi_i)+\frac{\pi(1-\pi)}{(2\pi-1)^2}}$$

（10）式中权趋于 0，而当π或$(1-\pi)$ 趋于 1/2 时，信息矩阵趋于 0 矩阵。

3、结论

针对微观计量经济建模的数据暴露问题及解决方法，本文考察无个性化处理的二值变量作为二值 probit 模型的因变量，而解释变量仍保持最初形式，从得到的因变量后随机化对统计推断的影响结果表明，当转移概率趋于 1/2 时，二值因变量随机化引起的效率损失，而当转移概率接近于 0 或 1 时，效率损失更小。这点揭示了在解决微观计量经济建模的数据暴露问题与数据保密性之间一种顾此失彼的权衡关系。

参考文献

[1] 艾春荣，冯帅章和吴玉玲，微观统计数据的公布及相应的保密方法，《统计研究》2007，6，75-79.

[2] Kooiman, P., Willenborg, L., Gouweleeuw, L., PRAM: a method for discloure limitation of micro data. http:// www.cbs.nl/sdc/tuis.htm. 1997

[3] Amemiya, T., Advanced Econometrics. Basil Blackwell. 1985

[4] Hausman, J. A., Abrevaya, J., Scott-Morton, F. M., Miscallsification of the dependent variable in a discrete-response. Journal of Econometrics 87, 239-269.

[5] Wooldridge, J.M., Econometric Analysis of Cross Section and Panel Data, MIT Press. 2002. 王忠玉译，《横截面与面板数据的经济计量分析》，中国人民大学出版社，2007。

[6] Cameron, C. A., and Trivedi, P. K., Microeconometrics: Methods and Applications, Cambridge University Press. 2005.

[7] Greene, W. H., *Econometric Analysis*. 2003. 费剑平译，《计量经济分析》，中国人民大学出版社，2007 年。