

## 截断非平衡似无关模型的极大仿真似然估计

陈永伟, 葛翔宇

(中南财经政法大学 统计与数学学院, 武汉 430073)

**摘要** 样本数据缺失和截断是现代统计调查中经常遇到的两个问题, 它们在一定程度上影响模型参数估计的准确性和有效性. 该研究首先提出了一个新的截断非平衡似无关回归模型, 这个模型能够同时考虑数据缺失和截断的特征; 然后基于 Geweke-Hajivassiliou-Keane (GHK) 的仿真算子, 建立了该模型的极大仿真似然估计方法; 蒙特卡罗实验结果表明, 在大样本和有限样本下这种估计方法在参数估计的准确性和有效性方面均具有良好表现.

**关键词** 似无关回归; 截断; 缺失; 极大仿真似然估计

## Maximum simulated likelihood estimation of censored seemingly unrelated regressions with missing observations

CHEN Yong-wei, GE Xiang-yu

(School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China)

**Abstract** In many survey data sets, multiple observations are frequently missed and censored at some threshold values, which cause the biased estimation based on such samples. In this paper, we develop a censored and unbalanced seemingly unrelated regression model, allowing for the censored dependent variables and missing observations. By using the Geweke-Hajivassiliou-Keane (GHK) simulator, we propose a procedure of maximum simulated likelihood estimators for this new model. A small scale Monte Carlo simulation demonstrates that the estimators can perform well both in finite and large sample.

**Keywords** seemingly unrelated regression; censored; missing observations; maximum simulated likelihood estimation

### 1 引言

数据缺失和截断是现代统计调查中经常遇到的问题. 所谓缺失, 是指调查者并不能观察到某一变量在给定时间的观测值. 截断, 是指某一变量的观测值在高于或低于一个已知阈值时被设定为该阈值 (Moeltner 和 Layton<sup>[1]</sup>). 例如, 在收入调查中, 由于个人对收入隐私权的充分重视, 调查者往往不愿意向调查者提供个人真实收入而导致数据出现缺失, 或者即使愿意提供, 有时候也只能获得被调查者的收入区间而不是具体的数值, 如在美国健康访问调查中 (Schenker 等<sup>[2]</sup>), 调查者为了最大限度获得数据信息, 对个人收入采用区间式调查法 (即你的月收入是 2000 元以上还是 2000 元以下), 由此导致调查数据出现截断. 在数据截断中, 一个比较著名的例子是, Heckman 和 MaCurdy<sup>[3]</sup> 对美国密西根州已婚女性的劳动供给情况进行调查, 其中未参加工作的女性的劳动供给被截断为 0, 由此开启了人们对数据截断问题的广泛关注.

样本数据的缺失和截断会导致模型的有偏和无效估计, 进而对模型总体性质推断产生错误结论. 特别是在面板数据的统计调查中, 由于很难获得同一个人在不同年份的连续跟踪调查数据, 因此, 数据缺失和截断对模型参数估计的影响就显得尤为严重. 现有文献对截断面板模型的参数估计做了广泛讨论, 如 Charlier 等<sup>[4]</sup> 针对含有两期样本观测值的截断面板模型, 提出了一种条件矩估计思想. Greene<sup>[5]</sup> 对含有受限因变量的面板模型进行分析, 发现截断面板模型的极大似然估计方法依赖于对模型扰动项的方差的准确估计. Chen<sup>[6]</sup> 对一般面板转换模型进行研究, 提出了该模型在包含截断样本时的一种  $\sqrt{n}$  半参数一致估计方法.

收稿日期: 2012-04-10

资助项目: 国家自然科学基金 (71001107); 国家社会科学基金 (10BJY104); 中央高校基本科研经费项目 (2722013JC086)

作者简介: 陈永伟 (1981-), 男, 浙江金华人, 副教授, 博士, 研究方向: 应用微观计量经济学, E-mail: yongwei@znu.edu.cn.

上述文献对截断样本所产生的参数估计问题进行了广泛研究, 但却并未将其扩展到包含数据缺失时的情形。特别是, 在对多个截面方程进行参数估计时, 不同截面之间往往存在一定相关性, 而现有文献并未对这种相关性引起足够重视。由于似无关回归模型已经成为研究一般截面相关问题的方法论基础, 并在实证研究中得到了广泛应用 (Phillips<sup>[7]</sup>; Geweke<sup>[8]</sup>), 因此, 本文将以似无关回归模型为基础, 构建一个包含数据缺失和截断样本的似无关回归模型。为表述方便, 当似无关回归模型包含数据缺失样本时, 本文称之为非平衡似无关回归模型<sup>1</sup>; 而当似无关回归模型包含数据缺失和截断样本时, 本文称之为截断非平衡似无关回归模型。

对于非平衡似无关回归模型, Schmidt<sup>[10]</sup> 以嵌套型数据缺失<sup>2</sup>为例, 提出了一种两步广义最小二乘估计思想, 即首先利用数据缺失样本估计模型扰动项的协方差矩阵, 然后再利用估计的协方差矩阵加权估计各回归系数。但是 Schmidt 的方法仅仅适用于含有两个截面方程的嵌套型数据缺失模型。Baltagi 等<sup>[11]</sup> 还指出, 在两步广义最小二乘估计中, 即使在第一步获得了扰动项协方差矩阵的最优估计, 也无法保证在第二步能够获得回归系数的有效估计。基于此, Hwang 和 Schulman<sup>[12]</sup> 提出了似无关回归模型在包含多个截面方程且具有非嵌套型数据缺失时的极大似然估计方法。然而, Hwang 和 Schulman 的方法也并未考虑数据截断对模型参数估计的影响。

本文对 Hwang 和 Schulman 模型进行扩展, 构建了一个同时包含数据缺失和截断特征的截断非平衡似无关回归模型。由于在估计包含截断样本的似无关回归模型时涉及到计算一个高维累积正态分布函数的概率积分, 因此, 本文将采用不同于目前普遍运用的研究方法, 以 Geweke-Hajivassiliou-Keane (GHK) 仿真算子为基础, 提出一种极大仿真似然估计方法。蒙特卡罗实验结果表明, 本文建立的参数估计方法在大样本和有限样本下对参数估计的准确性和有效性均具有良好表现。

本文余下部分安排为: 第二部分简要介绍经典似无关回归模型的极大似然估计方法; 第三部分构建一个截断非平衡似无关回归模型并给出该模型的极大仿真似然估计方法; 第四部分是蒙特卡罗实验, 该实验结果验证了本文提出的参数估计方法的准确性和可靠性; 第五部分为结论。

## 2 似无关回归模型及其极大似然估计

假定一个含有  $G$  个截面方程和  $T$  期观测值的似无关回归模型为:

$$\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t, \quad t = 1, 2, \dots, T \quad (1)$$

其中,  $\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{Gt})'$  和  $\mathbf{x}_t = (x_{1t}, x_{2t}, \dots, x_{Kt})'$  分别表示  $G \times 1$  和  $K \times 1$  维向量,  $\mathbf{B}$  是与解释变量  $\mathbf{x}_t$  相对应的  $K \times G$  维系数矩阵,  $\mathbf{u}_t = (u_{1t}, u_{2t}, \dots, u_{Gt})'$  为模型扰动项。为了反映不同截面方程之间的同期相关性, 假设  $\mathbf{u}_t$  的均值为  $\mathbf{0}$ , 方差为  $\Sigma$ , 即  $E(\mathbf{u}_t \mathbf{u}_t') = \Sigma$ 。显然, 在这一假设下不同截面方程的误差项之间是同期相关的。模型 (1) 被称为是经典似无关回归模型 (Zellner<sup>[13]</sup>)。

出于一般性考虑, 假设模型 (1) 中的系数矩阵  $\mathbf{B}$  满足一系列的线性约束条件<sup>3</sup>:

$$\mathbf{R}\beta = \mathbf{r} \quad (2)$$

其中,  $\beta = \text{vec}(\mathbf{B})$ ,  $\mathbf{R}$  和  $\mathbf{r}$  是由已知常数构成的参数约束矩阵和向量。

为说明似无关回归模型 (1) 的极大似然估计思想, 进一步假设模型中扰动项  $\mathbf{u}_t$  服从独立相同正态分布, 即  $\mathbf{u}_t \sim \text{i.i.d.} N(\mathbf{0}, \Sigma)$ 。在这一假设条件下,  $\mathbf{y}_t$  也服从均值为  $\mathbf{B}'\mathbf{x}_t$ 、方差为  $\Sigma$  的正态分布。因此, 按照极大似然估计思想,  $\mathbf{y}_t$  的对数似然函数  $\ln L_t$  为:

$$\ln L_t = -\frac{1}{2}\{G \ln(2\pi) + \ln|\Sigma| + (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)' \Sigma^{-1} (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)\} \quad (3)$$

而基于全部样本观测值的整体对数似然函数  $\ln L$  为  $\ln L_t$  的和:

$$\ln L = \sum_{t=1}^T \ln L_t = -\frac{T}{2}\left\{G \ln(2\pi) + \ln|\Sigma| + \frac{1}{T} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)' \Sigma^{-1} (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)\right\} \quad (4)$$

1. 这一概念借鉴于非平衡面板模型的定义, 具体可见 Baltagi 和 Song<sup>[9]</sup>。

2. 数据缺失分为嵌套型缺失和非嵌套型缺失两类。嵌套型数据缺失是指某一截面方程变量的数据缺失完全包含在另一截面方程对应变量的数据缺失情形中。非嵌套型数据缺失则指某一截面方程变量的数据缺失并不一定是另一截面方程对应变量数据缺失的子集。嵌套型数据缺失可以看成是非嵌套型数据缺失的特例。

3. 例如, 如果方程之间含有不相同的解释变量, 这一条件可以通过对矩阵  $\mathbf{B}$  施加 0 约束来实现。当然, 如果系数矩阵  $\mathbf{B}$  之间没有约束, 那么矩阵  $\mathbf{R}$  和  $\mathbf{r}$  对应位置的元素即为 0。

因此, 参数矩阵  $\mathbf{B}$  的极大似然估计量是在条件(2)的约束下对方程(4)求最大化结果, 即是对下列目标函数  $\ln L^R$  求最大化:

$$\begin{aligned}\ln L^R &= -\frac{T}{2} \left\{ G \ln (2\pi) + \ln |\Sigma| + \frac{1}{T} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{B}' \mathbf{x}_t)' \Sigma^{-1} (\mathbf{y}_t - \mathbf{B}' \mathbf{x}_t) \right\} - \lambda' (\mathbf{R}\beta - \mathbf{r}) \\ &= -\frac{T}{2} \left\{ G \ln (2\pi) + \ln |\Sigma| + \frac{1}{T} \text{trace} [(\mathbf{Y} - \mathbf{X}\mathbf{B}) \Sigma^{-1} (\mathbf{Y} - \mathbf{X}\mathbf{B})'] \right\} - \lambda' (\mathbf{R}\beta - \mathbf{r})\end{aligned}\quad (5)$$

其中,  $\mathbf{Y} = (y_1, y_2, \dots, y_T)'$ ,  $\mathbf{X} = (x_1, x_2, \dots, x_T)'$ ,  $\lambda$  为拉格朗日乘子向量. 对方程(5)求解得到关于参数  $\mathbf{B}$ ,  $\lambda$  和  $\Sigma$  的一阶条件为:

$$\partial \ln L^R / \partial \beta = \text{vec} (\mathbf{X}' (\mathbf{Y} - \mathbf{X}\mathbf{B}) \Sigma^{-1}) - \mathbf{R}' \lambda = \mathbf{0} \quad (6)$$

$$\partial \ln L^R / \partial \lambda = -(\mathbf{R}\beta - \mathbf{r}) = \mathbf{0} \quad (7)$$

$$\partial \ln L^R / \partial \text{vec} (\Sigma) = -1/2 \text{vec} (T\Sigma^{-1} - \Sigma^{-1} (\mathbf{Y} - \mathbf{X}\mathbf{B})' (\mathbf{Y} - \mathbf{X}\mathbf{B}) \Sigma^{-1}) = \mathbf{0} \quad (8)$$

因此, 参数矩阵  $\mathbf{B}$  和  $\Sigma$  的极大似然估计量是满足上述方程(6)~(8)的解.

### 3 截断非平衡似无关回归模型及其极大仿真似然估计

#### 3.1 截断非平衡似无关回归模型的设定

以上对似无关回归模型的参数估计要求所使用的样本必须是完整的, 并且不存在数据截断的问题. 然而, 从已有研究来看, 现代统计调查数据往往存在数据缺失和截断, 这一问题在微观计量的实证研究中尤为突出. 因此, 对似无关回归模型进行扩展, 使之能够消除数据缺失和截断对参数估计的影响, 就成为当前理论研究和实证分析的发展需要.

为此, 对于上述一个含有  $G$  个方程和  $T$  期观测值的似无关回归模型(1), 当含有数据缺失时, 可将其重新设定为:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U} \quad (9)$$

其中,  $\mathbf{Y} = (y_1, y_2, \dots, y_T)'$  和  $\mathbf{X} = (x_1, x_2, \dots, x_T)'$  分别是  $T \times G$  和  $T \times K$  维观测值矩阵,  $\mathbf{U} = (u_1, u_2, \dots, u_T)'$  为模型扰动项矩阵. 模型(9)中的参数矩阵  $\mathbf{B}$  仍然满足线性约束条件(2). 与方程(1)不同的是, 现在方程(9)给出了包含所有观测值的回归形式, 其目的在于欲说明数据缺失对参数估计的影响.

假设矩阵  $\mathbf{Y}$  中的某些观测值存在数据缺失<sup>4</sup>. 需要注意的是, 矩阵  $\mathbf{Y}$  包含了  $G$  个方程的因变量, 而在每一个方程之间, 数据缺失的情形可能并不相同, 由此导致矩阵  $\mathbf{Y}$  中每一列都可能含有不同的数据缺失形式. 换言之, 对于第  $i$  列,  $\mathbf{Y}$  可能含有  $T_i$  个观测值 ( $T_i < T$ ), 对于第  $j$  列,  $\mathbf{Y}$  可能含有  $T_j$  个观测值 ( $T_j < T$ ), 且第  $i$  列和第  $j$  列缺失的观测值之间可能并不存在包含或嵌套等特殊关系. 然而, 为表述方便, 我们仍然假设各个截面方程均含有  $T$  个观测值, 因为对于其中缺失的观测值, 我们可以用缺失符号“.”表示.

进一步, 若在  $\mathbf{Y}$  中同时存在数据截断, 例如, 当  $y_{it} < c$  时,  $y_{it}$  的观测值为  $c$ ; 当  $y_{it} > c$  时,  $y_{it}$  的观测值为  $y_{it}$  本身<sup>5</sup>. 沿用截断回归模型的表示方法, 可将  $y_{it}$  的观测值设定为

$$z_{it} = \max (y_{it}, c) \quad (10)$$

其中,  $y_{it}$  为潜在变量,  $z_{it}$  为  $y_{it}$  的可观测变量,  $c$  为已知阈值.

上述方程(9)和(10)是包含了数据缺失和截断的似无关回归模型, 本文称之为截断非平衡似无关回归模型. 常用的极大似然估计方法对于该模型而言将产生有偏的估计结果, 其原因是在方程(10)的约束下,  $z_{it}$  服从一个非线性的分布函数. 基于这个函数, 我们需要重新构造似然函数, 并利用新的仿真近似算法得到参数估计结果. 以下将着重说明截断非平衡似无关回归模型的估计思想.

#### 3.2 极大仿真似然估计

本文首先将截断非平衡的似无关回归模型转化为平衡的截断似无关回归模型, 然后再利用极大仿真似然方法进行估计. 根据观测值矩阵  $\mathbf{Y}$  中的数据缺失形式, 本文按照 Hwang 和 Schulman<sup>[12]</sup> 的思想将  $\mathbf{Y}$  划分为  $N$  个不同的组 ( $N \leq T$ ), 使得在每一个分组  $n$  中的  $T_n$  个样本具有相同的观测值或数据缺失形式. 即在第  $n$  个分组中, 可能有  $G_n$  个变量具有完整的观测值, 其余  $(G - G_n)$  个变量均为缺失值.

4. 当  $\mathbf{X}$  中数据缺失时, 我们可将其设定为 0. 对这一问题的具体表述可参见 Hwang 和 Schulman<sup>[12]</sup>.

5. 本文实际上给出了一种左截断的情形, 对于右截断的情形可依此类推. 实证中更为常见的一种情形是令  $c = 0$ .

上述划分思想可以用矩阵语言进一步明确为: 记  $\mathbf{P}_n$  为第  $n$  个分组中的  $T_n \times T$  维观测值选择矩阵, 其中  $\mathbf{P}_n$  的元素由  $T$  维单位矩阵中的  $T_n$  个不同行向量构成. 记  $\mathbf{Q}_n$  为第  $n$  个分组中对应于具有完整观测值变量的配置矩阵, 其中  $\mathbf{Q}_n$  中的元素由  $G$  维单位矩阵中的  $G_n$  个不同列向量构成. 也就是说,  $G$  维单位矩阵中剩余的  $(G - G_n)$  个列向量对应于具有缺失值的变量, 记由这些列向量构成的矩阵为  $\bar{\mathbf{Q}}_n$ . 基于此, 第  $n$  个分组中具有完整观测值的回归模型和具有缺失值的回归模型可以分别表述为:

$$\mathbf{Y}_n \mathbf{Q}_n = \mathbf{X}_n \mathbf{B}_n + \mathbf{V}_n \quad (11)$$

$$\mathbf{Y}_n \bar{\mathbf{Q}}_n = \mathbf{X}_n \bar{\mathbf{B}}_n + \bar{\mathbf{V}}_n \quad (12)$$

其中,  $\mathbf{Y}_n = \mathbf{P}_n \mathbf{Y}$ ,  $\mathbf{X}_n = \mathbf{P}_n \mathbf{X}$ ,  $\mathbf{U}_n = \mathbf{P}_n \mathbf{U}$ ,  $\mathbf{V}_n = \mathbf{U}_n \mathbf{Q}_n$ ,  $\bar{\mathbf{V}}_n = \mathbf{U}_n \bar{\mathbf{Q}}_n$ ,  $\mathbf{B}_n = \mathbf{B} \mathbf{Q}_n$ ,  $\bar{\mathbf{B}}_n = \mathbf{B} \bar{\mathbf{Q}}_n$ . 至此, 截断非平衡的似无关回归模型被转化为平衡的截断似无关回归模型 (11). 模型中扰动项  $\mathbf{V}_n$  的协方差矩阵可以相应表述为  $\Sigma_n = \mathbf{Q}'_n \Sigma \mathbf{Q}_n$ . 由于方程 (12) 包含了样本缺失值, 因此, 该方程并不为模型参数估计提供信息.

进一步, 由于变量  $\mathbf{Y}_n$  存在数据截断, 为清晰说明这一问题, 以模型 (11) 中第  $t$  期样本观测值为例:

$$\mathbf{y}_t^{(n)} = \mathbf{B}'_n \mathbf{x}_t + \mathbf{u}_t^{(n)}, \quad t = 1, 2, \dots, T_n \quad (13)$$

从表面上看, 方程 (13) 似乎与经典似无关回归模型 (1) 具有相同的形式. 但是, 此时  $\mathbf{y}_t^{(n)}$  只是由不含数据缺失的  $G_n$  个因变量构成, 即  $\mathbf{y}_t^{(n)} = (y_{1t}^{(n)}, y_{2t}^{(n)}, \dots, y_{G_n t}^{(n)})'$ ,  $y_{it}^{(n)}$  表示第  $n$  个分组中被矩阵  $\mathbf{Q}_n$  选择出来的第  $i$  个分量. 且由于  $\mathbf{y}_t^{(n)}$  为潜在变量, 即  $z_{it}^{(n)} = \max(y_{it}^{(n)}, c)$ . 因此, 即使在  $\mathbf{u}_t^{(n)} \sim \text{i.i.d.} N(\mathbf{0}, \mathbf{Q}'_n \Sigma \mathbf{Q}_n)$  的条件下,  $\mathbf{y}_t^{(n)}$  的可观测变量  $z_t^{(n)}$  也不服从标准的正态分布, 其分布函数是由截断的正态概率密度函数和累积分布函数构成的分段函数<sup>6</sup>, 即:

$$f(z_t^{(n)}) = \begin{cases} \Phi(c\boldsymbol{\iota} - \mathbf{B}'_n \mathbf{x}_t), & z_t^{(n)} = c\boldsymbol{\iota} \\ \varphi(\mathbf{y}_t^{(n)} - \mathbf{B}'_n \mathbf{x}_t), & z_t^{(n)} > c\boldsymbol{\iota} \end{cases} \quad (14)$$

其中,  $\Phi(\cdot)$  和  $\varphi(\cdot)$  分别表示多维累积正态分布函数和概率密度函数,  $\boldsymbol{\iota}$  是常数为 1 的列向量. 故按照极大似然估计思想, 此时基于第  $n$  个分组样本的对数似然函数为:

$$\begin{aligned} \ln L_n = & \sum_{t=1}^{T_n} -\frac{1}{2} \left\{ G_n \ln(2\pi) + \ln |\mathbf{Q}'_n \Sigma \mathbf{Q}_n| + (\mathbf{y}_t^{(n)} - \mathbf{B}'_n \mathbf{x}_t)' \mathbf{Q}_n^{-1} \Sigma^{-1} \mathbf{Q}'_n^{-1} (\mathbf{y}_t^{(n)} - \mathbf{B}'_n \mathbf{x}_t) \right\} + \\ & \sum_{\substack{t=1 \\ z_t^{(n)} > c\boldsymbol{\iota}}}^{T_n} \ln \Phi(c\boldsymbol{\iota} - \mathbf{B}'_n \mathbf{x}_t) \end{aligned} \quad (15)$$

而基于全部  $N$  个组的对数似然函数  $\ln L$  是  $\ln L_n$  的和:

$$\ln L = \sum_{t=1}^{T_n} \ln L_n \quad (16)$$

因此, 参数矩阵  $\mathbf{B}$  和  $\Sigma$  的极大似然估计量是在条件 (2) 的约束下, 求方程 (16) 最大化, 这就涉及到计算一个多元累积正态分布函数  $\Phi(\cdot)$  的概率积分, 然而现有方法并不能通过对方程 (16) 求一阶条件而得到参数  $\mathbf{B}$  和  $\Sigma$  的具体解析解. 因此, 本文首先利用 GHK 方法仿真计算  $\Phi(\cdot)$  的概率积分值, 然后再对似然函数 (16) 进行迭代求解, 这种方法即为极大仿真似然估计.

具体来看, 在对一个服从均值为  $\mathbf{0}$ , 协方差为  $\Sigma$  的  $K$  维正态随机变量  $\mathbf{X}$  计算  $\Phi(\cdot)$  的概率积分<sup>7</sup> 时, 需计算其联合概率积分  $\text{Prob}(a_1 < x_1 < b_1, \dots, a_K < x_K < b_K)$ , 该联合概率积分可用单个随机变量概率积分的乘积的均值来近似, 即:

$$\text{Prob}(a_1 < x_1 < b_1, \dots, a_K < x_K < b_K) \approx \frac{1}{R} \sum_{r=1}^R \prod_{k=1}^K Q_{rk} \quad (17)$$

其中,  $Q_{rk}$  表示单个随机变量  $x_k$  落在积分区间  $(a_k, b_k)$  的概率, 而  $\prod_{k=1}^K Q_{rk}$  则表示  $K$  维随机变量  $\mathbf{X}$  同时落在区间  $(a_1 < x_1 < b_1, \dots, a_K < x_K < b_K)$  一次的概率,  $R$  表示仿真抽样的次数. 且  $Q_{rk}$  按如下递归方法计算: 首先对协方差矩阵  $\Sigma$  进行 Cholesky 分解, 即  $\Sigma = \mathbf{L} \mathbf{L}'$ , 记矩阵  $\mathbf{L}$  中的元素为  $l_{km}$ ; 然后计算第一个随机变量的概率积分  $Q_{r1} = \Phi(b_1/l_{11}) - \Phi(a_1/l_{11})$ , 并从区间  $(A_{r1}, B_{r1}) = (a_1/l_{11}, b_1/l_{11})$  中生成标准正态分布的随机数  $\varepsilon_{r1}$ , 这里  $\Phi(\cdot)$  指单个随机变量的正态分布函数; 最后, 按  $Q_{rk} = \Phi(B_{rk}) - \Phi(A_{rk})$

6. 方程 (14) 的  $z_t^{(n)}$  中可能有部分分量大于  $c$ , 部分分量等于  $c$ .

7. 对 GHK 仿真计算方法的描述具体可参见 Greene<sup>[14]</sup>.

等式递归计算第  $k = 2, 3, \dots, K$  个随机变量的概率积分, 其中  $B_{rk} = (b_k - \sum_{m=1}^{k-1} l_{km} \varepsilon_{rm})/l_{kk}$ ,  $A_{rk} = (a_k - \sum_{m=1}^{k-1} l_{km} \varepsilon_{rm})/l_{kk}$ .

#### 4 蒙特卡罗实验

本文利用蒙特卡罗实验方法检验极大仿真似然估计的可行性. 在数据生成时, 假设截断非平衡似无关回归模型由如下两个方程构成:

$$\begin{cases} y_1 = 0.5 + 2x_1 - 1.5x_2 + u_1 \\ y_2 = 3 + 0.5x_3 - 3x_4 + u_2 \end{cases} \quad (18)$$

并且, 潜在变量  $y_1$  和  $y_2$  的样本截断值为 0, 即  $z_1 = \max(y_1, 0)$ ,  $z_2 = \max(y_2, 0)$ . 为避免解释变量的正态分布对参数估计带来的“增进效果”(Chesher<sup>[15]</sup>), 本文在数据生成时令  $x$  服从区间  $(0, 1)$  上的均匀分布,  $u$  服从标准正态分布, 且  $u_1$  和  $u_2$  之间的相关系数为  $\rho$ . 在仿真实验时, 取  $\rho = 0.2, 0.5, 0.8$  依次表示模型的低度相关、中度相关和高度相关.

进一步, 本文沿用 Hwang 和 Schulman<sup>[12]</sup> 的思想设计模型 (18) 的样本缺失情况. 假设模型 (18) 中两个方程的样本总数均为  $T$ , 其中, 第一个方程的最初  $T_1$  个样本为缺失样本, 第二个方程的最后  $T_3$  个样本为缺失样本, 两个方程含有的共同样本为  $T_2$ . 在仿真实验时, 样本缺失数  $(T_1, T_3)$  均设计为样本总数  $(T)$  的 10%. 取  $T = 50, 100, 200, 500$ , 分别表示截断非平衡似无关回归模型的有限样本和大样本情况.

本文将蒙特卡罗实验重复 1000 次, 并计算各个系数估计值的均值 (Mean) 和均方误差根 (RMSE), 所得结果如表 1 所示. 由实验结果可以看出, 截断非平衡似无关回归模型的极大仿真似然估计能获得非常好的估计结果. 从整体来看, 各个系数估计值的均值都非常接近于参数真值, 并没有表现出向上或向下的系统性估计偏误, 且估计值的均方误差根也都比较小, 表明在 1000 次重复实验中, 每一次系数估计的结果均比较接近于参数真值. 具体来看, 当样本容量  $T = 50$  时, 各个系数估计值表现出了 0.04 左右的绝对偏差, 而当样本容量  $T = 100$  或更高时, 估计的偏误明显减少, 且估计的均方误差根也显著下降, 这说明截断非平衡似无关回归模型的极大仿真似然估计具有良好的一致性. 同时, 如表 1 所示的蒙特卡罗实验结果还进一步表明, 由于极大仿真似然估计方法考虑了模型扰动项之间的同期相关性, 因此, 截面相关系数  $\rho$  的高低并没有对系数估计结果产生显著影响.

表 1 极大仿真似然估计的蒙特卡罗仿真实验结果

$\alpha_0 = 0.5$		$\alpha_1 = 2$		$\alpha_2 = -1.5$		$\beta_0 = -0.5$		$\beta_1 = -1$		$\beta_2 = 2$		
均值	均方根	均值	均方根	均值	均方根	均值	均方根	均值	均方根	均值	均方根	
$(\rho = 0.2)$												
$T = 50$	0.462	0.512	2.038	0.721	-1.47	0.704	-0.531	0.649	-1.026	0.402	2.038	0.473
$T = 100$	0.499	0.344	1.997	0.405	-1.505	0.436	-0.515	0.36	-1.009	0.231	2.02	0.256
$T = 200$	0.493	0.205	2.019	0.267	-1.503	0.27	-0.5	0.231	-1.001	0.158	2.002	0.166
$T = 500$	0.502	0.12	2.005	0.161	-1.509	0.169	-0.502	0.132	-0.999	0.09	2.001	0.101
$(\rho = 0.5)$												
$T = 50$	0.453	0.48	2.046	0.674	-1.461	0.654	-0.527	0.59	-1.026	0.366	2.035	0.429
$T = 100$	0.502	0.324	1.993	0.377	-1.508	0.4	-0.515	0.349	-1.009	0.207	2.021	0.233
$T = 200$	0.495	0.19	2.018	0.243	-1.505	0.247	-0.497	0.209	-1.004	0.143	2.001	0.153
$T = 500$	0.502	0.11	2.003	0.148	-1.506	0.153	-0.503	0.121	-0.999	0.083	2.002	0.091
$(\rho = 0.8)$												
$T = 50$	0.46	0.39	2.031	0.531	-1.461	0.521	-0.52	0.453	-1.02	0.275	2.027	0.325
$T = 100$	0.503	0.274	1.995	0.294	-1.511	0.301	-0.513	0.293	-1.008	0.153	2.017	0.177
$T = 200$	0.497	0.155	2.011	0.184	-1.501	0.187	-0.496	0.16	-1.003	0.104	2	0.115
$T = 500$	0.501	0.086	2.002	0.112	-1.504	0.112	-0.501	0.092	-0.999	0.061	2	0.067

#### 5 结论

似无关回归模型是研究一般截面相关问题的方法论基础, 并在当前实证研究中具有广泛的应用. 本文对似无关回归模型进行扩展, 构建了一个截断非平衡的似无关回归模型, 使之能够消除样本数据缺失和截断对

模型参数估计的影响。在估计方法上, 本文首先基于 Hwang 和 Schulman<sup>[12]</sup> 的思想把截断非平衡似无关回归模型转化为平衡的截断似无关回归模型, 然后建立该模型的极大仿真似然估计。蒙特卡罗实验表明, 本文建立的参数估计方法无论在大样本或者是有限样本下均具有良好表现, 并且该方法对模型扰动项之间的同期相关性也表现出了较强的稳健性。本文建立的估计方法对一般截面相关模型的参数估计具有较好的借鉴作用。

## 参考文献

- [1] Moeltner K, Layton D F. A censored random coefficients model for pooled survey data with application to the estimation of power outage costs[J]. *The Review of Economics and Statistics*, 2002, 83(4): 552–561.
- [2] Schenker N, Raghunathan T E, Chiu P L, et al. Multiple imputation of missing income data in the national health interview survey[J]. *Journal of the American Statistical Association*, 2006, 101(475): 924–933.
- [3] Heckman J J, MaCurdy T E. A life cycle model of female labor supply[J]. *Review of Economic Studies*, 1980, 47(1): 47–74.
- [4] Charlier E, Melenberg B, Soest A V. Estimation of a censored regression panel data model using conditional moment restrictions efficiently[J]. *Journal of Econometrics*, 2000, 95(1): 25–56.
- [5] Greene W. The behavior of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects[J]. *Econometrics Journal*, 2004, 7(1): 98–119.
- [6] Chen S. Root- $N$ -consistent estimation of fixed-effect panel data transformation models with censoring[J]. *Journal of Econometrics*, 2010, 159(1): 222–234.
- [7] Phillips P C B. The exact distribution of the SUR estimator[J]. *Econometrica*, 1985, 53(4): 745–756.
- [8] Geweke J. *Contemporary Bayesian econometrics and statistics*[M]. Hoboken NJ: Wiley, 2003.
- [9] Baltagi B H, Song S H. Unbalanced panel data: A survey[J]. *Statistical Papers*, 2006, 47(4): 493–523.
- [10] Schmidt P. Estimation of seemingly unrelated regressions with unequal numbers of observations[J]. *Journal of Econometrics*, 1977, 5(3): 365–377.
- [11] Baltagi B H, Garvin S, Kerman S. Further Monte Carlo evidence on seemingly unrelated regressions with unequal number of observations[J]. *Annals of Economics and Statistics*, 1989, 14: 103–115.
- [12] Hwang H S, Schulman C. Estimation of SUR model with non-nested missing observations[J]. *Annals of Economics and Statistics*, 1996, 44: 219–240.
- [13] Zellner A. An efficient method of estimating seemingly unrelated regression and tests for aggregation bias[J]. *Journal of the American Statistical Association*, 1962, 57(298): 348–368.
- [14] Greene W. *Econometric analysis*[M]. 5th ed. Upper Saddle River, NJ: Prentice Hall, 2003.
- [15] Chesher A. A mirror image invariance for M-estimators[J]. *Econometrica*, 1995, 63(1): 207–213.