**Pure Mathematics** 理论数学, 2021, 11(7), 1320-1325
Published Online July 2021 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2021.117148

Hans汉斯

# 一种基于Tobit回归模型的序贯压缩估计方法研究

**鲁海波**

新疆师范大学数学科学学院，新疆 乌鲁木齐
Email: andyluhaibo@foxmail.com

收稿日期：2021年5月31日；录用日期：2021年7月1日；发布日期：2021年7月8日

## 摘　要

**Tobit回归模型在计量经济学等研究领域中有着广泛的应用。但是我们在处理面板数据以及时间序列数据时经常会遇到包含太多变量的数据集，而这些变量中只有少数变量对模型有贡献。为了去除这些“无效变量”的影响，在本文中，我们提出一种基于自适应压缩估计的序贯抽样策略来构造“有效”参数的固定长度的置信集，并在自适应设计下对所提出的序贯抽样策略进行数值模拟，最后数值模拟达到了预期的效果。**

## 关键词

**Tobit模型，样本量，序贯压缩估计，停止法则**

# A Sequential Shrinkage Estimate Based on Tobit Regression Model

**Haibo Lu**

School of Mathematics Science, Xinjiang Normal University, Urumqi Xinjiang
Email: andyluhaibo@foxmail.com

Received: May 31st, 2021; accepted: Jul. 1st, 2021; published: Jul. 8th, 2021

## Abstract

**In the applications of Tobit regression models we always encounter the data sets which contain too many variables, but only a few of them contribute to the model. Therefore, it will waste much**

文章引用: 鲁海波. 一种基于 Tobit 回归模型的序贯压缩估计方法研究[J]. 理论数学, 2021, 11(7): 1320-1325.
DOI: 10.12677/pm.2021.117148

**more samples to estimate the "non-effective" variables in the inference. In this paper, we use a sequential procedure for constructing the fixed size confidence set for the "effective" parameters to the model based on an adaptive shrinkage estimate such that the "effective" coefficients can be efficiently identified with the minimum sample size. Adaptive design is considered for numerical simulation.**

## Keywords

**Tobit Models, Sample Size, Shrinkage Estimate, Stopping Rule**



## 1. 引言

Tobit 回归模型[1]是一种因变量受限模型，被称作样本选择模型，或者删失回归模型。Tobit 回归模型被广泛应用于计量经济学等众多研究领域[2] [3] [4]，在面板数据和时间序列数据的分析中发挥着越来越重要的作用。假设 $a^+ = \max\{a, c\}$，我们可以如下定义 Tobit 回归模型

$$y_i^+ = \max\left\{x_i^T \beta_0 + \varepsilon_i, c\right\},\ i = 1, 2, \cdots, n \tag{1}$$

其中，$\beta_0$（$p$ 维向量)是回归系数，$x_i$ 是 $p$ 维协变量，$\varepsilon_i$ 是随机误差。然而，在计量经济学等领域对面板数据或者时间序列数据等分析研究中，常常会遇到数据集通常有大量的解释变量，但其中只有少数对模型有贡献。也就是说，在一个 $p$ 维的回归系数中只有 $p_0$（$p_0 < p$ 且 $p_0$ 未知)个分量是取非零值的，我们称之为有效变量[5]。目前有很多方法可以用来识别有效变量，如 LASSO [6]和 LARS [7]等等。但另外需要关注的问题是，用多少样本才能既识别出有效变量，同时又能使参数估计达到预定的精度。这对于计量经济学等领域需要考虑抽样成本的研究具有重要的意义。对于线性回归模型，Wang 和 Zhang (2013) [5]提出了一种序贯压缩估计方法来识别有效变量，从而达到参数估计的精度。数值模拟结果表明，与传统的序贯抽样方法相比，序贯压缩估计不仅可以从所有变量中识别出有效变量，而且可以节省大量样本。对于 Tobit 回归模型，如何提出相应的序贯估计方法以及在自适应设计下给出相关性质和数据模拟有待进一步的研究。本文针对 Tobit 回归模型提出了一种基于自适应压缩估计(ASE)来构造有效变量的固定窗宽的置信集的序贯抽样方法，使有效变量能以最小样本量快速识别。本文将在适应性设计(adaptive design)下研究所提出的自适应压缩估计(ASE)的大样本性质，同时在自适应性设计下通过数值模拟得到了很好的模拟结果。

## 2. 基于 Tobit 模型的序贯自适应压缩估计(ASE)

### 2.1. 最小一乘估计(LAD)

不失一般性在模型(1)中，令 $c = 0$。假设随机误差 $\varepsilon_i$，$i = 1, 2, \cdots, n$ 独立同分布且 $\varepsilon_i \sim N\left(0, \sigma^2\right)$，那么似然函数的形式为：

$$L = \prod_0 \left(1 - \Phi\left(\frac{x_i^{\mathrm{T}} \beta}{\sigma}\right)\right) \prod_1 \sigma^{-1} \phi\left(\frac{x_i^{\mathrm{T}} \beta}{\sigma}\right)$$

其中 $\Phi$ 和 $\phi$ 分别为标准正态分布的概率分布函数和密度函数，$\prod_0$ 为集合 $\{i: y_i \le 0\}$ 中若干元素的乘积，$\prod_1$ 为集合 $\{i: y_i > 0\}$ 中若干元素的乘积。记

$$Q_n(\beta) = \sum_{i=1}^{n} \left| y_i^+ - \max\left\{x_i^{\mathrm{T}}\beta, 0\right\} \right|$$

使 $Q_n(\beta)$ 达到最小的 $\beta$ 被称为回归参数 $\beta$ 的最小一乘估计[8]，记为 $\tilde{\beta}_n$。我们给定假设条件：

(A1) $\sup_i \|x_i\| < \infty$；

(A2) 若随机误差 $\varepsilon_i$ 的密度函数 $f(x)$ 满足 $f(0)=0$ 和 $med(\varepsilon_i)=0$，那么存在 $\delta > 0$ 使得 $\lim_{n\to\infty} \frac{\lambda}{\log n} \sum_{i=1}^{n} I\left(x_i^{\mathrm{T}}\beta > \delta\right) x_i x_i^{\mathrm{T}} = \infty$。

当 $\tilde{\beta}_n$ 满足(A1)和(A2)时，文献[9]给出了 $\tilde{\beta}_n$ 的相合性和渐近正态性：

$$\lim_{n\to\infty} \tilde{\beta}_n = \beta_0, \ a.s.$$

$$\left(2f(0)M_n^{1/2}\right) \cdot \sqrt{n}\left(\tilde{\beta}_n - \beta_0\right) \xrightarrow{d} N(0, I_n)$$

其中 $I_n$ 是单位阵，并且 $M_n = E\left(\frac{1}{n}\sum_i I\left(x_i^{\mathrm{T}}\beta_0 > 0\right) x_i x_i^{\mathrm{T}}\right)$。

## 2.2. 自适应压缩估计(ASE)

设 $\kappa = \kappa(n)$，当 $n \to \infty$ 时，存在 $0 < \delta < 1/2$ 和 $\gamma > 0$ 使得 $n^{\frac{1}{2}}\kappa \to 0$，$n^{\frac{1}{2}+\gamma\delta}\kappa \to \infty$。下面我们给出 Tobit 回归模型下回归系数的自适应压缩估计的定义：

**定义 2.2.1** 设 $\tilde{\beta}$ 为模型(1)的最小一乘估计，则称 $\hat{\beta}_n = In(\varepsilon)\tilde{\beta}_n$ 为回归系数 $\beta_0$ 的自适应压缩估计(ASE)，其中 $I_n(\varepsilon) = diag\left\{I_{n1}(\varepsilon), I_{n2}(\varepsilon), \cdots, I_{np}(\varepsilon)\right\}$ 是一个 $p \times p$ 维对角阵。同时可以证明 $\hat{\beta}_n = In(\varepsilon)\tilde{\beta}_n$ 满足相合性和渐进正态性。

## 2.3. 序贯抽样策略

依据文献[10] [11]中的结论我们可以证明 $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$, $n = 1, 2, \cdots$ 是依概率一致连续的，由此可得如下定理：

**定理 2.3.1** 设随机变量 N(t)取正整数值，当 $t \to \infty$ 有 $N(t)/t$ 依概率收敛于 1，且条件(A1)和(A2)成立，则当 $t \to \infty$ 时，

$$\sqrt{N(t)}\left(\hat{\beta}_{N(t)} - \beta_0\right) \to N\left(0, I_0 \Sigma I_0^{-1}\right)$$

由定理 2.3.1 我们可以构造 $\beta_0$ 的置信集和能够决定最小样本量的停止法则的序贯抽样策略。设 $\left\{(y_i, x_i): i = 1, 2, \cdots, k\right\}$ 是最先进入研究的 $k$ 个样本，用 $C_k$ 来表示。在任意给定小正数 $\varepsilon$ 下，

$$\hat{p}_0(k) = \sum_{j=1}^{p} I_{kj}(\varepsilon)$$

是回归系数 $p_0$ 基于条件 $C_k$ 的估计量。令 $a_k^2 \in R$ 对任意 $\alpha > 0$，有 $P\left(\chi^2_{\hat{p}_0(k)} \le a_k^2 \middle| C_k\right) = 1 - \alpha$ 成立。现在定义停时法则 $N_d$ 为

$$N = N_d \equiv \inf\left\{k : k \ge n_0 \text{ and } \frac{d^2}{a_k^2} \ge \nu_k\right\}, \tag{2}$$

其中 $\nu_k$ 是 $kI_k(\varepsilon)(\Sigma)^{-1}I_k(\varepsilon)$ 的最大特征值，$d$ 是置信集的预设精度。在本文的序贯估计策略中，一次只有一个新的观测进入研究直到满足(2)式的停止法则时就停止抽样，此时 $\beta_0$ 的置信集为

$$R_N=\left\{Z\in R^p:\frac{S_N}{N}\le\frac{d^2}{\nu_N}\text{且当}I_{N_j}(\varepsilon)=0\text{时},\ z_j=0,1\le j\le p\right\} \tag{3}$$

其中 $S_N=\left(Z_{N_1}-\hat{\beta}_{N_1}\right)^{\mathrm{T}}\tilde{\Sigma}_{11}\left(Z_{N_1}-\hat{\beta}_{N_1}\right)$。我们所提出的序贯抽样方法致力于找到有效变量的同时忽略无效变量的影响，这是和传统序贯方法相比我们能够节省大量样本的关键，在下面的定理中我们给出停时 $N_d$ 和置信集 $R_N$ 的相关性质。

**定理 2.3.2** 假定条件(A1)和(A2)都成立，设 $N$ 是满足(2)式的停时，则：

i) $\lim\limits_{d\to0}\dfrac{d^2N}{a^2\nu}=1$，a.s.；ii) $\lim\limits_{d\to0}P(\beta_0\in R_N)=1-\alpha$；

iii) $\lim\limits_{d\to0}\dfrac{d^2E(N)}{a^2\nu}=1$；iv) $\lim\limits_{d\to0}\hat{p}_0(N)=p_0$，a.s.且 $\lim\limits_{d\to0}E\left(\hat{p}_0(N)\right)=p_0$，

其中 $\nu$ 是矩阵 $I_0\Sigma^{-1}I_0$ 的最大特征值。

## 3. 数值模拟

在固定样本量下用所提方法对随机数据集合进行分析，以此来验证所提出的序贯压缩估计方法的性能。按照停止法则的定义，当抽样停止时，最终的置信集将满足预设精度和覆盖概率，因此我们可以比较分别基于 LAD 和 ASE 的序贯抽样方法的平均停时。由于序贯压缩估计方法忽略无效变量的影响，故理论上平均所需停时应该显著小于不考虑变量选择的序贯方法。如果事先已知有效变量为 $p_0$ 个同时无无效变量，那么只使用这 $p_0$ 个有效变量的序贯方法无疑是效率最高的。所以，为便于比较，我们将所有( $p_0$ 个)变量全部为有效变量的序贯估计方法作为基准线，在此情况下所获得的样本量应该是最小的。在自适应设计下，随机模拟数据集中的 $x_1$ 仍然由多元标准正态分布生成，$x_j\,(j>1)$ 由均值为 $\sum_{i=1}^{j-1}\left[x_j/(j-1)\right]$，方差协方差矩阵为单位阵的多元正态分布生成。不失一般性，选择模型(1)中的常数 $c=0$。回归系数真值取 $(-1.2,2.0,0,0,0,0,0,0,0,0)$，其中含有八个无效变量，回归系数置信集的预设精度 $d\in\{0.3,0.4,0.5,0.6\}$，取 $\alpha=0.05$，$\gamma=1$，$\delta=0.45$，$\theta=0.75$。另外当用 ASE 方法时我们用 BIC 方法来确定 $\varepsilon$。

表 1 描述了 Tobit 回归模型下的序贯抽样方法的数值模拟结果。在表 1 中我们列出了最终样本量 N (停时)，$\kappa^*=d^2N/\left(a^2\nu\right)$ 和 95%置信集的经验覆盖概率 $R_N$。所有三种情况( $\mathrm{LAD}_{p_0}$, ASE, LAD)下的 $\kappa$ 值都非常接近 1，并且当 $d$ 不断减小时经验覆盖概率 CP 越来越接近 95%，正如定理 2.3.2 描述的一样。然而，应用 LAD 方法所得的样本量 $N$ 比应用 ASE 方法和 $\mathrm{LAD}_{p_0}$ 都大得多。而应用 ASE 的抽样策略所需的样本量和应用 $\mathrm{LAD}_{p_0}$ 的抽样策略所需样本量差不多，这说明我们所提方法在变量选择的同时效率和回归参数中只有有效变量无无效变量的情况下的效率非常接近，而比不做变量选择情况下(即 LAD)的抽样效率提高很多。

表 2 比较了在估计 Tobit 回归模型的回归系数时分别应用 ASE 和 LAD 的抽样策略对识别回归系数中的有效变量和无效变量的效率。从结果可以看出应用 ASE 的抽样策略时不能被正确识别的零变量的平均个数几乎趋向于 0，而能被正确识别的非 0 变量的平均个数和模型中有效变量个数的真值非常接近(2 和 8)。结果表明基于 ASE 的序贯抽样策略下 $\hat{p}_0$ 是 $p_0$ 的优良估计。而基 LAD 的序贯抽样策略不能识别有效变量，因此无法获得 $N_c^*$ 和 $N_{ic}^*$ 的值。此外，所有参数的估计值和它们的真值都非常接近。

**Table 1.** Results of sequential sampling method based on ASE, LAD with all variables and LAD with only $p_0$ non-zero variables for Tobit regression model
**表 1.** Tobit 回归模型下分别应用 ASE，LAD 和 $\text{LAD}_{p_0}$ 的序贯抽样方法的结果分析

$\beta=(-1.2,\ 2.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0)$

| | | $\text{LAD}_{p_0}$ | | | ASE | | | LAD | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Design | d | $N$ | $\kappa^*$ | $CP^+$ | $N$ | $\kappa$ | $CP^+$ | $N$ | $\kappa$ | $CP$ |
| Adaptive | 0.6 | 93.58 (12.36) | 1.021 | 0.95 | 111.75 (19.04) | 1.040 | 0.935 | 292.71 (30.16) | 1.006 | 0.95 |
| | 0.5 | 127.55 (18.89) | 1.014 | 0.96 | 154.15 (23.01) | 1.031 | 0.92 | 371.82 (33.36) | 1.005 | 0.93 |
| | 0.4 | 192.02 (28.9) | 1.011 | 0.95 | 224.6 (24.90) | 1.019 | 0.935 | 598.85 (43.25) | 1.003 | 0.93 |
| | 0.3 | 349.08 (31.57) | 1.000 | 0.95 | 352.98 (39.08) | 1.016 | 0.93 | 983.90 (65.16) | 1.002 | 0.97 |

$\kappa^*=d^2N/(a^2\nu)$；$CP^+$ 是 95%置信集 $R_N$ 的经验覆盖概率；**经验标准差在括号内。

**Table 2.** Power of variable identification and estimation of nonzero components under sequential sampling method based on ASE and LAD with Tobit regression model
**表 2.** Tobit 回归模型下分别应用 ASE 和 LAD 的序贯抽样策略的变量识别和非零参数估计效率

$\beta_1=-1.2,\ \beta_2=2.0$

| | | ASE | | | | LAD | | | |
|---|---|---|---|---|---|---|---|---|---|
| Design | d | $N_{ic}^*$ | $N_c^*$ | $\beta_1$ | $\beta_2$ | $N_{ic}^*$ | $N_c^*$ | $\beta_1$ | $\beta_2$ |
| Adaptive | 0.6 | 0 | 7.76 | −1.28 (0.16) | 2.07 (0.18) | - | - | −1.258 (0.09) | 2.074 (0.117) |
| | 0.5 | 0 | 7.92 | −1.21 (0.19) | 2.104 (0.113) | - | - | −1.226 (0.07) | 2.031 (0.095) |
| | 0.4 | 0 | 7.97 | −1.23 (0.11) | 2.061 (0.015) | - | - | −1.213 (0.069) | 2.021 (0.079) |
| | 0.3 | 0 | 7.985 | −1.21 (0.07) | 2.013 (0.006) | - | - | −1.208 (0.044) | 2.01 (0.068) |

$N_{ic}^*$：$\beta$ 中零分量(无效变量)被错误识别的平均个数；$N_c^*$：$\beta$ 中非零分量(有效变量)被正确识别的平均个数。

## 4. 结论

在 Tobit 回归模型下基于自适应压缩估计(ASE)建立的序贯抽样方法不仅能够用最少的样本识别出回归参数中的有效变量，同时可以使回归参数的估计值达到预设的精度[12]。我们在自适应设计下对相关性质做数值模拟，结果表明和传统的序贯抽样方法相比，我们提出的方法能够节省大量样本。然而，本文中所提方法涉及到的变量维数是固定的，后期我们将研究当变量维数随样本量变化时的序贯抽样方法的相关性质。

## 基金项目

1) 新疆师范大学博士科研启动基金项目：“基于广义线性模型的序贯分析研究”XJNUBS1539；
2) 新疆维吾尔自治区高校科研计划项目：“基于 Cox 比例风险回归模型的序贯分析研究”(XJEDU2016I033)。

## 参考文献

[1] Tobin, J. (1958) Estimation of Relationships for Limited Dependent Variables. *Econometrica*, **26**, 24-36. https://doi.org/10.2307/1907382

[2] Adams, J.D. (1980) Personal Wealth Transfers. *Quarterly Journal of Economics*, **95**, 159-179. https://doi.org/10.2307/1885354

[3] Ashenfelter, O. and Ham, J. (1979) Education, Unemployment, and Earnings. *Journal of Political Economy*, **87**, S99-S116. https://doi.org/10.1086/260824

[4] Fair, R.C. (1978) A Theory of Extramarital Affairs. *Journal of Political Economy*, **86**, 45-61. https://doi.org/10.1086/260646

[5] Wang, Z.F. and Chang, Y.I. (2013) Sequential Estimate for Linear Regression Models with Uncertain Number of Effective Variables. *Metrika*, **76**, 949-978. https://doi.org/10.1007/s00184-012-0426-4

[6] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[7] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Journal of Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[8] Powell, J.L. (1984) Least Absolute Deviations Estimation for the Censored Regression Model. *Journal of Econometrics*, **25**, 303-325. https://doi.org/10.1016/0304-4076(84)90004-6

[9] Chen, X.R. and Wu, Y.H. (1994) Consistency of $l_1$ Estimates in Censored Linear Regression Models. *Communications in Statistics*, **23**, 1847-1858. https://doi.org/10.1080/03610929408831360

[10] Anscombe, F.J. (1952) Large Sample Theory of Sequential Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **48**, 600-607. https://doi.org/10.1017/S0305004100076386

[11] Woodroofe, M. (1982) Nonlinear Renewal Theory in Sequential Analysis. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611970302

[12] Chow, Y.S. and Robbins, H. (1965) On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean. *Annals of Mathematical Statistics*, **36**, 457-462. https://doi.org/10.1214/aoms/1177700156