Tobit 模型与税收稽查

李选举

ABSTRACT

The cases-choice for tax audit contains two main aspects: the first is to distinguish whether the taxpayer is tax evasion, the second is to estimate the quantum of tax evasion. The paper systematically have studied the estimating of the quantum of tax evasion and made progresses and results. The author first applied the Tobit model to estimate the quantum of tax evasion and fairly solved the problems in the estimations of the quantum of tax evasion.

关键词：税收稽查；Tobit 模型；估测；逃税额

一、Tobit 模型

Tobit 模型是经济学家、1981 年诺贝尔经济学奖获得者 J·托宾(James. Tobin)1958 年在研究耐用消费品需求时首先提出来的一个经济计量学模型。其基本结构如下：

设某一耐用消费品支出为 y_i(被解释变量)，解释变量为 X_i，则耐用消费品支出 y_i 要么大于 y_0(y_0 表示该耐用消费品的最低支出水平)，要么等于零。因此，在线性模型假设下，耐用消费品支出 y_i 和解释变量 X_i 之间的关系为：

$$y_i=\begin{cases}\beta^T X_i+e_i & \text{若 } \beta^T X_i+e_i>y_0\\ 0 & \text{其他}\end{cases} \quad (1)$$

$$e_i\sim N(0,\sigma^2),\ i=1,2,\cdots,n$$

其中，X_i 是(k+1)维的解释变量向量，β 是(k+1)维的未知参数向量。此模型称为截取回归模型(censored regression model)。假设 y_0 已知，模型两边同时减去 y_0，变换后模型的常数项是原常数减去 y_0，由此得到的模型标准形式称为“Tobit 模型”(tobit regression model)：

$$y_i=\begin{cases}\beta^T X_i+e_i & \text{若 } \beta^T X_i+e_i>0\\ 0 & \text{其他}\end{cases} \quad (2)$$

$$e_i\sim N(0,\sigma^2),\ i=1,2,\cdots,n$$

Tobit 模型还可表示为：

$$y_i=\max\{\beta^T X_i+e_i,0\} \quad (3)$$

或者

$$y_i^*=\beta^T X_i+e_i$$

$$y_i=\begin{cases}y_i^* & \text{若 } y_i^*>0\\ 0 & \text{若 } y_i^*\leqslant 0\end{cases} \quad (4)$$

Tobit 模型的一个重要特征是，解释变量 X_i 是可观测的(即 X_i 取实际观测值)，而被解释变量 y_i 只能以受限制的方式被观测到：当 $y_i^*>0$ 时，取 $y_i=y_i^*>0$，称 y_i 为“无限制”观测值；当 $y_i^*\leqslant 0$ 时，取 $y_i=0$，称 y_i 为“受限”观测值。即，“无限制”观测值均取实际的观测值，“受限”观测值均截取为 0。

本文得到国家自然科学基金资助，是国家自然科学基金资助项目“金融数学、金融工程、金融管理—税收系统工程”的部分成果。

二、Tobit 模型的性质及其估计[1],[2]

建立 Tobit 模型,就是要求在对 y_i 和 X_i 进行 $n(n>k)$次观测的基础上估计 β 和 σ^2。下面讨论 Tobit 模型的性质及其估计问题。设 n_0 是使 $y_i=0$ 的观测值个数, n_1 是使 $y_i>0$ 的观测值个数, $n=n_0+n1$。

如果将 $y_i=0$ 的 n_0 个观测值忽略不计,则剩余的 n_1 个观测值是完全观测值($y_i>0$),可用最小二乘法估计 β,但是最小二乘估计量在此范围内是有偏的,并且是不一致的。实际上,观测值 y_i 在 $y_i>0$ 下的条件期望为

$$E(e_i|y_i>0)=E(e_i|e_i>-\beta^T X_i)=\sigma\cdot f_i/F_i \qquad (5)$$

因此

$$Y_i=\beta^T X_i+e_i=\beta^T X_i+\sigma\frac{f_i}{F_i}+u_i \qquad (6)$$

其中, f_i 和 F_i 分别是在 $(\beta^T X_i/\sigma)$处计算的标准正态分布的概率密度函数和分布函数。由于最小二乘估计忽略了与 X_i 不独立的 $\sigma\frac{f_i}{F_i}$ 项,因此造成最小二乘估计量的有偏性和不一致性。又因为 $e_i\sim N(0,\sigma^2)$,所以

$$E(e_i|y_i>0)=\sigma\cdot f_i/F_i>0$$

即 β 的最小二乘估计量是有偏的。

若考虑全部 $n=n_0+n_1$ 个观测值,则观测值 y_i 的无条件期望为

$$E(y_i)=F_i\cdot(\beta^T X_i)+\sigma\cdot f_i$$

因此,对所有的 n 个观测值应用最小二乘法也不会产生 Tobit 模型的无偏估计量和一致估计量。

可以证明, β 和 σ^2 的最大似然估计量是一致估计量。因此估计 Tobit 模型的最好方法是最大似然估计。

三、Tobit 模型与税收稽查[7]

下面根据某地区商业企业的纳税资料作实证研究,讨论 Tobit 模型在税收稽查中的应用。不妨考虑商业企业的所得税。

假设要估算商业企业应纳所得税额(假设利润所得额等于利润总额),而

应纳所得税额=利润总额×适用税率

利润总额=销售收入净额×销售利润率

所以,如果估测出了销售利润率,就可推算出利润总额。因为税率是固定的,企业要逃税,只有瞒报利润总额。如果销售收入净额为如实申报,则可估测出真实的销售利润率,从而推算出真实的利润总额,因此,只需估测出企业的真实销售利润率即可(当然,也可考虑估测其他指标)。

根据以上所述及对该地区商业企业的纳税资料分析,选择销售利润率为被解释变量 y(第 i 个企业的销售利润率记为 y_i),选择其他 k 个有关指标 $X_1\cdots,X_k$ 作为解释变量,记为向量 $X=(X_1,X_2,\cdots,X_k)^T$(第 i 个企业的解释变量则记为 $X_i=(X_{i1},X_{i2},\cdots,X_{ik})^T$)。下面的讨论围绕如何估测企业的真实销售利润率进行。

假设已将纳税人按某种方法分成了"诚实申报" G_1 和"不诚实申报" G_2 两类。将"诚实申报"类 G_1 的被解释变量 y_i 看成是"无限制"观测值 $y_i=y_i^*>0$,即取实际申报值 $y_i=y_i^*$;将"不诚实申报"类 G_2 的被解释变量 y_i 看成是"受限"观测值,即将实际申报值截取为 $y_i=0$。

对于"诚实申报" G_1 的销售利润率 y 和解释变量 $X_1,\cdots,X_k$,取实际(申报)观测值:

y_i $X_{i1},\cdots,X_{ik}$

不受限观测值; 不受限观测值; ⟶

y_i $X_{i1},\cdots,X_{ik}$

取实际申报值; 取申报值;

对于"不诚实申报" G_2,其销售利润率 y 截取为 0,而解释变量 $X_1,X_2,\cdots,X_k$ 则均取实际申报值:

y_j $X_{j1},\cdots,X_{jk}$

受限观测值; 不受限观测值; ⟶

0 $X_{j1}, \cdots, X_{jk}$

截取为 0；取申报值

根据“诚实申报”G_1 的申报值、“不诚实申报”G_2 的申报值和截取后的值：

y_i $X_{i1}, \cdots, X_{ik}$

取实际观测值；取申报值 和

0 $X_{j1}, \cdots, X_{jk}$

截取为 0；取申报值

可求得

$$y^{*}=\beta^{T}X+e$$

的估计模型

$$\hat{y}^{*}=\beta^{T}X \quad (7)$$

此模型是“诚实申报”G_1 的估计模型。再将“不诚实申报”G_2 的解释变量 X_i 代入式(7)，求出“不诚实申报”G_2 的销售利润率 y 的估测值 $\hat{y}$ 以及 y 的置信区间，从而推算出逃税额。估测值 $\hat{y}$ 以及置信区间的含义是：对于“不诚实申报”企业的销售利润率 y，如果诚实申报的话，其销售利润率 y 的真实值应该是多少及其可能范围。这就是用 Tobit 模型估算逃税额的基本思想和核心。

根据某地区“诚实申报”商业企业 G_1 和部分“不诚实申报”商业企业 G_2 的纳税资料(选用 5 个指标作为解释变量)，应用 SAS 软件[5,6]建立了 Tobit 模型，模型参数的最大似然估计值见表 1。

表 1 **Tobit** 模型的参数估计值

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCPT	1	-3.2860302	0.152850	462.1811	0.0001
X1	1	-0.6015830	0.121609	24.47127	0.0001
X2	1	-0.1524340	0.029513	26.67695	0.0001
X3	1	-0.1174739	0.020366	33.27021	0.0001
X4	1	-0.5656758	0.231786	5.956078	0.0147
X5	1	23.3031818	1.158436	404.6563	0.0001
SCALE	1	0.03253027	0.00766		

即得

$$y^{*}=\beta^{T}X+e \quad (8)$$

的估计模型

$$\hat{y}^{*}=-3.286-0.602X_1-0.152X_2-0.117X_3-0.566X_4+23.303X_5 \quad (9)$$

为了检验模型的估测效果，应用模型(9)对“诚实申报”类 G_1 中的一家企业 $g_{1,0}$ 和“不诚实申报”类 G_2 中的一家企业 $g_{2,0}$ 进行估测。虽然 $g_{1,0}$ 确实属于“诚实申报”类，但仍然将 $g_{1,0}$ 和 $g_{2,0}$ 的销售利润率 y 均截取为 0，再估测两企业的真实销售利润率。表 2 第 3 列中括号内的数据为申报值，将其截取为 0，第 2 列为估测值。

表 2 **Tobit** 模型的估测值与申报值比较

样品企业	估测值 $\hat{y}$	申报值 y	90%置信区间	
			下限	上限
(1)	(2)	(3)	(4)	(5)
$g_{1,0}$	0.077201	0(0.074)	0.073337	0.081065
$g_{2,0}$	0.011648	0(0.002)	0.010564	0.012731

作为验证模型效果的“诚实申报”企业 $g_{1,0}$，其销售利润率的申报值 y=0.074 落在 90%置信区间(0.073337，0.081065)内，与点估计(测)值相当接近，与点估计值的相对误差为 4.32%。由此可见，模型的估测效果相当好。

“不诚实申报”企业 $g_{2,0}$ 申报值的销售利润率 y=0.002，通过 Tobit 模型得到的点估计值 $\hat{y}$=0.011648，90%置信区间为(0.010564，0.012731)，因此有 90%的把握断定，其真实的销售利润率至少应为 0.010564(取置信区间的下限)，销售利润率的估测值(仅取下限)是申报值的 5 倍多(0.010564÷0.002=5.282)。根据此点估计和区间估计，不难得到真实利润总额和真实应纳所得税额的估测值。

由此可见，Tobit 模型是解决逃税额估测问题的有效方法。

显然，Tobit 模型也有识别诚实申报与否的功能，如果申报值与点估计和区间估计相差较大，则有理由判定其属于“不诚实申报”。

Tobit 模型也可用于解决类似的需要估测和识别“真”与“伪”的问题。

非线性 GARCH 模型在中国股市波动预测中的应用研究

刘国旗

ABSTRACT

This paper studies the performance of the GARCH model and two of its non-linear modifications to forecast China's weekly stock market volatility. The models are the Quadratic GARCH and the Glosten, Jagannathan and Runkle models which have proposed to describe the often observed negative skewness in stock market indices. We find that the QGARCH model is best when the estimation sample does not contain extreme observations such as the stock market crash and that the GJR model cannot be recommended for forecasting.

关键词：中国股票市场；波动预测；非线性 GARCH 模型

股票价格频繁的波动是股票市场最明显的特征之一。股票价格的时间序列经常表现出一个时期的波动明显地大于另一时期的特征。尽管有大量证据表明，短期的金融资产价格及收益率是不可预测的[1]；但目前人们普遍认为，使用特定的时间序列技术可成功地预测金融资产收益率的方差。国外学者的研究结果表明，Bollerslev 提出的广义自回归条件异方差（GARCH）模型[2] 和 Engle 的自回归条件异方差（ARCH）模型[3]，在预测金融资产收益率方差方面是最为成功的。文献[4] 较全面地综述了 GARCH 模型的应用。简单地讲，GARCH 模型的建模过程是使用 ARMA 类模型来描述误差的方差。GARCH 模型的优势在于它可有效地排除资产收益率中的过度峰值（excess kurtosis）。

金融时间序列的另一显著特点是，金融资产收益率的分布可能是有偏的。例如，在概率分布图上，某些股票市场指数的收益率偏向左边，即负收益大于正收益；而另一些股票市场指数的收益率可能偏向右边，即正收益大于负收益。这样，使用对称的 GARCH

本文思路和方法得到张尧庭教授、李茂年教授的启发和悉心指导，在此表示衷心的感谢。

参考文献

[1] George G. Judge（1988）：Introduction to the Theory and Practice of Econometrics, John Wiley & Sons.

[2] Amemiya, T.（1973）“Regression Analysis when the Dependent Variable is Truncated Normal.” Econometrica, 42, 999—1012.

[3] Fair, R.（1977）“A Note on the Computation of the Tobit Estimator.” Econometrica, 45, 1723—1727.

[4] 贺铿等，《经济计量学原理与应用》，辽宁大学出版社，1987 年。

[5] 高惠璇等，《SAS 系统—Base SAS 软件使用手册》，中国统计出版社，1997 年。

模型就难以处理这类问题。为了解决这类问题，最近，一些学者提出了修正的 GARCH 模型。修正的 GARCH 模型的显著优点在于：它们不仅能描述资产收益率序列的有偏分布，而且保留了 GARCH 模型描述过度峰度的优势。本文使用两个 GARCH 修正模型：一个是二次 GARCH 模型（即 QGARCH 模型[5]）；另一个是 Glosten、Jagannathan 和 Runkle 于 1992 年提出的模型，即 GJR 模型[6]。还有一类可描述有偏时间序列的模型是，Nelson 于 1990 年提出的指数 GARCH 模型（即 EGARCH 模型）[7]。本文也将 E-GARCH 模型当作可供选择的模型之一。我们使用的模型选择标准是估计方法简单、参数收敛速度快。但计算结果却发现 E-GARCH 模型效果欠佳。因此，本文重点研究前两种非线性 GARCH 模型和标准 GARCH 模型对中国股市波动的预测能力，以及它们是否能优于随机游动模型。

全文组织如下：第一部分给出本文所使用的模型，第二部分讨论中国股市数据的统计特征，第三部分给出模型的估计结果，第四部分评价 GARCH、QGARCH、GJR 模型以及随机游动模型的预测效果。最后给出本文结论。

一、GARCH 模型

考虑一个综合指数为 I_t、收益率为 r_t 的股票市场。其中 $r_t = \ln(I_t) - \ln(I_{t-1})$，下标 t 记为观测值的周数。r_t 是具有 GARCH(1, 1)扰动的 P 阶自回归模型可表示为：

$$\phi_P(B) r_t = \mu + \varepsilon_t, \text{ with } \phi_P(B) = 1 - \phi_1 B - \cdots - \phi_P B^P \quad (1)$$

$$\varepsilon_t \sim N(0, h_t)$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1}$$

其中 B 是反向算子：$B^k x_t = x_{t-k}$，μ 是常数项。在实证分析中，μ 通常等于或接近于零；P 通常为零或很小的正整数，这说明从 r_t 自身的滞后来预测 r_t 通常是不可能的。假设特征方程 $\phi(z)=0$ 的根位于单位园之外，并且 ω，α，$\beta > 0$，$\alpha + \beta < 1$（模型的稳定性条件参考文献[2]）。

QGARCH 模型和 GJR 模型分别采用不同的形式对方程(1)中的 h_t 的表达式进行修正。QGARCH 模型中的 h_t 的表达式为

$$h_t = \omega + \alpha(\varepsilon_{t-1} - \gamma)^2 + \beta h_{t-1} \quad (2)$$

由模型(2)可见，如果 ε_{t-1} 的值为负，γ 取正值比取负值对 h_t 的影响大。GJR 模型也类似于模型(1)，但 h_t 过程由下式给定：

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \delta D_{t-1} \varepsilon_{t-1}^2 + \beta h_{t-1} \quad (3)$$

其中 D_{t-1} 是当 $\varepsilon_{t-1} < 0$ 时值为 1、当 $\varepsilon_{t-1} \geq 0$ 时值为 0 虚拟变量。类似于 QGARCH 模型，当 $\delta > 0$ 时，负冲击比正冲击对 h_t 影响大。本文给出的有关参考文献讨论了这些模型的平稳性和稳定性。QGARCH 模型和 GJR 模型能改进标准的 GARCH 模型，这是因为它们能处理有偏度（正或负）的分布问题，其改进效果取决于所附加的参数符号。

二、数据与研究方法

本文使用的数据是上证综合指数（HSEC）和深证成份指数（ZSEC）。数据时间跨度为 7

[6] 高惠璇等，《SAS 系统—SAS/Stat 软件使用手册》，中国统计出版社，1997 年。

[7] 李选举，《税收稽查选案统计研究（修改稿）》，博士论文，中南财经大学，1999 年。

[8] 杨英玲主编，《企业财务分析与评价》，中国金融出版社，1995 年。

[9] 童恒庆，《经济回归模型及计算》，湖北科学技术出版社，1997 年。

作者简介：李选举，男，45 岁，中南财经大学统计系教授，经济学博士。曾在《统计研究》、《财政研究》等刊物发表学术论文 20 余篇。主要研究方向：数理统计在经济统计中的应用；抽样调查。

（责任编辑：石庆焱）