泊松伪极大似然估计

嘉树

2026 年 1 月

说明：本文主要是对 Santos Silva and Tenreyro (2006) 的介绍，讲述了为什么在估计一个非线性模型时，「对数线性化」+「OLS」是不可取的，以及泊松伪极大似然估计的原理。此文并非百分百还原原文，而是根据自己的理解进行了一些改动。

1 对数线性化的问题：以引力模型为例

Santos Silva and Tenreyro (2006) 注意到一个重要问题，传统的引力模型通过估计一个对数线性化的方程得到弹性，然而这样做很有可能会引来偏误，特别是在有异方差的情况下。

在贸易理论中，引力模型是一个非常经典的模型，它被用来研究双边国际贸易流量的决定。具体而言，从国家 i 到国家 j 的贸易流量 T_{ij} 被建模为：

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \eta_{ij} \tag{1}$$

其中 Y_i 和 Y_j 分别表示国家 i 和 j 的 GDP，D_{ij} 表示国家 i 和 j 之间的距离，η_{ij} 是一个非负的随机扰动项，$\alpha_0, \alpha_1, \alpha_2, \alpha_3$ 是待估参数。通常，我们会假设 $\mathbb{E}[\eta_{ij}|Y_i, Y_j, D_{ij}] = 1$，因此有

$$\mathbb{E}[T_{ij}|Y_i, Y_j, D_{ij}] = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3}$$

在实际估计中，为了得到一个线性形式，人们将 (1) 两边取对数[1]，得到：

$$\log(T_{ij}) = \log(\alpha_0) + \alpha_1 \log(Y_i) + \alpha_2 \log(Y_j) + \alpha_3 \log(D_{ij}) + \log(\eta_{ij})$$

然后用 OLS 估计这个对数线性化的方程。可是——我们不禁要问——这样得到的估计量是一致的吗?

我们知道，OLS 估计量的一致性依赖于外生性条件，即误差项和解释变量不相关。在上述对数线性化的方程中，是否有

$$\mathbb{E}[\log(\eta_{ij})|Y_i, Y_j, D_{ij}] = 0$$

成立? 或者至少是一个常数（从而可以被截距项吸收)? 很遗憾，在现实的很多应用中，这个条件通常不成立。我们可以考虑一种简单情况，设 $\eta_{i,j}$ 条件于 Y_i, Y_j, D_{ij}，服从均值为 1，方差为 $\sigma_{i,j}^2 = f(Y_i, Y_j, D_{ij})$ 的对数正态分布（即取完对数后服从正态分布)，那么根据对数正态分布的性质，有

$$\mathbb{E}[\log(\eta_{i,j})|Y_i, Y_j, D_{ij}] = -\frac{1}{2}\log(1 + \sigma_{i,j}^2)$$

可以看到，对数线性化方程中的误差项的条件期望是一个关于 Y_i, Y_j, D_{ij} 的函数，这几乎意味着误差项和解释变量相关，因此 OLS 估计量是不一致的。

1. 为什么人们这么喜欢取对数? 因为对数变换可以将乘法模型转化为加法模型，从而变成线性的，而待估参数直接就是具有经济含义的弹性。简单的、易于解释的形式就是容易被研究者们青睐。

这个例子表明，如果原模型中的误差项有异方差，那么对数线性化方程的外生性条件不成立，会给 OLS 带来偏误。在现实的很多例子中，$\eta_{i,j}$ 的异方差问题确实普遍存在！所以「对数线性化」+「OLS」这种传统上被普遍采用的范式是不可取的。

取对数还会带来一个显而易见的问题，那就是如何处理零值。在实际数据特别是计数数据中，零值是一个非常普遍的现象。在某段时期，两个国家可能根本就没有贸易；或者，在数据统计过程中，如果两个国家的贸易额太小，也可能会被统计机构忽略而取零；又或者，当贸易数据缺失时，也可能会被错记为零。早期的实证研究者们大多采取的解决办法是，完全扔掉这些零值，或者以 $\log(T_{ij}+1)$ 代替 $\log(T_{ij})$ 作为被解释变量。前者是一个较为粗暴的做法，且不论这造成的信息损失，更严重的后果是，这还可能引入偏误，因为正如前面所说，被扔掉的这些零值可能正是那些经济规模相对较小的国家之间的贸易流量，即样本的选择本身和解释变量相关，带来了 *样本选择偏误*。而取 $\log(T_{ij}+1)$ 的做法看似规避了零值带来的数学上的麻烦而又保留了原始数据的信息，但这样做同样存在问题，最近一篇发表在 QJE 的文章 Chen and Roth (2024) 对这一点进行了系统阐述。

2 非线性最小二乘法

既然对数化会带来这样或那样的问题，为什么我们不能直接估计原始的乘法形式的方程呢？其实这也是可以的，就是通过「非线性最小二乘法」(NLS) 估计。

假设真实模型为[2]

$$y_i = \exp(x_i\beta) + \varepsilon_i$$

其中 $y_i \geq 0$，且 $\mathbb{E}[\varepsilon_i|x_i] = 0$。那么，$\beta$ 的 NLS 估计量就是

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{n} [y_i - \exp(x_i b)]^2$$

其一阶条件是

$$\sum_{i=1}^{n} [y_i - \exp(x_i\hat{\beta})] \exp(x_i\hat{\beta}) x_i = 0 \tag{2}$$

从矩估计的角度来看，β 的 NLS 估计量对应如下总体矩条件[3]：

$$\mathbb{E}[\varepsilon_i \exp(x_i\beta) x_i] = 0$$

直观来看，如果我们把 $\exp(x_i\beta)$ 视作赋予个体 i 的权重，那么显然 $x_i\beta$ 越大的个体的权重就越大。但这些个体往往也有着更多的噪声，于是 NLS 估计量的效率通常不那么高。当然，提升效率是可能的。譬如，如果我们知道了条件方差 $\text{var}(y_i|x_i)$ 的具体函数形式，那么可以估计它，然后再用加权最小二乘法来改进效率。这种做法由于多数情况下条件方差的具体函数形式未知（从而可能需要使用非参数的方法），且操作上更为复杂，因而未被广泛采用。实证研究人员想要的是一种既简单易用，又对一般异方差稳健的有效率的估计方法，把它作为日常使用的标准工具。「泊松伪极大似然估计」(Poisson pseudo maximum likelihood, PPML) 完美契合了这一需求，它并不要求真实数据必须服从泊松分布，甚至不要求是离散数据，而基本上只要求条件均值模型是正确的，因此零值问题甚至不再是一个问题。在这个例子中，它等价于添加了异方差设定的 NLS 估计量。

2. 设定乘法形式的误差项是等价的。如 $y_i = \exp(x_i\beta)\eta_i$，其中 $\mathbb{E}[\eta_i|x_i] = 1$。令 $\varepsilon_i = \exp(x_i\beta)[\eta_i - 1]$，则 $\mathbb{E}[\varepsilon_i|x_i] = 0$。

3. 我们知道由零条件均值 $\mathbb{E}[\varepsilon_i|x_i] = 0$ 可以推出 $\mathbb{E}[\varepsilon_i f(x_i)] = 0$，这里 $f(x_i)$ 可以是 x_i 的任意（可测）函数。NLS 相当于选定了特定的函数 $f(x_i) = \exp(x_i\beta)x_i$。

3 提升效率：异方差设定

为了提升效率，我们必须知道关于异方差形式的一些信息，这一般通过假设来实现。我们可以不必假设具体的函数形式，而只需假设

$$\mathbb{E}[y_i|x_i] = \exp(x_i\beta) \propto \text{var}(y_i|x_i)$$

即异方差和条件均值成正比（这个假设使我们联想到泊松分布的一个重要性质，即均值和方差是相等的）。这个假设的直觉是，对那些条件均值较大的个体，其观测值噪声更大，从而方差更大。这符合实际贸易数据的表现，对于那些贸易量很小的国家，贸易量的波动一般也是较小的。当把这个假设加入到 NLS 的估计框架中（通过加权最小二乘法），我们会得到如下一阶条件：

$$\sum_{i=1}^{n}[y_i - \exp(x_i\hat{\beta})]x_i = 0 \tag{3}$$

相比于 (2)，(3) 对每个观测是等权重的，这使得估计量更有效率。即便「异方差正比于条件均值」这一假设并不严格成立，只要异方差随着条件均值的增大而增大，那么 (3) 也比普通 NLS 估计量更有效率。

事实上，(3) 得到的估计量正好等价于 PPML 估计量，见下一节。

4 伪极大似然估计

所谓「伪极大似然」(pseudo maximum likelihood, PML)[4]，（在计量里）指的是所采用的似然函数并不一定来自于真实的密度函数（真实的密度函数是未知的），而是属于「线性指数分布族」(linear exponential family)。这一方法早在上世纪就由 Gourieroux et al. (1984) 提出，他们证明了在一些正则条件下，PML 估计量是一致的、渐进正态的。

这里我们特别讨论泊松伪极大似然估计。如果我们有条件均值模型

$$\lambda_i = \mathbb{E}[y_i|x_i] = \exp(x_i\beta)$$

那么以此为均值的泊松分布的密度函数为

$$f(y_i|x_i) = \frac{\lambda_i^{y_i}}{y_i!}\exp(-\lambda_i)$$

于是对数似然函数为

$$\begin{aligned} L(\beta) &= -\sum_{i=1}^{n}\lambda_i + \sum_{i=1}^{n} y_i \log(\lambda_i) - \sum_{i=1}^{n}\log(y_i!) \\ &= -\sum_{i=1}^{n}\exp(x_i\beta) + \sum_{i=1}^{n} y_i x_i\beta - \sum_{i=1}^{n}\log(y_i!) \end{aligned}$$

一阶条件为

$$\frac{dL}{d\beta} = -\sum_{i=1}^{n} x_i[\exp(x_i\beta) - y_i] = 0$$

这正是 (3) 所示的一阶条件。

4. 尽管在有些课本中（例如，Hayashi (2000) 的 *Econometrics*）拟极大似然和「准极大似然」(quasi maximum likelihood, QML) 指的是同一个概念，但这里我还是想做一下区分。QML 的似然函数也不来自于真实的密度函数，但它强调的是分布误设带来的后果，这个思想是由统计学家率先发展起来的，从这个角度讲，PML 属于 QML 的一个子类，即在特定的模型误设下极大似然估计量仍然具有良好的性质。在统计学中，其实也有 PML 的概念 Gong and Samaniego (1981)，但和计量里面 Gourieroux et al. (1984) 所发展的方法并不是一回事。

5 小结

我们探讨了异方差和零值问题给对数线性化方法带来的麻烦，并介绍了 PPML 是如何对它们保持稳健的。现在它成了国际贸易研究中一个标准的工具。原文还探讨了对异方差假设的检验方法，以及用模拟说明了 PPML 的优越性，并给出了一个实证例子，感兴趣的读者可以阅读原文。

参考文献

Chen, J. and J. Roth (2024): "Logs with Zeros? Some Problems and Solutions," *Quarterly Journal of Economics*, 139(2), 891–936.

Gong, G. and F. J. Samaniego (1981): "Pseudo Maximum Likelihood Estimation: Theory and Applications," *Annals of Statistics*, 9(4), 861–869.

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52(3), 681–700.

Hayashi, F. (2000): *Econometrics*. Princeton: Princeton University Press.

Santos Silva, J. M. C. and S. Tenreyro (2006): "The Log of Gravity," *Review of Economics and Statistics*, 88(4), 641–658.