

第 25 章 非线性回归与门限回归

25.1 非线性最小二乘法

对于非线性回归模型, 除了 MLE, 还可使用“非线性最小二乘法”(Nonlinear Least Square, 简记 NLS)。

考虑以下非线性回归模型:

$$y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i \quad (i = 1, \dots, n)$$

$\boldsymbol{\beta}$ 为 K 维参数向量, $g(\cdot)$ 是 $\boldsymbol{\beta}$ 的非线性函数, 且无法通过变量转换变为 $\boldsymbol{\beta}$ 的线性函数。

如果 $g(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i' \boldsymbol{\beta}$, 则回到古典线性回归模型。

记 $\tilde{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的一个假想值, 对应的残差为 $e_i = y_i - g(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})$ 。

非线性最小二乘法通过选择 $\tilde{\boldsymbol{\beta}}$, 使得残差平方和最小:

$$\min_{\tilde{\boldsymbol{\beta}}} \text{SSR}(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left[y_i - g(\mathbf{x}_i, \tilde{\boldsymbol{\beta}}) \right]^2$$

最小化的一阶条件为

$$\frac{\partial \text{SSR}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = -2 \sum_{i=1}^n [y_i - g(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})] \frac{\partial g(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{0}$$

可简化为

$$\sum_{i=1}^n [y_i - g(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})] \frac{\partial g(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{0}$$

$$\sum_{i=1}^n e_i \frac{\partial g(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{0}$$

这是一个 K 个方程、 K 个未知数的非线性方程组。

满足这个非线性方程组的估计量被称为“非线性最小二乘估计量”，记为 $\hat{\boldsymbol{\beta}}_{\text{NLS}}$ 。

残差向量 \mathbf{e} 与 $\frac{\partial g(\mathbf{x}, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$ 正交，而不是与 \mathbf{x} 正交(线性回归的情形)。

通常没有解析解，要用数值迭代方法求解，比如牛顿-拉夫森法。

例 考虑如下非线性回归模型：

$$y_i = \beta_1 + \beta_2 \exp(\beta_3 x_i) + \varepsilon_i$$

这个模型含有三个未知参数 $(\beta_1, \beta_2, \beta_3)$ ，即 $K = 3$ 。

使用 NLS 进行估计，残差平方和为

$$\min_{\tilde{\beta}} \text{SSR}(\tilde{\beta}) = \sum_{i=1}^n \left[y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i) \right]^2$$

NLS 估计量的一阶条件为

$$\frac{\partial \text{SSR}(\tilde{\beta})}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^n \left[y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i) \right] = 0$$

$$\frac{\partial \text{SSR}(\tilde{\beta})}{\partial \tilde{\beta}_2} = -2 \sum_{i=1}^n \left[y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i) \right] \exp(\tilde{\beta}_3 x_i) = 0$$

$$\frac{\partial \text{SSR}(\tilde{\beta})}{\partial \tilde{\beta}_3} = -2 \sum_{i=1}^n \left[y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i) \right] \tilde{\beta}_2 x_i \exp(\tilde{\beta}_3 x_i) = 0$$

NLS 的大样本性质

如果 $E(\varepsilon_i | \mathbf{x}_i) = 0$ ，再加上一些技术性条件，则 $\hat{\boldsymbol{\beta}}_{\text{NLS}}$ 为 $\boldsymbol{\beta}$ 的一致估计量，且 $\hat{\boldsymbol{\beta}}_{\text{NLS}}$ 服从渐近正态。

如果扰动项为球型扰动项，则 $\hat{\boldsymbol{\beta}}_{\text{NLS}}$ 是渐近有效的(asymptotically efficient)。

25.2 非线性回归的 **Stata** 命令及实例

25.3 门限回归

考察回归系数是否稳定：将样本分成若干子样本分别回归，能否得到相近的估计系数？

对于时间序列，经济结构是否随着时间推移而改变(Chow test)？

对于横截面数据，比如，样本中有男性与女性，可根据性别将样本一分为二，分别估计男性样本与女性样本。

如果用来划分样本的变量是连续型变量，比如，企业规模、人均国民收入，则需要给出一个划分的标准，即“门限(门槛)值”(threshold level)。

例 在应用研究中，人们常常怀疑大企业与小企业的投资行为不同，那么如何区分大企业与小企业呢？

例 受到流动性约束(liquidity constraint)的企业与没有流动性约束企业的投资行为也可能不同，如何通过债务股本比(debt to equity ratio)或其他指标来区分这两类企业？

例 发达国家与发展中国家的经济增长规律可能不同，如何通过人均国民收入这一指标来区分一个国家发达与否？

传统的做法由研究者主观确定一个门限值，把样本一分为二，既不对门限值进行参数估计，也不进行统计检验，结果并不可靠。

Hansen(2000)提出“门限(门槛)回归”(threshold regression)，以严格的统计推断方法对门限值进行参数估计与假设检验。

假设样本数据为 $\{y_i, \mathbf{x}_i, q_i\}_{i=1}^n$ 。

q_i 为用来划分样本的“门限变量” (threshold variable), q_i 可以是解释变量 \mathbf{x}_i 的一部分:

$$\begin{cases} y_i = \boldsymbol{\beta}'_1 \mathbf{x}_i + \varepsilon_i, \text{ 若 } q_i \leq \gamma \\ y_i = \boldsymbol{\beta}'_2 \mathbf{x}_i + \varepsilon_i, \text{ 若 } q_i > \gamma \end{cases}$$

其中, γ 为待估门限值, \mathbf{x}_i 为外生解释变量, 与 ε_i 不相关。将此分段函数合并写为

$$y_i = \boldsymbol{\beta}'_1 \underbrace{\mathbf{x}_i \cdot \mathbf{1}(q_i \leq \gamma)}_{=z_{i1}} + \boldsymbol{\beta}'_2 \underbrace{\mathbf{x}_i \cdot \mathbf{1}(q_i > \gamma)}_{=z_{i2}} + \varepsilon_i$$

可用 NLS 来估计。

如果 γ 已知, 可定义 $\mathbf{z}_{i1} \equiv \mathbf{x}_i \cdot \mathbf{1}(q_i \leq \gamma)$ 与 $\mathbf{z}_{i2} \equiv \mathbf{x}_i \cdot \mathbf{1}(q_i > \gamma)$, 将此方程转化为线性回归模型:

$$y_i = \boldsymbol{\beta}'_1 \mathbf{z}_{i1} + \boldsymbol{\beta}'_2 \mathbf{z}_{i2} + \varepsilon_i$$

实践中, 常分两步来最小化残差平方和。

首先, 给定 γ 的取值, 用 OLS 估计 $\hat{\boldsymbol{\beta}}_1(\gamma)$ 与 $\hat{\boldsymbol{\beta}}_2(\gamma)$ ($\hat{\boldsymbol{\beta}}_1$ 与 $\hat{\boldsymbol{\beta}}_2$ 依赖于 γ), 并计算残差平方和 $\text{SSR}(\gamma)$ (称为 Concentrated Sum of Squared Residuals), 也是 γ 的函数。

其次, 选择 γ 使得 $\text{SSR}(\gamma)$ 最小化。

给定 q_i ，由于示性函数 $\mathbf{1}(q_i \leq \gamma)$ 与 $\mathbf{1}(q_i > \gamma)$ 只能取值 0 或 1，故是 γ 的阶梯函数，而“阶梯的升降点”正好是 q_i (只有一级“台阶”)。

故 $\text{SSR}(\gamma)$ 也是 γ 的阶梯函数，而阶梯的升降点恰好在 $\{q_i\}_{i=1}^n$ 不重叠的观测值上，因为如果 γ 取 $\{q_i\}_{i=1}^n$ 以外的其他值，不会对子样本的划分产生影响，故不改变 $\text{SSR}(\gamma)$ 。

最多只需要考虑 γ 取 n 个值即可，即 $\gamma \in \{q_1, q_2, \dots, q_n\}$ 。

这使得 $\text{SSR}(\gamma)$ 的最小化计算得以简化。

记最后的参数估计量为 $(\hat{\beta}_1(\hat{\gamma}), \hat{\beta}_2(\hat{\gamma}), \hat{\gamma})$ 。

在一定的条件下，Hansen(2000)导出了 $\hat{\gamma}$ 的大样本渐近分布，在此基础上构造 $\hat{\gamma}$ 的置信区间，并对 $H_0: \gamma = \gamma_0$ 进行似然比检验。

类似地，可考虑包含“多个门限值”的门限回归。

比如，对于门限变量 q_i ，假设两个门限值为 $\gamma_1 < \gamma_2$ ，则门限回归模型为

$$y_i = \boldsymbol{\beta}'_1 \mathbf{x}_i \cdot \mathbf{1}(q_i \leq \gamma_1) + \boldsymbol{\beta}'_2 \mathbf{x}_i \cdot \mathbf{1}(\gamma_1 < q_i \leq \gamma_2) + \boldsymbol{\beta}'_3 \mathbf{x}_i \cdot \mathbf{1}(q_i > \gamma_2) + \varepsilon_i$$

25.4 面板数据的门限回归

对于面板数据 $\{y_{it}, \mathbf{x}_{it}, q_{it} : 1 \leq i \leq n, 1 \leq t \leq T\}$, Hansen (1999)考虑了如下的固定效应门限回归模型:

$$\begin{cases} y_{it} = \mu_i + \boldsymbol{\beta}'_1 \mathbf{x}_{it} + \varepsilon_{it}, \text{ 若 } q_{it} \leq \gamma \\ y_{it} = \mu_i + \boldsymbol{\beta}'_2 \mathbf{x}_{it} + \varepsilon_{it}, \text{ 若 } q_{it} > \gamma \end{cases}$$

其中, q_{it} 为门限变量(可以是 \mathbf{x}_{it} 的一部分), γ 为门限值, 扰动项 ε_{it} 为 iid。假设 \mathbf{x}_{it} 为外生变量 (不包含 y_{it} 的滞后值), 与 ε_{it} 不相关。

将模型更简洁地表示为

$$y_{it} = \mu_i + \boldsymbol{\beta}'_1 \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} \leq \gamma) + \boldsymbol{\beta}'_2 \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} > \gamma) + \varepsilon_{it}$$

假设 n 较大, T 较小(短面板), 故大样本的渐近理论基于“ $n \rightarrow \infty$ ”。

定义 $\boldsymbol{\beta} \equiv \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$, $\mathbf{x}_{it}(\gamma) \equiv \begin{pmatrix} \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} \leq \gamma) \\ \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} > \gamma) \end{pmatrix}$, 则方程简化为

$$y_{it} = \mu_i + \boldsymbol{\beta}' \mathbf{x}_{it}(\gamma) + \varepsilon_{it}$$

对于个体 i , 将方程两边对时间求平均:

$$\bar{y}_i = \mu_i + \boldsymbol{\beta}' \bar{\mathbf{x}}_i(\gamma) + \bar{\varepsilon}_i$$

将两方程相减，可得离差形式：

$$y_{it} - \bar{y}_i = \boldsymbol{\beta}' [\mathbf{x}_{it}(\gamma) - \bar{\mathbf{x}}_i(\gamma)] + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

记 $y_{it}^* \equiv y_{it} - \bar{y}_i$, $\mathbf{x}_{it}^*(\gamma) \equiv \mathbf{x}_{it}(\gamma) - \bar{\mathbf{x}}_i(\gamma)$, $\varepsilon_{it}^* \equiv \varepsilon_{it} - \bar{\varepsilon}_i$, 则可得

$$y_{it}^* = \boldsymbol{\beta}' \mathbf{x}_{it}^*(\gamma) + \varepsilon_{it}^*$$

仍使用两步法进行估计。首先，给定 γ 的取值，用 OLS 进行一致估计(组内估计量)，得到估计系数 $\hat{\boldsymbol{\beta}}(\gamma)$ 以及残差平方和 $\text{SSR}(\gamma)$ 。

其次，对于 $\gamma \in \{q_{it} : 1 \leq i \leq n, 1 \leq t \leq T\}$ (γ 最多有 nT 个可能取值)，选择 $\hat{\gamma}$ ，使得 $\text{SSR}(\hat{\gamma})$ 最小。最后得到估计系数 $\hat{\boldsymbol{\beta}}(\hat{\gamma})$ 。

如果不希望某个子样本中的观测值过少，可限制 γ 的取值，比如不考虑 $\{q_{it}\}$ 中最大 5% 或最小 5% 的取值。

对于是否存在“门限效应” (threshold effect)，可检验原假设：

$$H_0 : \beta_1 = \beta_2$$

如果此原假设成立，则不存在门限效应，模型简化为

$$y_{it} = \mu_i + \beta_1' x_{it} + \varepsilon_{it}$$

对于这个标准的固定效应面板模型，将其转化为离差形式，然后用 OLS 来估计(组内估计量)。

记在 “ $H_0: \beta_1 = \beta_2$ ” 约束下所得到的残差平方和为 SSR^* ，以区别于无约束的残差平方和 $SSR(\hat{\gamma})$ 。

显然， $SSR^* \geq SSR(\hat{\gamma})$ 。如果 $[SSR^* - SSR(\hat{\gamma})]$ 越大，加上约束条件后使得SSR增大越多，则越应该倾向于拒绝 “ $H_0: \beta_1 = \beta_2$ ”。

Hansen (1999)提出使用以下似然比检验(LR)统计量：

$$LR \equiv [SSR^* - SSR(\hat{\gamma})] / \hat{\sigma}^2$$

其中， $\hat{\sigma}^2 \equiv \frac{SSR(\hat{\gamma})}{n(T-1)}$ 为对扰动项方差的一致估计。

如果“ $H_0: \beta_1 = \beta_2$ ”成立，则不存在门限效应，也就无所谓门限值 γ 等于多少。

在 H_0 成立的情况下，无论 γ 取什么值，对模型都没有影响，故参数 γ 不可识别。

检验统计量LR的渐近分布并非标准的 χ^2 分布，而依赖于样本矩，无法将其临界值列表，但可用自助法得到临界值。

如果拒绝“ $H_0: \beta_1 = \beta_2$ ”，认为存在门限效应，可进一步对门限值进行检验，即检验“ $H_0: \gamma = \gamma_0$ ”。定义似然比检验统计量为

$$\text{LR}(\gamma) \equiv [\text{SSR}(\gamma) - \text{SSR}(\hat{\gamma})] / \hat{\sigma}^2$$

在“ $H_0: \gamma = \gamma_0$ ”成立的情况下， $LR(\gamma)$ 的渐近分布的累积分布函数为 $(1 - e^{-x/2})^2$ ，可直接算出临界值。可利用统计量 $LR(\gamma)$ 来计算 γ 的置信区间。

考虑多门限值的面板回归模型。以两个门限值为例：

$$y_{it} = \mu_i + \beta_1' x_{it} \cdot \mathbf{1}(q_{it} \leq \gamma_1) + \beta_2' x_{it} \cdot \mathbf{1}(\gamma_1 < q_{it} \leq \gamma_2) + \beta_3' x_{it} \cdot \mathbf{1}(q_{it} > \gamma_2) + \varepsilon_{it}$$

其中，门限值 $\gamma_1 < \gamma_2$ 。

将这个模型转换为离差形式，仍用两步法进行估计。首先，给定 (γ_1, γ_2) ，使用 OLS 估计离差模型，得到残差平方和 $SSR(\gamma_1, \gamma_2)$ 。其次，选择 (γ_1, γ_2) 使得 $SSR(\gamma_1, \gamma_2)$ 最小化。