

点估计

2025年11月

概览

- ① 经典矩估计
- ② 矩估计
- ③ 矩估计的大样本性质
- ④ 广义矩估计
- ⑤ 广义矩估计的大样本性质



长的参数估计方法，其思路是使用样
法做进一步推广。

- 接下来我们介绍一些经典的点估计方法。
- 矩估计 (method of moments) 是使用历史最长的参数估计方法, 其思路是使用样本矩代替总体矩对参数进行估计。
- 接下来我们将介绍经典的矩估计方法, 并对此方法做进一步推广。

同分布的样本

x_1, x_2, \dots, x_n

的总体的期望。

- 同分布的样本
- x_1, x_2, \dots, x_n
- 的总体的期望。

同分布的样本

x_1, x_2, \dots, x_n

的总体的期望。

同分布的样本

x_1, x_2, \dots, x_n

的总体的期望。

- 同分布的样本
- x_1, x_2, \dots, x_n
- 的总体的期望。

矩估计的思想

- 我们知道，在一定比较宽松的条件下，根据大数定律有：

$$m_1(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N x_i \xrightarrow{p} \mathbb{E}_{\theta_0}(x_i) = \mu_1(\theta_0)$$

- 如果 $\mu_1(\cdot)$ 是一个连续且可逆的函数，那么真实参数 θ_0 可以写为：

$$\theta_0 = \mu_1^{-1}(\mu_1(\theta_0))$$

- 那么我们可以使用样本矩 $m_1(\mathbf{x})$ 代替上式中的总体矩 $\mu_1(\theta_0)$ ，由于 $m_1(\mathbf{x}) \xrightarrow{p} \mu_1(\theta_0)$ ，而 $\mu_1^{-1}(\cdot)$ 为连续函数，从而估计量：

$$\hat{\theta} \triangleq \mu_1^{-1}(m_1(\mathbf{x})) \xrightarrow{p} \mu_1^{-1}(\mu_1(\theta_0)) = \theta_0$$

从而 $\hat{\theta}$ 是 θ_0 的一致估计。

确定的概率函数，因而可以计算在 θ 情
足够大的情况下应该是充分接近的
差距最小，从而得到对真值 θ_0 的估计。

- 形象的理解是，给定任何一个 θ ，总体 P_θ 是一个确定的概率函数，因而可以计算在 θ 情况下的样本矩 $\mathbb{E}_\theta(x_i)$ 。
- 理论上，样本矩 $m_1(\mathbf{x})$ 和总体矩 $\mu_1(\theta)$ 在样本量足够大的情况下应该是充分接近的
- 那么我们可以找到一个 $\hat{\theta}$ 使得 $\mu_1(\hat{\theta})$ 与 $m_1(\mathbf{x})$ 的差距最小，从而得到对真值 θ_0 的估计。
- 以上就是矩估计的思想。

矩估计

泊松分布的矩估计

- 如果样本 $x_i \sim \mathcal{P}(\lambda_0)$ *i.i.d.*, 我们知道样本矩 $m_1(x) = \bar{x}$, 比如, 如果我们的样本观测值为 $\mathbf{x} = (3, 5, 7, 2, 3)$, 那么样本矩为 $m_1(\mathbf{x}) = \bar{x} = \frac{3+5+7+2+3}{5} = 4$
- 假如任意给定一个 λ , 比如令 $\lambda = 2$, 总体的期望为 $\mathbb{E}_\lambda(x_i) = \lambda = 2 \neq 4$, 因而如果认为 $\lambda_0 = 2$, 那么总体 $P(2)$ 所产生的总体矩与样本矩仍然有差异。
- 只有当 $\lambda = 4$ 时, 总体矩 $\mathbb{E}_\lambda(x_i) = 4 = m_1(x)$, 总体矩与我们观察到的样本矩相等, 因而我们可以推断 $\hat{\lambda} = 4$ 。

矩估计

泊松分布的矩估计

一般的，对于泊松分布总体，我们可以直接令总体矩等于样本矩得到估计，即： $\hat{\lambda} = m_1(x) = \bar{x}$ 下面我们分别讨论该估计量的无偏性和一致性。

- 首先, 对于无偏性, 由于:

$$\mathbb{E}(\hat{\lambda}) = \mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{N} \sum_{i=1}^N x_i\right) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}(x_i) = \lambda_0$$

因而 $\hat{\lambda}$ 是 λ_0 的无偏估计。

- 而对于一致性, 根据大数定理: $\hat{\lambda} = \bar{x} \xrightarrow{P} E(x_i) = \lambda_0$ 因而 $\hat{\lambda}$ 是 λ_0 的一致估计。

矩估计

泊松分布的矩估计

- 当然，一致性还可以通过分析 $\hat{\lambda}$ 的偏差与方差来证明。
 - 根据以上讨论，该估计量的偏差为 $\text{Bias}(\hat{\lambda}) = \mathbb{E}(\hat{\lambda}) - \lambda_0 = 0$ ，而其方差为：

$$\mathbb{V}(\hat{\lambda}) = \mathbb{V}\left(\frac{1}{N} \sum_{i=1}^N x_i\right) = \frac{1}{N^2} \sum_{i=1}^N \mathbb{V}(x_i) = \frac{\lambda_0}{N}$$

从而

$$\mathbb{E}(\hat{\lambda} - \lambda_0)^2 = \mathbb{V}(\hat{\lambda}) + [\text{Bias}(\hat{\lambda})]^2 \rightarrow 0$$

从而 $\hat{\lambda} \xrightarrow{L^2} \lambda_0$ ，从而 $\hat{\lambda} \xrightarrow{P} \lambda_0$ 。

- 进一步，根据中心极限定理，有：

$$\sqrt{N}(\hat{\lambda} - \lambda_0) = \sqrt{N}(\bar{x} - \lambda_0) \overset{a}{\sim} \mathcal{N}(0, \lambda_0)$$

从而得到了渐近分布。

矩估计

对数正态分布的矩估计

- 如果样本 $x_i \sim \mathcal{LN}(\mu_0, 2)$ i.i.d, 即总体为对数正态分布, 且一个参数 $\sigma^2 = 2$ 已知。类似的, 样本矩 $m_1(\mathbf{x}) = \bar{x}$, 而总体矩 $\mathbb{E}_\mu(x_i) = e^{\mu+1}$ 。
- 根据矩估计的思想, 令总体矩等于样本矩, 即:

$$e^{\hat{\mu}+1} = m_1(\mathbf{x}) = \bar{x}$$

可以得到 μ_0 的矩估计值: $\hat{\mu} = \ln \bar{x} - 1$

矩估计

对数正态分布的矩估计

- 现在讨论该估计量的无偏性和一致性。首先根据Jensen不等式：

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(\ln \bar{x}) - 1 \leq \ln(\mathbb{E}\bar{x}) - 1 = \ln e^{\mu_0+1} - 1 = \mu_0$$

因而 $\hat{\mu}$ 并不是 μ_0 的无偏估计。

- 而根据大数定律, $\bar{x} \xrightarrow{P} \mathbb{E}x_i = e^{\mu_0+1}$, 由于 \ln 为连续函数, 从而:

$$\hat{\mu} = \ln \bar{x} - 1 \xrightarrow{P} \ln e^{\mu_0+1} - 1 = \mu_0$$

因而 $\hat{\mu}$ 是 μ_0 的一致估计。

矩估计

对数正态分布的矩估计

- 此外, 我们还可以使用delta方法计算 $\hat{\mu}$ 的极限分布。
- 根据中心极限定理:

$$\sqrt{N}(\bar{x} - e^{\mu_0+1}) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}(x_i))$$

其中 $\mathbb{V}(x_i) = e^{2(\mu_0+2)} - e^{2\mu_0+2}$ 。

- 因而, 对估计量在 e^{μ_0+1} 处进行泰勒展开:

$$\begin{aligned}\sqrt{N}(\hat{\mu} - \mu_0) &= \sqrt{N}(\ln \bar{x} - 1 - \ln e^{\mu_0+1} + 1) \\&= \sqrt{N}(\ln \bar{x} - \ln e^{\mu_0+1}) \\&= \sqrt{N}\left(\ln e^{\mu_0+1} + \frac{1}{e^{\mu_0+1}}(\bar{x} - e^{\mu_0+1}) + O\left((\bar{x} - e^{\mu_0+1})^2\right) - \ln e^{\mu_0+1}\right) \\&= \sqrt{N}\frac{1}{e^{\mu_0+1}}(\bar{x} - e^{\mu_0+1}) + o_p(1) \xrightarrow{D} \frac{1}{e^{\mu_0+1}}\sqrt{N}(\bar{x} - e^{\mu_0+1}) \\&\stackrel{a}{\sim} \mathcal{N}\left(0, e^{-2(\mu_0+1)}\mathbb{V}(x_i)\right)\end{aligned}$$

得到了未知参数的点估计，并使用
和delta方法计算了矩估计量的大样
分布中：
 $(0, \lambda_0)$
估计。比如在上面的例子中，我们知
= 95%

- 得到了未知参数的点估计，并使用
和delta方法计算了矩估计量的大样
分布中：
 $(0, \lambda_0)$
估计。比如在上面的例子中，我们知
= 95%

得到了未知参数的点估计，并使用
和delta方法计算了矩估计量的大样
分布中：
 $(0, \lambda_0)$
估计。比如在上面的例子中，我们知
= 95%

得到了未知参数的点估计，并使用
和delta方法计算了矩估计量的大样
分布中：
 $(0, \lambda_0)$
估计。比如在上面的例子中，我们知
= 95%

得到了未知参数的点估计，并使用
和delta方法计算了矩估计量的大样
分布中：
 $(0, \lambda_0)$
估计。比如在上面的例子中，我们知
= 95%

矩估计与区间估计

- 因而对以上不等式稍加改变，即得到：

$$P\left(\hat{\lambda} - 1.96 \times \sqrt{\frac{\lambda_0}{N}} \leq \lambda_0 \leq \hat{\lambda} + 1.96 \times \sqrt{\frac{\lambda_0}{N}}\right) = 95\%$$

因而 λ_0 的95%的置信区间应该为 $\left(\hat{\lambda} - 1.96 \times \sqrt{\frac{\lambda_0}{N}}, \hat{\lambda} + 1.96 \times \sqrt{\frac{\lambda_0}{N}}\right)$ 。

- 当然，上式中由于 λ_0 未知，不过由于 $\hat{\lambda} \xrightarrow{P} \lambda_0$ ，因而我们可以使用 $\hat{\lambda}$ 代替 λ_0 ，最终得到 λ_0 的区间估计：

$$\left(\hat{\lambda} - 1.96 \times \text{s.e.}(\hat{\lambda}), \hat{\lambda} + 1.96 \times \text{s.e.}(\hat{\lambda})\right)$$

其中 $\text{s.e.}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{N}}$ 。

可以联立前 K 个样本矩和总体矩对 θ 进

$$\begin{aligned} \mu_1(\theta) &= \mathbb{E}_{\theta}(x_i^1) \\ \mu_2(\theta) &= \mathbb{E}_{\theta}(x_i^2) \\ \vdots \\ \mu_K(\theta) &= \mathbb{E}_{\theta}(x_i^K) \end{aligned}$$

可以联立前 K 个样本矩和总体矩对 θ 进

$$\begin{aligned} \mu_1(\theta) &= \mathbb{E}_{\theta}(x_i^1) \\ \mu_2(\theta) &= \mathbb{E}_{\theta}(x_i^2) \\ \vdots \\ \mu_K(\theta) &= \mathbb{E}_{\theta}(x_i^K) \end{aligned}$$

可以联立前 K 个样本矩和总体矩对 θ 进

$$\begin{aligned} \mu_1(\theta) &= \mathbb{E}_{\theta}(x_i^1) \\ \mu_2(\theta) &= \mathbb{E}_{\theta}(x_i^2) \\ \vdots \\ \mu_K(\theta) &= \mathbb{E}_{\theta}(x_i^K) \end{aligned}$$

多个参数的矩估计

一般而言，如果有 K 个参数， $\theta = [\theta_1, \dots, \theta_K]'$ ，那么我们使用前 K 个矩，解方程：

$$\begin{cases} m_1(\mathbf{x}) = \mu_1(\hat{\theta}) \\ m_2(\mathbf{x}) = \mu_2(\hat{\theta}) \\ \vdots \\ m_K(\mathbf{x}) = \mu_K(\hat{\theta}) \end{cases}$$

如果该联立方程有解，即可得到参数 θ_0 的估计。

多个参数的矩估计

正态分布的矩估计

- 对于正态总体 $x_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$ i.i.d, 其中未知总体参数 $\theta = (\mu, \sigma^2)$, 其一阶样本矩为 $m_1(\mathbf{x}) = \bar{x}$, 二阶样本矩为 $m_2(\mathbf{x}) = \overline{x^2}$ 。
- 我们知道对于正态分布, $\mu_1(\theta) = \mu, \mu_2(\theta) = \mu^2 + \sigma^2$, 从而矩估计为:

$$\begin{cases} m_1(\mathbf{x}) = \bar{x} = \hat{\mu} \\ m_2(\mathbf{x}) = \overline{x^2} = \hat{\mu}^2 + \hat{\sigma}^2 \end{cases}$$

解得:

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2 \end{cases}$$

多个参数的矩估计

正态分布的矩估计

下面分析其无偏性和一致性。

- 根据之前的结论, $\mathbb{E}\hat{\mu} = \mu_0, \mathbb{E}(\hat{\sigma}^2) = \frac{N-1}{N}\sigma_0^2$, 因而 $\hat{\mu}$ 是无偏估计量而 $\hat{\sigma}^2$ 并非无偏估计量。
- 而由于 $\bar{x} \xrightarrow{p} \mu_0, \overline{x^2} \xrightarrow{p} \mu_0^2 + \sigma_0^2$, 从而 $\hat{\sigma}^2 \xrightarrow{p} \mu_0^2 + \sigma_0^2 - \mu_0^2 = \sigma_0^2$, 因而 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 都是一致估计量。

多个参数的矩估计

条件正态的矩估计

- 考虑如下模型： $y_i|v_i \sim \mathcal{N}(\mu, v_i)$ ，其中条件方差 $v_i \sim \mathcal{E}(0, \beta)$ ，而 v_i 不可观测。
- 此时，我们可以推算矩，其中

$$\mathbb{E}(y_i) = \mathbb{E}[\mathbb{E}(y_i|v_i)] = \mu$$

而方差为

$$\begin{aligned}\mathbb{V}(y_i) &= \mathbb{E}[\mathbb{V}(y_i|v_i)] + \mathbb{V}[\mathbb{E}(y_i|v_i)] \\ &= \mathbb{E}(v_i) + \mathbb{V}(\mu) \\ &= \beta\end{aligned}$$

从而 $\mathbb{E}(y_i^2) = \beta + \mu^2$ 。

多个参数的矩估计

条件正态的矩估计

- 从而矩估计为

$$\begin{cases} \bar{x} = \hat{\mu} \\ \overline{x^2} = \hat{\beta} + \hat{\mu}^2 \end{cases}$$

即

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\beta} = \overline{x^2} - \bar{x}^2 \end{cases}$$

更加一般的矩估计

接下来，我们将讨论矩估计的一般形式。

- 对于一个统计模型，记 $\mathbf{w} = [w_1, \dots, w_N]'$ 为独立同分布的随机样本，真实参数 $\theta_0 \in \Theta \subset \mathbb{R}^K$
- 只要我们可以找到 K 个矩条件，使得 θ_0 为矩条件（moment condition）方程：

$$\mathbb{E}[g(w_i, \theta)] = \mathbb{E} \left(\begin{bmatrix} g_1(w_i, \theta) \\ \vdots \\ g_K(w_i, \theta) \end{bmatrix} \right) = 0$$

的唯一解，那么我们就可以解以上总体矩方程组的样本方程等价形式：

$$\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) = 0$$

解得 $\hat{\theta}$

- 可以证明，在一些额外比较宽松的条件下， $\hat{\theta} \xrightarrow{P} \theta_0$ 。

更加一般的矩估计

对数正态分布的矩估计

- 之前我们已经计算了对数正态分布的矩估计，我们使用了矩条件 $\mathbb{E}_{\mu} x_i = e^{\mu+1}$ 进行估计
- 而实际上我们也可以使用 x_i 其他函数的期望进行估计
- 比如，我们可以令 $g(x_i, \mu) = \ln x_i - \mu$ ，使用矩条件 $\mathbb{E}(\ln x_i - \mu) = 0$ ，从而一个自然的估计为： $\hat{\mu} = \overline{\ln x} = \frac{1}{N} \sum_{i=1}^N \ln(x_i)$
- 可以验证以上估计是一个无偏、一致估计。

识别

- 注意我们要求 θ_0 为矩条件方程组的唯一解，即模型是可识别的（identifiable），这要求矩条件方程不仅有解，而且解唯一。
 - 如果矩条件方程有不只一组解，那么我们无法区分真实参数究竟是哪一组解，导致该统计问题无法准确回答。
- 识别（identification）问题是计量经济学的核心问题。

识别问题

Beta分布的识别问题

如果 $x_i \sim \text{Beta}(\alpha, \beta)$, 我们知道其前两阶矩的矩条件为:

$$\begin{cases} \mathbb{E}(x_i) = \frac{\alpha}{\alpha+\beta} \\ \mathbb{E}(x_i^2) = \frac{\alpha\beta + \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{cases}$$

因而其矩估计为:

$$\begin{cases} \frac{\alpha}{\alpha+\beta} = \bar{x} \\ \frac{\alpha\beta + \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} = \overline{x^2} \end{cases}$$

识别问题

Beta分布的识别问题

- 然而在估计以上问题之前，我们必须首先证明识别问题，即真值 (α_0, β_0) 为矩条件的唯一解，或者不存在另外的 (α_1, β_1) 使得：

$$\begin{cases} \frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\alpha_0}{\alpha_0 + \beta_0} \\ \frac{\alpha_1 \beta_1 + \alpha_1^2 (\alpha_1 + \beta_1 + 1)}{(\alpha_1 + \beta_1)^2 (\alpha_1 + \beta_1 + 1)} = \frac{\alpha_0 \beta_0 + \alpha_0^2 (\alpha_0 + \beta_0 + 1)}{(\alpha_0 + \beta_0)^2 (\alpha_0 + \beta_0 + 1)} \end{cases}$$

- 如果存在着这样的 (α_1, β_1) ，那么使用 (α_1, β_1) 和使用真值 (α_0, β_0) 所代表的总体有相同的一、二阶矩，因而我们无法判断哪一个是对的。

识别问题

Beta分布的识别问题

为了证明以上结论，用反证法，现在假设存在这样的 (α_1, β_1) ，根据第一个等式，那么必然有： $\frac{\alpha_1}{\beta_1} = \frac{\alpha_0}{\beta_0} \triangleq c_0$ 带入到第二个式子，有：

$$\begin{aligned} \frac{\alpha_1 \beta_1 + \alpha_1^2 (\alpha_1 + \beta_1 + 1)}{(\alpha_1 + \beta_1)^2 (\alpha_1 + \beta_1 + 1)} &= \frac{\beta_1 c_0 \beta_1 + \beta_1^2 c_0^2 (\beta_1 c_0 + \beta_1 + 1)}{(\beta_1 c_0 + \beta_1)^2 (\beta_1 c_0 + \beta_1 + 1)} \\ &= \frac{c_0 \beta_1^2 + \beta_1^2 c_0^2 (\beta_1 c_0 + \beta_1 + 1)}{\beta_1^2 (c_0 + 1)^2 (\beta_1 c_0 + \beta_1 + 1)} \\ &= \frac{c_0 + c_0^2 (\beta_1 c_0 + \beta_1 + 1)}{(c_0 + 1)^2 (\beta_1 c_0 + \beta_1 + 1)} \end{aligned}$$

识别问题

Beta分布的识别问题

同理

$$\frac{\alpha_0 \beta_0 + \alpha_0^2 (\alpha_0 + \beta_0 + 1)}{(\alpha_0 + \beta_0)^2 (\alpha_0 + \beta_0 + 1)} = \frac{c_0 + c_0^2 (\beta_0 c_0 + \beta_0 + 1)}{(c_0 + 1)^2 (\beta_0 c_0 + \beta_0 + 1)}$$

如果存在这样的 (α_1, β_1) ，那么自然有

$$\frac{c_0 + c_0^2 (\beta_1 c_0 + \beta_1 + 1)}{(c_0 + 1)^2 (\beta_1 c_0 + \beta_1 + 1)} = \frac{c_0 + c_0^2 (\beta_0 c_0 + \beta_0 + 1)}{(c_0 + 1)^2 (\beta_0 c_0 + \beta_0 + 1)}$$

然而上式如果成立，则必然有 $\beta_1 = \beta_0$ ，从而 $\alpha_1 = c_0 \beta_1 = c_0 \beta_0 = \alpha_0$ 。因而，以上矩条件有且仅有唯一解，使用前两阶矩是完全可以识别 (α_0, β_0) 的。

识别问题

外汇业务人数的识别问题

- 如果 $M|N \sim \text{Bi}(N, p)$, $N \sim \mathcal{P}(\lambda)$, (p, λ) 为未知参数。
- 现在, 如果我们只观察到 M , 而 N 观察不到, 如果我们试图使用 M 的前两阶矩估计 p 和 λ , 有:

$$\left\{ \begin{array}{l} \mathbb{E}(M) = \mathbb{E}[\mathbb{E}(M|N)] = \mathbb{E}(Np) = \lambda p \\ \mathbb{E}(M^2) = \mathbb{V}(M) + [\mathbb{E}(M)]^2 \\ \quad = \mathbb{V}[\mathbb{E}(M|N)] + \mathbb{E}[\mathbb{V}(M|N)] + \lambda^2 p^2 \\ \quad = \mathbb{V}(Np) + \mathbb{E}[Np(1-p)] + \lambda^2 p^2 \\ \quad = p^2 \lambda + p(1-p)\lambda + \lambda^2 p^2 \\ \quad = \lambda p + \lambda^2 p^2 \end{array} \right.$$

识别问题

外汇业务人数的识别问题

- 联立以上方程得到了一个恒等式： $\mathbb{E}(M^2) = \mathbb{E}(M) + [\mathbb{E}(M)]^2$ 或者 $\mathbb{V}(M) = \mathbb{E}(M)$ ，无法单独解出 λ 或者 p 。
- 实际上，在这种情况下，由于 $\mathbb{V}(M) = \mathbb{E}(M)$ ，因而第二个方程是多余的，而如果只使用第一个方程，一个方程无法解除两个参数，即有无穷多组解。因而如果只使用前两阶矩估计 p 和 λ ，是无法识别的。

一元线性回归

在保证识别的前提下，很多统计学问题都可以使用矩估计进行参数估计。

一元线性回归

- 给定数据 $w_i = (y_i, x_i)'$ ，我们希望估计条件期望： $\mathbb{E}(y_i|x_i)$ 。如果我们假定条件期望为线性函数形式，即

$$\mathbb{E}(y_i|x_i) = \alpha_0 + \beta_0 x_i$$

那么只要估计得到 α 和 β 就得到了条件期望的估计。

- 注意在这里我们并没有对 x_i 和 y_i 的联合分布做任何假定，只是假设条件期望 $\mathbb{E}(y_i|x_i)$ 为线性函数形式。

一元线性回归

一元线性回归

- 如果我们令 $u_i = y_i - \mathbb{E}(y_i|x_i) = y_i - \alpha_0 - \beta_0 x_i$ 那么根据条件期望的性质, 有:

$$\mathbb{E}(u_i|x_i) = \mathbb{E}(y_i - \alpha_0 - \beta_0 x_i|x_i) = \mathbb{E}(y_i|x_i) - (\alpha_0 + \beta_0 x_i) = 0$$

从而真值 α_0, β_0 满足:

$$\begin{cases} \mathbb{E}(y_i - \alpha_0 - \beta_0 x_i) = \mathbb{E}(u_i) = \mathbb{E}[\mathbb{E}(u_i|x_i)] = 0 \\ \mathbb{E}[x_i(y_i - \alpha_0 - \beta_0 x_i)] = \mathbb{E}(x_i u_i) = \mathbb{E}[\mathbb{E}(x_i u_i|x_i)] = \mathbb{E}[x_i \mathbb{E}(u_i|x_i)] = 0 \end{cases}$$

一元线性回归

一元线性回归

- 因而 (α_0, β_0) 是矩条件方程：

$$\begin{cases} \mathbb{E}(y_i - \alpha_0 - \beta_0 x_i) = 0 \\ \mathbb{E}[x_i (y_i - \alpha_0 - \beta_0 x_i)] = 0 \end{cases}$$

的解。

一元线性回归

一元线性回归

- 在使用以上矩条件之前还需要讨论识别问题，也就是虽然 (α_0, β_0) 是以上矩条件方程的解，但是我们还要求该解时唯一解。
- 使用上式解得

$$\begin{cases} \alpha_0 = \mathbb{E}(y_i) - \beta_0 \mathbb{E}(x_i) \\ \beta_0 = \frac{\mathbb{E}(x_i y_i) - \mathbb{E}(x_i) \mathbb{E}(y_i)}{\mathbb{E}(x_i^2) - [\mathbb{E}(x_i)]^2} = \frac{\mathbb{C}(x_i, y_i)}{\mathbb{V}(x_i)} \end{cases}$$

- 如果 $\mathbb{V}(x_i) \neq 0$ ，即 x_i 不是常数，那么方程组有唯一解；
- 如果 $x_i = c$ 为一个常数，只要满足 $\alpha - c\beta = \mathbb{E}(y_i)$ 的所有 (α, β) 都是解，因而以上问题是不可识别的。

一元线性回归

一元线性回归

- 假设 x_i 不是常数，那么我们可以用样本的等价形式

$$\begin{cases} \frac{1}{N} \sum_{i=1}^N (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\ \frac{1}{N} \sum_{i=1}^N [x_i (y_i - \hat{\alpha} - \hat{\beta}x_i)] = 0 \end{cases}$$

解得：

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} \end{cases}$$

使用大数定律可以得到，以上为一致估计量。

多元线性回归

多元线性回归

- 给定数据 $w_i = [y_i, x_i']'$, 其中 $x_i \in \mathbb{R}^K$, 即

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix}_{K \times 1}$$

- 如果令 $x_{i1} = 1$, 假设

$$\mathbb{E}(y_i | x_i) = x_i' \beta_0 = \beta_{10} + \beta_{20} x_{i2} + \cdots + \beta_{K0} x_{iK}$$

那么令 $u_i = y_i - x_i' \beta_0$, 有 $\mathbb{E}(u_i | x_i) = 0$

多元线性回归

多元线性回归

- 从而

$$\begin{aligned}\mathbb{E} \left(\begin{bmatrix} u_i \\ u_i x_{i2} \\ \vdots \\ u_i x_{iK} \end{bmatrix} \right) &= \mathbb{E} (u_i x_i) \\ &= \mathbb{E} [\mathbb{E} (u_i x_i | x_i)] \\ &= \mathbb{E} [x_i \mathbb{E} (u_i | x_i)] = 0\end{aligned}$$

多元线性回归

多元线性回归

- 将 $u_i = y_i - x_i' \beta_0$ 代入, 有

$$\mathbb{E}[x_i (y_i - x_i' \beta_0)] = \mathbb{E}[x_i y_i] - \mathbb{E}[x_i x_i' \beta_0] = 0$$

如果假设 $\mathbb{E}(x_i x_i')$ 可逆, 那么

$$\beta_0 = [\mathbb{E}(x_i x_i')]^{-1} [\mathbb{E}(x_i y_i)]$$

多元线性回归

多元线性回归

- 在这里，我们的矩条件方程为： $\mathbb{E}[x_i y_i] - \mathbb{E}[x_i x_i' \beta_0] = 0$ ，使用样本矩代替总体矩，即

$$\frac{1}{N} \sum_{i=1}^N [x_i y_i] - \frac{1}{N} \sum_{i=1}^N [x_i x_i' \hat{\beta}] = 0$$

解以上方程可得

$$\begin{aligned} \hat{\beta} &= \left[\frac{1}{N} \sum_{i=1}^N (x_i x_i') \right]^{-1} \left[\frac{1}{N} \sum_{i=1}^N (x_i y_i) \right] \\ &= \left[\sum_{i=1}^N (x_i x_i') \right]^{-1} \left[\sum_{i=1}^N (x_i y_i) \right] \end{aligned}$$

多元线性回归

多元线性回归

- 如果令

$$X = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}$$

以及：

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

多元线性回归

多元线性回归

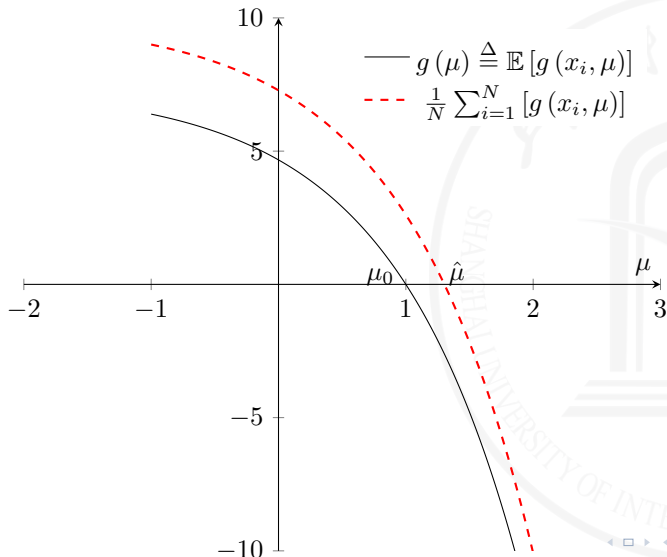
- 那么

$$\begin{cases} X'X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_N \end{bmatrix} = x_1x'_1 + \cdots + x_Nx'_N = \sum_{i=1}^N (x_ix'_i) \\ X'Y = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = x_1y_1 + \cdots + x_Ny_N = \sum_{i=1}^N (x_iy_i) \end{cases}$$

- 从而估计量可以写为 $\hat{\beta} = (X'X)^{-1} X'Y$ 。以上就是所谓的普通最小二乘 (ordinary least squares, OLS)。

矩估计的一致性

- 得到统计量以后，我们经常还需要讨论统计量的无偏、一致、有效等的统计性质。
- 一般而言，矩估计并不能够完全保证估计量是无偏的，但是一致性在很很大程度上都是满足的。
- 为了说明这点，考虑如下图所示的总体矩与样本矩函数。
- 该图画出了对数正态例子中（令 $\mu_0 = 1$ ）的：
 - 总体矩条件方程 $\mathbb{E}[g(x_i, \mu)] = \mathbb{E}(x_i - e^{\mu+1}) = e^{\mu_0+1} - e^{\mu+1} \triangleq g(\mu)$
 - 样本矩条件方程： $\frac{1}{N} \sum_{i=1}^N g(x_i, \mu) = \frac{1}{N} \sum_{i=1}^N (x_i - e^{\mu+1})$
 - 注意以上我们将样本矩和总体矩都视为 μ 的函数（而非 $\mu_0!$ ）。



矩估计的一致性

- 识别条件意味着总体矩条件方程 $g(\mu) = 0$ 的解必须为 μ_0
- 而样本矩条件方程 $\frac{1}{N} \sum_{i=1}^N g(x_i, \mu) = 0$ 的解即我们的估计量 $\hat{\theta}$ 。
- 根据大数定律，自然有

$$\frac{1}{N} \sum_{i=1}^N g(x_i, \mu) \xrightarrow{p} g(\mu)$$

- 可以想象，当样本量足够大时，图中样本矩和总体矩的曲线会无限接近，且由于 $g(x_i, \mu)$ 对于 μ 而言是一个连续函数，那么自然 $\hat{\mu}$ 也会收敛到 μ_0 。

矩估计的一致性

- 然而以上论证在某些特殊情况下也是不成立的。
- 比如，对于帕累托分布 ($x_i \sim \mathcal{Pa}(a, \beta)$ ，其中 a 为范围参数， β 为形状参数，且 $\beta > 0$)，其密度函数为

$$f(x|a, \beta) = \mathbb{1}\{x > a\} \beta \cdot a^\beta x^{-(\beta+1)}$$

从而 $\text{supp}(x) = (a, +\infty)$ 。

- 如果计算期望，有

$$\mathbb{E}(x_i) = \begin{cases} \frac{\beta a}{\beta-1} & \beta > 1 \\ \infty & 0 < \beta \leq 1 \end{cases}$$

矩估计的一致性

- 不妨令 $a = 1$, 那么估计 β 的总体矩条件为

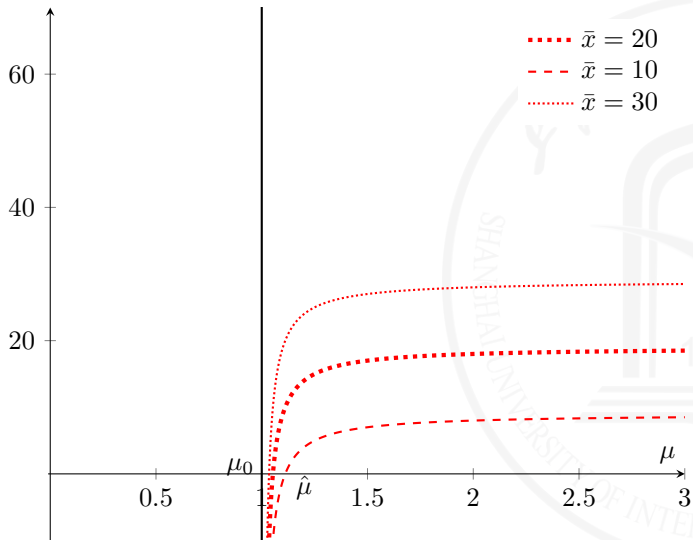
$$\mathbb{E}[g(x_i, \beta)] = \mathbb{E}\left(x_i - \frac{\beta}{\beta - 1}\right) = \frac{\beta_0}{\beta_0 - 1} - \frac{\beta}{\beta - 1} \triangleq g(\beta)$$

而样本矩条件为

$$\frac{1}{N} \sum_{i=1}^N g(x_i, \mu) = \frac{1}{N} \sum_{i=1}^N \left(x_i - \frac{\beta}{\beta - 1}\right)$$

- 下图展示了当 $\beta_0 = 1$ 时的总体和样本矩函数, 其中, 由于当 $\beta_0 = 1$ 时, $\mathbb{E}(x_i) = \infty$, 从而总体矩函数退化成了 $x = 1$ 的竖直线。
- 如果我们使用矩估计 $\frac{1}{N} \sum_{i=1}^N \left(x_i - \frac{\beta}{\beta - 1}\right) = 0$ 从而 $\hat{\beta} = \frac{\bar{x}}{\bar{x} - 1}$ 。
 - 虽然 $\mathbb{E}(x_i) = \infty$, 但是无论样本量多大, 样本均值 \bar{x} 不会正无穷。
 - 注意到 β 的取值范围为 $(0, \infty)$, 然而对于 $\beta_0 < 1$, 估计值 $\hat{\beta} > 1$ 永远成立, 从而我们无法得到一致估计。

矩估计的一致性



矩估计的一致性

- 为了防止出现以上的情况，我们不仅仅需要保证

$$\frac{1}{N} \sum_{i=1}^N g(w_i, \theta) \xrightarrow{p} \mathbb{E}[g(x_i, \theta)] \triangleq g(\theta)$$

而且需要保证对于任意的 θ ，两者之间都要足够的接近，而不能像帕累托分布中一样出现

$$\begin{aligned} g(\beta) - \frac{1}{N} \sum_{i=1}^N g(x_i, \mu) &= \frac{\beta_0}{\beta_0 - 1} - \frac{\beta}{\beta - 1} - \frac{1}{N} \sum_{i=1}^N \left(x_i - \frac{\beta}{\beta - 1} \right) \\ &= \frac{\beta_0}{\beta_0 - 1} - \bar{x} \end{aligned}$$

当 $\beta_0 = 1$ 时 $g(\beta_0) - \frac{1}{N} \sum_{i=1}^N g(x_i, \mu) = \infty$ 的情况。

矩估计的一致性

- 此时，我们需要比大数定律更强的一致大数定律（Uniform law of large numbers, ULLN），即要求对于任意的 $\theta \in \Theta$,

$$\sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^N g(x_i, \theta) - g(\theta) \right| \xrightarrow{p} 0$$

- 这就要求样本矩函数 $\frac{1}{N} \sum_{i=1}^N g(w_i, \theta)$ 和总体矩函数 $\mathbb{E}[g(x_i, \theta)]$ 作为 θ 的函数，两者之间的收敛不应是针对某一个 θ 的，而是针对所有的 $\theta \in \Theta$ ， $\frac{1}{N} \sum_{i=1}^N g(w_i, \theta)$ 和 $\mathbb{E}[g(x_i, \theta)]$ 之间的距离的上界也要收敛到0。

矩估计的一致性

- 一致大数定律可以通过以下几个条件得到：

- Θ 为紧集
- 对于所有的 x , $g(x, \theta)$ 对 θ 都是连续的
- 存在一个不依赖于 θ 的函数 $K(x)$ 满足 $\mathbb{E}[K(x)] < \infty$, 使得对于所有的 x 和 θ , 有: $|g(x, \theta)| \leq K(x)$



矩估计的一致性

矩估计的一致性

如果 $w_i \in \mathbb{R}^p$, $\mathbf{w} = [w_1, \dots, w_N]'$ 为一系列独立同分布的随机样本, $g(w, \theta) \in \mathbb{R}^K$ 为 Borel 可测函数, 假设:

- ① $\theta_0 \in \Theta \subset \mathbb{R}^K$, 其中 Θ 为紧集;
- ② (连续性条件) 对于任意的 w , $g(w, \theta)$ 在 Θ 上对 θ 为连续函数;
- ③ (收敛性条件) 存在一个函数 $K(w)$, 使得对于任意的 $\theta \in \Theta$, $|g(w, \theta)| \leq K(w)$, 其中 $\mathbb{E}[K(w)] < \infty$;
- ④ (识别条件) θ_0 为方程: $\mathbb{E}[g(w_i, \theta)] = 0$ 的唯一解

那么方程:

$$\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) = 0$$

的解 $\hat{\theta} \xrightarrow{p} \theta_0$.

矩估计的一致性

对数正态的矩估计的一致性

在对数正态的例子中，矩条件为： $\mathbb{E}(x_i - e^{\mu+1}) = 0$ ，因而 $g(x, \mu) = x - e^{\mu+1}$ ，可以检查：

- ① 取 $\Theta = [-C, C]$, 为一个紧集, 其中 C 为一个 (可以足够大的) 正常数;
- ② 对 μ 为连续 (且单调) 的函数;
- ③ 取 $K(x) = x + e^{C+1}$, 可知在 Θ 内, 有

$$|g(x, \mu)| = |x - e^{\mu+1}| \leq x + e^{\mu+1} \leq K(x)$$

而 $\mathbb{E}[K(x)] = e^{\mu_0+1} + e^{C+1} < \infty$;

- ④ 当 $\mathbb{E}[g(x_i, \mu)] = 0$ 时, 具有唯一解 $\mu = \mu_0$ 。

因而满足以上定理的条件, 从而矩估计量 $\hat{\mu} \xrightarrow{p} \mu_0$ 。

要知道估计量 $\hat{\theta}$ 的抽样分布，比如其大假设检验等任务。

我们经常诉诸于估计量的大样本分布

- 在得到未知参数 θ 的估计量 $\hat{\theta}$ 之后，我们经常还需要知道估计量 $\hat{\theta}$ 的抽样分布，比如其大样本分布，才能够在此基础上完成区间估计、假设检验等任务。
- 估计量的精确分布一般是非常难以计算的，因而我们经常诉诸于估计量的大样本分布进行近似。

- 首先考虑一维情形。我们假设 $\theta \in \mathbb{R}$ ，那么在矩条件： $\mathbb{E}[g(w_i, \theta_0)] = 0$ 成立的条件下，矩估计量即解如下方程： $\sum_{i=1}^N \left[g(w_i, \hat{\theta}) \right] = 0$
- 如果我们假设函数 $g(\cdot, \cdot)$ 对 $\hat{\theta}$ 是连续可微的，那么我们可以对其在 θ_0 处进行泰勒展开：

$$0 = \frac{1}{N} \sum_{i=1}^N [g(w_i, \theta_0)] + \frac{1}{N} \sum_{i=1}^N \frac{dg(w_i, \theta_0)}{d\theta} (\hat{\theta} - \theta_0) + O\left((\hat{\theta} - \theta_0)^2\right)$$

两边乘以 \sqrt{N} , 得到

$$\begin{aligned} -\sqrt{N} \frac{\sum_{i=1}^N [g(w_i, \theta_0)]}{N} &= \left[\frac{\sum_{i=1}^N \frac{dg(w_i, \theta_0)}{d\theta}}{N} \right] [\sqrt{N} (\hat{\theta} - \theta_0)] + O\left(\sqrt{N} (\hat{\theta} - \theta_0)^2\right) \\ &= \left[\frac{\sum_{i=1}^N \frac{dg(w_i, \theta_0)}{d\theta}}{N} \right] [\sqrt{N} (\hat{\theta} - \theta_0)] + o_p(1) \end{aligned}$$

矩估计的渐近正态性

- 等式左边, 根据中心极限定理, 由于 $\mathbb{E}[g(w_i, \theta_0)] = 0$, 因而有

$$\sqrt{N} \frac{1}{N} \sum_{i=1}^N [g(w_i, \theta_0)] \overset{a}{\sim} \mathcal{N}(0, \mathbb{V}[g(w_i, \theta_0)])$$

其中

$$\begin{aligned} \mathbb{V}[g(w_i, \theta_0)] &= \mathbb{E}[g(w_i, \theta_0)^2] - [\mathbb{E}(g(w_i, \theta_0))]^2 \\ &= \mathbb{E}[g(w_i, \theta_0)^2] \triangleq B \end{aligned}$$

- 等式右边, 根据大数定律, 有

$$\frac{1}{N} \sum_{i=1}^N \frac{dg(w_i, \theta_0)}{d\theta} \xrightarrow{p} \mathbb{E}\left[\frac{dg(w_i, \theta_0)}{d\theta}\right] \triangleq A$$

矩估计的渐近正态性

- 根据以上结论, 原式可以写为:

$$A \cdot \sqrt{N} (\hat{\theta} - \theta_0) + o_p(1) \stackrel{a}{\sim} \mathcal{N}(0, B)$$

从而:

$$\sqrt{N} (\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}\left(0, \frac{B}{A^2}\right)$$

- 注意到 A 和 B 都依赖于未知参数 θ_0 , 因而如果需要计算 $\hat{\theta}$ 的渐近方差, 可以将 $\hat{\theta}$ 带入到 A 和 B 的表达式中进行计算
 - 由于 $\hat{\theta}$ 是 θ_0 的一致估计量, 所以大样本条件下对渐近方差的估计仍然是准确的。

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ▶ ↺ 🔍 ↻

矩估计的渐近正态性

- 对于多元的情形，同样可以使用delta方法。
- 如果我们的总体矩条件方程为：

$$\mathbb{E}[g(w_i, \theta)] = \mathbb{E} \left(\begin{bmatrix} g_1(w_i, \theta) \\ \vdots \\ g_K(w_i, \theta) \end{bmatrix} \right) = 0$$

真值 θ_0 为以上方程的解，那么估计值 $\hat{\theta}$ 使得其样本矩条件方程

$$\frac{1}{N} \sum_{i=1}^N [g(w_i, \hat{\theta})] = 0$$

矩估计的渐近正态性

对以上方程在 θ_0 处进行泰勒展开，即对于 $g(w_i, \hat{\theta})$ 的 K 个分量都进行泰勒展开，有

$$\begin{aligned} 0 &= \frac{1}{N} \sum_{i=1}^N [g(w_i, \hat{\theta})] \\ &= \frac{1}{N} \sum_{i=1}^N [g(w_i, \theta_0)] + \frac{1}{N} \sum_{i=1}^N \left[\frac{\partial g(w_i, \theta_0)}{\partial \theta'} \right] (\hat{\theta} - \theta_0) + O\left(\|\hat{\theta} - \theta_0\|^2\right) \end{aligned} \quad (1)$$

其中：

$$\frac{\partial g(w_i, \theta_0)}{\partial \theta'} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(w_i, \theta_0) & \cdots & \frac{\partial}{\partial \theta_K} g_1(w_i, \theta_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} g_K(w_i, \theta_0) & \cdots & \frac{\partial}{\partial \theta_K} g_K(w_i, \theta_0) \end{bmatrix}_{K \times K}$$

矩估计的渐近正态性

根据大数定律, 令

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{\partial}{\partial \theta'} g(w_i, \theta_0) \right] \xrightarrow{p} \mathbb{E} \left[\frac{\partial}{\partial \theta'} g(w_i, \theta_0) \right] \triangleq \mathcal{G}_0$$

由于 $\mathbb{E}[g(w_i, \theta_0)] = 0$, 而协方差矩阵:

$$\begin{aligned} \mathbb{V}[g(w_i, \theta_0)] &= \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] - \mathbb{E}[g(w_i, \theta_0)] \mathbb{E}[g(w_i, \theta_0)'] \\ &= \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] \end{aligned}$$

因而根据中心极限定理:

$$\sqrt{N} \frac{1}{N} \sum_{i=1}^N [g(w_i, \theta_0)] \overset{a}{\sim} \mathcal{N}(0, \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'])$$

矩估计的渐近正态性

因而方程(1)可以写为

$$\mathcal{G}_0 \sqrt{N} (\hat{\theta} - \theta_0) + o_p(1) \stackrel{a}{\sim} \mathcal{N}(0, \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'])$$

最后得到结论:

$$\sqrt{N} (\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}(0, \mathcal{G}_0^{-1} \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] (\mathcal{G}_0^{-1})')$$

矩估计的渐近正态性

OLS估计量的渐近分布

在OLS中, $g(w_i, \beta) = x_i u_i = x_i (y_i - x_i' \beta)$, 因而

$$G_0 = \mathbb{E} \left[\frac{\partial}{\partial \beta'} g(w_i, \beta) \right] = -\mathbb{E}(x_i x_i')$$

以及:

$$\mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] = \mathbb{E}[x_i u_i u_i' x_i'] = \mathbb{E}[u_i^2 x_i x_i']$$

从而

$$\sqrt{N}(\hat{\beta} - \beta_0) \stackrel{a}{\sim} \mathcal{N}(0, [\mathbb{E}(x_i x_i')]^{-1} \mathbb{E}[u_i^2 x_i x_i'] [\mathbb{E}(x_i x_i')]^{-1})$$

矩估计的渐近正态性

OLS估计量的渐近分布

为了估计渐近方差，可以直接计算 $\hat{u}_i = y_i - x_i' \hat{\beta}$ ，然后将所有的期望用平均进行估计，得到渐近方差的估计

$$\begin{aligned}\mathbb{V}(\hat{\beta}) &= \frac{\left[\frac{1}{N} \sum_{i=1}^N (x_i x_i') \right]^{-1} \frac{1}{N} \sum_{i=1}^N (\hat{u}_i^2 x_i x_i') \left[\frac{1}{N} \sum_{i=1}^N (x_i x_i') \right]^{-1}}{N} \\ &= \left[\sum_{i=1}^N (x_i x_i') \right]^{-1} \sum_{i=1}^N (\hat{u}_i^2 x_i x_i') \left[\sum_{i=1}^N (x_i x_i') \right]^{-1} \\ &= (X'X)^{-1} \sum_{i=1}^N (\hat{u}_i^2 x_i x_i') (X'X)^{-1}\end{aligned}$$

即怀特异方差标准误 (White, 1980)。

矩估计的渐近正态性

OLS估计量的渐近分布（同方差）

如果假设同方差，即 $\mathbb{E}(u_i^2|x_i) = \sigma^2$ ，那么

$$\begin{aligned}\mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] &= \mathbb{E}[u_i^2 x_i x_i'] \\ &= \mathbb{E}(\mathbb{E}[u_i^2 x_i x_i' | x_i]) \\ &= \sigma^2 \mathbb{E}(x_i x_i')\end{aligned}$$

因而根据以上结论，有

$$\begin{aligned}\mathcal{G}_0^{-1} \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] (\mathcal{G}_0^{-1})' &= [-\mathbb{E}(x_i x_i')]^{-1} \sigma^2 \mathbb{E}(x_i x_i') [-\mathbb{E}(x_i x_i')]^{-1} \\ &= \sigma^2 [\mathbb{E}(x_i x_i')]^{-1}\end{aligned}$$

矩估计

- 以上介绍了矩估计，即使用总体矩条件式

$$\mathbb{E}[g(w_i, \theta_0)] = \mathbb{E} \left(\begin{bmatrix} g_1(w_i, \theta_0) \\ \vdots \\ g_K(w_i, \theta_0) \end{bmatrix} \right) = 0$$

的样本对应式

$$\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^N g_1(w_i, \hat{\theta}) \\ \vdots \\ \frac{1}{N} \sum_{i=1}^N g_K(w_i, \hat{\theta}) \end{bmatrix} = 0$$

解方程得到估计

- 当然前提是可识别性，即总体矩条件式有且仅有唯一解。

泊松分布的矩条件

- $$\begin{cases} \mathbb{E}(x_i - \lambda) = 0 \\ \mathbb{E}(x_i^2 - \lambda - \lambda^2) = 0 \end{cases}$$

- $$\mathbb{E} \left(x_i^2 \right) = \mathbb{E} \left(x_i \right) + \left[\mathbb{E} \left(x_i \right) \right]^2$$

- 所以以上虽然有两个矩条件，但是是有唯一解 $\lambda = \mathbb{E}(x_i) = \mathbb{V}(x_i)$ 的。

，即有唯一解，那么我们能不能从样
数呢？

用更多的矩条件有可能可以改善估计
个数，是不能保证样本的矩条件一定

- 如果我们使用了 $G > K$ 个方程还能够保证识别性，即有唯一解，那么我们能不能从样本的角度综合使用 G 个方程一起估计 $K < G$ 个参数呢？
- 由于额外的矩条件可能带来更多的信息，从而使用更多的矩条件有可能可以改善估计量的有效性，但是一旦矩条件个数多于未知参数个数，是不能保证样本的矩条件一定有解的。

矩条件个数

泊松分布的样本矩条件

- 考虑上例，其样本矩为：

$$\begin{cases} \frac{1}{N} \sum_{i=1}^N (x_i - \lambda) = 0 \\ \frac{1}{N} \sum_{i=1}^N (x_i^2 - \lambda - \lambda^2) = 0 \end{cases}$$

根据第一个方程 $\hat{\lambda} = \bar{x}$ ，将其带入到第二个方程，得到 $\overline{x^2} = \bar{x} + \bar{x}^2$ ，然而这一关系在样本中很难满足，从而以上方程组也很难有解。

矩估计的其他形式

- 注意以上例子中，总体矩条件是有解的，仅仅是由于样本的抽样误差而导致样本矩条件没有解，那么有没有可能在样本中也充分使用所有矩条件得到估计呢？
- 考虑总体矩条件的样本对应式：

$$\sum_{i=1}^N g(w_i, \hat{\theta}) = \sum_{i=1}^N \begin{bmatrix} g_1(w_i, \hat{\theta}) \\ \vdots \\ g_K(w_i, \hat{\theta}) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N g_1(w_i, \hat{\theta}) \\ \vdots \\ \sum_{i=1}^N g_K(w_i, \hat{\theta}) \end{bmatrix} = 0$$

由于乘以 $\frac{1}{N}$ 不改变方程的解，方便起见我们将其省略。

方程实际上等价于最小化每个方程

$$\left[\sum_{i=1}^n \ell(w_i, \theta) \right]^2$$

$\ell(w_i, \hat{\theta}) = 0$, 那么 $\hat{\theta}$ 也会使

函数等于0。

完全等价的。

- $$\hat{\theta} = \arg \min_{\theta} \sum_{k=1}^K \left[\sum_{i=1}^N g_k(w_i, \theta) \right]^2$$

- 如果方程组有唯一解，即存在唯一的 $\hat{\theta}$ 使得 $\sum_{i=1}^N g(w_i, \hat{\theta}) = 0$ ，那么 $\hat{\theta}$ 也会使得 $\sum_{k=1}^K \left[\sum_{i=1}^N g_k(w_i, \hat{\theta}) \right]^2$ 达到最小值0；
- 其他的 θ' 都会使得 $\sum_{k=1}^K \left[\sum_{i=1}^N g_k(w_i, \theta') \right]^2 > 0$
- 从而 $\hat{\theta}$ 是以上最优化问题的唯一解，且恰好使目标函数等于0。
- 从而，解以上的最优化问题与解样本矩的方程组是完全等价的。

矩估计的其他形式

- 比如，在正态分布的矩估计中，其样本矩条件为

$$\begin{cases} \frac{1}{N} \sum_{i=1}^N (x_i - \mu) = 0 \\ \frac{1}{N} \sum_{i=1}^N (x_i^2 - \mu^2 - \sigma^2) = 0 \end{cases}$$

如果使用最优化的方法，即最小化

$$\min_{\mu, \sigma^2} \left[\sum_{i=1}^N (x_i - \mu) \right]^2 + \left[\sum_{i=1}^N (x_i^2 - \mu^2 - \sigma^2) \right]^2$$

可以验证， $\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2$ 刚好使得以上目标函数等于0，即最小化了以上的目标函数。

矩估计的Stata实现

Beta分布的矩估计

- 其中gmm命令即为（广义）矩估计命令
- winit(identity)即将两个样本矩条件的平方和相加，稍后介绍这个选项的含义；
- 第一次使用“{alpha}”、“{beta}”时大括号中的“=1”的作用是设置一个初始值。我们可以使用模拟的方法验证以上矩估（MM_beta.do）

广义矩估计

- 将以上想法进行推广，如果矩条件个数 $G > K$ ， $\theta \in \mathbb{R}^K$ ，总体矩条件为

$$\mathbb{E}[g(w_i, \theta_0)] = \mathbb{E} \left(\begin{bmatrix} g_1(w_i, \theta_0) \\ \vdots \\ g_G(w_i, \theta_0) \end{bmatrix} \right) = 0$$

- 样本矩条件

$$\sum_{i=1}^N g(w_i, \hat{\theta}) = \begin{bmatrix} \sum_{i=1}^N g_1(w_i, \hat{\theta}) \\ \vdots \\ \sum_{i=1}^N g_G(w_i, \hat{\theta}) \end{bmatrix} = 0$$

可能没有解

广义矩估计

- 但是我们可以最小化目标函数

$$\hat{\theta} = \arg \min_{\theta} \sum_{g=1}^G \left[\sum_{i=1}^N g_g(w_i, \theta) \right]^2$$

- 此时, $\hat{\theta}$ 并不能保证每个 $\sum_{i=1}^N g_g(w_i, \theta)$ 都等于0, 但是每个样本矩都会足够贴近于0, 即 $\sum_{i=1}^N g_g(w_i, \theta) \approx 0$, 而与0的误差可能仅仅因为抽样误差所致, 此时 $\hat{\theta}$ 也是一个足够好的估计。

广义矩估计

- 此外，注意到

$$\begin{aligned}\sum_{g=1}^G \left[\sum_{i=1}^N g_g(w_i, \theta) \right]^2 &= \begin{bmatrix} \sum_{i=1}^N g_1(w_i, \theta) \\ \vdots \\ \sum_{i=1}^N g_G(w_i, \theta) \end{bmatrix}' \begin{bmatrix} \sum_{i=1}^N g_1(w_i, \theta) \\ \vdots \\ \sum_{i=1}^N g_G(w_i, \theta) \end{bmatrix} \\ &= \left[\sum_{i=1}^N g(w_i, \theta) \right]' \left[\sum_{i=1}^N g(w_i, \theta) \right]\end{aligned}$$

从而以上的目标函数也可以写为

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

广义矩估计的Stata实现

Poisson分布的矩估计

- 在泊松分布的例子中，如果使用两个矩条件，该估计可以通过最小化

$$\min_{\lambda} \left[\sum_{i=1}^N (x_i - \lambda) \right]^2 + \left[\sum_{i=1}^N (x_i^2 - \lambda - \lambda^2) \right]^2$$

得到。使用gmm命令对其进行估计：

```
1 gen `x2'=`varlist'^2
2 gmm (`varlist'-{lambda=`lambda'}) (`x2'-({lambda}+{lambda}^2)), winit(
    identity) onestep
```

- 同样使用winit(identity)实现了两个样本矩条件求和再最小化，而onestep为一步估计，接下来会详细解释

广义矩估计

- 在上例中，winit(identity)中的identity实际上代指单位阵，即在目标函数中乘以一个单位阵

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N g(w_i, \theta) \right]' I \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

其中 I 为 $G \times G$ 维的单位阵，其结果并不改变。

- 上式只是将两个矩条件的平方简单的加起来，一个很自然的想法是，两个矩条件所带有的信息可能是不同的，甚至是相关的，那么是否可以将其做一个加权平均呢？

广义矩估计

- 现在，如果我们找到一个 $G \times G$ 的（对称的）正定矩阵 \mathbf{W} ，代替单位阵，即最小化

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \mathbf{W} \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

- 那么如果 $G = K$ ，我们仍然可以保证矩估计 $\hat{\theta}$ 是以上最优化问题的唯一解
- 而当 $G > K$ 时，以上目标函数对不同的矩条件进行了“加权”，直觉上我们可以给信息量更大的矩条件以更大的权重，这可能可以提高估计的精度
- 从而，矩阵 \mathbf{W} 也被称为权重矩阵（weighting matrix）
- 而最小化以上目标函数得到的估计量称为广义矩估计（generalized method of moments, GMM）。
 - 根据以上论述，矩估计是当矩条件个数与未知参数维数相等（ $G = K$ ）时的特例。

最优权重矩阵

- 可以想象，当我们使用不同的权重矩阵时，得到的GMM估计量 $\hat{\theta}^w$ 也是不相等的，且其精度（方差、标准误）也会有区别
- 此时就会存在一个特殊的权重矩阵 W^* ，使用该矩阵得到的GMM估计量 $\hat{\theta}^{W^*}$ 所能达到的标准误最小，即对于所有的正定矩阵 W ，有

$$\mathbb{V}(\hat{\theta}^{W^*}) \preceq \mathbb{V}(\hat{\theta}^W)$$

从而我们称 W^* 为最优权重矩阵（optimal weighting matrix）或者有效权重矩阵（efficient weighting matrix）

- 在下一节中我们将证明，最优权重矩阵应该为：

$$W^* = (\mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'])^{-1}$$

最优权重矩阵

- 注意该矩阵是一个有关真值 θ_0 的函数，从而该矩阵并不可观测，为此我们需要对该矩阵进行估计：

$$\hat{\mathcal{W}}^* = \left(\mathbb{E} \left[g(w_i, \hat{\theta}) g(w_i, \hat{\theta})' \right] \right)^{-1}$$

- 需要注意的是，我们不能最小化以下目标函数估计 θ ：

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \left(\frac{1}{N} \sum_{i=1}^N [g(w_i, \theta) g(w_i, \theta)'] \right)^{-1} \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

即权重矩阵 \mathcal{W} 应为事先给定的，而不能作为最优化的一部分。

最优权重矩阵

- 为了解决最优权重矩阵的估计问题，我们可以使用两步法 (twostep) :
 - ① 给一个初始的正定矩阵 $\hat{\mathcal{W}}_0$ ，比如 $\hat{\mathcal{W}}_0 = I$ 即单位阵，最小化目标函数，得到一个初始估计 $\hat{\theta}^{(1)}$;
 - ② 使用 $\hat{\theta}^{(1)}$ 计算最优权重矩阵

$$\hat{\mathcal{W}}_1 = \left(\frac{1}{N} \sum_{i=1}^N \left[g(w_i, \hat{\theta}^{(1)}) g(w_i, \hat{\theta}^{(1)})' \right] \right)^{-1}$$

并最小化

$$\hat{\theta}^{(2)} = \arg \min_{\theta} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \hat{\mathcal{W}}_1 \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

- 由于给定任意的正定矩阵，GMM估计都是一致估计，从而以上第2步实际是事先估计了最优权重矩阵，再进行“两步估计”。
 - 当然，对于 $G = K$ 的情况，第二步是可以忽略的，因为给定任意的正定矩阵都可以使目标函数达到最小值0。

广义矩估计的Stata实现

Poisson分布的矩估计

- 在泊松分布的例子中，如果将onestep去掉：

```
1 clear
2 set obs 100
3 set seed 8889
4 gen x=rpoisson(10)
5 gen x2=x^2
6 gmm (x-{\lambda=1}) (x2-({\lambda}+{\lambda}^2)), winit(identity)
```

或者加入twostep选项（twostep为gmm的默认选项），就可以进行以上的两步估计了，其中winit(identity)选项设定了第一步的权重矩阵使用了单位阵。

最优权重矩阵

- 当然，以上步骤还可以继续，即给定 $\hat{\theta}^{(s)}$ 计算最优权重矩阵

$$\hat{W}_s = \left(\frac{1}{N} \sum_{i=1}^N \left[g(w_i, \hat{\theta}^{(s)}) g(w_i, \hat{\theta}^{(s)})' \right] \right)^{-1}$$

再将其带入到目标函数中最优化，得到 $\hat{\theta}^{(s+1)}$

- 取一个足够小的数字 $\epsilon > 0$ ，直到 $|\hat{\theta}^{(s)} - \hat{\theta}^{(s+1)}| < \epsilon$ 时才停止迭代。这种方法被称为迭代GMM (iterative GMM)。

广义矩估计的Stata实现

Poisson分布的矩估计

- 在泊松分布的例子中，迭代GMM可以使用igmm选项：

```
1 || gmm (x-{\lambda=1}) (x2-({\lambda}+{\lambda}^2)), winit(identity) igmm  
    igmmeps(1e-4)
```

其中igmmeps(1e-4)即设定 $\epsilon = 10^{-4}$ 。

广义矩估计的Stata实现

ET Tobit

- 考虑如下模型：

$$\ln(a + y_i) = \alpha + \beta x_i + u_i$$

其中 a 为未知参数，且 $\mathbb{E}(u_i|x_i) = 0$ 。

- 如果 a 可以观测，那么我们可以计算出左手边的变量，然后使用OLS方法就可以估计 α, β 了
- 但是， a 是不可观测的，这意味着我们无法计算出左手边的变量。

ET Tobit

- 为了估计未知参数，注意到

$$\mathbb{E}(u_i|x_i) = 0 \Rightarrow \mathbb{E}[g(x_i) \cdot u_i] = 0$$

其中 $g(x_i)$ 为 x_i 的任意函数

- 从而我们可以选取很多的 x_i 的函数，比如 $1, x_i, x_i^2, x_i^3, e^{x_i}, \sin(x_i), \dots$ ，并使用矩条件：

$$\begin{cases} \mathbb{E}(1 \cdot u_i) &= \mathbb{E}\{1 \cdot [\ln(a + y_i) - \alpha + \beta x_i]\} = 0 \\ \mathbb{E}(x_i \cdot u_i) &= \mathbb{E}\{x_i \cdot [\ln(a + y_i) - \alpha + \beta x_i]\} = 0 \\ \mathbb{E}(x_i^2 \cdot u_i) &= \mathbb{E}\{x_i^2 \cdot [\ln(a + y_i) - \alpha + \beta x_i]\} = 0 \\ &\vdots \end{cases}$$

注意到以上矩条件存在唯一解，从而可识别性条件满足，从而可以使用广义矩估计对其进行估计。

广义矩估计的Stata实现

ET Tobit

- 如下代码产生了如此数据生成过程的数据：

```
1 set obs 500
2 set seed 8997
3 gen x=rnormal(3,2)
4 local a=10
5 local alpha=5
6 local beta=3
7 local sigma=2
8 gen y = exp(`alpha'+`beta'*x+rnormal()*sqrt(`sigma'))-`a'
```

广义矩估计的Stata实现

ET Tobit

- 如下代码使用GMM对其进行了估计：

```
1 // 生成工具
2 gen x2=x^2
3 gen x3=x^3
4 gen x_e=exp(x)
5 gen x_sin=sin(x)
6 // 计算的初始值a
7 su y
8 local a_init=-1*r(min)+5
9 // 估计GMM
10 gmm (log({a=`a_init'+y})-{alpha=0}-{beta=0}*x) ((log({a}+y)-{alpha}-{beta}*x)*x) ((log({a}+y)-{alpha}-{beta}*x)*x2) ((log({a}+y)-{alpha}-{beta}*x)*x3) ((log({a}+y)-{alpha}-{beta}*x)*x_e) ((log({a}+y)-{alpha}-{beta}*x)*x_sin), winit(identity)
```

广义矩估计的Stata实现

ET Tobit

- 注意到我们在对参数 a 设定初始值时，使用了 $a_{init} = -1 \times \min\{y\} + 5$ ，其目的是为了在初始值时就让 $\ln(a + y_i)$ 对于所有的 i 都有意义，即 $a_{init} + y_i > 0$ （在这里数字5的作用实际上使得 $a_{init} + y_i \geq 5$ ）
- 否则如果初始值使得部分 $a_{init} + y_i \leq 0$ ，那么第一次估计时使用的样本就不包含那些 $a_{init} + y_i \leq 0$ 的样本，估计就会有很大偏差。
- 所以在计算时初始值的选取是非常重要的，至少应该关注到参数的取值范围以及数据的取值范围，否则可能会犯很严重的错误。
- 可见估计的结果与真实值无异。

工具变量

- 注意上例中的矩条件，都是

$$\mathbb{E}[g(x_i) \cdot u_i] = 0$$

的形式，不同的矩条件仅仅是选择了不同的 $g(x_i)$ 。

- 我们可以把这些与 u_i 不相关的变量 $z_i = g(x_i)$ 称为“工具变量” (instrumental variables, Reiersöl, 1945, 以及Sargan, 1958)
- 而gmm命令可以使用工具变量选项简化以上估计命令，从而上例代码中的gmm命令可以直接改写为：

```
1 gmm (log({a=`a_init'+y})-{alpha=0}-{beta=0}*x), instruments(x*)
```

- 其中gmm命令后面的括号里是计算 $u_i = \ln(a + y_i) - \alpha + \beta x_i$ 的表达式，而instruments()选项中的 x^* 代表了用所有以 x 开始的变量名作为工具变量进行估计。
- 注意我们在这条命令中没有使用winit()选项，是因为instruments()选项有其默认的初始权重矩阵选项，所以无需额外提供。也正是由于初始矩阵的这一区别，导致了两种写法的估计结果有些许差别。

广义矩估计的一致性

而对于广义矩估计

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \hat{W} \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

的一致性，考虑到权重矩阵 \hat{W} 经常是一个估计的结果，所以我们需要在矩估计一致性的条件中额外加入一个假设：

最优权重矩阵的极限假定

假设 $\hat{W} \xrightarrow{P} W_0$ ，其中 W_0 为 $G \times G$ 的正定矩阵。

那么可以保证广义矩估计也是一致估计量。

广义矩估计的渐近分布

现在考虑 $\{w_i\}$ 独立同分布条件下广义矩估计的渐近分布。由于权重矩阵 \hat{W} 是在优化之前事先给定的，所以当我们考虑最优化问题的一阶条件时，无需对 \hat{W} 求导，从而目标函数的一阶导为

$$\frac{\partial \left[\sum_{i=1}^N g(w_i, \theta) \right]'}{\partial \theta} \hat{W} \left[\sum_{i=1}^N g(w_i, \theta) \right] = 2 \left[\sum_{i=1}^N \frac{\partial g(w_i, \theta)}{\partial \theta} \right]' \hat{W} \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

其中

$$\frac{\partial g(w_i, \theta)}{\partial \theta'} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(w_i, \theta) & \cdots & \frac{\partial}{\partial \theta_K} g_1(w_i, \theta) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} g_G(w_i, \theta) & \cdots & \frac{\partial}{\partial \theta_K} g_G(w_i, \theta) \end{bmatrix}_{G \times K}$$

从而 $\left[\sum_{i=1}^N \frac{\partial g(w_i, \theta)}{\partial \theta'} \right]' \hat{W} \left[\sum_{i=1}^N g(w_i, \theta) \right]$ 为 $K \times 1$ 的列向量。一阶条件即

$$\left[\sum_{i=1}^N \frac{\partial g(w_i, \hat{\theta})}{\partial \theta'} \right]' \hat{W} \left[\sum_{i=1}^N g(w_i, \hat{\theta}) \right] = 0$$

即可解得 $\hat{\theta}$ 。

广义矩估计的渐近分布

现在我们可以仿照矩估计大样本性质的做法，对以上一阶条件进行泰勒展开。我们可以首先将一阶条件写为

$$\left[\frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \hat{\theta})}{\partial \theta'} \right]' \hat{W} \left[\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) \right] = 0$$

其中根据一致性结论 $\hat{\theta} \xrightarrow{p} \theta_0$ ，从而如果 $g(w_i, \theta)$ 连续可微，有

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \hat{\theta})}{\partial \theta'} = \frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \theta_0)}{\partial \theta'} + o_p(1) \xrightarrow{p} \mathbb{E} \left[\frac{\partial g(w_i, \theta_0)}{\partial \theta'} \right] \triangleq \mathcal{G}_0$$

以及 $\hat{W} \xrightarrow{p} W_0$ ，从而

$$\left[\frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \hat{\theta})}{\partial \theta'} \right]' \hat{W} \left[\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) \right] = \mathcal{G}_0' W_0 \left[\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) \right] + o_p(1)$$

广义矩估计的渐近分布

对 $\mathcal{G}'_0 \mathcal{W}_0 \left[\frac{1}{N} \sum_{i=1}^N g(w_i, \hat{\theta}) \right]$ 进行泰勒展开, 有

$$0 = \mathcal{G}'_0 \mathcal{W}_0 \left[\frac{1}{N} \sum_{i=1}^N g(w_i, \theta_0) \right] + \mathcal{G}'_0 \mathcal{W}_0 \left[\frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \theta_0)}{\partial \theta'} \right] (\hat{\theta} - \theta_0) + o_p(1)$$

其中:

- $\frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \theta_0)}{\partial \theta} \xrightarrow{p} \mathcal{G}_0$
- 由于总体矩条件为 $\mathbb{E}[g(w_i, \theta_0)] = 0$, 根据中心极限定理:

$$\sqrt{N} \frac{1}{N} \sum_{i=1}^N g(w_i, \theta_0) \overset{a}{\sim} \mathcal{N}(0, \mathbb{V}[g(w_i, \theta_0)]) \quad (2)$$

其中

$$\mathbb{V}[g(w_i, \theta_0)] = \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)']$$

广义矩估计的渐近分布

从而原式可以改写为

$$\begin{aligned} -\mathcal{G}_0' \mathcal{W}_0 \mathcal{G}_0 \sqrt{N} (\hat{\theta} - \theta_0) &= \mathcal{G}_0' \mathcal{W}_0 \left[\sqrt{N} \frac{1}{N} \sum_{i=1}^N g(w_i, \theta_0) \right] + o_p(1) \\ &\stackrel{a}{\sim} \mathcal{N}(0, \mathcal{G}_0' \mathcal{W}_0 \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] \mathcal{W}_0 \mathcal{G}_0) \end{aligned}$$

如果记

$$A_0 = \mathcal{G}_0' \mathcal{W}_0 \mathcal{G}_0$$

$$B_0 = \mathcal{G}_0' \mathcal{W}_0 \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] \mathcal{W}_0 \mathcal{G}_0$$

从而 A_0 为 $K \times K$ 的对称矩阵, 那么

$$\sqrt{N} (\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}(0, A_0^{-1} B_0 A_0^{-1})$$

- $\hat{\theta}$ 的渐近方差为 $\mathbb{V}(\hat{\theta}) = \frac{A_0^{-1} B_0 A_0^{-1}}{N}$
- 上面的结论中我们要求 A_0 为可逆矩阵, 要求 $\text{rank}(\mathcal{G}_0) = K$ 。

GMM估计 $\hat{\theta}^w$
不相同的。
差中， A_0 和 B_0 都与 w_0 有关，那么怎样

- 前面提到，使用不同的权重矩阵 W 会得到不同的GMM估计 $\hat{\theta}^W$
- 虽然他们都是一致的，但是估计量的渐近方差是不相同的。
- 注意到在以上推导得到的GMM估计量的渐近方差中， A_0 和 B_0 都与 W_0 有关，那么怎样的 W_0 会使得以上渐近方差达到最小呢？

最优权重矩阵

观察 B_0 的结构, 如果我们取

$$\mathcal{W}^* = (\mathbb{V}[g(w_i, \theta_0)])^{-1} = (\mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'])^{-1}$$

那么

$$\begin{aligned} B_0 &= \mathcal{G}_0' \mathcal{W}^* \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] \mathcal{W}^* \mathcal{G}_0 \\ &= \mathcal{G}_0' \mathcal{W}^* \mathcal{W}^{*-1} \mathcal{W}^* \mathcal{G}_0 \\ &= \mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0 \end{aligned}$$

从而

$$\begin{aligned} A_0^{-1} B_0 A_0^{-1} &= (\mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0)^{-1} (\mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0) (\mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0)^{-1} \\ &= (\mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0)^{-1} \end{aligned}$$

此时, 渐近方差应该为

$$\mathbb{V}(\hat{\theta}) = \frac{(\mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0)^{-1}}{N}$$

最优权重矩阵

- 现在，取任意的 \mathcal{W} ，如果我们可以证明

$$(\mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0)^{-1} \preceq (\mathcal{G}_0' \mathcal{W} \mathcal{G}_0)^{-1} \mathcal{G}_0' \mathcal{W} \mathcal{W}^{*-1} \mathcal{W} \mathcal{G}_0 (\mathcal{G}_0' \mathcal{W} \mathcal{G}_0)^{-1}$$

就可以证明使用 \mathcal{W}^* 所得到的渐近方差是最小的，从而 $\hat{\theta}$ 的每个分量 $\hat{\theta}_k$ 的方差都更小，从而标准误差更小。

- 对于实对称矩阵 A, B ， $A \succeq B$ 当且仅当 $A^{-1} \preceq B^{-1}$ ，证明见Horn和Johnson（2013）的7.7节。
- 为了证明上式，只需要证明

$$\mathcal{G}_0' \mathcal{W}^* \mathcal{G}_0 \succeq (\mathcal{G}_0' \mathcal{W} \mathcal{G}_0) (\mathcal{G}_0' \mathcal{W} \mathcal{W}^{*-1} \mathcal{W} \mathcal{G}_0)^{-1} (\mathcal{G}_0' \mathcal{W} \mathcal{G}_0)$$

或者

$$\mathcal{G}_0' \left[\mathcal{W}^* - (\mathcal{W} \mathcal{G}_0) (\mathcal{G}_0' \mathcal{W} \mathcal{W}^{*-1} \mathcal{W} \mathcal{G}_0)^{-1} (\mathcal{G}_0' \mathcal{W}) \right] \mathcal{G}_0 \succeq 0$$

等价的

$$\mathcal{W}^* - (\mathcal{W} \mathcal{G}_0) (\mathcal{G}_0' \mathcal{W} \mathcal{W}^{*-1} \mathcal{W} \mathcal{G}_0)^{-1} (\mathcal{G}_0' \mathcal{W}) \succeq 0$$

上式成立（幂等矩阵）

最优权重矩阵

工具变量的最优权重矩阵

- 考虑工具变量的矩条件

$$\mathbb{E}[z_i \cdot u_i(\theta_0)] = 0$$

其中 $u_i(\theta) : \mathbb{R}^K \rightarrow \mathbb{R}$ 为未知参数的函数

- 比如在ET Tobit例子中 $u_i(\theta) = \ln(a + y_i) - \alpha - \beta x_i$
- 记 $u_i = u_i(\theta_0)$ 为真实的误差项, z_i 为 $G \times 1$ 的工具变量。
- 我们可以计算, 最优权重矩阵应该为

$$\begin{aligned}\mathcal{W}^{*-1} &= \mathbb{E}[g(w_i, \theta_0) g(w_i, \theta_0)'] \\ &= \mathbb{E}([z_i \cdot u_i(\theta_0)] [z_i \cdot u_i(\theta_0)]') \\ &= \mathbb{E}(u_i^2 z_i z_i')\end{aligned}$$

最优权重矩阵

工具变量的最优权重矩阵

- 如果假设 $\mathbb{E}(u_i^2 | z_i) = \sigma^2$ 那么

$$\begin{aligned}\mathcal{W}^{*-1} &= \mathbb{E}(u_i^2 z_i z_i') \\ &= \mathbb{E}[\mathbb{E}(u_i^2 z_i z_i' | z_i)] \\ &= \sigma^2 \mathbb{E}(z_i z_i')\end{aligned}$$

实际上, σ^2 为常数, 不影响最优化, 从而可以直接选取

$$\mathcal{W}^* = [\mathbb{E}(z_i z_i')]^{-1}$$

- 而实践中可以使用

$$\hat{\mathcal{W}}^* = \left[\sum_{i=1}^N (z_i z_i') \right]^{-1}$$

最优权重矩阵

工具变量的最优权重矩阵

- 如果令

$$Z = \begin{bmatrix} z'_1 \\ z'_2 \\ \vdots \\ z'_N \end{bmatrix}_{N \times G}$$

那么最优权重矩阵可以写为 $\hat{W}^* = (Z'Z)^{-1}$ 。

- 在Stata中，如果使用了instruments()选项，默认就是使用该权重矩阵作为初始权重矩阵，从而无需额外使用winit()选项。

最优权重矩阵

工具变量的最优权重矩阵

- 但是同时也要注意，这一权重矩阵只有在同方差的假设下才成立
- 如果使用两步法或者迭代GMM，可以首先使用 $\hat{W}_0 = (Z'Z)^{-1}$ 作为初始权重矩阵得到估计 $\hat{\theta}^{(1)}$ ，然后计算

$$\hat{W}_1 = \left(\sum_{i=1}^N \left[u_i^2 \left(\hat{\theta}^{(1)} \right) z_i z_i' \right] \right)^{-1}$$

作为最优权重矩阵的估计。

最小卡方统计量

- 现在我们讨论广义矩估计的目标函数

$$\gamma(\hat{\theta}; w_i) = \left[\sum_{i=1}^N g(w_i, \hat{\theta}) \right]' \hat{W} \left[\sum_{i=1}^N g(w_i, \hat{\theta}) \right]$$

的渐近分布。

- 同样的，对于不同的权重矩阵，以上分布都是不一样的，我们只讨论使用了最优权重矩阵

$$W^* = (\mathbb{E} [g(w_i, \theta_0) g(w_i, \theta_0)'])^{-1}$$

的情况。

最小卡方统计量

- 注意到, 式(2)意味着

$$\sqrt{N} \frac{1}{N} \sum_{i=1}^N g(w_i, \theta_0) \overset{a}{\sim} \mathcal{N}(0, \mathcal{W}^{*-1})$$

从而

$$\mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \overset{a}{\sim} \mathcal{N}(0, I_G)$$

其中 I_G 为 $G \times G$ 的单位阵。

- 当然 $\mathcal{W}^{*1/2}$ 是不可观测的, 我们可以将其替换成其估计值 $\hat{\mathcal{W}}^{*1/2} = \mathcal{W}^{*1/2} + o_p(1)$, 根据Slutsky定理, 不改变渐近分布, 从而

$$\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \overset{a}{\sim} \mathcal{N}(0, I_G)$$

最小卡方统计量

根据上述结论，有

$$\begin{aligned} & \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right]' \hat{\mathcal{W}}^* \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right] \\ &= \left[\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right]' \left[\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right] \\ &\stackrel{a}{\sim} \chi^2(G) \end{aligned}$$

从而广义矩估计的目标函数在 θ_0 处应该渐近服从一个自由度为 G 的卡方分布：

$$\frac{1}{N} \mathcal{R}(\theta_0; w_i) \stackrel{a}{\sim} \chi^2(G)$$

最小卡方统计量

- 然而，以上结论仍然不是我们最终需要的。
- 注意到，该目标函数是关于 θ_0 而非 $\hat{\theta}$ 的函数，而 θ_0 是未知的，我们只能在 $\hat{\theta}$ 处计算目标函数值
- 那么 $\frac{1}{N}\Upsilon(\hat{\theta}; w_i)$ 与 $\frac{1}{N}\Upsilon(\theta_0; w_i)$ 的分布相同吗？
- 答案是否定的。
- 比如，当 $G = K$ 时，广义矩估计就变为矩估计，从而 $\Upsilon(\hat{\theta}; w_i) \equiv 0$ ，那么 $\frac{1}{N}\Upsilon(\hat{\theta}; w_i)$ 自然不会是一个 $\chi^2(K)$ 分布。

最小卡方统计量

我们可以从 $\mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0)$ 的渐近分布出发, 由于 $\hat{\theta} = \theta_0 + o_p(1)$, 从而如果将其带入, 可得

$$\begin{aligned}\mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta}) &= \mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N [g(w_i, \theta_0) + o_p(1)] \\ &= \mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) + \frac{1}{\sqrt{N}} \sum_{i=1}^N o_p(1) \\ &= \mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) + o_p(\sqrt{N})\end{aligned}$$

这里可以看到, $\mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta})$ 与 $\mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0)$ 之间并不是相差 $o_p(1)$, 那么

$$\mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta}) \overset{d}{\approx} \mathcal{N}(0, I_G)$$

从而 $\frac{1}{N} \mathcal{Y}(\hat{\theta}; w_i) \overset{d}{\approx} \chi^2(G)$ 。

最小卡方统计量

为了分析 $\frac{1}{N} \Upsilon(\hat{\theta}; w_i)$ 的渐近分布, 我们可以将 $\mathcal{W}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta})$ 进行泰勒展开:

$$\begin{aligned} & \hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta}) \\ &= \hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) + \hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\partial g(w_i, \theta_0)}{\partial \theta'} (\hat{\theta} - \theta_0) + o_p(1) \\ &= \hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) + \hat{\mathcal{W}}^{*1/2} \frac{1}{N} \sum_{i=1}^N \frac{\partial g(w_i, \theta_0)}{\partial \theta'} \sqrt{N} (\hat{\theta} - \theta_0) + o_p(1) \\ &= \hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) + \hat{\mathcal{W}}^{*1/2} \mathcal{G}_0 \sqrt{N} (\hat{\theta} - \theta_0) + o_p(1) \end{aligned}$$

注意到其中第二项, 有:

$$\sqrt{N} (\hat{\theta} - \theta_0) = -(\mathcal{G}_0' \hat{\mathcal{W}}^* \mathcal{G}_0)^{-1} \mathcal{G}_0' \hat{\mathcal{W}}^* \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right] + o_p(1)$$

从而

$$\begin{aligned}& \hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta}) \\&= \hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \\&\quad - \hat{\mathcal{W}}^{*1/2} \mathcal{G}_0 \left(\mathcal{G}_0' \hat{\mathcal{W}}^* \mathcal{G}_0 \right)^{-1} \mathcal{G}_0' \hat{\mathcal{W}}^* \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right] + o_p(1) \\&= \left[\hat{\mathcal{W}}^{*1/2} - \hat{\mathcal{W}}^{*1/2} \mathcal{G}_0 \left(\mathcal{G}_0' \hat{\mathcal{W}}^* \mathcal{G}_0 \right)^{-1} \mathcal{G}_0' \hat{\mathcal{W}}^* \right] \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right] + o_p(1) \\&= \left[I_G - \hat{\mathcal{W}}^{*1/2} \mathcal{G}_0 \left(\mathcal{G}_0' \hat{\mathcal{W}}^* \mathcal{G}_0 \right)^{-1} \mathcal{G}_0' \hat{\mathcal{W}}^{*1/2} \right] \hat{\mathcal{W}}^{*1/2} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right] + o_p(1)\end{aligned}$$

最小卡方统计量

令 $\mathcal{D} = \hat{\mathcal{W}}^{*1/2} \mathcal{G}_0$ 为 $G \times K$ 的矩阵, 那么

$$I_G - \hat{\mathcal{W}}^{*1/2} \mathcal{G}_0 \left(\mathcal{G}_0' \hat{\mathcal{W}}^* \mathcal{G}_0 \right)^{-1} \mathcal{G}_0' \hat{\mathcal{W}}^{*1/2} = I_G - \mathcal{D} (\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}'$$

其中

$$\mathcal{D} (\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}' \times \mathcal{D} (\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}' = \mathcal{D} (\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}'$$

从而 $\mathcal{D} (\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}'$ 是一个幂等矩阵, 且

$$\text{tr} \left(\mathcal{D} (\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}' \right) = \text{tr} \left((\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}' \mathcal{D} \right) = \text{tr} (I_K) = K$$

由此得出 $\text{tr} \left(I_G - \mathcal{D} (\mathcal{D}' \mathcal{D})^{-1} \mathcal{D}' \right) = G - K$ 。

最小卡方统计量

现在，使用类似的方法，我们得到

$$\begin{aligned}\frac{1}{N} \Upsilon(\hat{\theta}; w_i) &= \left[\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta}) \right]' \left[\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \hat{\theta}) \right] + o_p(1) \\ &= \left[\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right]' (I_G - \mathcal{D}(\mathcal{D}'\mathcal{D})^{-1}\mathcal{D}') \left[\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \right]\end{aligned}$$

其中 $\hat{\mathcal{W}}^{*1/2} \frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0) \overset{a}{\sim} \mathcal{N}(0, I_G)$ ，从而有

$$\frac{1}{N} \Upsilon(\hat{\theta}; w_i) \overset{a}{\sim} \chi^2(G - K)$$

即GMM的目标函数渐近服从自由度为 $G - K$ 的卡方分布 (Hanson, 1982)。

- 当 $G = K$ ，即GMM退化为矩估计的情况下， $\frac{1}{N} \Upsilon(\hat{\theta}; w_i) = 0$ ，同时其分布自由度也变为0。

最小卡方统计量

- 需要特别注意的是，只有当使用最优权重矩阵 \mathcal{W}^* 时，以上结论才是成立的！
- 而实际应用中，需要将 \mathcal{W}^* 替换为其估计

$$\hat{\mathcal{W}}^* (\hat{\theta}) = \left(\frac{1}{N} \sum_{i=1}^N \left[g(w_i, \hat{\theta}) g(w_i, \hat{\theta})' \right] \right)^{-1} = N \left(\sum_{i=1}^N \left[g(w_i, \hat{\theta}) g(w_i, \hat{\theta})' \right] \right)^{-1}$$

其中 $\hat{\theta}$ 为要给事前的GMM估计。

- 当使用最优权重矩阵估计时，GMM估计量最小化

$$\Upsilon(\theta; w_i) = \left[\sum_{i=1}^N g(w_i, \theta) \right]' \hat{\mathcal{W}}^* (\hat{\theta}) \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

最小卡方统计量

- 或者等价的, 最小化

$$\begin{aligned}
 \frac{1}{N} Y(\theta; w_i) &= \frac{1}{N} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \hat{\mathcal{W}}^* (\hat{\theta}) \left[\sum_{i=1}^N g(w_i, \theta) \right] \\
 &= \frac{1}{N} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \hat{\mathcal{W}}^* (\hat{\theta}) \left[\sum_{i=1}^N g(w_i, \theta) \right] \\
 &= \frac{1}{N} \left[\sum_{i=1}^N g(w_i, \theta) \right]' N \left(\sum_{i=1}^N \left[g(w_i, \hat{\theta}) g(w_i, \hat{\theta})' \right] \right)^{-1} \left[\sum_{i=1}^N g(w_i, \theta) \right] \\
 &= \left[\sum_{i=1}^N g(w_i, \theta) \right]' \left(\sum_{i=1}^N \left[g(w_i, \hat{\theta}) g(w_i, \hat{\theta})' \right] \right)^{-1} \left[\sum_{i=1}^N g(w_i, \theta) \right] \\
 &\triangleq \left[\sum_{i=1}^N g(w_i, \theta) \right]' \hat{\Xi} \left[\sum_{i=1}^N g(w_i, \theta) \right]
 \end{aligned}$$

最小卡方统计量

- 以上目标函数最优化后得到的估计量

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N g(w_i, \theta) \right]' \hat{\Xi} \left[\sum_{i=1}^N g(w_i, \theta) \right]$$

即使用最优权重矩阵的估计量，可以成为最优GMM估计量（**optimal GMM estimator**）

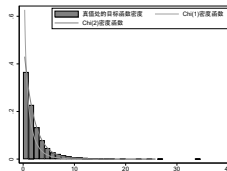
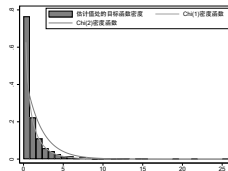
- 同时由于以上目标函数在估计值处服从一个自由度为 $G - K$ 的卡方分布：

$$\left[\sum_{i=1}^N g(w_i, \hat{\theta}) \right]' \hat{\Xi} \left[\sum_{i=1}^N g(w_i, \hat{\theta}) \right] \stackrel{a}{\sim} \chi^2(G - K)$$

从而以上估计量也被称为最小卡方估计量（**minimum chi-square estimator**）。

GMM目标函数的模拟

GMM_mini_chi2.do



- 10.2
- 10.3
- 10.5
- 10.6

