Econometrics by Prof. Ju (Continue)

> Author: `Qinzhu Sun`
> Updated: `2021/2/18`

Chapter 11: IV and GMM

11.1 IV Estimation

How to solve endogeneity?

- Proxy
- IV
- Structural Modelling
- Panel Method: differencing to erase FE

$$y_i = X'_{i_{m\times 1}}\alpha + u_i, i = 1, 2, ..., n, \quad E(u_i|X_i) \neq 0$$
$$IV : Z_{i_{l\times 1}}, \quad l \geq m$$

Two requirement of IV:

- Relevance condition: $\frac{1}{n}Z'X = \frac{1}{n}\sum_{i=1}^{n} Z_i X_i' \to E(ZX) \neq 0$
- Exogeneity condition: $\frac{1}{n}Z'u \to_a 0$

1st stage:

$$X = Z\gamma + \mu \quad \Rightarrow \quad X = Z\hat{\gamma} + \hat{u}$$

2nd stage:

$$y = X\alpha + u = \hat{X}\alpha + (X - \hat{X})\alpha + u, \quad (X - \hat{X})\alpha + u \equiv \varepsilon$$

Note that

$$\begin{gathered} \hat{X} \perp X - \hat{X}, \\ \hat{X}'u = X'Z(Z'Z)^{-1}(Z'u) \to_a 0 \Rightarrow \hat{X} \perp u, \\ \Rightarrow \hat{X} \perp \varepsilon. \end{gathered}$$

Thus, we can use the **least square method** to estimate:

$$\begin{gathered} \hat{\alpha}_{2SLS} = \hat{\alpha}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y \\ \overset{\hat{X}=P_Z X}{=} (X'P_Z X)^{-1} X' P_Z y \\ = \alpha + (X'P_Z X)^{-1} X' P_Z \varepsilon \\ = \alpha + (X'P_Z X)^{-1} X' P_Z u \\ = \alpha + (\frac{X'Z}{n} \frac{(Z'Z)^{-1}}{n} \frac{Z'X}{n})^{-1} \frac{X'Z}{n} \frac{(Z'Z)^{-1}}{n} \frac{Z'u}{n} \\ \to_a \alpha \end{gathered}$$

The last equation uses LLN (each term converges to finite matrix & $\frac{Z'u}{n} = o_p(1)$).

Assume that $Var(u|Z, X) = \Omega$ is known. We have

$$Var(\hat{\alpha}_{2SLS}|Z, X) = (X'P_Z X)^{-1} X' P_Z \Omega P_Z X (X'P_Z X)^{-1}.$$

In reality, we don't estimate Ω directly alone; instead, we estimate $Z'\Omega Z$ (lower dimension, so that less parameters are needed to estimate).

If $l = m$:

$Z'X$是方阵。

$$\begin{gathered} \hat{\alpha} = (\hat{X}'\hat{X})^{-1}\hat{X}'y \\ = (X'Z(Z'Z)Z'X)^{-1} X'(Z(Z'Z)Z')y \\ = (Z'X)^{-1}Z'y \end{gathered}$$

is SAE (sample analogue estimator). Why?

Proof:

$$
\begin{gathered}
y_i = X_i'\alpha + u_i \\
Z_i y_i = Z_i X_i'\alpha + Z_i u_i \\
\overset{E(\cdot)}{\Rightarrow} E(Z_i y_i) = E(Z_i X_i')\alpha + E(Z_i u_i) = E(Z_i X_i')\alpha \\
\Rightarrow \hat{\alpha} = (E(Z_i X_i'))^{-1} E(Z_i y_i) \\
\Rightarrow \hat{\alpha}_{SAE} = (\frac{1}{n}\sum_{i=1}^{n} Z_i X_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} Z_i y_i') \\
= (Z'X)^{-1}Z'y
\end{gathered}
$$

If $l > m$:

$$Z'y = Z'X\alpha + Z'u$$

要求$\hat{\alpha}$，左乘$m \times l$阶矩阵

$$
\begin{gathered}
\Rightarrow \frac{X'Z}{n}\frac{(Z'Z)^{-1}}{n}\frac{Z'y}{n} = \frac{X'Z}{n}\frac{(Z'Z)^{-1}}{n}\frac{Z'X\alpha}{n} + \frac{X'Z}{n}\frac{(Z'Z)^{-1}}{n}\frac{Z'u}{n} \\
\to \frac{X'Z}{n}\frac{(Z'Z)^{-1}}{n}\frac{Z'X\alpha}{n} \\
\Rightarrow \hat{\alpha}_{SAE} = (X'P_Z X)^{-1}X'P_Z y
\end{gathered}
$$

但是左乘$m \times l$阶矩阵不是唯一的，怎么找到最好的取法?

$$X_{n\times l} \to X_{n\times l}T_{l\times m},$$

$T_{l\times m}$ is called **"selection matrix"**.

In 2SLS,

$$T = (Z'Z)^{-1}Z'X.$$

But what is the **most efficient** one (consistent & the least variance)? Let's find it!

$$Z'y = Z'X\alpha + Z'u$$

引入 selection matrix T, 转化成 square matrix 求逆:

$$T'Z'y = T'Z'X\alpha + T'Z'u$$

$$\Rightarrow \quad \hat{\alpha} = (T'Z'X)^{-1}T'Z'y = \alpha + (T'Z'X)^{-1}T'Z'u$$

$$\Rightarrow Var(\hat{\alpha}) = (T'Z'X)^{-1} \cdot T'Z'\Omega ZT \cdot (X'ZT)^{-1}$$

夹心估计量的优化技巧：当$T'Z'X$，$T'Z'\Omega ZT$，$(X'ZT)^{-1}$三者相等时，方差最小。
凑得

$$T^* = (Z'\Omega Z)^{-1} Z'X$$

is the most efficient.

$$Var(\hat{\alpha}) = (X'Z(Z'\Omega Z)^{-1}Z'X)^{-1}$$

In order to prove it, we introduce below

Theorem 11.1

A and B are p.s.d.. Then (using eigen decomposition)

$$A \geq B \quad \Leftrightarrow \quad A^{-1} \leq B^{-1}.$$

Thus, we only need to prove

$$X'Z(Z'\Omega Z)^{-1}Z'X \geq Var(\alpha)$$

which is true. (We don't present the proof here.)

11.2 MME

$$y = X'\beta + u_i$$

Moment condition: $E(u_i|X_i) = 0$ 的约束性强于 $E(X_iu_i) = 0$

Method of moment solves an equation system where # of restrictions = # of unknowns.

Example 1:

$$y_i = g(z_i, \theta_{m\times 1}) + u_i, \quad E(u_i|z_i) = 0$$

引入IV w_i:

$$E(w_iu_i) = E(w_i(y_ii - g(z_i, \theta))) = 0$$
$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - g(z_i, \theta)) = 0 \Rightarrow \hat{\theta}_{MME}$$

- 需要注意的是，用MME方法估计IV时，工具变量个数与内生变量个数只能相同，无法解决过度识别的情形(会无解)。

Example 2:

MLE is also MME. Suppose we have $\{(x_i, y_i)\}_{i=1,\dots,n}$ $i.i.d.$

$$\mathcal{L} = \sum_{i=1}^{n} \ln f(y_i | X_i, \theta)$$

F.O.C:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{n} g_i(\theta) = 0 \quad \Rightarrow \quad \hat{\theta}_{MLE},$$

where $g_i(\theta)$ is the score function.

Alternatively,

$$Eg_i(\theta) = 0 \Rightarrow \sum_{i=1}^{n} g_i(\theta) = 0 \quad \Rightarrow \quad \hat{\theta}_{MME}.$$

11.3 GMM

Suppose there are l population moments.

$$\mu_t(\theta_0) = Em(z_t, \theta_0) = E \begin{pmatrix} m_1(z_t, \theta_0) \\ m_2(z_t, \theta_0) \\ \vdots \\ m_l(z_t, \theta_0) \end{pmatrix} = 0$$

$l \geq m$ identification conditions are needed. (identifiable means "unique $\hat{\theta}_0$")

We construct sample moment conditions:

$$\bar{m}(\theta) = \frac{1}{T} \sum_{i=1}^{T} m(z_t, \theta_0) \xrightarrow{p} 0.$$

直接的想法：让 $\bar{m} = 0$. 但问题在于：$\bar{m} \to 0$ 是$T \to \infty$ 时的结果(此时$l \geq m$ 也无妨)。但是在finite sample 下，$\bar{m} = 0$不一定是对的，$l \geq m$ 会导致无解。所以不能用MME ($l = m$ 时MME仍然适用)。

GMM的方法具有一般性，可以囊括上述所有情形，也可以解决过度识别的问题。GMM一开始设计时针对时间序列数据(故下标为t)，后来也用在了截面数据上。

> E.g. in IV cases: (Z is IV)
>
> $$y = X\alpha + u.$$

> $$\hat{\alpha}_{2SLS} = (X'P_Z X)^{-1}X'P_Z y$$
>
> is equivalent to
>
> $$\min_\alpha (y - X\alpha)'P_Z(y - X\alpha) \quad \Leftrightarrow$$
> $$\min_\alpha \sqrt{T}\left[\frac{Z'(y-X\alpha)}{T}\right]'\left(\frac{Z'Z}{T}\right)^{-1}\sqrt{T}\left[\frac{Z'(y-X\alpha)}{T}\right]$$
>
> Here, $W = (Z'Z)^{-1}$ is inefficient. The efficient W is
>
> $$W^* = (Z'\Omega Z)^{-1}$$

Go back to GMM

$$\min_\theta J(\theta) = \left(\sqrt{T}\bar{m}(\theta)'\right)W\left(\sqrt{T}\bar{m}(\theta)\right)$$

- write as $\sqrt{T}\bar{m}$ so that we can use CLT.
- At the limit, this is equivalent to MME; under finite sample condition, this is calculable.

For any W, we have consistent estimates

$$\hat{\theta}_{GMM} \xrightarrow{p} \theta.$$

So the problem now is how to choose the **most efficient estimator**.

Define

$$D(\theta) \triangleq \frac{\partial J}{\partial \theta'}_{l\times m} = \begin{pmatrix} \frac{\partial \bar{m}_1(\theta)}{\partial \theta_1} & \cdots & \frac{\partial \bar{m}_1(\theta)}{\partial \theta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial \bar{m}_l(\theta)}{\partial \theta_1} & \cdots & \frac{\partial \bar{m}_l(\theta)}{\partial \theta_m} \end{pmatrix}_{l\times m}$$

F.O.C.

$$\frac{\partial J}{\partial \theta} = 2TD'(\hat{\theta})W\bar{m}(\hat{\theta}) = 0$$

Consider Taylor equation and plug in

$$\bar{m}(\hat{\theta}) = \bar{m}(\theta) + D(\theta_1)(\hat{\theta} - \theta),$$

we have

$$0 = D'(\hat{\theta})W\bar{m}(\theta) + D'WD(\theta_1)(\hat{\theta} - \theta)$$

$$
\hat{\theta}-\theta=-[D'(\hat{\theta})WD(\theta_1)]^{-1}D'(\hat{\theta})W\bar{m}(\theta)
$$

$$
\begin{aligned}
\sqrt{T}(\hat{\theta}-\theta)&=-[D'(\hat{\theta})WD(\theta_1)]^{-1}D'(\hat{\theta})W[\sqrt{T}\bar{m}(\theta)]\\
&\triangleq Q\sqrt{T}\bar{m}(\theta),
\end{aligned}
$$

where

$$
Q\triangleq -[D'(\hat{\theta})WD(\theta_1)]^{-1}D'(\hat{\theta})W
$$

According to CLT, we have

$$
\sqrt{T}\bar{m}(\theta)\underset{d}{\rightarrow}N(0,S_0).
$$

$$
\sqrt{T}(\hat{\theta}-\theta)\underset{d}{\rightarrow}N(0,\operatorname{plim}Q\cdot S_0\cdot(\operatorname{plim}Q)')\triangleq N(0,V)
$$

- Under finite sample condition, Q and W may be random. So we use the plim instead.

$$
\operatorname{plim}Q\triangleq -(D_0'W_0D_0)^{-1}D_0'W_0
$$

We now want to optimize

$$
\min_{W_0}V=(D_0'W_0D_0)^{-1}(D_0'W_0S_0W_0D_0)(D_0'W_0D_0)^{-1}.
$$

The trick of sandwich form:

$$
\begin{aligned}
D_0'W_0^*D_0&=D_0'W_0^*S_0W_0^*D_0\\
\Rightarrow W_0^*&=S_0^{-1}
\end{aligned}
$$

We can show that $\forall$ p.d. symmetric W_0

$$
V^*=(D_0'S_0^{-1}D_0)'\leq V(W_0)
$$

Proof:

$$
\begin{aligned}
&(V^*)^{-1}-V^{-1}\\
=&D_0'S_0^{-1}D_0-D_0'W_0D_0(D_0'W_0S_0W_0D_0)^{-1}D_0'W_0D_0\\
=&D_0'S_0^{-1/2}[I-P_{S_0^{1/2}W_0D_0}]S_0^{-1/2}D_0\geq 0
\end{aligned}
$$

However, we do not know S_0, and we need to first estiamte it!

Feasible GMM: Two-step Estimation

> 1. Get a consistent estimate of θ

> $$\hat{\theta} = \arg\min J(\theta) = \left(\sqrt{T}\bar{m}(\theta)'\right) W \left(\sqrt{T}\bar{m}(\theta)\right)$$
>
> Arbitrarily choose a W (E.g. $W = I$) to get $\hat{\theta}$.
>
> $$\left\{m(z_t, \hat{\theta})\right\}_{t=1}^{T} \quad \Rightarrow \quad \hat{S}_0 = Cov(\sqrt{T}\bar{m}(\hat{\theta}))$$
>
> 2. Plug in $\hat{S}_0$ to get $\hat{\theta}_{GMM}$
>
> $$\min_{\theta} J(\theta) = \left(\sqrt{T}\bar{m}(\theta)'\right) \left[\hat{S}_0\right]^{-1} \left(\sqrt{T}\bar{m}(\theta)\right)$$
> $$\Rightarrow \quad \hat{\theta}_{GMM}$$
>
> - We can even go back to step 1 and use the $\hat{\theta}_{GMM}$ to update $\hat{S}_0$ -- do iteration. But usually no need to do so.

The essence of GMM is to do optimization instead of solving the equations.

GMM estimators of a linear model

See the lecture notes given by Prof. Ju.

- How to estimate feasibly? To estimate $S_0 = \text{plim}\, \frac{1}{T} X'\Omega X$, we estimate the whole $X'\Omega X$, instead of estimate Ω alone.

Feasible Efficient GMM estimator

$$\sqrt{T}\bar{m}(\theta_0) \xrightarrow{d} N(0, S_0)$$

Let $Q_t = m(z_t, \theta_0)$. $\bar{Q} \triangleq \frac{1}{T}\sum_{t=1}^{T} Q_t = \bar{m}(z_t, \theta_0)$.

Denote $\Gamma_j = E(Q_t Q_{t-j}'), \Gamma_{-j} = E(Q_{t-j} Q_t') = \Gamma_j'$.

$$S_0 = Cov\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Q_t\right) = \sum_{j=-(T-1)}^{T-1} \left(1 - \frac{|j|}{T}\right)\Gamma_j$$

If it is cross-section data, $\Gamma_j = 0 \ (j \neq 0)$.

$$S_0 = Cov\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Q_t\right) = \Gamma_0$$
$$\Rightarrow \quad \hat{S}_0 = \hat{\Gamma}_0$$

But for time series data, define two indicators:

$$\hat{S}_{HW} = \hat{\Gamma}_0 + \sum_{j=1}^{q} \left(\hat{\Gamma}_j + \hat{\Gamma}_{-j}\right) = \hat{\Gamma}_0 + \sum_{j=1}^{q} \left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right)$$

- Idea: The larger $|j|$, the weaker relavance, so that we can omit it. And, if $T \gg q$, we don't need to time $\left(1 - \frac{|j|}{T}\right)$.

$$\hat{S}_{NW} = \hat{\Gamma}_0 + \sum_{j=1}^{q} \left(1 - \frac{j}{q+1}\right) \left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right)$$

- The larger $|j|$, the smaller size of obs can be used to estimate, and thus the weaker precision. So assign it a smaller weight.
- How to choose the optimal q ? Since the final goal is to minimize variance of $\hat{\theta}$,

$$q = \arg\min Var(\hat{\theta})$$

Test of Overidentifying Moment Restrictions

$$\mu(\theta_0) = E(m(z_t, \theta_0)) = 0$$

$$H_0 : \mu(\theta_0) = 0, \qquad H_1 : \mu(\theta_0) \neq 0$$

If we reject H_0, it means that at least 1 restriction is denied.

$$\sqrt{T}\left(\bar{m}(\theta_0) - \mu(\theta_0)\right) \xrightarrow{d} N(0, S_0).$$

Under H_0, we have

$$\sqrt{T}\bar{m}(\theta_0)' S_0^{-1} \sqrt{T}\bar{m}(\theta_0) \xrightarrow{d} \chi^2(l).$$

But we don't know S_0 and θ_0.

> **Hansen's J-test (Hansen, 1982)**
>
> $$\sqrt{T}\bar{m}(\hat{\theta})' \hat{S}_0^{-1} \sqrt{T}\bar{m}(\hat{\theta}) \xrightarrow{d} \chi^2(l-k)$$
>
> where k is # of parameters.

Chapter 12: Simultaneous Equations Model

$$\begin{pmatrix} y_{t1}, & y_{t2} \end{pmatrix} \begin{pmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{pmatrix} = \begin{pmatrix} x_{t1}, & x_{t2} \end{pmatrix} \begin{pmatrix} \beta_{11} & \beta_{12} \\ 0 & \beta_{22} \end{pmatrix} + \begin{pmatrix} u_{t1}, & u_{t2} \end{pmatrix}$$

Generally speaking,

$$Y_{t_{1\times G}}\Gamma_{G\times G} = X_{t_{1\times K}}B_{K\times G} + U_{t_{1\times G}}$$

$$\Rightarrow (Y_t, X_t)(\Gamma_{G\times G}, -B_{K\times K})' = U_{t_{1\times G}}$$

Define

$$A_{(G+K)\times G} \equiv (\Gamma_{G\times G}, -B_{K\times K})'.$$

$$\Rightarrow (Y_t, X_t)A_{(G+K)\times G} = U_{t_{1\times G}}$$

Assumptions:

- Γ is invertible $\Rightarrow$ complete system
- $EU_t = 0, \quad Cov(U_t') = E(U_t'U_t) = \Sigma_{G\times G}$. U_t is structural residual, not reduced-form residual.

Reduced-form:

$$Y_{t_{1\times G}}\Gamma_{G\times G} = X_{t_{1\times K}}B_{K\times G} + U_{t_{1\times G}}$$
$$\Rightarrow$$
$$Y_t = X_tB\Gamma^{-1} + U_t\Gamma^{-1} \triangleq X_t\Pi + V_t$$

And we can run least square estimation.

$$EV_t = 0, \quad \Omega \equiv Cov(V_t') = Cov((V_t\Gamma^{-1})') = (\Gamma^{-1})'\Sigma\Gamma^{-1}$$

12.1 Identification

The question is: whether we can recover the structural coefficients from reduced-form estimators.$\Rightarrow$ **Identification**!

In reduced-form, we estimate $K\times G$ parameters; in structural form, there are $(G^2-G)+K\times G$ parameters. We must **impose restrictions** so as to recover.

Write $A=(\alpha_1, A_2)$. We identify one column (corresponding to one equation) by one column.

Impose restriction $R\alpha_1 = 0$.

Theorem 12.1

α_1 is identifiable iff $rank(R_{J\times(G+K)}A) = G-1$.

Corollary 12.1 (Order Condition)

$$J \geq G - 1$$

Corollary 12.2 (Exclusion Condition)

of excluded exogenous variables $\geq$ # of included endogenous variables.

> **An example:**
>
> # of IV $\geq$ # of endogenous variables.

12.2 Limited Information Estimation of a Single Equation

Semi-structural model:

$$\begin{cases} y & = Y_1\gamma_1 + X_1\beta_1 + u_1 \equiv Z_1\delta_1 + u_1 & (1) \\ Y_1 & = X_1\Pi_{12} + X_2\Pi_{22} + V_1 \equiv X\Pi_2 + V_1 & (2) \end{cases}$$

We don't care about (2), so we just write in reduced-form.

For equation (1),

$$y = (Y_1, X_1)\begin{pmatrix}\gamma_1 \\ \beta_1\end{pmatrix} + u_1 \equiv Z_1\delta_1 + u_1.$$

Let $X = (X_1, X_2)$.

Define

> $$A_1 = I - P_{X_1}$$
>
> $$A_3 = I - P_X$$
>
> $$A_2 = A_1 - A_3 = P_X - P_{X_1} = P_{(I-P_{X_1})X_2}$$

12.2.1 2SLS

- Step 1:

$$\hat{Z}_1 = P_X Z_1$$

- Step 2:

$$y = \hat{Z}_1\delta_1 + u_1 + (Z_1 - \hat{Z}_1)\delta_1$$

$$\Rightarrow \hat{\delta}_1 = (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1 y = \begin{pmatrix}\hat{\gamma}_1 \\ \hat{\beta}_1\end{pmatrix}$$

$$\Rightarrow \begin{cases} \hat{\gamma}_1 = (Y_1' A_2 Y)^{-1} Y_1' A_2 y \\ \hat{\beta}_1 = (X_1' X_1)^{-1} X_1' (y - Y_1 \hat{\gamma}_1) \end{cases}$$

12.2.2 GLS estimator on a transfered equation

$$\begin{aligned} \hat{\delta}_{GLS} &= \arg\min (X'u_1)'(X'X)^{-1} X'u_1 \\ &= (y - Z_1\delta_1)' X (X'X)^{-1} X' (y - Z_1\delta_1) \end{aligned}$$

The idea of GLS: transfer error term into **spheric** error term.

It can be found that

$$\hat{\delta}_{GLS} = \hat{\delta}_{2SLS}.$$

12.2.3 Test of Independence: Hausman Test

$$\begin{cases} y &= Y_1\gamma_1 + X_1\beta_1 + u_1 \equiv Z_1\delta_1 + u_1 \\ Y_1 &= X_1\Pi_{12} + X_2\Pi_{22} + V_1 \equiv X\Pi_2 + V_1 \end{cases}$$

$$H_0 : Cov(Y, u_1) = 0, \quad H_1 : Cov(Y, u_1) \neq 0$$

Hausman Test:

$$H_0 : OLS \rightarrow truth; \quad H_1 : OLS \overset{\times}{\rightarrow} truth$$

What we do is to see the distance between $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$.

$$\begin{aligned} TH = \frac{(\hat{\gamma}_{12} - \hat{\gamma}_{11})' \left[(Y_1' A_2 Y_1)^{-1} - (Y_1' A_1 Y_1)^{-1} \right]^{-1} (\hat{\gamma}_{12} - \hat{\gamma}_{11})}{\hat{\sigma}_{11}} \\ \sim \chi^2(G_1) \end{aligned}$$

12.3 Full Information Estimation of a Single Equation

每个方程的 identification information 都要用到。

3SLS的实质是用了两个GLS。第一个GLS其实就是IV estimation，在IV后加了一个GLS，所以称作3SLS。

$$\begin{aligned} \hat{\delta}_{3SLS} &= [Z^{*\prime}(\Sigma \otimes I_K)^{-1} Z^*]^{-1} Z^{*\prime}(\Sigma \otimes I_K)^{-1} y^* \\ &= \ldots \\ &= \delta + \left[Z'\Sigma^{-1} \otimes X(X'X)^{-1}X'Z \right]^{-1} Z'(\Sigma^{-1} \otimes X(X'X)^{-1}X')u \end{aligned}$$

$$\Rightarrow Cov(\hat{\delta}_{3SLS}) = \left[Z'(\Sigma^{-1} \otimes P_X) Z \right]^{-1}$$

$\Sigma = Cov(V_t')$ can be estimated using 2SLS residuals.

Chapter 16: Limited Dependent Variable Models

> *All the models in this chapter can be estimated using MLE.*

$$y_t^* = X_t\beta + u_t, \quad E(u_t|X_t) = 0$$

We observe $y_t = T(y_t^*)$. Different kinds of $T(\cdot)$ yields different models:

- Binary response model
- Multiple choice model
  - Ordered qualitative model
  - Unordered qualitative model
- Censored model
- Truncated model

> E.g. BRM
>
> $$y_t = \begin{cases} 1, & y_t^* = X_t\beta + u_t > 0 \\ 0, & y_t^* = X_t\beta + u_t \neq 0 \end{cases}$$
>
> If we misspecify
>
> $$y_t = X_t\beta + \varepsilon_t$$
>
> the coefficient β will be **biased**. Specifically, note that
>
> $$E(\varepsilon_t|X_t) = 0 \Leftrightarrow E(y_t|X_t) = X_t\beta.$$
>
> In LPM, we have
>
> $$E(y_t|X_t) = P(y_t = 1|X_t) = F(X_t\beta) \neq X_t\beta,$$
>
> which means
>
> $$E(\varepsilon_t|X_t) \neq 0. \quad \Rightarrow \quad biased\ results!$$

16.1 Unordered Qualitative Response Models

$$u_{ij} = X_i\beta_j + Z_{ij}\gamma + \varepsilon_{ij}, \quad i = 1, ..., N, j, 1, ..., J.$$

Define $m_{ij} \equiv X_i\beta_j + Z_{ij}\gamma$. So

$$U_{ij} = m_{ij} + \varepsilon_{ij}.$$

Set $\varepsilon_{ij} \sim$ **standard extreme value distribution (Gumbell distribution)**. CDF $F(z) = \exp(-\exp(-z))$, PDF $f(z) = \exp(-z)F(z)$

$$\Rightarrow \quad P_{ij} = \ldots = \frac{e^{m_{ij}}}{\sum_{k=1}^{J} e^{m_{ik}}}.$$

16.1.1 Interpretation of Multinomial Logit Model

Impose restriction $\beta_1 = 0$, and we have the log odds ratio

$$\ln(\frac{P_{ij}}{P_{i1}}) = X_i \beta_j.$$

So β_j can be interpreted as the marginal effect of X_i on choice j 's log odds ratio.

16.2 Tobit Model

16.2.1 Type I Tobit

$$y_t^* = X_t \beta + u_t, \quad u_t | X_t \sim N(0, \sigma^2)$$

$$y_t = \begin{cases} y_t^*, & y_t^* > 0 \\ 0, & y_t^* \leq 0 \end{cases}$$

If we neglect this problem and do least square estimation, there will be sample selection.

$$\begin{aligned} E(y_t | y_t > 0, X_t) &= E(y_t^* | y_t^* > 0, X_t) \\ &= X_t \beta + \sigma \frac{\phi(X_t \beta / \sigma)}{\Phi(X_t \beta / \sigma)}. \end{aligned}$$

Heckman suggests that we can directly estimate

$$y_t = X_t \beta + \sigma \frac{\phi(X_t \beta / \sigma)}{\Phi(X_t \beta / \sigma)} + \varepsilon_t.$$

$$E(\varepsilon_t | y_t > 0, X_t) = 0 \Rightarrow \textit{no endogeneity}!$$

How to estimate $\frac{\phi(X_t \beta / \sigma)}{\Phi(X_t \beta / \sigma)}$? **Use Probit to get β / σ !**

16.2.2 Type II Tobit (Incidental Truncation Model/ Self-Selection Model)

$$
\begin{aligned}
y_1^* &= X_1\beta_1 + u_1,\\
y_2^* &= X_1\beta_2 + u_2.
\end{aligned}
$$

y_1^* is working hours; while y_2^* is desired/reservation wage. That is,

$$
y_2 = \begin{cases} 1, & y_2^* > 0 \\ 0, & y_2^* \le 0 \end{cases}
$$

$$
y_1 = \begin{cases} y_1^* \quad , & y_2 = 1, \\ 0 \quad , & y_2 = 0. \end{cases}
$$

Assume

$$
\begin{pmatrix} u_1|X_1, X_2 \\ u_2|X_1, X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right]
$$

Estimate directly creates bias, so we need adjustion (调整项).

$$
E(y_1^*|y_2^*, X_1, X_2) = X_1\beta_1 + \sigma_{12}\sigma_{22}^{-1}(y_2^* - X_2\beta_2).
$$

> 对该公式的证明需回顾 joint distribution 部分章节内容。
> X_1, X_2 are normal $\Rightarrow$ conditional distribution $X_1|X_2$ is also normal.

$$
y_1^* = X_1\beta_1 + \sigma_{12}\sigma_{22}^{-1}(y_2^* - X_2\beta_2) + \eta_1
$$

$\Rightarrow$

$$
\begin{gathered}
E(\eta_1|y_2^*, X_1, X_2) = 0,\\
Var(\eta_1|y_2^*, X_1, X_2) = Var(y_1^*|y_2^*, X_1, X_2) = \sigma_{11} - \sigma_{12}\sigma_{22}^{-1}\sigma_{21}
\end{gathered}
$$

$$
E(y_1^*|y_2^* > 0, X_1, X_2) = \ldots = X_1\beta_1 + (\sigma_{12}/\sigma_2)\lambda_2
$$

where

$$
\lambda_2 = \frac{\phi(X_2\beta_2/\sigma_2)}{\Phi(X_2\beta_2/\sigma_2)}.
$$

We can run OLS with observations $y = y_1^*(y_2^* > 0)$

$$
y_1^* = X_1\beta_1 + \frac{\sigma_{12}}{\sigma_2}\lambda_2 + \xi_1
$$

- Considering heteroskedasticity, GLS is more efficient!

How to estimate first λ_2? **1st stage Probit!**

> Although we can estimate using pure MLE, Heckman two-step method is still more popular.

Chapter 15: Panel Data

15.1 RE

RE is much easier because there is no endogeneity problem, so OLS estimte is consistent. Considering the heteroskedasticity, GLS is more efficient.

> Assumption RE1:
>
> - Strict exogeneity assumption
>
> $$E(u_{it}|x_i, c_i) = 0, \forall t \in \{1, 2, ..., T\}$$
>
> - Orthogonality
>
> $$E(c_i|x_i) = 0$$

$$y_i = x_i\beta + c_i j_T + u_i \equiv x_i\beta + v_i, \quad j_T = (1, ..., 1)'_{T\times 1}$$

$$\Omega = E(v_i v_i'|x_i)$$

> Assumption RE2:
>
> - Homoskedasticity (this assumption is unimportant; we can estimate under heteroskedasticity)
>
> $$E(u_i u_i'|x_i, c_i) = \sigma_u^2 I, \quad E(c_i^2|x_i) = \sigma_c^2$$

$$E(v_i v_i'|x_i) = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \cdots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \cdots & \sigma_c^2 + \sigma_u^2 \end{pmatrix} \equiv \Omega = \sigma_u^2 I_T + \sigma_c^2 j_T j_T'$$

Error term is not spheric. So OLS is consistent, but not efficient. Consider GLS!

$$\hat{\beta}_{RE} = \beta + \left[\sum_{i=1}^n x_i'\Omega^{-1}x_i\right]^{-1}\left[\sum_{i=1}^n x_i'\Omega^{-1}v_i\right]$$

> Assumption RE3:
>
> - $E(x_i'\Omega^{-1}x_i)$ has a full rank.

$$\sqrt{N}(\hat{\beta}_{RE}-\beta)=\left[\frac{1}{n}\sum_{i=1}^{n}x_i'\Omega^{-1}x_i\right]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i'\Omega^{-1}v_i\right]$$
$$\xrightarrow{d} N\left(0, E(x_i'\Omega^{-1}x_i)^{-1}\right)$$

15.1.1 Feasible GLS

But we need to first estimate σ_u^2, σ_c^2 to determine Ω.

Idea: 先利用OLS回归找到consistent估计量 $\hat{\sigma}_v^2, \hat{\sigma}_u^2, \hat{\sigma}_c^2$，再回过头去用GLS重新估一遍。

$$\hat{\sigma}_v^2=\frac{1}{NT-K}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2\to\sigma_v^2$$

$$\hat{\sigma}_c^2=\frac{1}{NT(T-1)/2-K}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{v}_{it}\hat{v}_{is}\to\sigma_c^2$$

$$\hat{\sigma}_u^2=\hat{\sigma}_v^2-\hat{\sigma}_c^2$$

15.1.2 Robust Variance matrix estimator

If assumption RE2 does not hold,

$$\hat{\Omega}=\frac{1}{N}\sum_{i=1}^{N}\hat{v}_i\hat{v}_i'.$$

$$\hat{\beta}_{RE}=\beta+\left[\sum_{i=1}^{n}x_i'\hat{\Omega}^{-1}x_i\right]^{-1}\left[\sum_{i=1}^{n}x_i'\hat{\Omega}^{-1}v_i\right].$$

15.2 FE

$$y_{it}=x_{it}'\beta+c_i+u_{it}$$

> Assumption FE1:
>
> $$E(u_{it}|c_i,x_i)=0,\quad E(c_i|x_i)\neq 0$$

15.2.1 Time-Demeaning

$$y_{it}-\bar{y}_i=(x_{it}-\bar{x}_i)\beta+(u_{it}-\bar{u}_i)$$

Denote as

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{u}_{it}.$$

$$\Rightarrow$$

$$\begin{aligned}
\hat{\beta}_{FE} &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}\ddot{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}\ddot{y}_{it}\right) \\
&= \beta + \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}\ddot{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}\ddot{u}_{it}\right) \\
&= \beta + \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}\ddot{x}_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}u_{it}\right)
\end{aligned}$$

> Assumption FE2:
>
> Homoskedasticity:
>
> $$E(u_i u_i'|x_i, c_i) = \sigma_u^2 I$$

$$\sqrt{N}(\hat{\beta}_{FE} - \beta) \xrightarrow{d} N\left(0, \sigma_u^2\left(\frac{1}{N}\sum_{i=1}^{N}\ddot{x}_i'\ddot{x}_i\right)^{-1}\right)$$

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)-K}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\ddot{u}}_{it}^2$$

is consistent and unbiased.

15.2.2 First Differencing

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}$$

> Assumption FD1:
>
> $$E(\Delta x_{it}'\Delta u_{it} = 0), t = 2, 3, ..., T.$$
>
> - 只需要考虑相邻期的相关性，比time-demeaning 的外生性要求低。
>
> Assumption FD2:
>
> $$rank\sum_{t=1}^{T}E(\Delta x_{it}'\Delta x_{it}) \quad is\ full$$

OLS is consistent, but not efficient.

$$\hat{\beta}_{FD} = \beta + \left[\sum_{i=1}^{N} \Delta x_i' \Delta x_i\right]^{-1} \left[\sum_{i=1}^{N} \Delta x_i' \Delta u_i\right]$$

$$\sqrt{N}(\hat{\beta}_{FD} - \beta) \xrightarrow{d} N\left(0, V\right)$$

$$\hat{V} = \left(\frac{1}{N}\sum_{i=1}^{N} \Delta x_i' \Delta x_i\right)^{-1} \cdot \frac{1}{N}\sum_{i=1}^{N} \Delta x_i' \Delta u_i \Delta u_i' \Delta x_i \left(\frac{1}{N}\sum_{i=1}^{N} \Delta x_i' \Delta x_i\right)^{-1}$$

15.3 Large N, Large T

Consider Panel Data Models with Interactive Fixed Effects

$$y_{it} = x_{it}\beta + \lambda_i' F_t + \varepsilon_{it}.$$

F_t is called common factor.

$$E(\varepsilon_{it}|x_{it}, \lambda_i, F_t) = 0$$

We care about β. We need large N large T panel to solve λ_i, F_t.

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, x_i = \begin{pmatrix} x_{i1}' \\ \vdots \\ x_{iT}' \end{pmatrix}, F = \begin{pmatrix} F_1' \\ \vdots \\ F_T' \end{pmatrix}, \varepsilon_i = \begin{pmatrix} \varepsilon_{i1}' \\ \vdots \\ \varepsilon_{iT}' \end{pmatrix}, \Lambda = \begin{pmatrix} \Lambda_1' \\ \vdots \\ \Lambda_N' \end{pmatrix}_{N\times r}.$$

$$\Rightarrow y_i = x_i\beta + F\lambda_i + \varepsilon_i$$

15.3.1 Identification

Impose r^2 restrictions

$$\begin{cases} F'F/r = I_r \\ \Lambda'\Lambda \ is \ diagonal. \end{cases}$$

15.3.2 Estimation

$$SSR(\beta, F, \Lambda) = \sum_{i=1}^{N} (y_i - x_i\beta - F\lambda_i)'(y_i - x_i\beta - F\lambda_i)$$

Given F,

$$\hat{\beta}(F) = \left(\sum_{i=1}^{N} x_i' M_F x_i\right)^{-1} \left(\sum_{i=1}^{N} x_i' M_F y_i\right),$$

where $M_F = I - F(F'F)^{-1}F' = I - FF'/T$.

Given β, let $W_i = y_i - x_i\beta$ be the new dependent variable.

$$\Rightarrow W_i = F\lambda_i + \varepsilon_i, \quad or \quad W = F\Lambda + \varepsilon$$

$$\min SSR(F,\Lambda) = tr\left[\left(W - F\Lambda'\right)\left(W - F\Lambda'\right)'\right] \qquad (1)$$

Given F,

$$\Lambda' = (F'F)^{-1}F'W \qquad (2)$$

Plug (2) in (1), we have the objective function

$$\begin{aligned} \min SSR(F,\Lambda) &= tr\left[\left(M_F W\right)\left(M_F W\right)'\right] \\ &= tr\left[\left(M_F W\right)'\left(M_F W\right)\right] \\ &= tr\left[W' M_F W\right] \\ &= tr\left(W'W\right) - tr\left(W' \frac{FF'}{T} W\right) \\ &= tr\left(W'W\right) - tr\left(F'WW'F\right)/T \end{aligned}$$

Given β, $W_i = y_i - x_i\beta$ is also given. So

$$\begin{gathered} \min SSR(F,\Lambda) \Leftrightarrow \\ \max_F tr\left(F'WW'F\right), \quad s.t. F'F/T = I_r \end{gathered}$$

The latter is standard PCA process.

Generally, the idea is that: given F solves β; given β solves F, which solves β -- iteration.

> **Algorithm I: (Bai, 2009, ECMA)**
> Arbitrarily choose β_0
>
> $$\beta_0 \Rightarrow F_0 \Rightarrow \beta_1 \Rightarrow F_1 \ldots$$
>
> until convergence. Then we can work out
>
> $$\hat{\lambda}_i = \frac{1}{T}\hat{F}'(y_i - x_i\hat{\beta}_i)$$
>
> **Algorithm II: (Su and Ju, 2016, JOE)**

> $$\hat{\beta} = \min \frac{1}{T} \sum_{k=r+1}^{T} \mu_k \left[\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i\beta)(y_i - x_i\beta)' \right],$$
>
> where μ_k is the k-th largest eigen values by counting eigen values multiple times.
>
> Given $\hat{\beta}$, we can solve F by computin the eigen vectors of
>
> $$\left[\frac{1}{NT} \sum_{i=1}^{N} (y_i - x_i\beta)(y_i - x_i\beta)' \right]$$
>
> which corresponds to the r-largest eigen values of the matrix.
> Given $\hat{\beta}$ and $\hat{F}$,
>
> $$\hat{\lambda}_i = \frac{1}{T} \hat{F}'(y_i - x_i\hat{\beta}_i)$$

Chapter 17: Nonparametric Density Estimation

17.1 Frequency Estimator

17.1.1 Estimation of $F(x) = P(x_i \leq x)$

Define

$$z_i = 1(x_i \leq x) = \begin{cases} 1, & x_i \leq x \\ 0, & x_i > x \end{cases}$$

$$F(x) = P(x_i \leq x) = P(z_i = 1)$$

$$\Rightarrow \hat{F}(x) = \frac{\sum_{i=1}^{n}}{n} \xrightarrow{p} F(x)$$

$$E(\hat{F}(x)) = F(x), \quad Var(\hat{F}(x)) = \frac{F(x)[1 - F(x)]}{n}$$

17.1.2 Estimation of $f(x)$

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

$$\Rightarrow \hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}$$

$$= \frac{\# \, of \, x_i \in (x-h, x+h]}{2nh}$$

17.2 Kernel Estimator

Requirement of Kernel Function $k(v)$:

- $$k(\cdot) \geq 0$$
- $$k(\cdot) \leq M$$
- $$\int k(v)dx = 1$$
- $$k(v) = k(-v)$$
- $$\int v^2 k(v)dv = \kappa_2 > 0$$

Theorem 17.1

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)$$

is consistent, and

$$MSE[\hat{f}(x)] = \frac{h^4}{4}\left[\kappa_2 f''(x)\right]^2 + \frac{\kappa f(x)}{nh} + o(h^4 + \frac{1}{nh})$$

$$F.O.C. \Rightarrow h_{opt} = \left[\frac{\kappa f(x)}{[\kappa_2 f''(x)]^2}\right]^{1/5} n^{-1/5}$$

h is called bandwidth/smoothing parameter. Note that h_{opt} correlates with x. We wanna choose h for all x.

$$\min_h IMSE = \int E\left[\hat{f}(x) - f(x)\right]^2 dx$$

$$F.O.C. \Rightarrow h_{opt} = c_0 n^{-1/5}, \quad where \; c_0 = \left[\frac{\kappa}{\kappa_2^2 \int [f''(x)]^2 dx}\right]^{1/5}.$$

17.3 How to Choose h in Practice

17.3.1 Rule of thumb

$$h = 1.06 x_{sd} n^{-1/5}$$

Idea: We assume F is normal distribution, and estimate roughly c_o. Usually, the error is not big.

17.3.2 Pilot h

Idea: First, we use the rule of thumb to estimate h, and get a nonparametric estimate of $f(x)$. Calculate $f''(x)$ and plug in $h_{opt} = c_o n^{-1/5}$.

17.3.3 Data Driven Bandwidth Selection (most popular)

It uses least square cross-validation methods.

$$\min ISE = \int \hat{f}(x)dx - 2\int \hat{f}(x)f(x)dx,$$

where

$$\int \hat{f}(x)dx = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i).$$

$$\hat{f}_{-i}(x_i) = \frac{1}{(n-1)h}\sum_{j\neq i} k\left(\frac{x_j - x_i}{h}\right)$$

is called leave-one-out estimator. (-- cross-validation)

The second term is much more difficult. Not present here.

Chapter 18: Nonparametric Regression Estimation

$$\begin{gathered} y_i = g(x_i) + u_i, \quad E(u_i|x_i = 0) \\ \Rightarrow g(x) = E(y_i|x_i = x) \end{gathered}$$

$$\Rightarrow g(x) = \dots = \frac{\int y f(x,y)dy}{f(x)}$$

18.1 Local Constant Kernel Estimation

$$\hat{f}(x) = \frac{1}{nh_1h_2\cdots h_q}\sum_{i=1}^{n} \mathrm{K}\left(\frac{x_i - x}{h}\right)$$

where

$$\mathrm{K}\left(\frac{x_i - x}{h}\right) \equiv \prod_{m=1}^{q} \kappa\left(\frac{x_{im} - x_m}{h_m}\right).$$

Add the dimension of y, and we have

$$\hat{f}(x,y) = \frac{1}{nh_0h_1\cdots h_q}\sum_{i=1}^{n}\mathrm{K}\left(\frac{x_i - x}{h}\right)\kappa\left(\frac{y - y_i}{h_0}\right)$$

Then, we can define

$$\hat{g}(x) = \hat{E}(y_i|x_i = x) = \frac{\int y\hat{f}(x,y)dy}{\hat{f}(x)}.$$

$$\Rightarrow \hat{g}(x) = \ldots = \frac{\sum_{i=1}^{n}\mathrm{K}\left(\frac{x_i - x}{h}\right)y_i}{\sum_{i=1}^{n}\mathrm{K}\left(\frac{x_i - x}{h}\right)}$$

is weighted average of y_i.

Theorem 18.1 Consistency of $\hat{g}(x)$

$$\hat{g}(x) - g(x) = O_p(\sum_{s=1}^{q} h_s^2 + (nh_1 \ldots h_q)^{-1/2})$$

18.2 Data Driven Method of Bandwidth Selection

$$\min_h \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{g}_{-i}(x_i)]^2$$

where

$$\hat{g}_{-i}(x_i) = \frac{\sum_{j\neq i}\mathrm{K}\left(\frac{x_j - x_i}{h}\right)y_j}{\sum_{j\neq i}\mathrm{K}\left(\frac{x_j - x_i}{h}\right)}$$

Optimazing over h is less common. Note that $h_s = c_s x_{s,sd} n^{-1/(q+4)}$, we usually optimize over c_s.

Chapter 19: Semi-parametric Model

- Parametric Model: $y_i = g(x_i, \beta) + u_i$
  - Convergence rate is faster: $\sqrt{n}$;
  - The probability of misspecification is large.

    - Easy to interpret
- Non-parametric Model: $y_i = g(x_i) + u_i$
    - Convergence rate is much slower: curse of dimensionality;
    - The probability of misspecification is much smaller.

19.1 Partially Linear Models

19.2 Varying Coefficient Models

19.3 Single Index Models

> **Mathematical Supplement**
>
> $$\frac{\partial}{\partial \beta}\alpha'\beta = \alpha, \quad \frac{\partial}{\partial \beta}\beta'\alpha = \alpha$$
>
> $$\frac{\partial}{\partial \beta}\beta' A\beta = (A + A')\beta$$