KEYPOINT SUMMARY II

I. PRELIMINARIES

1. Biasness: $\hat{\theta}$ is unbiased if $\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = 0$.
2. Efficiency: $\hat{\theta}$ is efficient if $\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta})$ for all unbiased estimator $\tilde{\theta}$.
3. Consistency: $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{p} \theta$.
4. Mean square error: $\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$
5. Mean square convergence: $\hat{\theta} \xrightarrow{m.s.} \theta$ if $\text{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$.
   - $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}, \theta))^2$
   - $\hat{\theta} \xrightarrow{m.s.} \theta$ iff $\text{Bias}(\hat{\theta}) \to 0$ and $\text{Var}(\hat{\theta}) \to 0$
6. Small o (Convergence in probability): $X_n = o_p(n^k)$ if $\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_n}{n^k}\right| > \epsilon\right) = 0$
7. Big O (Stochastic boundedness): $X_n = O_p(n^k)$ if $\forall \epsilon > 0$, $\exists K > 0, N > 0$ s.t. $\forall n > N$, $\mathbb{P}\left(\left|\frac{X_n}{n^k}\right| > K\right) < \epsilon$

II. SIMPLE LINEAR REGRESSION

Let $y_i = \beta x_i + \epsilon_i$. How to find an estimator for β?

Derive an estimator by solving

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

The first-order condition gives

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \beta + \frac{\frac{1}{n}\sum_{i=1}^n x_i \epsilon_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}$$

By Law of Large Numbers (LLN),

- $\frac{1}{n}\sum_{i=1}^n x_i \epsilon_i \xrightarrow{p} \mathbb{E}(x_i \epsilon_i)$
- $\frac{1}{n}\sum_{i=1}^n x_i^2 \xrightarrow{p} \mathbb{E}(x_i^2)$

$\hat{\beta}$ is *consistent* if $\mathbb{E}(x_i\epsilon_i) = 0$ and $\mathbb{E}(x_i^2) \neq 0$.

To derive the *asymptotic normality*, consider

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n} \cdot \frac{\frac{1}{n}\sum_{i=1}^n x_i \epsilon_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}$$

Assume $\mathbb{E}(x_i^2 \epsilon_i^2)$ is finite, by Central Limit Theorem,

- $\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^n x_i \epsilon_i \xrightarrow{d} \mathcal{N}(0, \text{Var}(x_i\epsilon_i))$

Since $\frac{1}{n}\sum_{i=1}^n x_i^2 \xrightarrow{p} \mathbb{E}(x_i^2)$, by Slutsky's Theorem,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}(x_i\epsilon_i)}{[\mathbb{E}(x_i^2)]^2}\right)$$

III. MULTIPLE LINEAR REGRESSION

A. Notation

Suppose the model is specified by

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \epsilon_i$$

Let $\boldsymbol{x}_i, \boldsymbol{\beta}$ be $k \times 1$ vectors,

$$\boldsymbol{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Then the model can be written as $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$.

Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\epsilon}$ be matrices containing all observations,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_k' \end{bmatrix} = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Therefore, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

B. Assumptions

(A1) Linearity: $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i = x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \epsilon_i$

(A2) Full rank: $\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$ is nonsingular.

(A3) Exogeneity: $\mathbb{E}(\epsilon_i|\boldsymbol{x}_j) = 0$, for $i, j = 1, \ldots, n$

(A4) Homoskedasticity and nonautocorrelation: $\text{Var}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$ ($\text{Var}(\epsilon_i) = \sigma^2$ and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$)

(A5) Independent and identical data: $\{(y_i, \boldsymbol{x}_i)\}$ are i.i.d.

If the regressors can be treated as nonstochastic (as they would be in an experiment situation in which the analyst choose the values in $\boldsymbol{X}$), $\boldsymbol{X}$ can be treated as contant matrix. If $\boldsymbol{X}$ is stochastic (random variables), the anasysis should be done conditioned on the observed $\boldsymbol{X}$. For notation simplicity, in the following text, $\boldsymbol{X}$ is treated as constant wherever possible.

C. The OLS Estimator

Minimizing the sum of squared residuals:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

or equivalently,

$$\min_{\boldsymbol{\beta}}(\boldsymbol{Y}-\boldsymbol{X\beta})(\boldsymbol{Y}-\boldsymbol{X\beta})'$$

The first-order condition gives

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})=\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right)$$

D. Properties of OLS Estimator

1. Unbiasedness

$$\mathbb{E}[\hat{\boldsymbol{\beta}}]=\boldsymbol{\beta}+\mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}]=\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{\epsilon})=\boldsymbol{\beta}$$

2. Efficiency

OLS is the most efficient unbiased linear estimator.

Proof. Let $\tilde{\boldsymbol{\beta}}=\boldsymbol{C}'\boldsymbol{y}$ be another unbiased linear estimator. Since $\tilde{\boldsymbol{\beta}}$ is unbiased, $\mathbb{E}(\tilde{\boldsymbol{\beta}})=\mathbb{E}(\boldsymbol{C}'\boldsymbol{y})=\mathbb{E}(\boldsymbol{C}'(\boldsymbol{X\beta}+\boldsymbol{\epsilon}))=\boldsymbol{C}'\boldsymbol{X\beta}=\boldsymbol{\beta}$, which implies $\boldsymbol{C}'\boldsymbol{X}=\boldsymbol{I}$. Therefore, $\tilde{\boldsymbol{\beta}}=\boldsymbol{C}'\boldsymbol{Xb}+\boldsymbol{C}'\boldsymbol{\epsilon}=\boldsymbol{\beta}+\boldsymbol{C}'\boldsymbol{\epsilon}$.

$$\begin{aligned}\mathrm{Var}(\tilde{\boldsymbol{\beta}})&=\mathrm{Var}(\boldsymbol{C}'\boldsymbol{\epsilon})=\mathbb{E}(\boldsymbol{C}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{C})=\sigma^2\boldsymbol{C}'\boldsymbol{C}\\&=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}+\sigma^2\boldsymbol{Z}\boldsymbol{Z}'\\&\geq\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned}$$

where $\boldsymbol{Z}=\boldsymbol{C}'-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Gauss-Markov Theorem: the least square estimator $\hat{\boldsymbol{\beta}}$ is the minimal variance (most efficient) linear unbiased estimator.

3. Consistency

$$\begin{aligned}&\mathrm{Var}(\hat{\boldsymbol{\beta}})=\mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})']\\&=\mathbb{E}\left[\left(\textstyle\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_i \boldsymbol{x}_i\epsilon_i\right)\left(\sum_i \boldsymbol{x}_i\epsilon_i\right)'\left(\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right]\\&=\mathbb{E}\left[\left(\textstyle\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_i\sum_j \boldsymbol{x}_i\epsilon_i\epsilon_j\boldsymbol{x}_j'\right)\left(\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right]\\&=\mathbb{E}\left[\left(\textstyle\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\sigma^2\right)\left(\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right]\\&=\sigma^2\left(\textstyle\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}=\frac{\sigma^2}{n}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\\&\to\boldsymbol{0}\text{ as }n\to\infty\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}\xrightarrow{m.s.}\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$.

4. Asymptotic normality

Multiply by the "stabler" $\sqrt{n}$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\epsilon_i\right)$$

The following properties hold as $n\to\infty$,

- $\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\xrightarrow{p}[\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$ by LLN;
- $\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\epsilon_i\xrightarrow{d}\mathcal{N}(\boldsymbol{0},\mathrm{Var}(\boldsymbol{x}_i\epsilon_i))$ by CLT;
- $\mathrm{Var}(\boldsymbol{x}_i\epsilon_i)=\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i'\epsilon_i^2)=\sigma^2\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$

Therefore, $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}\mathcal{N}(\boldsymbol{0},\sigma^2[\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1})$.

In practice, $\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$ is estimated by $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$, and σ^2 is estimated by either

- $\hat{\sigma^2}=\frac{1}{n}\sum_{i=1}^n e_i^2$, or
- $s^2=\frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}$.

where e_i and $\boldsymbol{e}$ both stand for residuals.

E. Violation of Assumptions

1. Multicollinearity

If $\boldsymbol{X}'\boldsymbol{X}$ is "close" to singular, i.e. $\det(\boldsymbol{X}'\boldsymbol{X})\approx 0$, then $\mathrm{Var}(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ will be very large, which leads to imprecise estimator.

2. Heteroskedasticity

If $\mathbb{E}(\epsilon_i^2|\boldsymbol{x}_i)=\sigma_i^2$ different for each i. Assume $\mathbb{E}(\epsilon_i\epsilon_j)=0$ for $i\neq j$. Reevaluate the properties of OLS estimator:

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}})&=\boldsymbol{\beta}+\left(\textstyle\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\sum_{i=1}^n \boldsymbol{x}_i\mathbb{E}(\epsilon_i)=\boldsymbol{\beta}\\\mathrm{Var}(\hat{\boldsymbol{\beta}})&=\frac{1}{n}\left(\frac{1}{n}\textstyle\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\sigma_i^2\right)\left(\frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\\&\to\boldsymbol{0}\text{ as }n\to\infty\end{aligned}$$

$\hat{\boldsymbol{\beta}}$ is still unbiased and consistent.

Asymptotic normality:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\epsilon_i\right)$$

- $\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\xrightarrow{p}[\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$ by LLN;

· Though ϵ_i is heteroskedastic, we still have $\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\epsilon_i \xrightarrow{d} \mathcal{N}(0, \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\boldsymbol{x}_i\epsilon_i))$ under some conditions.

If we define:

· $\boldsymbol{Q} = \mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$

· $\boldsymbol{R} = \frac{1}{n}\sum_i \mathrm{Var}(\boldsymbol{x}_i\epsilon_i) = \frac{1}{n}\sum_i \mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')\sigma_i^2$

Then, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}\boldsymbol{R}\boldsymbol{Q}^{-1})$.

Therefore, in heteroskedastic case, $\hat{\boldsymbol{\beta}}$ still conforms to asymptotic normality. But the variance is no longer $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. So traditional statistical inference based on $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ will be misleading.

Robust standard error:

$$\tilde{\boldsymbol{R}}_{\boldsymbol{n}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' e_i^2$$

where e_i is the residual. It can be shown, under some conditions, $\tilde{\boldsymbol{R}}_{\boldsymbol{n}} \xrightarrow{p} \boldsymbol{R}$.

3. Problem of dependency

If $\{\boldsymbol{x}_i\}$ are not independent, $\hat{\boldsymbol{\beta}}$ is still unbiased, but it might not be consistent, because LLN and CLT no longer hold.

4. Endogeneity

Endogeneity problem rises when $\mathbb{E}(\boldsymbol{x}_i\epsilon_i) \neq \boldsymbol{0}$. Two sources of endogeneity: (a) measurement error; (b) ommited variable.

a. *Measurement error*

$$y_i = x_i^*\beta + \epsilon_i$$

Suppose x_i^* stands for the measurement error free value of x_i. Let $x_i^u = x_i^* + v_i$ where v_i is the measurement error. For simplicity, assume $\mathbb{E}(x_i^*) = \mathbb{E}(v_i) = 0$, $\mathrm{Var}(v_i) = \sigma_v^2$. And assume the best scenario when there is a measurement error: $x_i^* \perp \epsilon_i, x_i^* \perp v_i, v_i \perp \epsilon_i$.

$$y_i = (x_i^u - v_i)\beta + \epsilon_i^u = x_i^u\beta + \epsilon_i - \beta v_i = x_i^u\beta + \epsilon_i^u$$

where $\epsilon_i^u = \epsilon_i - \beta v_i$.

$$\mathbb{E}(x_i^u\epsilon_i^u) = \mathbb{E}((x_i^* + v_i)(\epsilon_i - \beta v_i)) = \mathbb{E}(-\beta v_i^2) = -\beta\sigma_v^2 \neq 0$$

So we have endogeneity problem. If we regress y_i on x_i^u,

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n} x_i^u y_i}{\sum_{i=1}^{n}(x_i^u)^2} = \beta + \frac{\sum_{i=1}^{n} x_i^u\epsilon_i^u}{\sum_{i=1}^{n}(x_i^u)^2} \\ &\xrightarrow{p} \beta + \frac{\mathbb{E}(x_i^u\epsilon_i^u)}{\mathrm{Var}(x_i^u)} = \beta\left(1 - \frac{\sigma_v^2}{\mathrm{Var}(x_i^*) + \sigma_v^2}\right)\end{aligned}$$

Several observations:

1. If $\sigma_v^2 = 0$, $\hat{\beta}$ is consistent.
2. The larger the measurement error σ_v^2, the larger the bias.
3. $\hat{\beta}$ always has the same sign as β.
4. Attenuation bias: $|\hat{\beta}| \leq |\beta|$. Therefore, if the result is significant in measurement error cases, it is also significant in measurement-error free case.

b. *Omitted variable*

$$y_i = x_i\beta + z_i\gamma + \epsilon_i$$

Suppose $\mathbb{E}(x_i\epsilon_i) = 0$, $\mathrm{Cov}(x_i, z_i) \neq 0$. If the variable z_i is omitted in the model,

$$y_i = x_i\beta + \delta_i$$

where $\delta_i = z_i\gamma + \epsilon_i$. Then δ_i is corrleted with x_i.
To resolve the omitted variable bias, we need to use instrumental variable (IV) estimation.

IV. IV ESTIMATION

A. The IV Estimator

Suppose the structural model is

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$$

where e_i is correleted with $\boldsymbol{x}_i$.

Suppose $\boldsymbol{z}_i$ are instruments satisfying:

· $\boldsymbol{z}_i$ has the same dimension as $\boldsymbol{x}_i$;

· $\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$ has full rank;

· $\mathbb{E}(\boldsymbol{z}_i\epsilon_i) = 0$.

Then we have

$$\mathbb{E}(\boldsymbol{z}_i\epsilon_i) = \mathbb{E}(\boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})) = \mathbb{E}(\boldsymbol{z}_i y_i) - \mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')\boldsymbol{\beta} = 0$$

Therefore, $\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i'))^{-1}\mathbb{E}(\boldsymbol{z}_i y_i)$.

Define the IV estimator:

$$\hat{\boldsymbol{\beta}}_{IV} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i y_i\right)$$

1. Consistency

$$\hat{\boldsymbol{\beta}}_{IV} = \boldsymbol{\beta} + \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\epsilon_i\right)$$

By Law of Large Numbers,

- $\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1} \xrightarrow{p} [\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}$;
- $\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\epsilon_i \xrightarrow{p} \mathbb{E}(\boldsymbol{z}_i\epsilon_i) = 0$.

Therefore, $\hat{\boldsymbol{\beta}}_{IV} \xrightarrow{p} \boldsymbol{\beta}$.

Note: IV estimator is generally biased, but consistent.

2. Asymptotic normality

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\epsilon_i\right)$$

- $\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1} \xrightarrow{p} [\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}$ by LLN;
- $\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\epsilon_i \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \mathbb{E}(\boldsymbol{z}_i\boldsymbol{z}_i'\epsilon_i^2))$ by CLT.

Therefore, by Slutsky's Theorem,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, [\mathbb{E}(\boldsymbol{x}_i\boldsymbol{z}_i')]^{-1}\mathbb{E}(\boldsymbol{z}_i\boldsymbol{z}_i'\epsilon_i^2)[\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1})$$

B. 2SLS Estimator

Suppose in a more general case,

$$\boldsymbol{Y} = \underset{n\times k}{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1, \dots, \boldsymbol{X}_k\end{bmatrix}$, and $\mathbb{E}(\boldsymbol{\epsilon}|\boldsymbol{X}) \neq 0$.

Suppose $\underset{n\times l}{\boldsymbol{Z}}$ are instruments, where $l \geq k$, i.e. there could be more instruments than independent variables.

The following two procedures are equivalent:
a. IV estimation

1) Regress $\boldsymbol{X}$ on $\boldsymbol{Z}$, get fitted $\hat{\boldsymbol{X}} = \boldsymbol{Q}$;
2) Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ using $\boldsymbol{Q}$ as the instrument.

b. 2SLS estimation

1) Regress $\boldsymbol{X}$ on $\boldsymbol{Z}$, get fitted $\hat{\boldsymbol{X}} = \boldsymbol{Q}$;
2) Regress $\boldsymbol{Y}$ on $\hat{\boldsymbol{X}}$.

Proof. Regressing $\boldsymbol{X}$ on $\boldsymbol{Z}$:

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}_{\boldsymbol{1}} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}_1, & \boldsymbol{Q}_1 &= \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\boldsymbol{1}}\\
\hat{\boldsymbol{\gamma}}_{\boldsymbol{2}} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}_2, & \boldsymbol{Q}_2 &= \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\boldsymbol{2}}\\
&\vdots & &\vdots\\
\hat{\boldsymbol{\gamma}}_{\boldsymbol{k}} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}_k, & \boldsymbol{Q}_k &= \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_{\boldsymbol{k}}
\end{aligned}$$

Then,

$$\begin{aligned}
\hat{\boldsymbol{X}} = \boldsymbol{Q} &= \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_k\end{bmatrix}\\
&= \begin{bmatrix}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}_1 & \cdots & \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}_k\end{bmatrix}\\
&= \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\begin{bmatrix}\boldsymbol{X}_1 & \cdots & \boldsymbol{X}_k\end{bmatrix}\\
&= \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}
\end{aligned}$$

If we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ using $\boldsymbol{Q}$ as IV, then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{IV} &= (\boldsymbol{Q}'\boldsymbol{X})^{-1}(\boldsymbol{Q}'\boldsymbol{Y})\\
&= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y})
\end{aligned}$$

If we regress $\boldsymbol{Y}$ directly on $\hat{\boldsymbol{X}}$,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{2SLS} &= (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}(\hat{\boldsymbol{X}}'\boldsymbol{Y})\\
&= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\cdot\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\\
&\quad\ (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y})\\
&= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y})
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}_{IV} = \hat{\boldsymbol{\beta}}_{2SLS}$.

V. MAXIMUM LIKELIHOOD EXTIMATION

A. The Likelihood Function

Let $\{z_i\}$ be i.i.d. $f(z_i|\boldsymbol{\theta})$ is the *pdf* for z_i conditioned on a set of parameters $\underset{k\times 1}{\boldsymbol{\theta}}$.
The likelihood function is the joint density function:

$$L(\boldsymbol{\theta}|\boldsymbol{Z}) = f(z_1, \dots, z_n|\boldsymbol{\theta}) = \Pi_{i=1}^n f(z_i|\boldsymbol{\theta})$$

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\boldsymbol{\theta}|\boldsymbol{Z}) = \sum_{i=1}^n \ln f(z_i|\boldsymbol{\theta})$$

The maximum likelihood estimator (MLE) is

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{i=1}^n \ln f(z_i|\boldsymbol{\theta})$$

B. The Likelihood Inequality

Suppose $z_i \sim f(z|\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is the true parameter. Then for any $\boldsymbol{\theta}$, the following inequality holds

$$\mathbb{E}[\ln f(z|\boldsymbol{\theta}_0)] \geq \mathbb{E}[\ln f(z|\boldsymbol{\theta})]$$

Proof.

$$\mathbb{E}[\ln f(z|\boldsymbol{\theta})] - \mathbb{E}[\ln f(z|\boldsymbol{\theta}_0)] = \mathbb{E}[\ln f(z|\boldsymbol{\theta}) - \ln f(z|\boldsymbol{\theta}_0)] =$$
$$\mathbb{E}\left[\ln\frac{f(z|\boldsymbol{\theta})}{f(z|\boldsymbol{\theta}_0)}\right] \leq \ln\mathbb{E}\left[\frac{f(z|\boldsymbol{\theta})}{f(z|\boldsymbol{\theta}_0)}\right] = \ln\int\frac{f(z|\boldsymbol{\theta})}{f(z|\boldsymbol{\theta}_0)}f(z|\boldsymbol{\theta}_0)dz = 0$$

C. Assumptions

(A1) $\{z_i\}$ are i.i.d.

(A2) $\boldsymbol{\theta}_0$ is the true parameter, Θ is a compact set;

(A3) $\text{Var}(\nabla_{\boldsymbol{\theta}} \ln f(z_i|\boldsymbol{\theta}_0))$ is nonsingular;

(A4) First, second and third own and cross derivatives of $\ln f(z_i|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are all bounded;

(A5) Let $\Omega_{\boldsymbol{Z}}$ be the support of $\boldsymbol{Z}$, then either

  (a) $\Omega_{\boldsymbol{Z}}$ does not depend on $\boldsymbol{\theta}$, or

  (b) $f(\boldsymbol{Z}|\boldsymbol{\theta}) = 0$ on the boundary of $\Omega_{\boldsymbol{Z}}$.

D. Score Function

Define the score function:

$$s(z_i|\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln f(z_i|\boldsymbol{\theta})$$

Properties of the score function:

1. $\{s(z_i|\boldsymbol{\theta})\}$ are i.i.d. because $\{z_i\}$ are i.i.d.
2. If (A5) holds, then $\mathbb{E}[s(z_i|\boldsymbol{\theta}_0)] = 0$.

Proof. By definition,

$$\int_{\Omega_{\boldsymbol{Z}}} f(z_i|\boldsymbol{\theta}) dz = 1$$

Therefore, $\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\Omega_{\boldsymbol{Z}}} f(z_i|\boldsymbol{\theta}) dz = 0$. By Leibniz rule,

$$\int_{A(\boldsymbol{\theta})}^{B(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(z_i|\boldsymbol{\theta}) dz + B'(\boldsymbol{\theta}) f(B(\boldsymbol{\theta})|\boldsymbol{\theta}) - A'(\boldsymbol{\theta}) f(A(\boldsymbol{\theta})|\boldsymbol{\theta}) = 0$$

Under (A5), either

1. $A'(\boldsymbol{\theta}) = B'(\boldsymbol{\theta}) = 0$, or
2. $f(A(\boldsymbol{\theta})|\boldsymbol{\theta}) = f(B(\boldsymbol{\theta})|\boldsymbol{\theta}) = 0$

In either case, we conclue

$$\int_{A(\boldsymbol{\theta})}^{B(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(z_i|\boldsymbol{\theta}) dz = 0$$

Therefore,

$$\begin{aligned}
\mathbb{E}[s(z_i|\boldsymbol{\theta}_0)] &= \int_{\Omega_{\boldsymbol{Z}}} s(z_i|\boldsymbol{\theta}_0) f(z_i|\boldsymbol{\theta}_0) dz \\
&= \int_{\Omega_{\boldsymbol{Z}}} \frac{\ln f(z_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} f(z_i|\boldsymbol{\theta}_0) dz \\
&= \int_{\Omega_{\boldsymbol{Z}}} \frac{1}{f(z_i|\boldsymbol{\theta}_0)} \frac{f(z_i|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} f(z_i|\boldsymbol{\theta}_0) dz \\
&= \int_{\Omega_{\boldsymbol{Z}}} \frac{\partial}{\partial \boldsymbol{\theta}} f(z_i|\boldsymbol{\theta}_0) dz = 0
\end{aligned}$$

E. Properties of MLE

1. Consistency

Under some regularity conditions, we have $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$.

2. Asymptotic normality

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln f(z_i|\boldsymbol{\theta})$$

The first-order condition is

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(z_i|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = 0 \Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} s(z_i|\hat{\boldsymbol{\theta}}) = 0$$

Let $\mathbf{H}(z_i|\boldsymbol{\theta}) = \frac{\partial s(z_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial^2 s(z_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$.

Taylor expand the FOC around $\boldsymbol{\theta}_0$:

$$\frac{1}{n} \sum_{i=1}^{n} s(z_i|\boldsymbol{\theta}_0) + \frac{1}{n} \sum_{i=1}^{n} \mathbf{H}(z_i|\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = 0$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left(-\frac{1}{n} \sum_{i=1}^{n} \mathbf{H}(z_i|\boldsymbol{\theta}_0)\right)^{-1} \left(\sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} s(z_i|\boldsymbol{\theta}_0)\right)$$

· By Law of Large Numbers,

$$\begin{aligned}
\left(-\frac{1}{n} \sum_{i=1}^{n} \mathbf{H}(z_i|\boldsymbol{\theta}_0)\right)^{-1} &\xrightarrow{p} \mathbb{E}[-\mathbf{H}(z_i|\boldsymbol{\theta}_0)]^{-1} \\
&\to \mathbb{E}[s(z_i|\boldsymbol{\theta}_0) s(z_i|\boldsymbol{\theta}_0)']^{-1}
\end{aligned}$$

· By Central Limit Theorem,

$$\begin{aligned}
\sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} s(z_i|\boldsymbol{\theta}_0) &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \text{Var}(s(z_i|\boldsymbol{\theta}_0))) \\
&\to \mathcal{N}(\mathbf{0}, \mathbb{E}[s(z_i|\boldsymbol{\theta}_0) s(z_i|\boldsymbol{\theta}_0)'])
\end{aligned}$$

By Slutsky's Theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{E}[s(z_i|\boldsymbol{\theta}_0) s(z_i|\boldsymbol{\theta}_0)']^{-1})$$

3. Asymptotic efficiency

$\hat{\boldsymbol{\theta}}$ is asymptotic efficient, meaning $\hat{\boldsymbol{\theta}}$ has an asymptotic covariance matrix that is not larger than the asymptotic covariance of any other consistent, asymptotically normally distributed estimator.

4. Invariance

If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, $g(\hat{\boldsymbol{\theta}})$ is the MLE for $g(\boldsymbol{\theta})$ for g a continuously differentiable function.

F. Delta Method

If there is a sequence of random variables $\{x_n\}$ satisfying

$$\sqrt{n}(x_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Then for any g satisfying the perperty that $g'(\theta)$ exists and non-zero valued, we have

$$\sqrt{n}(g(x_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$$

G. Quasi-MLE

Consider the maximum likelihood estimation of the linear model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$$

Assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (ϵ_i may not be normal, but we assume it anyway). Then the log likelihood function is

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \ln f(y_1, \boldsymbol{x}_1, \ldots, y_n, \boldsymbol{x}_n | \boldsymbol{\beta}, \sigma^2) \\ &= \ln f(y_1, \ldots, y_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\beta}, \sigma^2) \\ &\quad + \ln f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\beta}, \sigma^2) \end{aligned}$$

We assume $\boldsymbol{\beta}, \sigma^2$ do not depend on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, therefore $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. We can drop it from the likelihood function without altering the result.

$$\begin{aligned} (\boldsymbol{\beta}, \sigma^2) &= \operatorname*{argmax}_{\boldsymbol{\beta}, \sigma^2} \ln f(y_1, \ldots, y_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\beta}, \sigma^2) \\ &= \operatorname*{argmax}_{\boldsymbol{\beta}, \sigma^2} \sum_{i=1}^n \ln f(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2) \end{aligned}$$

Since $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$.

$$f(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right)$$

Therefore, the likelihood function can be instantiated as

$$L(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^n \left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

The first-order condition gives

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \\ \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2 \end{aligned}$$

which is exactly the same as the OLS estimator.