Week 2: Nov. 10, 2022

Note Writer: Yu-Chieh Kuo†

†Department of Information Management, National Taiwan University

Asymptotics (Large-Sample Theory)

Typically, in stats or econometrics, we derive the properties of estimators by taking expectations and taking sample size goes to infinity. For example, given *i.i.d.* data $y_1, \cdots, y_n$ and the corresponding expectation $\mathbb{E}[y_i] = \mu$, we are able to estimate

$$\hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \mathbb{E}[\hat{\mu}] = \mu.$$

Another example yeilds

$$\begin{aligned}\hat{\beta}_{OLS} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) \\ &\xrightarrow{p} \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1} \mathbb{E}[x_i y_i]\end{aligned}$$

Law of Large Numbers

Given $z_1, z_2, \cdots, z_n$ are *i.i.d.* (not necessary), we have

$$\bar{z}_n \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \quad \text{and} \quad \bar{z}_n \xrightarrow{p} \mathbb{E}[z_i]$$

Note that it is Weak Law of Large Number (WLLN) and almost-sure convergence here.

Central Limit Theorem

Given $z_1, z_2, \cdots, z_n$ are *i.i.d.* (not necessary) and $\mathbb{E}[z_i] \equiv \mu$, where z_i are $k \times 1$ vectors, we have

$$\sqrt{n}(\bar{z}_n - \mu) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[(z_i - \mu)(z_i - \mu)']\right),$$

where $\mathbb{E}[(z_i - \mu)(z_i - \mu)'] \equiv Var(z_i)$.

Least Square

Given the data

$$\begin{array}{lll} \underset{1\times 1}{y_1}, & y_2, \cdots, y_n & \text{dependent variables} \\ \underset{k\times 1}{x_1}, & x_2, \cdots, x_n & \text{independent variables,} \end{array}$$

we define

$$Y \equiv \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad X \equiv \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}.$$

Theorem. Suppose $g(x_i)$ is some function of x_i. Then, the conditional mean of y_i, $\mathbb{E}[y_i \mid x_i] \equiv \mu_i$, minimize $\mathbb{E}[y_i - g(x_i)]^2$. That is, $g(x_i) = \mu_i$ is the minimizer. □

Denote the predicted y_i as $\hat{y}_i$ and define $\hat{\mathbb{E}}[\cdot] \equiv \frac{1}{n}\sum_{i=1}^{n}(\cdot)$, we want to minimize

$$Q_\infty(\beta) \equiv \mathbb{E}[y_i - \hat{y}_i]^2 \quad \text{and} \quad Q_n(\beta) \equiv \hat{\mathbb{E}}[y_i - \hat{y}_i]^2$$

by using linear curve $\hat{y}_i = x_i'\beta$, where x_i' and β are $1 \times k$ and $k \times 1$ vectors, respectively. Note that econometrisians call Q as the objective function, and statistisians call it as the criterion function.

Theorem. The minimizer of $\mathbb{E}\left[y_i - x_i'\beta\right]^2$ is

$$\beta_\infty = \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}(\mathbb{E}[x_i y_i]).$$

The minimizer of $\hat{\mathbb{E}}[y_i - \hat{y}_i]^2$ is

$$\hat{\beta} = \left(\hat{\mathbb{E}}\left[x_i x_i'\right]\right)^{-1}\left(\hat{\mathbb{E}}[x_i y_i]\right).$$

□

Here, if we define

$$e_i \equiv y_i - x_i'\beta_\infty \quad \text{and} \quad \hat{e}_i \equiv y_i - x_i'\hat{\beta}$$
$$E \equiv \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \qquad \hat{E} \equiv \begin{pmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_n \end{pmatrix}.$$

then

$$\mathbb{E}[x_i e_i] = 0 \quad \text{and} \quad \hat{\mathbb{E}}[x_i \hat{e}_i] = 0.$$

Assume observations are *i.i.d.* since

$$\hat{\mathbb{E}}\left[x_i x_i'\right] \xrightarrow{p} \mathbb{E}\left[x_i x_i'\right] \quad \text{and} \quad \hat{\mathbb{E}}[x_i y_i] \xrightarrow{p} \mathbb{E}[x_i y_i]$$

therefore, we obtain $\hat{\beta} \xrightarrow{p} \beta_\infty$.

Remark. $x_i'\beta_\infty$ may not to be the true μ_i but we know $\hat{\beta}$ converges to β_∞. □

Remark.

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \xrightarrow{p} Q_\infty(\beta) = \mathbb{E}\left[y_i - x_i'\beta\right]^2$$
$$\hat{\beta} \equiv \arg\min_\beta Q_n(\beta) \xrightarrow{p} \beta_\infty \equiv \arg\min_\beta Q_\infty(\beta).$$

Typically, in econometrics textbook, β_∞ is the true parameters. That is, consistency means that estimators converge to true parameters in probability. □

Finite sample properties

Given the model $Y = X\beta_\infty + E$, we have

$$\hat{\beta} = (X'X)^{-1}(X'Y) \quad \text{and} \quad \hat{\beta} = \beta_\infty + (X'X)^{-1}X'E.$$

Note that X and Y are $n \times k$ and $n \times 1$ matrix and vector.

- We say the parameter as unbiasedness if $\mathbb{E}\left[\hat{\beta} \mid X\right] = \beta_\infty$ by assuming $\mathbb{E}[E \mid X] = 0$.
- We obtain
$$\mathbb{E}\left[\left(\hat{\beta} - \beta_\infty\right)\left(\hat{\beta} - \beta_\infty\right)' \mid X\right] = \sigma^2(X'X)^{-1}$$
by assuming $\mathbb{E}[EE' \mid X] = \sigma^2 I_n$.
- If $E \sim \mathcal{N}(0, \sigma^2 I_n)$, we obtain
$$\hat{\beta} \mid X \sim \mathcal{N}\left(\beta_\infty, \sigma^2(X'X)^{-1}\right).$$

Asymptotic properties (Large-Sample properties)

Given the model $y_i = x_i'\beta_\infty + e_i$, we have

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right) = \beta_\infty + \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i e_i\right).$$

The last part is sometimes called the sampling error. Note that since $\frac{1}{n}\sum_{i=1}^n x_i e_i \xrightarrow{p} \mathbb{E}[x_i e_i] = 0$, we have the consistency property

$$\hat{\beta} \xrightarrow{p} \beta_\infty.$$

Next, by re-scaling and the substraction, the estimators turns to

$$\sqrt{n}\left(\hat{\beta} - \beta_\infty\right) = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^n x_i e_i\right).$$

By CLT,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n x_i e_i - \mathbb{E}[x_i e_i]\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[x_i x_i' e_i^2\right]\right)$$

since $\mathbb{E}[x_i e_i] = 0$, therefore, it alters to

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta} - \beta_\infty\right) &= \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^n x_i e_i\right) \\
&\xrightarrow{d} \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}\mathcal{N}\left(0, \mathbb{E}\left[x_i x_i' e_i^2\right]\right) \\
&\rightarrow \mathcal{N}\left(0, \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}\mathbb{E}\left[x_i x_i' e_i^2\right]\left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}\right).
\end{aligned}$$

Assume that $\mathbb{E}\left[x_i x_i' e_i^2\right] = \mathbb{E}\left[x_i x_i'\right]\sigma^2$, where $\sigma^2 \equiv \mathbb{E}\left[e_i^2\right]$, the asymptotic covariance matrix is $\sigma^2\left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}$.

Remark. The existence of inverse $(X'X)^{-1}$ and $\left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}$ means that there is no perfect multi-collinearity. □

Theorem. $(X'X)^{-1}$ exists if and only if the columns of X are linearly independent. To elaborate, the eigenvalues of $X'X$ are not equal to 0. □

Note that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} X'X = \mathbb{E}\left[x_i x_i'\right].$$

The existence issues mentioned in the remark and the theorem above reveals the identification; that is, we can identify the unknown parameters.

Identification

These equations are identical:

$$\begin{aligned} \mathbb{E}\left[x_i x_i'\right]\beta &= \mathbb{E}[x_i y_i] \\ \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)\beta &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i \\ (X'X)\beta &= X'Y \end{aligned}$$

Here if the inverse of $X'X$ exists; that is, $X'X$ has k equations and we have k unknown β.

Note: chi-square are the square of normal distribution.

Projection and residual

The projection matrix is defined as $P \equiv X(X'X)^{-1}X'$. To represent the projection matrix mathematically, it projects vectors into the subspace spanned by columns of X. To elaborate, for any vector V, PV is the linear combination of columns of X.

To be more econometrics, it comes from

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y \equiv PY.$$

Now, we define another matrix $M \equiv I_n - P$, where M is $n \times n$. We have the following properties for M and P:

- P, M are symmetric.
- $PP = P$.
- $MM = M$.
- $PM = 0$ (important in the calculation of prediction errors).
- $trace(P) = k$ and $trace(M) = n-k$. For any square matrix A, $trace(A)$ is the sum of diagnol entries of A. Moreover, $trace(AB) = trace(BA)$.

Prediction error

Suppose we have the true in-sample data and model $y_i^{in} = \mu_i + e_i^{in}$ generated and estimated by in-sample data x_i, and there exists an out-sample data y_i^{out} generated by same in-sample x_i, i.e., $y_i^{out} = \mu_i + e_i^{out}$. Note that $y_i = \mu_i + e_i$ is really the true model given $\mathbb{E}[y_i \mid x_i] \equiv \mu_i \iff e_i \equiv y_i - \mu_i$. Now, we want to calculate the expected square errors

$$\mathbb{E}\left[\left(Y^{in} - \hat{Y}\right)'\left(Y^{in} - \hat{Y}\right) \mid X\right] \quad \text{and} \quad \mathbb{E}\left[\left(Y^{out} - \hat{Y}\right)'\left(Y^{out} - \hat{Y}\right) \mid X\right]$$

to specify the prediction power of the model. Observe that

$$
\begin{aligned}
Y^{in} - X\hat{\beta} &= \mu + E^{in} - X\hat{\beta} \\
&= \mu + E^{in} - X(X'X)^{-1}X'(\mu + E^{in}) \\
&= (I-P)\mu + (I-P)E^{in} \\
Y^{out} - X\hat{\beta} &= \mu + E^{out} - X\hat{\beta} \\
&= \mu + E^{out} - X(X'X)^{-1}X'(\mu + E^{in}) \\
&= (I-P)\mu + E^{out} - PE^{in}.
\end{aligned}
$$

Hence, we can take the expectation of the square error

$$
\begin{aligned}
\mathbb{E}\Big[\big(Y^{in} - \hat{Y}\big)'\big(Y^{in} - \hat{Y}\big) \mid X\Big] &= \mathbb{E}\Big[\mu'M'M\mu + E^{in\prime}M'ME^{in} + \mu'M'ME^{in} + E^{in\prime}M'M\mu\Big] \\
&= (n-k)\sigma^2 + \mu'(I-P)\mu \\
&\quad \text{(Since } \mu'M'ME^{in} = E^{in\prime}M'M\mu = 0) \\
\mathbb{E}\Big[\big(Y^{out} - \hat{Y}\big)'\big(Y^{out} - \hat{Y}\big) \mid X\Big] &= \mathbb{E}\Big[\mu'M'M\mu + \mu'M'E^{out} - \mu'M'PE^{in} + E^{out}M\mu + E^{out\prime}E^{out} \\
&\quad - E^{out\prime}PE^{in} + E^{in\prime}PM\mu - E^{in\prime}P'E^{out} + E^{in\prime}P'PE^{in} \mid X\Big] \\
&= (n+k)\sigma^2 + \mu'(I-P)\mu.
\end{aligned}
$$

Only the highlighted terms remain, and others go to 0 after taking the expectation. The reasons for being 0 include

Independence: μ and E^{out}; E^{in} and E^{out} are independent. Therefore, the expectation term goes to 0.

PM Matrix: $PM = 0$ by definition.

By dividing into n, prediction errors alter to

$$
\begin{aligned}
\frac{1}{n}\mathbb{E}\Big[\big(Y^{in} - \hat{Y}\big)'\big(Y^{in} - \hat{Y}\big) \mid X\Big] &= \sigma^2 - \frac{k}{n}\sigma^2 + \frac{1}{n}\mu'(I-P)\mu \\
\frac{1}{n}\mathbb{E}\Big[\big(Y^{out} - \hat{Y}\big)'\big(Y^{out} - \hat{Y}\big) \mid X\Big] &= \sigma^2 + \frac{k}{n}\sigma^2 + \frac{1}{n}\mu'(I-P)\mu.
\end{aligned}
$$

Note that when $k > n$ (k is the number of variables), $X'X$ is not invertible. Consequently, it is not a case.

Now, comparing in-sample and out-sample prediction errors yeilds

$$
\mathbb{E}\Big[\big(Y^{out} - \hat{Y}\big)'\big(Y^{out} - \hat{Y}\big) \mid X\Big] - \mathbb{E}\Big[\big(Y^{in} - \hat{Y}\big)'\big(Y^{in} - \hat{Y}\big) \mid X\Big] = 2k\sigma^2 \text{ (treated as a penalty).}
$$

There are many choices of the variable sets. Conventionally, people use the biggest approximating linear model to estimate σ^2.

Remark.

- The in-sample prediction error always suggests to use more complex models.
- However, the out-sample prediction error exhibits a trade-off between bias and variance. It penalizes too much variables.

□

Remark. For any matrix A,

$$\begin{aligned}\mathbb{E}[E'AE] &= trace(\mathbb{E}[E'AE]) \\ &= \mathbb{E}[trace(E'AE)] \\ &= \mathbb{E}[trace(AEE')] \\ &= trace(A\,\mathbb{E}[EE']) \\ &= trace\Big(A\sigma^2 I_n\Big) \\ &= \sigma^2 trace(A).\end{aligned}$$

□

Model Selection Theory

Mallows CP

Mallows CP calculates

$$\mathbb{E}\Big[(\mu-\hat{\mu})'(\mu-\hat{\mu}) \mid X\Big] = k\sigma^2 + \mu'(I-P)\mu,$$

where $\hat{\mu} = \hat{Y} = \hat{X}\hat{\beta}$. The result is similar to the out-sample prediction error.

Nonlinear Least Square (NLS)

Given *i.i.d.* data

$$\begin{array}{lll} \underset{1\times 1}{y_1,} & y_2, \cdots, y_n & \text{dependent variables} \\ \underset{k\times 1}{x_1,} & x_2, \cdots, x_n & \text{independent variables,} \end{array}$$

and the model $y_i = f(x_i;\beta) + e_i$. The objective function is

$$Q_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i;\beta))^2 \ \xrightarrow{p}\ Q_\infty(\beta) \equiv \mathbb{E}[y_i - f(x_i;\beta)]^2.$$

Here, econometrisians impose some restrictions:

Identification assumption: $\beta_\infty \equiv \arg\min_\beta Q_\infty(\beta)$ uniquely exists.

Probability convergence assumption: $Q_n(\beta) \xrightarrow{p} Q_\infty(\beta)$ uniformly.

Then, we have

$$\hat{\beta} \equiv \arg\min_\beta Q_n(\beta) \xrightarrow{p} \beta_\infty \equiv \arg\min_\beta Q_\infty(\beta),$$

i.e., a consistent estimator.

Statistical properties

FOC results in the estimated parameter (here $\hat{\beta}$). Clearly,

$$0 = \frac{\partial Q_n(\hat{\beta})}{\partial \beta} = \frac{-2}{n}\sum_{i=1}^{n}\Big(y_i - f(x_i;\hat{\beta})\Big)\frac{\partial f(x_i;\hat{\beta})}{\partial \beta}.$$

Therefore, we need to use numerical methods to solve the nonlinear problems.

Mean value theorem

However, we can still estimate the nonlinear function by using the mean value theorem.

$$0 = \frac{\partial Q_n(\hat{\beta})}{\partial \beta} = \frac{\partial Q_n(\beta)}{\partial \beta} + \frac{\partial^2 Q_n(\beta_m)}{\partial \beta \partial \beta'}(\hat{\beta} - \beta),$$

where $\beta_m \in [\hat{\beta}, \beta]$. Since $\hat{\beta} \xrightarrow{p} \beta$, therefore, it gives $\beta_m \xrightarrow{p} \beta$. Moreover, re-writing the NLS problem as an asymptotic form gives

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{\partial^2 Q_n(\beta_m)}{\partial \beta \partial \beta'}\right)^{-1}\left(-\sqrt{n}\frac{\partial Q_n(\beta)}{\partial \beta}\right).$$

If

$$-\sqrt{n}\frac{\partial Q_n(\beta)}{\partial \beta} \xrightarrow{d} \mathcal{N}\left(0, \operatorname*{plim}_{n\to\infty}\left(n\frac{\partial Q_n(\beta)}{\partial \beta}\frac{\partial Q_n(\beta)}{\partial \beta'}\right)\right),$$

then the distribution asymptotes to

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \left(\operatorname*{plim}_{n\to\infty}\frac{\partial^2 Q_n(\beta)}{\partial \beta \partial \beta'}\right)^{-1}\left(\operatorname*{plim}_{n\to\infty} n\frac{\partial Q_n(\beta)}{\partial \beta}\frac{\partial Q_n(\beta)}{\partial \beta'}\right)\left(\operatorname*{plim}_{n\to\infty}\frac{\partial^2 Q_n(\beta)}{\partial \beta \partial \beta'}\right)^{-1}\right).$$

In addition,

$$\frac{\partial Q_n(\beta)}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial f_i(\beta)}{\partial \beta}e_i,$$

which leads to

$$\operatorname*{plim}_{n\to\infty}\left(n\frac{\partial Q(\beta)}{\partial \beta}\frac{\partial Q(\beta)}{\partial \beta'}\right) = \mathbb{E}\left[\frac{\partial f_i(\beta)}{\partial \beta}\frac{\partial f_i(\beta)'}{\partial \beta}\right] \quad \text{and} \quad \operatorname*{plim}_{n\to\infty}\frac{\partial^2 Q_n(\beta)}{\partial \beta \partial \beta'} = \mathbb{E}\left[\frac{\partial f_i(\beta)}{\partial \beta}\frac{\partial f_i(\beta)'}{\partial \beta}\right].$$

As a result, the asymptotic covariance of the NLS is

$$\sigma^2\left(\mathbb{E}\left[\frac{\partial f_i(\beta)}{\partial \beta}\frac{\partial f_i(\beta)}{\partial \beta'}\right]\right)^{-1}.$$