Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 9: Recap of OLS

Instructor: Ma, Jun

Renmin University of China

November 18, 2021

Linear causal model

- Suppose we have a random sample $\left\{\left(Y_i, X_i^\top\right) : i = 1, 2, ..., n\right\}$, where $X_i = \left(X_{i,1}, X_{i,2}, ..., X_{i,k}\right)$ with $k < n$. $X_{i,j}$: the j-th variable for the i-th observation. By convention, $X_{i,1} = 1$. Its coefficient corresponds to the intercept.
- Assume the data is i.i.d.: $\left(Y_i, X_i^\top\right)$ has the same distribution as $\left(Y_j, X_j^\top\right)$ and is independent of $\left(Y_j, X_j^\top\right)$, $\forall i \neq j$.
- Linear model: $Y = X^\top \beta + U$. X: observed explanatory variables; U: unobserved explanatory factor.
- $\left(Y_i, X_i^\top\right)$ is generated by the model: $Y_i = X_i^\top \beta + U_i$ for some U_i.
- Strong exogeneity: $\mathrm{E}\left[U \mid X\right] = 0$ (implies $\mathrm{E}\left[U\right] = 0$).
- Weak exogeneity: $\mathrm{E}\left[U\right] = \mathrm{E}\left[UX\right]$ (= $\mathrm{Cov}\left[U, X\right]$) = 0.
- OLS estimator of β:

$$\widehat{\beta} = \underset{b_1,...,b_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - b_1 X_{i,1} - b_2 X_{i,2} - \cdots - b_p X_{i,k}\right)^2 .$$

- We should give an interpretation of the linear part $X_i^\top \beta$ as a feature of the population (the distribution of $(Y, X^\top)$).
- Under strong exogeneity, $\mathrm{E}[Y \mid X] = X^\top \beta$.
- Under weak exogeneity, $X^\top \beta$ is the best linear approximation of $\mathrm{E}[Y \mid X]$: $\beta = (\mathrm{E}[XX^\top])^{-1} \mathrm{E}[XY]$ and

  $$\beta = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \mathrm{E}\left[\left(\mathrm{E}[Y \mid X] - X^\top b\right)^2\right].$$

  β is called projection coefficients.
- Homoskedastic model: $\mathrm{E}\left[U^2 \mid X\right] = \sigma^2 > 0$.
- Heteroskedastic model: $\mathrm{E}\left[U^2 \mid X\right]$ is a function of X.

Matrix notations

- We can stack these n equations together

$$
\begin{aligned}
Y_1 &= X_1^\top \beta + U_1 \\
Y_2 &= X_2^\top \beta + U_2 \\
\vdots \quad & \vdots \quad \vdots \\
Y_n &= X_n^\top \beta + U_n.
\end{aligned}
$$

- Define

$$
\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_3^\top \end{pmatrix}, \ \mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}.
$$

  $\mathbf{Y}$ and $\mathbf{U}$ are $n \times 1$ vectors and $\mathbf{X}$ is an $n \times k$ matrix. The (i, j) element of $\mathbf{X}$ is the i-th observation on the j-th regressor.
- The system of n equations can be written as $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$.
- No multicollinearity: $\text{rank}(\mathbf{X}) = k$.

- For a homoskedastic model,

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\beta + \mathbf{U} \\ \mathrm{E}[\mathbf{U} \mid \mathbf{X}] &= \mathbf{0} \\ \mathrm{Var}[\mathbf{U} \mid \mathbf{X}] &= \sigma^2 \mathbf{I}_n,\end{aligned}$$

where $\mathbf{I}_n$ denotes the n-dimensional identity matrix.

OLS in matrix notations

- The OLS estimator of β is obtained by solving $\widehat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^k} \|\mathbf{Y} - \mathbf{X}b\|$.
- Then, $\widehat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$ and the fitted residuals are $\widehat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\widehat{\beta}$.
- $\widehat{\mathbf{U}}$ satisfies $\mathbf{X}^\top \widehat{\mathbf{U}} = \mathbf{0}$.
- The OLS is unbiased: $\mathrm{E}\left[\widehat{\beta} \mid \mathbf{X}\right] = \beta$.
- Under homoskedasticity, $\mathrm{Var}\left[\widehat{\beta} \mid \mathbf{X}\right] = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.

Projection matrices

- Let $\mathbf{X}$ be $n \times k$ with rank $(\mathbf{X}) = k$. Then define $\boldsymbol{P}_{\mathbf{X}} = \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}$. $\boldsymbol{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{Y} = \mathbf{X}\widehat{\beta}$ gives the fitted values.
- The fitted residuals are

$$\widehat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = \mathbf{Y} - \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{Y} = \left(\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\right)\mathbf{Y}.$$

- We define $\boldsymbol{M}_{\mathbf{X}} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top} = \mathbf{I}_n - \boldsymbol{P}_{\mathbf{X}}$. $\boldsymbol{M}_{\mathbf{X}}\mathbf{Y}$ gives the fitted residuals.
- Properties of $\boldsymbol{P}_{\mathbf{X}}$ and $\boldsymbol{M}_{\mathbf{X}}$:
  - $\boldsymbol{P}_{\mathbf{X}}$ and $\boldsymbol{M}_{\mathbf{X}}$ are symmetric;
  - $\boldsymbol{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$ and $\boldsymbol{M}_{\mathbf{X}}\mathbf{X} = \mathbf{0}$;
  - $\boldsymbol{P}_{\mathbf{X}}$ and $\boldsymbol{M}_{\mathbf{X}}$ are orthogonal: $\boldsymbol{M}_{\mathbf{X}}\boldsymbol{P}_{\mathbf{X}} = \mathbf{0}$ and $\boldsymbol{P}_{\mathbf{X}}\boldsymbol{M}_{\mathbf{X}} = \mathbf{0}$;
  - $\boldsymbol{P}_{\mathbf{X}}$ and $\boldsymbol{M}_{\mathbf{X}}$ are idempotent: $\boldsymbol{P}_{\mathbf{X}}\boldsymbol{P}_{\mathbf{X}} = \boldsymbol{P}_{\mathbf{X}}$ and $\boldsymbol{M}_{\mathbf{X}}\boldsymbol{M}_{\mathbf{X}} = \boldsymbol{M}_{\mathbf{X}}$;
  - rank $(\boldsymbol{P}_{\mathbf{X}}) = k$ and rank $(\boldsymbol{M}_{\mathbf{X}}) = n - k$.

Partitioned regression

- Partition $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$, $\beta = \left(\beta_1^\top, \beta_2^\top\right)^\top$ and the model as

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{U},$$

  where $\mathbf{X}_1$ is $n \times k_1$ and $\mathbf{X}_2$ is $n \times k_2$ ($k_1 + k_2 = k$).
- Partition $\widehat{\beta} = \left(\widehat{\beta}_1^\top, \widehat{\beta}_2^\top\right)^\top$.
- Denote $\boldsymbol{M}_1 = \boldsymbol{M}_{\mathbf{X}_1}$ and $\boldsymbol{M}_2 = \boldsymbol{M}_{\mathbf{X}_2}$. Then,

$$\begin{aligned} \widehat{\beta}_1 &= \left(\mathbf{X}_1^\top \boldsymbol{M}_2 \mathbf{X}_1\right)^{-1} \left(\mathbf{X}_1^\top \boldsymbol{M}_2 \mathbf{Y}\right) \\ \widehat{\beta}_2 &= \left(\mathbf{X}_2^\top \boldsymbol{M}_1 \mathbf{X}_2\right)^{-1} \left(\mathbf{X}_2^\top \boldsymbol{M}_1 \mathbf{Y}\right) \end{aligned}$$

  and

$$\begin{aligned} \operatorname{Var}\left[\widehat{\beta}_1 \mid \mathbf{X}\right] &= \sigma^2 \left(\mathbf{X}_1^\top \boldsymbol{M}_2 \mathbf{X}_1\right)^{-1} \\ \operatorname{Var}\left[\widehat{\beta}_2 \mid \mathbf{X}\right] &= \sigma^2 \left(\mathbf{X}_2^\top \boldsymbol{M}_1 \mathbf{X}_2\right)^{-1}. \end{aligned}$$

Omitted variable bias

- Suppose the researcher estimates β_1 by regressing $\mathbf{Y}$ on $\mathbf{X}_1$ only. Let $\widetilde{\beta}_1 = \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1^\top \mathbf{Y}\right)$ denote the OLS estimates.
- Then,

$$\begin{aligned}\widetilde{\beta}_1 &= \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1^\top \mathbf{Y}\right) \\ &= \beta_1 + \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \beta_2 + \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{U}\end{aligned}$$

  and $\mathrm{E}\left[\widetilde{\beta}_1 \mid \mathbf{X}\right] = \beta_1 + \left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \beta_2$.
- $\left(\mathbf{X}_1^\top \mathbf{X}_1\right)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \beta_2$ is the omitted variable bias.

Effects of covariates

- In practical applications, we often have a long list of potential explanatory variables. It is possible that k is close to n.
- Cross-country growth regression estimates the effect of initial GDP on future growth rates, with more than 50 other explanatory variable including institutional and technological factors and a sample of less than 100 observations.
- In addition, to capture the nonlinear effects and interaction effects, we may expand the linear model by incorporating higher order polynomials and interaction terms.
- While only few of the potential covariates may have non-zero coefficients in the true model, unfortunately we do not know which ones.
- Covariates with zero coefficients are called irrelevant.
- To avoid the omitted variables bias, the researcher may attempt to include all potential covariates. Unfortunately, that results in large variances and standard errors on the main parameters of interest.

- Partition the regression model:

$$\mathbf{Y} = \beta_1 \mathbf{X}_1 + \mathbf{X}_2 \beta_2 + \mathbf{U},$$

  where $\mathbf{X}_1$ is an $n \times 1$ vector which contains the observations on the main explanatory variable for research.
- $\mathbf{X}_2$ is an $n \times (k-1)$ matrix which includes observations on $k-1$ other potential explanatory variables (control variables).
- The variance of the OLS estimator:

$$\mathrm{Var}\left[\widehat{\beta}_1 \mid \mathbf{X}\right] = \frac{\sigma^2}{\mathbf{X}_1^\top \boldsymbol{M}_2 \mathbf{X}_1}.$$

- Since $\mathbf{X}_1^\top \boldsymbol{M}_2 \mathbf{X}_1 = \mathbf{X}_1^\top \boldsymbol{M}_2^\top \boldsymbol{M}_2 \mathbf{X}_1 = \widetilde{\mathbf{X}}_1^\top \widetilde{\mathbf{X}}_1$, where

$$\widetilde{\mathbf{X}}_1 = \boldsymbol{M}_2 \mathbf{X}_1 = \mathbf{X}_1 - \mathbf{X}_2 \left(\mathbf{X}_2^\top \mathbf{X}_2\right)^{-1} \mathbf{X}_2^\top \mathbf{X}_1 = \mathbf{X}_1 - \mathbf{X}_2 \widehat{\gamma}.$$

- $\widehat{\gamma}$ is the OLS coefficient from the regression of $\mathbf{X}_1$ against $\mathbf{X}_2$.
- $\widetilde{\mathbf{X}}_1$ is the vector of OLS residuals from OLS regression of $\mathbf{X}_1$ against $\mathbf{X}_2$ and $\widetilde{\mathbf{X}}_1^\top \widetilde{\mathbf{X}}_1$ is the sum of the squared residuals.
- When we include more control variables, a bigger portion of $\mathbf{X}_1$ is removed resulting in a smaller sum of the squared residuals.
- When we include irrelevant control variables, the variance of the OLS estimator increases. One would see larger standard errors, smaller t-statistics, larger p-values and wider confidence intervals.
- Two wrong practices: (1) include only significant regressors; (2) data snooping/p-hacking.

Include only significant regressors?

- If a subset of the coefficients in the linear model

  $$Y_i = \beta_1 X_{i,1} + \ldots + \beta_k X_{i,k} + U_i$$

  are exactly zero, we wish to find the smallest sub-model consisting of only explanatory variables with non-zero coefficients.
- Estimate the full model with all variables. Let $T_j = \widehat{\beta}_j / SE\left(\widehat{\beta}_j\right)$ denote the t-statistic for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.
- What if we run a second regression with only statistically significant coefficients in the first stage?
- Such a practice would typically result in exclusion of relevant covariates and the omitted variables bias.
  - Hypothesis testing controls for the probability of Type I error but leaves the probability of Type II error uncontrolled.
  - You find a coefficient to be non-significant, possibly due to a high probability of Type II error.
  - Failure to reject $H_0 : \beta_j = 0$ cannot be used as a reliable evidence that the true coefficient is zero.

Data snooping

- Data snooping or p-hacking occurs when the researcher uses the same data in order to produce statistically significant estimates with large t-statistics or small p-values.
- Data snooping destroys the validity of t-statistics and p-values and makes the empirical results less convincing.
- You may try dropping different combinations of potential explanatory variables from the regression to get a statistically significant estimate for the main variable of interest.
- Suppose that the researcher can construct J independent estimators for θ such that $\widehat{\theta}_j \sim \mathrm{N}\left(\theta, \sigma_j^2\right)$, $j = 1, 2, ..., J$, where σ_j^2 is known.
- The researcher conducts J tests with significance level 5% of $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$.

- The researcher concludes that $\theta \neq 0$ if one of the J tests rejects $\theta = 0$.
- Suppose that in fact $\theta = 0$. The probability of concluding that $\theta \neq 0$ (known as false discovery) is given by

$$
\begin{aligned}
\Pr\left[\max_{1 \le j \le J} \left|\frac{\widehat{\theta}_j}{\sigma_j}\right| > 1.96\right] &= 1 - \Pr\left[\max_{1 \le j \le J} \left|\frac{\widehat{\theta}_j}{\sigma_j}\right| \le 1.96\right] \\
&= 1 - \prod_{i=1}^{J} \Pr\left[\left|\frac{\widehat{\theta}_j}{\sigma_j}\right| \le 1.96\right] \\
&= 1 - (0.95)^J .
\end{aligned}
$$

- The false discovery probability quickly grows as $J \uparrow \infty$. E.g., $1 - (0.95)^{10} \approx 40\%$.
- When the researcher performs many of tests, the Type I error probability is not controlled and may be much larger than the nominal significance level.

- In practice, estimators are rarely independent, the same relationship holds qualitatively.
- If the researcher searchers long enough, with a high probability they would find a significant estimate.
- A procedure that automatically detects the smallest sub-model consisting of only relevant explanatory variables guards against data snooping and makes the empirical results more convincing to readers.

One classical approach to model selection

- Order $T_1, ..., T_k$ in absolute value:

$$\left|T_{(1)}\right| \geq \left|T_{(2)}\right| \geq \cdots \geq \left|T_{(k)}\right|.$$

- Let $\hat{j}$ denote the value of j that minimizes $RSS\,(j) + j s^2 \log\,(n)$, where $RSS\,(j)$ is the residual sum of squares from the model with j variables corresponding to the j largest absolute t-statistics and $s^2 = (n-k)^{-1} \sum_{i=1}^{n} \widehat{U}_i^2$.
- The selected model is the model with $\hat{j}$ variables corresponding to the $\hat{j}$ largest absolute t-statistics.
- When n is large, with high probability, this selected model is the same as the smallest sub-model with only nonzero coefficients.
- Disadvantages:
  - Assume homoskedasticity;
  - Break down in high-dimensional regression $k > n$ ($s^2 = 0$).

Convergence in probability

- Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random variables. Let X be random or non-random.
- We will consider non-random sequences with the following typical elements: 1. $\mathrm{E}\left[|X_n - X|^r\right]$; 2. $\Pr\left[|X_n - X| > \varepsilon\right]$ for some $\varepsilon > 0$.
  - Convergence in r-th mean. X_n converges to X in r-th mean if $\mathrm{E}\left[|X_n - X|^r\right] \to 0$ as $n \to \infty$.
  - Convergence in probability. X_n converges in probability to X if for all $\varepsilon > 0$, $\Pr\left[|X_n - X| \geq \varepsilon\right] \to 0$ as $n \to \infty$. It is denoted as $X_n \to_p X$. If $X_n \to_p 0$, we denote $X_n = o_p(1)$.

- Convergence in r-th mean implies convergence in probability.
- (Markov's Inequality) Let X be a random variable. For $\varepsilon > 0$ and $r > 0$,
$$\Pr[|X| \geq \varepsilon] \leq \frac{\mathrm{E}[|X|^r]}{\varepsilon^r}.$$
- Suppose that X_n converges to X in r-th mean, $\mathrm{E}[|X_n - X|^r] \to 0$. Then,
$$\begin{aligned} \Pr[|X_n - X| \geq \varepsilon] &\leq \frac{\mathrm{E}[|X_n - X|^r]}{\varepsilon^r} \\ &\to 0. \end{aligned}$$
- Let $X_1, \ldots X_n$ be a sample of i.i.d. random variables such that $\mathrm{E}[|X_1|] < \infty$. Then, $n^{-1} \sum_{i=1}^n X_i \to_p \mathrm{E}[X_1]$ as $n \to \infty$.
- Due to i.i.d. assumption, we have that $\mathrm{E}[X_i] = \mathrm{E}[X_1]$ for all $i = 1, \ldots, n$.

Suppose that $X_n \to_p a$ and $Y_n \to_p b$, where a and b are some finite constants. Let c be another constant.

- $cX_n \to_p ca$.
- $X_n + Y_n \to_p a + b$.
- $X_n Y_n \to_p ab$.
- $X_n / Y_n \to_p a/b$, provided that $b \neq 0$.
- If $0 \leq X_n \leq Y_n$ and $Y_n \to_p 0$, then $X_n \to_p 0$.
- $X_n \to_p 0$ if and only if $|X_n| \to_p 0$.

Continuous mapping theorem (CMT)

- Suppose that $X_n \rightarrow_p c$, a constant, and let $h(\cdot)$ be a continuous function at c. Then, $h(X_n) \rightarrow_p h(c)$.
- suppose that $\widehat{\beta}_n \rightarrow_p \beta$. Then $\widehat{\beta}_n^2 \rightarrow_p \beta^2$, and $1/\widehat{\beta}_n \rightarrow_p 1/\beta$, provided $\beta \neq 0$.

Convergence of random vectors

- The random vectors/matrices converge in probability if their elements converge in probability.
- Consider the vector case. Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random k-vectors. $X_n - X \rightarrow_p 0$ element-by-element, where X is a possibly random k-vector, if and only if $\|X_n - X\| \rightarrow_p 0$, where $\|\cdot\|$ denotes the Euclidean norm.
- The rules for manipulation of probability limits in the vector/matrix case are similar to those in the scalar case.
- The CMT is valid in vector/matrix case as well.
- OLS estimator is a consistent estimator of the coefficients: $\widehat{\beta} \rightarrow_p \beta$.

Convergence in distribution

- Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random variables.
- Let $F_n(x)$ denote the marginal CDF of X_n, i.e. $F_n(x) = \Pr(X_n \leq x)$. Let $F(x)$ be another CDF.
- We say that X_n converges in distribution if $F_n(x) \to F(x)$ for all x where $F(x)$ is continuous.
- In this case, we write $X_n \to_d X$, where X is any random variable with the distribution function $F(x)$.
- Note that while we say that X_n converges to X, the convergence in distribution is not convergence of random variables, but of the distribution functions.

- The extension to the vector case is straightforward. Let X_n and X be two random k-vectors.
- We say that $X_n \rightarrow_d X$ if the joint CDF of X_n converges to that of X at all continuity points, i.e.

$$\begin{aligned} F_n(x_1, \ldots, x_k) &= \Pr\left[X_{n,1} \leq x_1, \ldots, X_{n,k} \leq x_k\right] \\ &\rightarrow \Pr\left[X_1 \leq x_1, \ldots, X_k \leq x_k\right] \\ &= F(x_1, \ldots, x_k), \end{aligned}$$

  for all points $(x_1, \ldots, x_k)$ where F is continuous.
- In this case, we say that the elements of X_n, $X_{n,1}, \ldots X_{n,k}$, jointly converge in distribution to $X_1, \ldots X_k$, the elements of X.

Rules of convergence in distribution

- (Cramer Convergence Theorem) Suppose that $X_n \rightarrow_d X$ and $Y_n \rightarrow_p c$. Then,
  - $X_n + Y_n \rightarrow_d X + c$.
  - $Y_n X_n \rightarrow_d cX$.
  - $X_n / Y_n \rightarrow_d X/c$, provided that $c \neq 0$.
- If $X_n \rightarrow_p X$, then $X_n \rightarrow_d X$. Converse is not true with one exception: If $X_n \rightarrow_d c$, a constant, then $X_n \rightarrow_p c$.
- If $X_n - Y_n \rightarrow_p 0$, and $Y_n \rightarrow_d Y$, then $X_n \rightarrow_d Y$.

Continuous mapping theorem

- Suppose that $X_n \to_d X$, and let $h(\cdot)$ be a function continuous on a set $\mathcal{X}$ such that $\Pr[X \in \mathcal{X}] = 1$. Then, $h(X_n) \to_d h(X)$.
- Note that contrary to convergence in probability, $X_n \to_d X$ and $Y_n \to_d Y$ does not imply that, for example, $X_n + Y_n \to_d X + Y$, unless a joint convergence result holds.

The central limit theorem

- Let $X_1, \ldots, X_n$ be a sample of iid random variables such that $\mathrm{E}[X_1] = 0$ and $0 < \mathrm{E}\left[X_1^2\right] < \infty$. Then, as $n \to \infty$, $n^{-1/2} \sum_{i=1}^n X_i \to_d \mathrm{N}\left(0, \mathrm{E}\left[X_1^2\right]\right)$.
- Let $X_1, \ldots X_n$ be a sample of iid random variables with $\mathrm{E}[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2 < \infty$. Define

$$\overline{X}_n = n^{-1} \sum_{i=1}^n X_i.$$

- Consider $n^{-1/2} \sum_{i=1}^n (X_i - \mu)$. We have that $(X_1 - \mu), \ldots, (X_n - \mu)$ are i.i.d. with the mean $\mathrm{E}[(X_1 - \mu)] = 0$, and the variance $\mathrm{E}\left[(X_1 - \mu)^2\right] = \sigma^2 < \infty$. Therefore, by the CLT,

$$\begin{aligned} n^{1/2}\left(\overline{X}_n - \mu\right) &= n^{-1/2} \sum_{i=1}^n (X_i - \mu) \\ &\to_d \mathrm{N}\left(0, \sigma^2\right). \end{aligned}$$

- Let X_n be a random k-vector. Then, $X_n \to_d X$ if and only if $\lambda^\top X_n \to_d \lambda^\top X$ for all non-zero $\lambda \in \mathbb{R}^k$.
- Let $X_1, \ldots, X_n$ be a sample of i.i.d. random k-vectors such that $\mathrm{E}[X_1] = 0$ (denote $X_i = \left(X_{i,1}, ..., X_{i,k}\right)^\top$) and $\mathrm{E}\left[X_{1,j}^2\right] < \infty$ for all $j = 1, \ldots, k$, and $\mathrm{E}\left[X_1 X_1^\top\right]$ is positive definite. Then, $n^{-1/2} \sum_{i=1}^n X_i \to_d \mathrm{N}\left(0, \mathrm{E}\left[X_1 X_1^\top\right]\right)$.

Asymptotic normality of OLS

- Denote $\mathbf{V} = \mathrm{E}\left[U_i^2 X_i X_i^\top\right]$ and $\mathbf{G} = \mathrm{E}\left[X_i X_i^\top\right]$. Then,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i U_i\right)$$

  and $n^{-1}\sum_{i=1}^{n} X_i X_i^\top \to_p \mathbf{G}$ and $n^{-1/2}\sum_{i=1}^{n} X_i U_i \to_d \mathrm{N}(\mathbf{0}, \mathbf{V})$.
- Then,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \to_d \mathrm{N}\left(\mathbf{0}, \mathbf{G}^{-1}\mathbf{V}\mathbf{G}^{-1}\right).$$

- In the homoskedastic model, $\mathbf{V} = \sigma^2\mathbf{G}$ and $\mathbf{G}^{-1}\mathbf{V}\mathbf{G}^{-1} = \sigma^2\mathbf{G}^{-1}$.

Bounded in probability

- Suppose that $\lambda_n = \sqrt{n}\left(\widehat{\theta} - \theta\right) \to_d \mathrm{N}\left(0, \sigma^2\right)$. We say that the sequence $\{\lambda_n\}_{n=1}^{\infty}$ is bounded in probability and denote $\lambda_n = O_p(1)$.
- Suppose that $\xi_n \to_p 0$ ($\xi_n = o_p(1)$). Then, $\xi_n \lambda_n = o_p(1)\, O_p(1) = o_p(1)$.
- We also write

  $$\widehat{\theta} = \theta + \frac{1}{\sqrt{n}} \cdot \lambda_n = \theta + \frac{1}{\sqrt{n}} \cdot O_p(1) = \theta + O_p\left(\frac{1}{\sqrt{n}}\right).$$

  $\widehat{\theta}$ converges to θ at the rate $n^{-1/2}$.
- More generally, we write $X_n = O_p(\alpha_n)$ for some non-random sequence α_n, if $X_n / \alpha_n = O_p(1)$.