Introductory Econometrics

Lecture 27: Resampling Methods

Instructor: Ma, Jun

Renmin University of China

December 16, 2021

Asymptotic normality

- In previous lectures, we have so many estimators with the property

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to_d \mathrm{N}\left(0, \sigma^2\right)$$

  and equivalently we can write $\hat{\theta}_n \overset{a}{\sim} \mathrm{N}\left(\theta, \sigma^2/n\right)$.
- We use $\mathrm{N}\left(\theta, \sigma^2/n\right)$ as approximation to the unknown true (often called finite-sample) distribution of $\hat{\theta}_n$.
- To estimate σ^2 based on the analogue principle (i.e., replace population moments/unknown quantities by their sample moments/estimates), we need knowledge of the expression(formula) of σ^2. Very often the expression is very complicated.
- There are two computation-intensive approaches that do the estimation without requiring knowledge of the expression of σ^2.

Jackknife standard errors

- Consider a sample with $i = 1, ..., n$ independent observations of an explained variable Y and k explanatory variables $X_1, ..., X_k$. Suppose our data is $(Y_i, X_{i1}, ..., X_{ik})$ for $i = 1, ..., n$. Denote $Z_i = (Y_i, X_{i1}, ..., X_{ik})$.
- Suppose the estimator $\hat{\theta}$ can be written as $\hat{\theta}_n = \varphi_n(Z_1, ..., Z_n)$, e.g., $\varphi_n(z_1, ..., z_n) = n^{-1} \sum_{i=1}^{n} z_i$.
- Now denote $\hat{\theta}_{-j} = \varphi_{n-1}\left(Z_1, ...Z_{j-1}, Z_{j+1}, ..., Z_n\right)$, i.e., $\hat{\theta}_{-j}$ is an estimator obtained by removing the j-th observation from the entire sample. The variation in $\left\{\hat{\theta}_{-j} : j = 1, ..., n\right\}$ should be informative about the population variance of $\hat{\theta}_n$.
- Denote $\overline{\hat{\theta}} = n^{-1} \sum_{j=1}^{n} \hat{\theta}_{-j}$. The Jackknife standard error is

$$\widehat{se}_{jk} = \sqrt{\frac{n-1}{n} \sum_{j=1}^{n} \left(\hat{\theta}_{-j} - \overline{\hat{\theta}}\right)^2}.$$

- A 95% confidence interval is $\left[\hat{\theta}_n - 1.96 \cdot \widehat{se}_{jk}, \hat{\theta}_n + 1.96 \cdot \widehat{se}_{jk}\right]$.

- Indeed one can show

$$(n-1)\sum_{j=1}^{n}\left(\hat{\theta}_{-j}-\overline{\hat{\theta}}\right)^{2}\to_{p}\sigma^{2}.$$

- Consider the following simple example: for i.i.d. random variables $X_1, ..., X_n$, we use the sample average $\overline{X}$ as an estimator of $\mu = \mathrm{E}[X_1]$. It is known that $\sqrt{n}\left(\overline{X}-\mu\right)\to_{d}\mathrm{N}\left(0,\sigma^{2}\right)$, where $\sigma^2 = \mathrm{Var}(X_1)$ in this case.
- For this case,

$$\hat{\theta}_{-j}=\frac{1}{n-1}\left(n\overline{X}-X_{j}\right),$$

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{-j} =& \frac{1}{n(n-1)}\sum_{j=1}^{n}\left(n\overline{X}-X_{j}\right)\\ =& \overline{X}.\end{aligned}$$

- For this simple case,

$$\hat{\theta}_{-j} - \overline{\hat{\theta}} = \frac{1}{n-1}\left(n\overline{X} - X_j\right) - \overline{X} = \frac{1}{n-1}\left(\overline{X} - X_j\right).$$

- We have

$$(n-1)\sum_{j=1}^{n}\left(\hat{\theta}_{-j} - \overline{\hat{\theta}}\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2,$$

  which is the sample variance that is a consistent and unbiased estimator for σ^2.

Bootstrap

- The bootstrap takes the sample (the values of the realized explanatory and explained variables) as the population.
- The bootstrap is an alternative way to produce approximations for the true distribution of $\hat{\theta}_n$.
- Note that both asymptotic theory and the bootstrap only provide approximations for finite-sample properties.
- A bootstrap sample is obtained by independently drawing n pairs $(Y_i, X_{1i}, ..., X_{ik})$ from the observed sample with replacement.
- The bootstrap sample has the same number of observations as the original sample, however some observations appear several times and others never.

Bootstrap Standard Errors

- Step 1: Draw B independent bootstrap samples. B can be as large as possible. We can take $B = 1000$.
- Step 2: Estimate θ with each of the bootstrap samples, $\hat{\theta}_b^*$ for $b = 1, \ldots, B$.
- Step 3: Estimate the standard deviation of $\hat{\theta}$ by

$$\widehat{se}_{bs} \equiv \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}^* \right)^2}$$

  where $\hat{\theta}^* = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b^*$.
- Step 4: The bootstrap standard errors can be used to construct approximate confidence intervals and to perform asymptotic tests based on the normal distribution, e.g. if the coverage probability is 95%, a 95% confidence interval is $\left[\hat{\theta}_n - 1.96 \cdot \widehat{se}_{bs}, \hat{\theta}_n + 1.96 \cdot \widehat{se}_{bs} \right]$.

Bootstrap percentile confidence intervals

- Step 1: Draw B independent bootstrap samples. B can be as large as possible. We can take $B = 1000$.
- Step 2: Estimate θ with each of the bootstrap samples, $\hat{\theta}_b^*$ for $b = 1, \ldots, B$.
- Step 3: Order the bootstrap replications such that

$$\hat{\theta}_1^* \leq \cdots \leq \hat{\theta}_B^*.$$

- Step 4: The lower and upper confidence bounds are $B(\alpha/2)$-th and $B(1-\alpha/2)$-th ordered elements. For $B = 1000$ and $\alpha = 5\%$, these are the 25th and 975th ordered elements. The estimated $1-\alpha$ confidence interval is $\left[\hat{\theta}_{B(\alpha/2)}^*, \hat{\theta}_{B(1-\alpha/2)}^*\right]$.
- Bootstrap percentile confidence intervals often have more accurate coverage probabilities (i.e. closer to the nominal coverage probability $1-\alpha$) than the usual confidence intervals based on standard normal quantiles and estimated variance.

Bootstrap-t test

- We consider testing $H_0 : \theta = \theta_0$.
- We can conduct a bootstrap-based hypothesis testing based on the bootstrap percentile confidence interval: we simply reject H_0 if θ_0 fails to be an element of the bootstrap percentile confidence interval.
- We can show that $T = \sqrt{n}\left(\hat{\theta} - \theta_0\right)/\hat{\sigma} \to_d N(0, 1)$ under H_0. We use the standard normal distribution as approximation to the true distribution of T and define critical values based on standard normal quantile.
- For each bootstrap sample $b = 1, ..., B$, we can calculate $\hat{\sigma}^*$ using the bootstrap sample.

- Step 1: Draw B independent bootstrap samples. B can be as large as possible. We can take $B = 1000$.
- Step 2: Estimate θ and σ with each of the bootstrap samples, $\hat{\theta}_b^*$, $\hat{\sigma}_b^*$ for $b = 1, ..., B$ and the t-value for each bootstrap sample:

$$t_b^* = \frac{\sqrt{n}\left(\hat{\theta}_b^* - \hat{\theta}\right)}{\hat{\sigma}_b^*}$$

  Notice that $\hat{\theta}$ is used instead of θ_0 in the construction.
- Step 3: Order the bootstrap replications of t such that $t_1^* \leq \cdots \leq t_B^*$. The lower critical value and the upper critical value are then the $B(\alpha/2)$-th and $B(1-\alpha/2)$-th ordered elements. For $B = 1000$ and $\alpha = 5\%$, these are the 25th and 975th ordered elements. The bootstrap lower and upper critical values generally differ in absolute values.

- In finite samples (fixed n), for neither the bootstrap-t test nor the usual t-test that uses ± 1.96 as critical values , the true probability of making type-I error is exactly equal to α (e.g., 0.05).
- In almost all cases, the true probability of making type-I error is greater than α, i.e., we always "over-reject" the null hypothesis.
- One can show that for bootstrap-t test, in finite samples, the true probability of making type-I error is closer to the nominal significance level α than the standard t-test that uses ± 1.96 as critical values.

- In finite samples (fixed n), for neither the bootstrap-t test nor the usual t-test that uses ± 1.96 as critical values , the true probability of making type-I error is exactly equal to α (e.g., 0.05).
- In almost all cases, the true probability of making type-I error is greater than α, i.e., we always "over-reject" the null hypothesis.
- One can show that for bootstrap-t test, in finite samples, the true probability of making type-I error is closer to the nominal significance level α than the standard t-test that uses ± 1.96 as critical values.

Why does the bootstrap work?

- Suppose $X_1, \ldots, X_n$ is our random sample and we have an estimator $\hat{\theta}$ of some parameter θ. Notice that we can write $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ as a function of the data.
- The bootstrap sample $X_1^*, \ldots, X_n^*$ can be viewed as a new (i.i.d.) random sample such that the marginal distribution of X_i^* is the discrete distribution with $X_i^* = X_j$ with probability $1/n$, for $j = 1, \ldots, n$.
- Notice that conditionally on $X_1, \ldots, X_n$ being observed, we draw X_i^*, $i = 1, \ldots, n$. Therefore, we can write

$$\Pr\left[X_i^* = X_j \mid X_1, \ldots, X_n\right] = \frac{1}{n}, \text{ for } j = 1, \ldots, n.$$

- Let $F_n(t) = \Pr\left[\sqrt{n}\left(\hat{\theta} - \theta\right) \le t\right]$ be the distribution function of $\sqrt{n}\left(\hat{\theta} - \theta\right)$. If we knew F_n, we could easily construct a confidence interval

$$\left[\hat{\theta} - \frac{t_{1-\alpha/2}}{\sqrt{n}}, \hat{\theta} - \frac{t_{\alpha/2}}{\sqrt{n}}\right],$$

  where t_α is the α-quantile of F_n: $t_\alpha = F_n^{-1}(\alpha)$.
- In reality, we do not know F_n and we can often show that F_n can be approximated by the distribution function of some centralized normal random variable $\mathrm{N}\left(0, \sigma^2\right)$.
- The normal approximation with $\mathrm{N}\left(0, \sigma^2\right)$ requires that σ^2 can be estimated consistently.

- Consider an alternative approximation, the conditional distribution

$$\hat{F}_n(t) = \Pr\left[\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right) \leq t \mid X_1, \ldots, X_n\right],$$

  where $\hat{\theta}^*$ is the "bootstrap analogue" of $\hat{\theta}$, i.e., $\hat{\theta}^* = \hat{\theta}\left(X_1^*, \ldots, X_n^*\right)$.
- Notice that $\hat{F}_n$ is known to us since the distribution of the bootstrap sample is known. $\hat{F}_n$ can be approximated by computer simulations.

A simple example

- Suppose X_i has mean μ and variance σ^2. We want to construct a confidence interval for μ.
- Let $\hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i$ and $F_n(t) = \Pr\left[\sqrt{n}(\hat{\mu} - \mu) \leq t\right]$. The central limit theorem implies that F_n is approximately Φ_σ, the CDF of a $\mathrm{N}\left(0, \sigma^2\right)$ random variable.
- We want to show that

$$\hat{F}_n(t) = \Pr\left[\sqrt{n}(\hat{\mu}^* - \hat{\mu}) \leq t \mid X_1, \ldots, X_n\right]$$

is close to F_n.

Berry-Esseen theorem

- Berry-Esseen Theorem: Let $X_1, ..., X_n$ be i.i.d. with mean μ and variance σ^2. Denote $\mu_3 = \mathrm{E}\left[|X_1 - \mu|^3\right]$. Let $Z_n = \sqrt{n}\left(\overline{X}_n - \mu\right)$. Then
$$\max_{t \in \mathbb{R}} |\Pr[Z_n \le t] - \Phi_\sigma(t)| \le \frac{33}{4} \frac{\mu_3}{\sigma^3 \sqrt{n}}.$$
- Berry-Esseen Theorem is a refinement of the CLT, which only gives the conclusion that $\Pr[Z_n \le t] - \Phi_\sigma(t) \to 0$ as $n \to \infty$.

- Let $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$. It is true but somewhat hard to see that $\hat{\sigma}^2$ is the "population" conditional variance of X_i^* given $X_1, ..., X_n$. Let $\hat{\mu}_3 = n^{-1} \sum_{i=1}^{n} |X_i - \hat{\mu}|^3$.
- Now by the triangle inequality,

$$\max_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F_n(t) \right| \leq \max_{t \in \mathbb{R}} |F_n(t) - \Phi_\sigma(t)| + \max_{t \in \mathbb{R}} |\Phi_\sigma(t) - \Phi_{\hat{\sigma}}(t)| + \max_{t \in \mathbb{R}} \left| \hat{F}_n(t) - \Phi_{\hat{\sigma}}(t) \right|.$$

- The Berry-Esseen Theorem implies that

$$\max_{t \in \mathbb{R}} |F_n(t) - \Phi_\sigma(t)| \leq \frac{33}{4} \frac{\mu_3}{\sigma^3 \sqrt{n}}.$$

- Since $\hat{\sigma}^2 \to_p \sigma^2$, it can be shown

$$\max_{t \in \mathbb{R}} |\Phi_{\sigma}(t) - \Phi_{\hat{\sigma}}(t)| \to_p 0.$$

- The magic is that Berry-Esseen theorem can be applied to the last term:

$$\max_{t \in \mathbb{R}} \left|\hat{F}_n(t) - \Phi_{\hat{\sigma}}(t)\right| \le \frac{33}{4} \frac{\hat{\mu}_3}{\hat{\sigma}^3 \sqrt{n}}.$$

- Notice that $\hat{\mu}_3 \to_p \mu_3 > 0$ and $\hat{\sigma} \to_p \sigma > 0$. So we have

$$\max_{t \in \mathbb{R}} \left|\hat{F}_n(t) - \Phi_{\hat{\sigma}}(t)\right| \to_p 0.$$

  This implies $\max_{t \in \mathbb{R}} \left|\hat{F}_n(t) - F_n(t)\right| \to_p 0$. F_n, which is unknown, can be well-approximated by $\hat{F}_n$, which is known given the data.

STATA implementation

- In STATA, we can use the command

  bootstrap, reps(###): stata command

  The number ### specifies the number of bootstrap replications (B). For example, "bootstrap, reps(100): regress y x".
- This command can be applied to instrumental variable estimation, binary choice models, multinomial choice models, censored regression, the treatment effect estimator...
- We can use a post estimation command "estat bootstrap, percentile" to ask STATA to report bootstrap percentile confidence intervals for the parameters.