Intro to Econometrics: Recitation 10

Gustavo Pereira

December 9, 2019

This recitation

- This recitation is about boostrapping; based on Christoph Rothe's lecture notes
- Also our last meeting: if you haven't done so already, please submit a course evaluation

Notation

- Sample $\{x_1, \ldots, x_n\}$ drawn iid from an unknown CDF $F_0 \in \mathcal{F}$

Notation

- Sample $\{x_1, \ldots, x_n\}$ drawn iid from an unknown CDF $F_0 \in \mathcal{F}$
- Statistic: $T_n = T_n(x_1, \ldots, x_n; F_0)$

Notation

- Sample $\{x_1, \ldots, x_n\}$ drawn iid from an unknown CDF $F_0 \in \mathcal{F}$
- Statistic: $T_n = T_n(x_1, \ldots, x_n; F_0)$
  - For example:

$$T_n = \sqrt{n}\left(\frac{\bar{x}_n - \int x dF_0(x)}{\hat{\sigma}}\right)$$

Notation

- Sample $\{x_1, \ldots, x_n\}$ drawn iid from an unknown CDF $F_0 \in \mathcal{F}$
- Statistic: $T_n = T_n(x_1, \ldots, x_n; F_0)$
  - For example:

$$T_n = \sqrt{n}\left(\frac{\bar{x}_n - \int x dF_0(x)}{\hat{\sigma}}\right)$$

- We're interested in the CDF of T_n. We call it G_n:

$$G_n(u, F_0) = \mathbf{P}_{F_0}(T_n \leq u)$$

Notation

- Sample $\{x_1, \ldots, x_n\}$ drawn iid from an unknown CDF $F_0 \in \mathcal{F}$
- Statistic: $T_n = T_n(x_1, \ldots, x_n; F_0)$
  - For example:

$$T_n = \sqrt{n} \left(\frac{\bar{x}_n - \int x dF_0(x)}{\hat{\sigma}} \right)$$

- We're interested in the CDF of T_n. We call it G_n:

$$G_n(u, F_0) = \mathbf{P}_{F_0}(T_n \leq u)$$

- The statistic is *pivotal* if $G_n(u, F)$ does not depend on $F \in \mathcal{F}$

Notation

- Sample $\{x_1, \ldots, x_n\}$ drawn iid from an unknown CDF $F_0 \in \mathcal{F}$
- Statistic: $T_n = T_n(x_1, \ldots, x_n; F_0)$
  - For example:

$$T_n = \sqrt{n}\left(\frac{\bar{x}_n - \int x dF_0(x)}{\hat{\sigma}}\right)$$

- We're interested in the CDF of T_n. We call it G_n:

$$G_n(u, F_0) = \mathbf{P}_{F_0}(T_n \leq u)$$

- The statistic is *pivotal* if $G_n(u, F)$ does not depend on $F \in \mathcal{F}$
  - Our example T_n above is pivotal if $\mathcal{F}$ is the set of normal distributions (why?)

Notation

- Sample $\{x_1, \ldots, x_n\}$ drawn iid from an unknown CDF $F_0 \in \mathcal{F}$
- Statistic: $T_n = T_n(x_1, \ldots, x_n; F_0)$
  - For example:

$$T_n = \sqrt{n}\left(\frac{\bar{x}_n - \int x dF_0(x)}{\hat{\sigma}}\right)$$

- We're interested in the CDF of T_n. We call it G_n:

$$G_n(u, F_0) = \mathbf{P}_{F_0}(T_n \leq u)$$

- The statistic is *pivotal* if $G_n(u, F)$ does not depend on $F \in \mathcal{F}$
  - Our example T_n above is pivotal if $\mathcal{F}$ is the set of normal distributions (why?)
  - Note that as F changes, the subtracted term $\int x dF$ changes

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n\to\infty} G_n(u, F_0)$$

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n\to\infty} G_n(u, F_0)$$

  - Typically G_∞ will be a normal distribution

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n\to\infty} G_n(u, F_0)$$

  - Typically G_∞ will be a normal distribution
  - Find a consistent estimator F_n of F_0 (e.g., the empirical CDF)

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n\to\infty} G_n(u, F_0)$$

  - Typically G_∞ will be a normal distribution
  - Find a consistent estimator F_n of F_0 (e.g., the empirical CDF)
  - Make the approximation:

$$G_n(u, F_0) \approx G_\infty(u, F_n)$$

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n\to\infty} G_n(u, F_0)$$

  - Typically G_∞ will be a normal distribution
  - Find a consistent estimator F_n of F_0 (e.g., the empirical CDF)
  - Make the approximation:

$$G_n(u, F_0) \approx G_\infty(u, F_n)$$

- Example. Let $\mathcal{F}$ collect distributions with finite second moment

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n \to \infty} G_n(u, F_0)$$

  - Typically G_∞ will be a normal distribution
  - Find a consistent estimator F_n of F_0 (e.g., the empirical CDF)
  - Make the approximation:

$$G_n(u, F_0) \approx G_\infty(u, F_n)$$

- Example. Let $\mathcal{F}$ collect distributions with finite second moment
  - Suppose $T_n(\mathbf{x}, F_0) = \sqrt{n}(\bar{x}_n - \int x dF_0(x))$

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n \to \infty} G_n(u, F_0)$$

  - Typically G_∞ will be a normal distribution
  - Find a consistent estimator F_n of F_0 (e.g., the empirical CDF)
  - Make the approximation:

$$G_n(u, F_0) \approx G_\infty(u, F_n)$$

- Example. Let $\mathcal{F}$ collect distributions with finite second moment
  - Suppose $T_n(\mathbf{x}, F_0) = \sqrt{n}(\bar{x}_n - \int x dF_0(x))$
  - Then $G_\infty(u, F_0)$ is the CDF of $N(0, \sigma_0^2)$

Usual asymptotics

- The usual way of conducting inference:

$$G_\infty(u, F_0) = \lim_{n\to\infty} G_n(u, F_0)$$

  - Typically G_∞ will be a normal distribution
  - Find a consistent estimator F_n of F_0 (e.g., the empirical CDF)
  - Make the approximation:

$$G_n(u, F_0) \approx G_\infty(u, F_n)$$

- Example. Let $\mathcal{F}$ collect distributions with finite second moment
  - Suppose $T_n(\mathbf{x}, F_0) = \sqrt{n}(\bar{x}_n - \int x dF_0(x))$
  - Then $G_\infty(u, F_0)$ is the CDF of $N(0, \sigma_0^2)$
  - Procedure: inference based on $G_\infty(u, F_n)$; this means $N(0, \hat{\sigma}^2)$

Bootstrap inference

- Bootstrap inference: instead of $G_\infty(u, F_n)$, use $G_n^*(u) := G_n(u, F_n)$.
  - In our T_n example,

$$G_n(u, F_n) = \mathbf{P}_{F_n}\left\{\sqrt{n}(\bar{x}_n^* - \bar{x}_n) \leq u\right\}$$

  - Here $\bar{x}_n = \int x dF_n(x)$ and x_i^* are drawn iid from F_n
  - Note: the distribution of $\bar{x}_n^*$ is known (given F_n)
  - We use computational methods because the distribution is often not tractable

Estimating F_n

- Possibilities:

Estimating F_n

- Possibilities:
  - Empirical CDF

Estimating F_n

- Possibilities:
  - Empirical CDF
  - Parametric bootstrap (if we make parametric assumptions on $\mathcal{F}$)

Estimating F_n

- Possibilities:
  - Empirical CDF
  - Parametric bootstrap (if we make parametric assumptions on $\mathcal{F}$)
  - Wild bootstrap (in the linear regression context)

Algorithm

- For $b \in \{1, 2, \ldots, B\}$, do:

Algorithm

- For $b \in \{1, 2, \ldots, B\}$, do:
  - Draw sample $\{x^*_{1,b}, \ldots, x^*_{n,b}\}$ from F_n

Algorithm

- For $b \in \{1, 2, \ldots, B\}$, do:
  - Draw sample $\{x^*_{1,b}, \ldots, x^*_{n,b}\}$ from F_n
  - compute $T^*_{n,b} = T_n(\mathbf{x}^*_b; F_n)$

Algorithm

- For $b \in \{1, 2, \ldots, B\}$, do:
  - Draw sample $\{x^*_{1,b}, \ldots, x^*_{n,b}\}$ from F_n
  - compute $T^*_{n,b} = T_n(\mathbf{x}^*_b; F_n)$
- Approximate

$$G^*_n(u) \approx \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}(T^*_{n,b} \leq u)$$

Applications

1. Variance estimation:

Applications

1. Variance estimation:
   - Idea: suppose variance of $\hat{\theta}_n = \theta(F_n)$ is hard to compute

Applications

1. Variance estimation:
   - Idea: suppose variance of $\hat{\theta}_n = \theta(F_n)$ is hard to compute
   - Adapt algorithm above to estimate

$$V_n^* = \mathbf{E}^* \left[\left(\hat{\theta}_n^* - \mathbf{E}^*(\hat{\theta}_n^*) \right)^2 \right]$$

Applications

1. Variance estimation:
   - Idea: suppose variance of $\hat{\theta}_n = \theta(F_n)$ is hard to compute
   - Adapt algorithm above to estimate

$$V_n^* = \mathbf{E}^* \left[\left(\hat{\theta}_n^* - \mathbf{E}^*(\hat{\theta}_n^*) \right)^2 \right]$$

     - ⋆ If approximation of F_n is empirical distribution, this is simply average variance from bootstrap samples

Applications

2. Confidence intervals/hypothesis testing

Applications

2. Confidence intervals/hypothesis testing
   - Interested in population parameter $\theta_0 = \theta(F_0)$, have estimator $\hat{\theta} = \theta(F_n)$

Applications

2. Confidence intervals/hypothesis testing
   - Interested in population parameter $\theta_0 = \theta(F_0)$, have estimator $\hat{\theta} = \theta(F_n)$
   - Want to conduct two sided test of $\mathrm{H}_0 : \theta = \theta_0$

Applications

2. Confidence intervals/hypothesis testing
   - Interested in population parameter $\theta_0 = \theta(F_0)$, have estimator $\hat{\theta} = \theta(F_n)$
   - Want to conduct two sided test of $\mathrm{H}_0 : \theta = \theta_0$
   - Procedure:

Applications

2. Confidence intervals/hypothesis testing
   - Interested in population parameter $\theta_0 = \theta(F_0)$, have estimator $\hat{\theta} = \theta(F_n)$
   - Want to conduct two sided test of $\mathrm{H}_0 : \theta = \theta_0$
   - Procedure:
     - Construct the test statistic

$$T_n = \frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})}$$

Applications

2. Confidence intervals/hypothesis testing
    - Interested in population parameter $\theta_0 = \theta(F_0)$, have estimator $\hat{\theta} = \theta(F_n)$
    - Want to conduct two sided test of $\mathrm{H}_0 : \theta = \theta_0$
    - Procedure:
        - Construct the test statistic

$$T_n = \frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})}$$

        - Get the bootstrap approximation $G_n^*(u)$ of CDF of T_n

Applications

2. Confidence intervals/hypothesis testing
   - Interested in population parameter $\theta_0 = \theta(F_0)$, have estimator $\hat{\theta} = \theta(F_n)$
   - Want to conduct two sided test of $\mathrm{H}_0 : \theta = \theta_0$
   - Procedure:
     - Construct the test statistic
     $$T_n = \frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})}$$
     - Get the bootstrap approximation $G_n^*(u)$ of CDF of T_n
     - One possibility is to have rule that rejects when
     $$T_n > {G_n^*}^{-1}(1 - \alpha/2) \text{ or } T_n < {G_n^*}^{-1}(\alpha/2)$$
     (aka equal-tailed testing)

Applications

2. Confidence intervals/hypothesis testing
   - Interested in population parameter $\theta_0 = \theta(F_0)$, have estimator $\hat{\theta} = \theta(F_n)$
   - Want to conduct two sided test of $\mathrm{H}_0 : \theta = \theta_0$
   - Procedure:
     - Construct the test statistic
       $$T_n = \frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})}$$
     - Get the bootstrap approximation $G_n^*(u)$ of CDF of T_n
     - One possibility is to have rule that rejects when
       $$T_n > G_n^{*\,-1}(1 - \alpha/2) \text{ or } T_n < G_n^{*\,-1}(\alpha/2)$$
       (aka equal-tailed testing)
   - Q: why not construct $H_n^*(t)$ for $\tilde{T}_n = \hat{\theta}$, and reject based on
     $$\tilde{T}_n > H_n^{*\,-1}(1 - \alpha/2) \text{ or } \tilde{T}_n < H_n^{*\,-1}(\alpha/2)\ ?$$

Bootstrap consistency

- Bootstrap is consistent if $G_n^*(u)$ is uniformly consistent for $G_\infty(u, F_0)$.
- That is, for any $F_0 \in \mathcal{F}$ and $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbf{P}\left(\sup_u \left|G_n^*(u) - G_\infty(u, F_0)\right| > \epsilon\right) = 0$$