Introduction to Econometrics 2: Recitation 8

Seung-hun Lee

Columbia University

April 3rd, 2020 (Finally!)

Bootstrap

Refresher on the properties of estimators

- Assume that we have data $(y_1, ..., y_n)$ distributed according to some probability measure P.
- Normally, we do not have an exact knowledge of what P did the data $(y_1, ..., y_n)$ came from.
- We may know that the data is generated by a distribution F_θ for some $\theta \in \Theta$. However, there is a clear virtue in pinpointing a particular θ (denoted as θ_0).
  - Knowing this allows us to discover the population from which our data is from.
  - For instance, if $F_\theta = N(\theta, 1)$ and if we can pinpoint a particular θ_0, we know the mean and the variance of our data.
- Therefore, if we do not know the exact θ_0 yet, we estimate what this might be.

Bootstrap

Refresher on the properties of estimators

- An **estimator** is a function of the observations $(y_1, ..., y_n)$, often denoted as $\hat{\theta}_n$
- Ideally, we would like the estimators to satisfy these properties
  - **Unbiased**: $E(\hat{\theta}_n) = \theta_0$, where $E(\cdot)$ is the expectation under F_{θ_0}
  - **Consistent**: $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$
  - **Asymptotically Normal**: $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma_0)$ as $n \to \infty$. (Σ_0 is PD matrix)
  - **CAN**: Consistent+Asymptotically Normal
- In addition, we would like to carry out inference on our estimators reliably. To do this, we need to find a good way to approximate the finite-sample distribution of $\hat{\theta}_n$.

Bootstrap

Case of classical linear models

- Classical linear models is specified as

$$y = Xb_0 + u \text{ where } u \sim N(0, \sigma_0^2)$$

- Assuming that the observations are IID conditional on $X = x$, the model has distribution $y \sim N(xb_0, \sigma_0^2)$
- The OLS estimator has a finite-sample distribution that is characterized as follows

$$\begin{aligned} \hat{b}_n &= b_0 + (X'X)^{-1}X'u \\ \implies \sqrt{n}(\hat{b}_n - b_0) &= \left(\frac{X'X}{n}\right)^{-1}\frac{X'u}{\sqrt{n}} \\ &\sim N(0, \sigma_0^2(X'X)^{-1}) \end{aligned}$$

- If we are unsure of what σ_0^2 exactly is, we need to estimate for it using

$$\hat{\sigma}_n^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{u}_i^2 \text{ where } p = \dim(X) \text{ and } \hat{u}_i = y_i - x_i\hat{b}_n$$

Bootstrap

If you are interested in the variance estimator...

$$\begin{aligned}\hat{u} &= y - X\hat{b}_n \\ &= y - X(X'X)^{-1}X'y \\ &= Xb_0 + u - Xb_0 - X(X'X)^{-1}X'u = Mu\end{aligned}$$

where $M = I - X(X'X)^{-1}X$ and is idempotent and symmetric. Then,

$$\begin{aligned}E[\hat{u}'\hat{u}] &= E[u'Mu] \\ &= E[tr(u'Mu)] \text{ } (\because u'Mu \text{ is scalar.}) \\ &= E[tr(Muu')] \text{ } (tr(AB) = tr(BA)) \\ &= tr[E(Muu')] \text{ } (E(tr(A)) = tr(E(A))) \\ &= tr[ME(uu')] \text{ } (\text{If we condition } X = x, \ M \text{ is not stochastic}) \\ &= tr[M\sigma_0^2 I_n] = \sigma_0^2 tr(M) \\ &= \sigma_0^2 tr(I - X(X'X)^{-1}X') = \sigma_0^2[tr(I) - tr(X(X'X)^{-1}X')] \\ &= \sigma_0^2(n - tr((X'X)^{-1}X'X)) = \sigma_0^2(n - p)\end{aligned}$$

Bootstrap

Confidence Intervals

- A more honest way of inference
- We can estimate the coverage probability for b_0 of a region $B \in \mathbb{R}^p$ by

$$\Pr(N(\hat{b}_n, \hat{\sigma}_n^2 (X'X)^{-1}) \in B)$$

  In words, it estimates the proportion of times that the region B contains the distribution of b_0.
- Suppose we are finding the 95% confidence interval for $\hat{b}_{kn}$
  - Applying above setup, the relevant standard error is $\hat{\sigma}_{kn} = \hat{\sigma}_n \sqrt{(X'X)_k^{-1}}$ and the interval would have upper and lower bounds of $\hat{b}_{kn} \pm 1.96 \hat{\sigma}_{kn}$
  - We can test the null hypothesis that $b_{k0} = c$ with

$$\frac{|\hat{b}_{kn} - c|}{\hat{\sigma}_{kn}} > 1.96$$

    which gives a test with size (reject true null hypothesis wrongly by) 5% and power (correctly reject wrong hypothesis) that converges to 1.

Beyond Classical Setup

- In most cases, we only know the asymptotics.
- Suppose that $\hat{\theta}_n$ is a CAN estimator s.t.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N(0, \Sigma_0)$$

- Finite sample context: We need to rely on approximation. Let $\widehat{\Sigma}_n$ be a consistent estimator for Σ_0. Then we can write the above equation as

$$\hat{\theta}_n = \theta_0 + \frac{1}{\sqrt{n}} \widehat{\Sigma}_n^{-1/2} N(0, I_n)$$

- There are cases where above approximation can be very misleading. (e.g. White Misspecification test)

Bootstrap

Bootstrap

- In doing the bootstrap, we are looking for alternative ways to assign standard errors (and ultimately the test statistics and confidence intervals) that are *at least as good as* and ideally *better than* the typical asymptotic approximation that we normally do.
    - **Bootstrap** estimators are usually as good as asymptotic approximations in the sense that bootstrap estimators are also *consistent*
    - It is better in the sense that it has *smaller approximation error*.

Bootstrap

Bootstrap Procedure

- The procedure works as follows for the *t*-statistic :
  1. Start with an IID sample $(X_1, ..., X_n)$. Estimate the model and compute the *t*-statistic
  2. Make n draws *with replacement*. Re-estimate the model and compute the new *t*-statistic.
  3. Repeat step 2 many times
  4. Since there are different *t*-statistics for each sample, there will be a distribution of *t*-statistics. Use this *t*-distribution to get a reliable critical value

Bootstrap

General Discussion

- We have a model and a data $(X_1, ..., X_n)$ generated from an unknown distribution F_0.
- We can characterize the test statistic as $T_n(X_1, ..., X_n)$.
- Define the finite sample distribution of the test statistic under F_0 as

$$G_n(t, F_0) = \Pr(T_n \leq t)$$

- Two properties below are nice properties for T_n to have:
  - **Pivotal**: We say T_n is pivotal if G_n itself does not depend on the unknown parameters of F_0.
  - **Asymptotically Pivotal**: We say T_n is asymptotically pivotal if

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

    does not depend on F_0. (so that F_0 can be omitted in the arguments.)

Bootstrap

General Discussion

- Problem: We do not know F_0, so we need an estimator for it
  - The **empirical distribution function** (or EDF) for the data $X_1, .., X_n$ is defined as
$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$$
- Using the EDF, we can estimate $G_n(t, F_n)$ in this way
  1. Select a large B, say 10,000.
  2. For each $b \in \{1, .., B\}$ we do as follows
     - For each $i = 1, .., n$ pick $j = 1, ..., n$ randomly with replacement and let $X_i^b = X_j$. In other words, draw n times with replacement from $(X_1, .., X_n)$ and redefine the 'shuffled' sample as $(X_1^b, ..., X_n^b)$.
     - Then, compute the new test statistic $T_n^b = T_n(X_1^b, ..., X_n^b)$.
  3. Then, we will end up with B numbers of T_n^b. Then $G_n(t, F_n)$ can be written as
$$G_n(t, F_n) = \frac{1}{B} \sum_{b=1}^{B} 1(T_n^b \leq t)$$

Bootstrap

General Discussion

- Once we are done with this, we find the bootstrapped critical values.
- Suppose that our test is of size 5% and we are doing a two-sided test.
- We find lower(upper) bounds for our critical value c^- (c^+) s.t. $0.025B$ values of T_n^b are smaller(larger) than that critical value.
- We then reject the null hypothesis if the T_n from the original $(X_1, .., X_n)$ does not belong within the bounds set by this procedure.
  - Alternatively, we can check whether the observed estimator belongs in the confidence interval using the critical value from this procedure and the bootstrapped standard errors.
- Bootstrapped confidence intervals are obtained by inverting the test - we take all values not rejected by the above procedure and construct an interval.

Bootstrap

General Discussion

- Now we show why bootstrapped estimators are *at least as good as* and sometimes ideally *better than* normal asymptotic approximations.
- Using the technic of Edgeworth Expansion, we can write

$$G_n(t,F_0) = G_\infty(t,F_0) + \frac{1}{\sqrt{n}} g_1(t,F_0) + \frac{1}{n} g_2(t,F_0) + \frac{1}{n\sqrt{n}} g_3(t,F_0) + O\left(\frac{1}{n^2}\right)$$
$$G_n(t,F_n) = G_\infty(t,F_n) + \frac{1}{\sqrt{n}} g_1(t,F_n) + \frac{1}{n} g_2(t,F_n) + \frac{1}{n\sqrt{n}} g_3(t,F_n) + O\left(\frac{1}{n^2}\right)$$

  uniformly over t almost surely. (From Horowitz, 2011)
- Suppose that $G_\infty(t,\cdot)$ is continuous at F_0. Then

$$\begin{aligned} G_n(t,F_n) - G_n(t,F_0) = {} & G_\infty(t,F_n) - G_\infty(t,F_0) + \frac{1}{\sqrt{n}}(g_1(t,F_0) - g_1(t,F_n)) \\ & + \frac{1}{n}(g_2(t,F_0) - g_2(t,F_n)) + \frac{1}{n\sqrt{n}}(g_3(t,F_0) - g_3(t,F_n)) + O\left(\frac{1}{n^2}\right) \end{aligned}$$

General Discussion

- Notice that $G_\infty(t, F_n) - G_\infty(t, F_0)$ is $O\left(\frac{1}{\sqrt{n}}\right)$ since $F_n - F_0 = O\left(\frac{1}{\sqrt{n}}\right)$ ($\because$ Dvoretzky-Kleifer-Wolfowitz inequality).
- Therefore, the size of the bootstrap approximation error is $O\left(\frac{1}{\sqrt{n}}\right)$ and the bootstrap error is consistent.
- Note that normal asymptotic approximation also has the same size of approximation error.
- If T_n is asymptotically pivotal, $G_\infty(t, F_n) - G_\infty(t, F_0)$ term vanishes. The leading term is $\frac{1}{\sqrt{n}}(g_1(t, F_0) - g_1(t, F_n))$ and the size of the approximation error is $O\left(\frac{1}{n}\right)$ - improvement.

Bootstrap

Uses of Bootstraps

- Parametric: Assume that you have faith in the form of your model.
  - Then, either draw from estimated parametric distribution of u
  - Or you can hold x_i's fixed and draw upon residuals
- Bootstrapped standard errors: The bootstrap variance of the estimator $\hat{\theta}_n$ is defined as

$$\tilde{V}_n = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_n^b - \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_n^i \right)^2$$

- Block Bootstrap: This is used when we suspect that the data is serially correlated.
  - What we do is to use m-size blocks of consecutive observations.
  - Then we can use $B = n + 1 - m$ blocks in total. This works when $m/n \to 0, m \to \infty$ as $n \to \infty$

Bootstrap

Uses of Bootstraps

- Wild Bootstrap: This is useful with heteroskedastic error terms.
  - Suppose we have sample b, and there is i'th observation in that sample.
  - The trick is to use $u_i^* = \epsilon_i \hat{u}_i$ where $E[\epsilon_i] = 0, E[\epsilon_i^2] = 1$. As in class, you can use $\epsilon_i = \pm 1$
- Bias-Corrected Bootstrap: Typical estimation $\hat{\theta}_n$ could have some biases.
  - Instead, we define

$$\theta_n^* = 2\theta_n - E^*[\theta_n^*]$$

    to minimize the bias, at least asymptotically

Bootstrap

Uses of Bootstraps

- Subsampling: Instead of resampling with the size n, we resample with size m and do a bootstrap for B times. (without replacement)
  - So there are a total of $N_{n,m} = \binom{n}{m}$ total number of subsets
  - Compute $T_m(x_1^b, .., x_m^b)$
  - We are trying to estimate $G_n(t, F_n)$ with T_m, but on a rescaled version
  - Define

    $$J_n(t, F_0) = \Pr(r_n(T_n - T_\infty) \leq t) \text{ (You can define } J_m(t, F_0) \text{ similarly)}$$

    where r_n is a scaling parameter
  - Then, as $m/n \to 0, r_m/r_n \to 0$ and $m \to \infty$ as $n \to \infty$, we have

    $$r_n(T_n - T_\infty) \simeq r_m(T_m - T_n)$$

    which works better than bootstrap as long as $J_\infty(t, F_0)$ is continuous with respect to t

Multiple Testing

Motivation

- Suppose that we are testing S hypothesis where S is some large number and each hypothesis is mutually independent.
- Let $S = 1000$. Then, the probability of rejecting any one of the 1000 hypothesis is calculated by $1 - (0.95)^{1000} \simeq 1$
- In other words, you get something significant just out of pure luck.
- In doing so, the most dangerous consequence is running in to a **false discovery** - a rejection of a true null hypothesis
- The best approach that prevents this mistake is finding either an empirical or theoretical motivations for the hypothesis. In a context with larger datasets, we need to come up with a technical method for testing multiple hypothesis.

Multiple Testing

Setup

- We are running S simultaneous tests. For a particular test $s \leq S$, we are testing the null H_{0s} versus the alternative H_{1s}, we know that the p-value is

  $$p_s = \Pr(\text{Reject } H_{0s} | H_{0s} \text{ is true})$$

  This is also the mathematical definition of the false discovery. Some other key terms are

  - **False discovery proportion** (FDP): It is the proportion of the false discoveries out of all discoveries, or
  - **False discovery rate** (FDR): $E[FDP]$
  - **Familywise error rate** (FWE): It is the probability of one or more false discoveries. A k-FWE calculates probability of k or more false discoveries.

- There are two strands of methods. One controls for the FWE. The other controls for the FDR.

Multiple Testing

Bonferroni Method

- **Bonferroni method** is a classical method based on Bonferroni inequalities (Boole's inequality for $k = 1$)
- To see how this works, let A_s be the event that $p_s < w$ (and for which H_{0s} is rejected) for some value w associated with the significance level α, where $s = 1, ..., S$.
- Let I_0 be indices for the true hypothesis, with $n(I_s) = S_0 \leq S$.
- The goal is to determine the value of w that makes the $FWE \leq \alpha$. Then the following logic works

$$FWE = \Pr(\cup_{s \in I_0} A_s) \leq \sum_{s \in I_0} \Pr(A_s) = S_0 w \leq Sw$$

So for the FWE, we can let $w = \alpha/S$.

Multiple Testing

Holm Method

- **Holm method** is a step-down method in the sense that ordering of the different p-values are involved.
- The procedure is as follows
  1. Order the p-values s.t. $p_1 < ... < p_S$, and denote each null hypothesis as $H_{01}, ..., H_{0S}$.
  2. Reject H_{0j} if $p_j \leq \frac{\alpha}{S-j+1}$ for $j = 1, 2, ..., S$

  That is, we start with the hypothesis with lowest p-value and compare its p-value against α/S. If this is rejected, move on to the next hypothesis and compare its p-value against $\alpha/(S-1)$

Multiple Testing

Holm Method: How does this work?

- Let k be the index of the first true null hypothesis that is rejected.
- Since we ordered hypothesis according to the size of the p-value, the first $k-1$ rejected hypotheses are false null hypothesis.
- Since there are total of $S-S_0$ false hypotheses, we can claim $k-1 \leq S-S_0$. This equation implies $\frac{1}{S-k+1} \leq \frac{1}{S_0}$.
- Since kth hypothesis is still rejected, $p_k \leq \frac{\alpha}{S-k+1} \leq \frac{\alpha}{S_0}$.
- So, there exists a true hypothesis whose p-value is bounded above by α/S_0. Then

$$
\begin{aligned}
FWE &= \Pr(\text{Reject at least one true hypothesis}) \\
&\leq \Pr(p_s \leq \alpha/S_0 \text{ for some true hypothesis}) \\
&\leq \sum_{s \in I_0} \Pr(p_s \leq \alpha/S_0) = S_0 \times (\alpha/S_0) = \alpha
\end{aligned}
$$

where I use the fact that if null hypothesis is true and the statistic is continuous, p-value has $U[0,1]$ distribution

FWE controls

- For both methods, we did not need to make any assumptions on the dependence structure of the hypotheses
- Bonferroni method tends to be the more conservative out of the two - It rejects less in general.
- There are ways to control for k-FWE where $k > 1$. I have put this in the recitation notes. For more reference, look at Lehman, Romano (2005)

Multiple Testing

Benjamini-Hochberg Method

- The goal of this approach is to set the false discovery rate to be below some $\gamma \in [0, 1]$.
- The procedure is as follows
  1. Order p-values from the lowest to the highest: $p_1 < ..., < p_S$
  2. Make the following adjustment:

$$\tilde{p}_s = \min \left\{ \frac{Sp_j}{j} | j \geq s \right\}$$

  3. We move from the Sth hypothesis. If $\tilde{p}_S \leq \gamma$, then all S hypothesis are rejected and the process ends. Otherwise, move to $S - 1$th hypothesis.
  4. At some point, you will find i s.t. $\tilde{p}_i \leq \gamma$. Then reject the first i hypotheses.

Multiple Testing

Benjamini-Hochberg Method: How does it work?

- We do not reject H_{0s} iff for all $j \geq s, p_j \geq \frac{j\gamma}{S} \left(\geq \frac{s\gamma}{S}\right)$.
- So what we can do is to plot the ordered p-values and the threshold line $p(s) = \frac{s\gamma}{S}$.
- If a p-value of the sth hypothesis is larger than $\frac{s\gamma}{S}$, then that hypothesis can not be rejected.
- In this manner, we can find s^*, the largest index s.t. the p-value is below the $p(s)$ plot. Then, we reject all H_{0s} s.t. $s \leq s^*$.
- We can show that in suitable conditions, this method yields $FDR = \frac{S_0}{S}\gamma$. This is still conservative method in that $S_0 \leq S$. We can do better by estimating what S_0 would be and using $\frac{S}{S_0}\gamma$ as cutoff instead.