SFERS DATA Seminar with Python: 3-1

Yitae Kwon

SFERS of SNU

2024-1

Contents

Books

Recommended Books

- 모두를 위한 컨벡스 최적화: https://convex-optimization-for-all.github.io/
- 비모수통계학 with R: https://product.kyobobook.co.kr/detail/S000001762578
- An Introduction to Statistical Learning: https://www.statlearning.com/
- The Elements of Statistical Learning: https://hastie.su.domains/Papers/ESLII.pdf
- Bootstrap Methods in Econometrics: https://cpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/5/239/files/2020/04/EC11CH08_Horowitz193-224-2_Corrected.pdf

Bootstrap

Bootstrap

우리는 가끔 통계량의 표본분포를 구해야 할 때가 있습니다. 예를 들어 정규분포가 가정된 모집단에서 표본을 뽑아 그 표본평균을 보는 상황에서, 이 표본평균으로써 모평균의 95% 신뢰구간을 구하려면 표본평균이라는 통계량의 표본분포를 알아야만 합니다. 모수적 가정이 있는 상태라면 표본평균이 정규분포를 따르며 평균이 모평균과 같고 분산은 모분산을 표본의 크기로 나눈 값임을 알지만, 비모수통계에서는 이것이 불가능합니다. 애초에 해당 표본평균의 실제 분포가 유한 개의 모수로 표현되지 않을 수도 있습니다. 비모수적으로 표본평균의 분포를 근사하는 방법이 바로 **붓스트랩**(Bootstrap)입니다.

Bootstrap

Bootstrap의 기본 가정부터 먼저 세팅해 보도록 하겠습니다.

- $X_1, X_2, \cdots, X_n \sim_{i.i.d} F$
- $\hat{\theta}_n = T(\mathbf{X}_n) = T(X_1, \cdots, X_n)$은 표본으로부터 얻은 모수 θ의 추정량

즉 모집단 F에서 크기가 n인 표본 $\mathbf{X}_n$을 뽑아냈고 그것을 이용해 추정량 $\hat{\theta}_n$을 구한 상황입니다. 그런데 우리는 대부분의 상황에서 F는 모르기 때문에, 수리통계적인 계산을 통해 표본분포를 구하기가 어렵습니다. 혹은 F가 너무 복잡해서 표본분포를 계산, 혹은 기술하기 어려운 상황입니다. 그렇다면 이러한 표본을 여러 개 새로 뽑아 $(\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \cdots, \mathbf{X}_n^{(B)}$처럼 크기가 n인 표본을 B개) 몬테카를로 방법으로써 표본분포를 근사할 수도 없는 노릇입니다. 표본을 새로 뽑는 일이 가능할지도 의문이고, 비용도 많이 드니까요.

Bootstrap

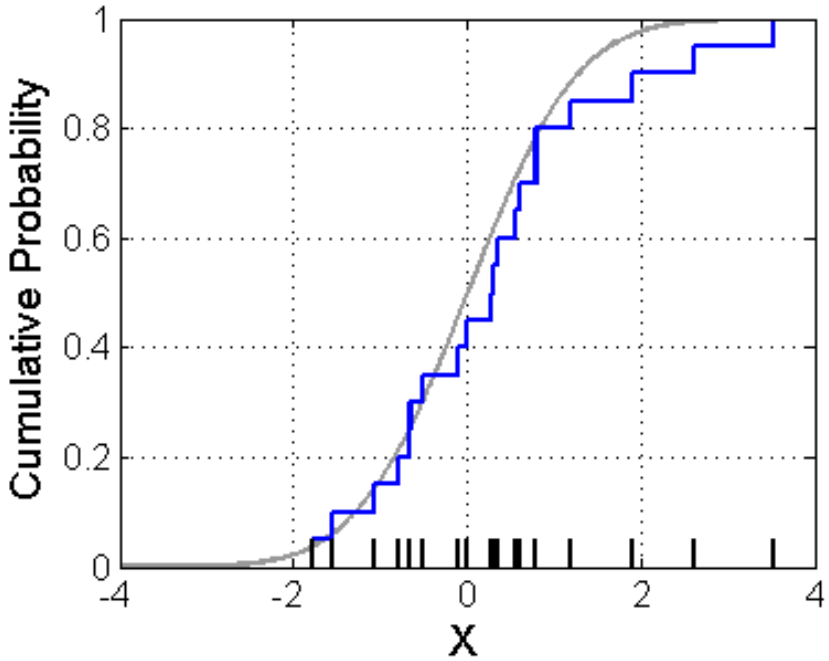
여기서 Bootstrap의 아이디어는, 우리가 F를 모르기 때문에 F에서 표본을 뽑는 게 불가능하니까, 표본 X_n으로부터 모집단의 분포 F를 추정하여 그 추정된 분포 F_n을 가상의 모집단으로 두고 여기서 B개의 크기가 n인 표본을 추출해보자는 것입니다. **경험적 분포함수**

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

는 표본 X_n에서 추정할 수 있는 $F(x)$의 좋은 추정량입니다. 일반적으로 둘의 차이가 대략 $1/\sqrt{n}$임이 알려져 있습니다. 즉 표본의 크기가 충분히 큰 상황에서는 F에서 뽑으나, F_n에서 뽑으나 사실 별 차이가 없다는 의미입니다.

Bootstrap

F_n을 조금 더 잘 들여다 보겠습니다. 이는 사실 x의 값이 특정 X_i를 지날 때마다 $1/n$씩 증가하는 계단형 함수입니다. 이러한 형태의 CDF를 가지는 함수는 바로 이산분포입니다. 즉 F_n은 관측값 X_i에 질량 $1/n$이 배정된 이산분포와 같습니다.



Bootstrap

우리는 이제 이 이산분포를 기반으로, X_n에서 서로 독립인 n개의 관측값을 단순복원추출하여 **붓스트랩 표본** $X_n^{*(i)}$를 얻어내고, 이로부터 통계량 $\hat{\theta}_n^{*(i)}$를 $i = 1, 2, \cdots, B$에 대해 계산합니다. 굳이 n개의 관측값을 뽑는 이유는, 표본분포 역시도 n개의 관측값으로부터 뽑힌 표본으로 얻은 통계량들이 만드는 분포이기 때문입니다. 이제 우리는 B개의 $\hat{\theta}_n^{*(i)}$로부터 얻어낸 경험분포를 $\hat{\theta}_n$의 **붓스트랩 표본분포**라 부르며, $\hat{\theta}_n$의 표본분포의 추정값으로 사용할 것입니다. 이때에도 경험분포를 사용해 추정하는 과정에서 오차 $1/\sqrt{B}$가 생성됩니다. 우리는 이를 G^*라고 부르겠습니다. (G는 원래 추정량 $\hat{\theta}_n$의 분포)

Bootstrap

$$\text{Population} : F$$

$$\Downarrow \text{ Choose } n \text{ observations to build sample with sample size } n \Downarrow$$

$$\text{Sample} : \mathbf{X}_n = (x_1, x_2, \cdots, x_n)$$

$$\Downarrow \text{ Under } F_n \text{ instead of } F, \text{ make B bootstrap samples with sample size } n \Downarrow$$

$$\text{Error} : \frac{1}{\sqrt{n}}$$

$$\text{Bootstrap sample } 1 : \mathbf{X}_n^{*(1)} = (x_1^{*(1)}, \cdots, x_1^{*(n)}), \quad \hat{\theta_n}^{*(1)} = T(\mathbf{X}_n^{*(1)})$$

$$\vdots$$

$$\text{Bootstrap sample } B : \mathbf{X}_n^{*(B)} = (x_1^{*(B)}, \cdots, x_1^{*(B)}), \quad \hat{\theta_n}^{*(B)} = T(\mathbf{X}_n^{*(B)})$$

Bootstrap

$$\text{Bootstrap sample } 1 : \mathbf{X}_n^{*(1)} = (x_1^{*(1)}, \cdots, x_1^{*(n)}), \quad \hat{\theta_n}^{*(1)} = T(\mathbf{X}_n^{*(1)})$$

$$\vdots$$

$$\text{Bootstrap sample } B : \mathbf{X}_n^{*(B)} = (x_1^{*(B)}, \cdots, x_1^{*(B)}), \quad \hat{\theta_n}^{*(B)} = T(\mathbf{X}_n^{*(B)})$$

$$\Downarrow \text{ Estimate boostrap sample distribution using } (\hat{\theta_n}^{*(1)}, \cdots, \hat{\theta_n}^{*(B)}) \Downarrow$$

$$\text{Error} : \frac{1}{\sqrt{B}}$$

$$\text{Estimate: } G^*(x) = \frac{1}{B} \sum_{i=1}^{B} I(\hat{\theta_n}^{*(i)} \leq x)$$

Estimation with Bootstrap

붓스트랩의 가장 기본적인 아이디어는

$$\text{모집단} \Leftrightarrow \text{표본}$$

과

$$\text{표본} \Leftrightarrow \text{붓스트랩 표본}$$

사이의 상동성을 들여다 보는 것입니다.

Estimation with Bootstrap

먼저 표본분포의 평균 $\mu_{\hat{\theta}_n}$은 아래와 같이 추정할 수 있겠습니다. 즉 B개의 붓스트랩 표본에서 얻은 B개의 추정량 $\hat{\theta}_n^*$들의 표본평균, 혹은 적률추정량, 혹은 몬테카를로 추정량을 구한다고 생각할 수 있습니다.

$$\hat{\mu}_{\hat{\theta}_n} = \mu_{\hat{\theta}_n^*} = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_n^{*(i)} = \int \hat{\theta}_n^* dG^*$$

Estimation with Bootstrap

다음으로는 편향 bias($\hat{\theta}_n$)을 추정해 보고자 합니다. 조심해야 할 것은, 여기서 구하는 편향은 붓스트랩을 시행하기 전, 크기가 n인 표본 X_n에서 얻은 $\hat{\theta}_n$의 편향입니다. 편향은 만약 우리가 F를 알았다면 그냥 계산을 통해 구할 수 있었겠지만, 지금의 우리는 F도 모르고 표본을 더 뽑을 수도 없는 상황입니다. 사실 추정하려는 모수 θ도 모르는 상황입니다. 따라서 편향 계산에 사용 표본분포의 기댓값과 모수 θ를 사용할 수 없습니다.

$$\begin{aligned}
\text{bias}(\hat{\theta}_n) &= E(\hat{\theta}_n) - \theta \\
&= \mu_{\hat{\theta}_n} - \theta \\
&\approx \mu_{\hat{\theta}_n^*} - \hat{\theta}_n \quad \text{(상동성!!)} \\
&= \mu_{\hat{\theta}_n^*} - \hat{\theta}_n = \widehat{\text{bias}}_{\text{Boot}}(\hat{\theta}_n)
\end{aligned}$$

따라서 위에서처럼 $\mu_{\hat{\theta}_n}$의 좋은 추정량이 $\mu_{\hat{\theta}_n}{}^*$이고, θ의 좋은 추정량이 $\hat{\theta}_n$임을 이용하여 편향을 추정하게 됩니다.

Estimation with Bootstrap

한편 이제는 표준오차를 추정해 보도록 하겠습니다. 이 경우 $(\hat{\theta}_n^{*(1)}, \cdots, \hat{\theta}_n^{*(B)})$를 G^*로부터 비롯된 크기가 B인 표본이라고 보고 그 표본표준편차를 구하여 이를 표준오차의 추정량으로 이용합니다. B가 크다면 G^*이 G와 충분히 유사하기 때문에, $\hat{\theta}_n^*$들을 G로부터 비롯된 B인 표본이라 보고 구하면 근사적으로 G의 표준오차를 추정할 수 있게 되는 것입니다. 따라서

$$\widehat{\mathrm{se}}_{\mathrm{Boot}}(\hat{\theta}_n) = \left(\frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_n^{*(i)} - \mu_{\hat{\theta}_n^*})^2 \right)^{\frac{1}{2}} \left(\approx \left(\int (\hat{\theta}_n - \mu_{\hat{\theta}_n})^2 dG \right)^{\frac{1}{2}} \right)$$

입니다. 이처럼 붓스트랩을 이용하면 표본분포를 근사할 수 있는 것에서 그치지 않고, 근사한 표본분포를 이용해 원래 추정량인 $\hat{\theta}_n$의 편향과 표준오차를 파악할 수 있습니다.

Bootstrap Confidence Interval

붓스트랩은 점추정량만을 구하는 데 국한되지 않습니다. 붓스트랩으로는 추정량의 신뢰구간도 구할 수 있고, 다양한 방식이 개발되어 있습니다.

정규근사 붓스트랩 신뢰구간(normal-approximated bootstrap confidence interval)은 표본의 크기 n이 충분히 크거나 분포가 대칭이라 중심극한정리에 의한 정규분포로의 수렴이 잘 일어난 상황에서 사용할 수 있습니다. 즉, 추정량의 편향 $\text{bias}(\hat{\theta}_n)$에 대하여

$$\frac{\hat{\theta}_n - \text{bias}(\hat{\theta}_n) - \theta}{\text{se}(\hat{\theta}_n)} \dot{\sim} N(0,1)$$

가정이 잘 성립하는 상황을 고려합니다.

Bootstrap Confidence Interval

그렇다면 기존 통계학에서 사용하던 것과 비슷하게, $1-\alpha$ 신뢰구간을

$$\theta \in \hat{\theta}_n - \text{bias}(\hat{\theta}_n) \pm z_{\alpha/2}\text{se}(\hat{\theta}_n)$$

로 구성할 수 있습니다. 한편 우리는 일반적인 경우 편향과 표준오차를 알지 못하므로, 이를 붓스트랩으로 구하자는 것입니다. 따라서 이를 붓스트랩 추정량으로 대체한

$$\theta \in \hat{\theta}_n - \widehat{\text{bias}}_{\text{Boot}}(\hat{\theta}_n) \pm z_{\alpha/2}\widehat{\text{se}}_{\text{Boot}}(\hat{\theta}_n)$$

가 정규근사 붓스트랩 신뢰구간입니다. 한편 이는 정규근사 가정 하에서 이루어진 것이기에, 표본의 수가 적어 정규근사가 불가능하거나, 추정량의 점근분포가 정규분포가 되지 않는 경우에는 부정확합니다. 일반적으로 이 경우 포함확률은 $1-\alpha+O(n^{-1/2})$ 수준임이 알려져 있습니다. 이런 상황을 **일차 정확도**를 가진다고 표현합니다.

Bootstrap Confidence Interval

다른 방법으로 신뢰구간을 구할 수도 있습니다.

- 표준 붓스트랩 신뢰구간: pivot H의 분포를 이용합니다.
- 붓스트랩-t 신뢰구간: 스튜던트화된 피벗이 있다고 가정하고 그 분포를 이용합니다. 이차 정확도를 가짐이 알려져 있습니다.(포함확률 $1-\alpha+O(n^{-1})$)
- 편향 조정 및 가속 신뢰구간(bias-corrected and accelerated confidence interval; BCa CI): 적당히 편향과 왜도를 조정하여 피벗(이차 정확도)
- 붓스트랩 퍼센타일 신뢰구간(Bootstrap percentile confidence interval)

Bootstrap Confidence Interval

붓스트랩 퍼센타일 신뢰구간은 매우 간단합니다. 모수 θ의 추정량 $\hat{\theta}_n$과 붓스트랩 표본분포 G^*을 붓스트랩으로써 구했다면, 신뢰수준 $1-\alpha$에서 이는 G^*의 분위수함수 $Q_{G^*}(p)$에 대하여

$$\theta \in \left(Q_{G^*}(\frac{\alpha}{2}), Q_{G^*}(1-\frac{\alpha}{2}) \right)$$

로써 주어집니다. 증명은 생략하겠습니다. 이는 적당한 가정 하에서 일차 정확도를 가짐이 알려져 있습니다. 다만 이는 가장 간단하게 구할 수 있는 신뢰구간 중 하나이므로, 가장 많이 사용되고 있는 붓스트랩 신뢰구간 중 하나입니다.

Hypothesis Testing with Boostrap

가설을 검정하는 방법 중 하나는, 가설검정의 유의수준을 α라 할 때 모수에 대한 $1-\alpha$ 신뢰구간을 만든 후 귀무가설 하에서의 모수가 이 신뢰구간에 들어가는지 보는 것입니다. 이는 통계적으로 문제가 없습니다. 따라서 우리는 **신뢰구간을 뒤집음으로써** 가설검정을 수행할 수 있습니다. 즉 근사적으로 포함확률이 $100(1-\alpha)\%$인 신뢰구간이

$$\theta \in (L(G^*), U(G^*))$$

의 형태로 주어진다면, $H_0 : \theta = \theta_0$의 유의수준 α에서의 근사적 검정은

$$\text{reject if } \theta_0 \leq L(G^*) \text{ or } \theta_0 \geq U(G^*)$$

입니다.

Take-Home Messages

- 붓스트랩 관련해서는 정말 간단히만 소개드렸습니다. 궁금하신 내용이 있으시면 위키...(https://en.wikipedia.org/wiki/Bootstrapping_(statistics)) 읽어보세요.
- 붓스트랩은 간단하면서도 쉽게, 추정량의 분포를 모르는 상황에서도 점근적으로 옳은 통계적 추론을 할 수 있다는 장점을 가지고 있습니다. 만약 분포 가정을 하고 싶지 않거나, 데이터가 모양이 이상한 경우 유용하게 이용됩니다.
- 반면 어찌하였든 $B = 1000$번 정도 다시 표집을 하므로, Bn번 수준의 표본추출이 더 필요합니다. 계산량이 엄청나게 증가합니다. 더불어 사실 $n^{-1/2}$나 $B^{-1/2}$ 정도면 그리 만족스러운 수렴 결과가 나오기는 어려운 차수입니다. 다르게 말하면 n이 작으면 애초에 붓스트랩을 사용하기 어렵습니다. 이에 따라 수렴 속도를 빠르게 하기 위한 다양한 해결법들이 제시되고 있습니다.