Estymator wariancji i bootstrap

- Najwcześniejsze użycie bootstrap (wtórnego próbkowania) Efron(1979)

- Z próby n obserwacji wybieramy B sztucznych próbek (boostrap samples) o liczebności n losując z orginalnej próby *ze zwracaniem*

- Dla każdej z tak wylosowanych prób liczymy estymator $\widehat{\boldsymbol{\theta}}_j^*$

- Liczymy średnią z tak policzonych estymatorów $\widehat{\boldsymbol{\theta}}^*$

- Obliczamy estymator wariancji ze standardowego wzoru:

$$\widehat{\Sigma} = \left(\frac{1}{B-1} \sum_{j=1}^{B} \left(\widehat{\boldsymbol{\theta}}_j^* - \widehat{\boldsymbol{\theta}}^* \right) \left(\widehat{\boldsymbol{\theta}}_j^* - \widehat{\boldsymbol{\theta}}^* \right)' \right)$$

- Ten sposób użycia bootstrapu obecnie wyszedł raczej z użycia - zazwyczaj nie da się uzyskać w ten sposób estymatorów efektywniejszych niż standardowe estymatory zgodne

Korekta obciążenia i bootstrap

- Szacujemy estymator boostrap parametrów

$$\widehat{\boldsymbol{\theta}}^* = \frac{1}{B} \sum_{j=1}^{B} \widehat{\boldsymbol{\theta}}_j^*$$

- Szacujemy obciążenie estymatora

$$b^* \left(\widehat{\boldsymbol{\theta}} \right) = \widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}}$$

- Tworzymy oszacowanie skorygowane o obciążenie

$$\widetilde{\theta} = \widehat{\boldsymbol{\theta}} - b^* \left(\widehat{\boldsymbol{\theta}} \right) = 2\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^*$$

- Stosując tę procedurę zakładamy, że obciążenie estymatora *nie zależy* od wektora nieznanych parametrów

- Jeśli obciążenie zależy od wektora nieznanych parametrów, wtedy stosując boostrap możemy często uzyskać redukcję obciążenia ale za cenę spadku efektywności esymacji.

- Niekiedy spadek tej efektywności jest tak duży, że rośnie błąd średniokwadratowy $MSE = \mathrm{E}\left(\widehat{\theta} - \theta\right)$.

Przedziały ufności i boostrap

- Jednym z bardzo popularnych zastoswań boostrap jest tworzenie przedziałów ufności

- Teoretycznie najlepszą metodą jest metoda procentowego t, przedział ufności
$$\left[\widehat{\theta} - \widehat{s}_\theta t^*_{1-\alpha/2}, \widehat{\theta} + \widehat{s}_\theta t^*_{\alpha/2}\right]$$
gdzie $\widehat{s}_\theta$ jest błędem standardowym $\widehat{\theta}$ a t^*_δ jest kwantylem δ bootstrapowanej statystyki
$$t^*_j = \frac{\widehat{\theta}^*_j - \widehat{\theta}^*}{se\left(\widehat{\theta}^*_j\right)}$$

- Inną metodą jest prosty boostrapowy przedział ufności

$$\left[\widehat{\theta} - se^{*}t_{1-\alpha/2}, \widehat{\theta} + se^{*}t_{\alpha/2}\right]$$

  w którym wykorzystujemy standardowy rozkład t-studenta ale oszacowania błędów standardowych są z bootstrap

- Można w tym przypadku zastosować także korektę obciążania

Testowanie hipotez i bootstrap

- Obecnie najpopularniejszym zastosowaniem boostrap jest testowanie hipotez

- Boostrapowa wartość p dla statystyki τ jest równa

$$p^*\left(\widehat{\tau}\right) = \frac{1}{B}\sum_{j=1}^{B} I\left(r_j^* > \widehat{\tau}\right)$$

  gdzie $I\left(\cdot\right)$ jest funkcją indykatorową (charakterystyczną)

- Można to także zapisać jako

$$p^*\left(\widehat{\tau}\right) = 1 - \widehat{F}^*\left(\widehat{\tau}\right)$$

gdzie $\widehat{F}^{*}(\cdot)$ jest dyystrybuantą empiryczną uzyskaną z boostrapu

Warunki poprawności testów boostrap

- Dla szczególnego przypadku można pokazać, że policzone na postawie boostrap α jest dokładnie rowne prawdziemu α (test jest dokładny) w małych próbach
- Warunki:
  - rozkład statystyki nie zależy od parametrów (*pivotal statistics*)
  - B jest takie, że $\alpha(B+1)$ jest liczbą naturalną
- Przykładami testów o tych własnościach jest test Durbina-Watsona, test ARCH, test Jarque-Bera
- W większości przypadków mamy jednak do czynienia z testami, których

rozkłady statystyk są jedynie asymptotycznie niezależne od wielkości parametrów

- Boostrapowane wartości krytyczne w tym przypadku dążą do prawdziwych wartości krytycznych dla $n \to \infty$ ale nie dla $B \to \infty$
- jeśli dystrybyanta testu $F(\cdot)$ nie zlaeży silnie od nieznanych wielkości parametrów wtedy bootstrap powinien dobrze działać
- jeśli $F(\cdot)$ silnie zależy od nieznanych parametrów bądź parametry są estymowane nieefektywnie lub z dużym obciążeniem, wtedy bootstrap może działać źle
- jeśli mamy do wyboru kilka statystyk (np LM, LR, W), to powinnismy użyć tej, która najsłabiej zależy od nieznanych parametrów
- różne metody wtórnego próbkowania (bootstrap) mogą mieć bardzo różne własności w małych próbach
- testy bazowane na bootstrapie nie są słabsze od testów bazowanych na wynikach asymptotycznych przy założeniu, że B jest wystarczająco duże

Bootstrapowanie modelu regresji

- Założenie

$$
\begin{aligned}
y_t &= \boldsymbol{x}_t\,\boldsymbol{\beta} + \varepsilon_t \\
\mathrm{E}\left(\varepsilon_t \middle|\, \boldsymbol{x}_t\right) &= 0 \\
\varepsilon_t &= IID\left(0, \sigma^2\right)
\end{aligned}
$$

- Bootstrap na resztach

$$
\begin{aligned}
y_t^* &= \boldsymbol{x}_t\,\boldsymbol{\beta} + \varepsilon_t^* \\
\varepsilon_t^* &\sim EDF\left(\ddot{\varepsilon}_t\right)
\end{aligned}
$$

gdzie $\ddot{\varepsilon}_t = \left(\frac{n}{n-k}\right)^{\frac{1}{2}} e_t$ (korekta na liczbę stopni swobody) a EDF jest empiryczną dystrybuantą $\ddot{\varepsilon}_t$.

- Boostrap parametryczny. Identyczny jak w poprzednim przypadku ale

$$\varepsilon_t^* \sim NID\left(0, s^2\right)$$

- Wild bootstrap

$$y_t^* = \boldsymbol{x}_t \, \boldsymbol{\beta} + f\left(\widehat{\varepsilon}_t\right) v_t^*$$

v_t jest zmienną losową o wartości oczekiwanej 0 i wariancji 1(np zmienna binarna -1, 1)

$$f\left(\widehat{\varepsilon}_t\right) = \frac{\widehat{\varepsilon}_t}{\left(1 - h_t\right)^{\frac{1}{2}}}$$

gdzie h_t jest t-wartością diagonalną macierz $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$.

Metoda ta jest odporna na heteroskedastyczność

- Boostrap na parach:
  - wybiaramy losowo pary $\left[\begin{array}{cc} y_t & \boldsymbol{x}_t \end{array}\right]$

Boostrap w modelach ze skorelowanymi błędami losowymi

- Sitowy boostrap (sieve boostrap): zakładamy że błędy losowe zachowują się zgodnie z procesem $AR\,(p)$

1. Oszacuj model $y_t = \boldsymbol{x}_t\,\boldsymbol{\beta} + \varepsilon_t$
2. Oszacuj model $AR\,(p)$

$$\widehat{\varepsilon}_t = \sum_{i=1}^{p} \rho_i \widehat{\varepsilon}_{t-i} + u_t \tag{1}$$

3. Wygeneruj błędy boostrapowane

$$\varepsilon_t^* = \sum_{i=1}^{p} \rho_i \varepsilon_{t-i}^* + u_t^*$$

gdzie u_t^* pochodzą z przeskalowanej dystrybuanty empirycznej reszt 1

- Boostrap blokowy. Dzielimy wielkości, które mamy wtórnie próbkować na s bloków b następujących po sobie obserwacji. Mogą to być zarówno reszty jak i pary $\left[\begin{array}{cc} y_t & \boldsymbol{x}_t \end{array}\right]$. Następnie losujemy bloki.

Jacknife

- Jacknife jest inną metodą wtórnego próbkowania
- Wtórne próbkowanie polega na stworzeniu n próbek poprzez usunięcie z pierwotnej próby j-tej obserwacji
- Estymatory $\widehat{\boldsymbol{\theta}}_j^*$ tworzymy dla tak powstałych próbek
- Używając tej metody możemy uzyskać skonstruować podobnie jak w przypadku boostrap korektę obciążenia i estymator wariancji