Programowanie narzędzi analitycznych Z11

Rafał Woźniak

Faculty of Economic Sciences, University of Warsaw

Warszawa, 20-01-2022

Bootstrap vs Monte Carlo

- Bootstrap
  a) nie znamy rozkładu
  b) jeden wektor realizacji
  c) wiele podpróbek z jednego wektora realizacji (próbki)
- Monte Carlo
  a) znamy rozkład albo proces generujący dane
  b) generowanie wielu realizacji procesu lub danych

Bootstrap dla jednej populacji ([1] str. 101)

Załóżmy, że mamy N-elementową próbkę z populacji.

1. Wylosować ze zwracaniem N-elementową podpróbkę z próbki, którą dysponujemy
2. Powtórzyć podpróbkowanie wiele razy, np. 10000
3. Wyznaczyć bootstrapowy rozkład statystyki

Idea bootstrapu ([1] str. 100)

- Oryginalna próbka przybliża populację z której została wylosowana
- Podpróbki przybliżają to co otrzymalibyśmy losując wiele próbek z populacji
- Bootstrapowy rozkład statystyki bazujący na wielu podpróbkach przybliża rozkład statystyki bazujący na wielu próbkach

Bootstrap

The population is to the sample as the sample is to the bootstrap samples

Ćwiczenie

Dysponujemy próbką z nieznanego rozkładu
-0.15,-0.22,0.09,-2.47,-1.02,-0.69,0.52,-0.04,-0.24,-2.74,-0.18,0.39,
-1.14, 0.69,-0.99,-0.20,-0.40,0.33,1.02,-0.22,0.14,0.33,-1.26,0.23,
-1.12,-1.65,-0.36,2.18,-0.39,-0.98,-0.22,1.63,-2.18,0.22,1.13,0.24,
0.54,-0.85,-1.07,0.25. Interesuje nas oszacowanie średniej rozkładu.

Cele bootstrapu

- Rozkład próbkowy
- Rozkład bootstrapowy

Cele bootstrapu

- Rozkład próbkowy
- Rozkład bootstrapowy

Własności

Dla większości statystyk rozkład bootstrapowy przybliża kształt i obciążenie rozkładu próbkowego

Przedziały ufności - [2] str. 292-301

- Metoda percentylowa (Percentile Method)
- Accelerated Bias-Corrected Percentile Method, BC_a
- The Bootstrap t
- Empirical Variance Stabilization
- Nested Bootstrap and Prepivoting

Bootstrap precentile confidence interval

The interval between 2.5 and 97.5 precentiles of the bootstrap distribution of a statistic is a 95% *bootstrap percentile confidence interval* for the corresponding parameter.

Testowanie hipotez

Peter Hall and Susan R. Wilson, *Two Guidelines for Bootstrap Hypothesis Testing*, Biometrics, Vol. 47, No. 2 (Jun., 1991), pp. 757-762.

First Guideline: Resample $\hat{\theta}^* - \hat{\theta}$, not $\hat{\theta}^* - \theta_0$.

Second Guideline: Base the test on the bootstrap distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$, not on the bootstrap distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ or of $(\hat{\theta}^* - \hat{\theta})$.

Bootstraping Regression - [2] str. 290-291

- Bootstrapping the residuals
  1. Start by fitting the regression model to the observed data and obtaining the fitted responses $\hat{y}_i$ and residuals $\hat{\varepsilon}_i$.
  2. Sample a bootstrap set of residuals, $\{\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*\}$, from the set of fitted residuals, completety at random with replacement.
  3. Create a set of pseudo-responses, $Y_i^* = \hat{y}_i + \hat{\varepsilon}_i^*$ dla $i = 1, \ldots, n$.
  4. Regress Y_i^* on x to obtain a bootstrap estimate $\hat{\beta}^*$.
  5. Repeat this process many times to build an empirical distribution for $\hat{\beta}^*$

Bootstraping Regression - [2] str. 290-291

- Paired bootstrap
  1. Suppose, that the data from an observational study, where both response and predictors are measured from a collection of individuals selected at random.
  2. In this case, the data pairs $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ can be viewed as values observed for i.i.d.random variables $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ drawn from a joint response-predictor distribution.
  3. To bootstrap, sample $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_n^*$ completely at random with replacement from the set of observed data pairs, $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$.
  4. Apply the regression model to the resulting pseudo-dataset to obtain a bootstrap parameter estimate $\hat{\beta}^*$.
  5. Repeat these steps many times, then proceed to inference...

Bootstraping Regression - [2] str. 290-291

- Paired bootstrap
  1. Suppose, that the data from an observational study, where both response and predictors are measured from a collection of individuals selected at random.
  2. In this case, the data pairs $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ can be viewed as values observed for i.i.d.random variables $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ drawn from a joint response-predictor distribution.
  3. To bootstrap, sample $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_n^*$ completely at random with replacement from the set of observed data pairs, $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$.
  4. Apply the regression model to the resulting pseudo-dataset to obtain a bootstrap parameter estimate $\hat{\beta}^*$.
  5. Repeat these steps many times, then proceed to inference...

Bootstrapping the residuals vs Paired bootstrap

If you have doubts about the adequacy of the regression model, the constancy of the resiudal variance, or other regression assumptions, the paired bootstrap will be less sensitive to violations in the assumptions than will bootstrapping the residuals.

Literatura

[1] Laura Chihara, Tim Hesterberg, *Mathematical Statistics with Resampling and R*, John Wiley&Sons, 2011.

[2] Geof H. Givens, Jennifer A. Hoeting, *Computational Statistics 2nd Ed.*, John Wiley&Sons, 2013.

[3] Peter Hall and Susan R. Wilson, *Two Guidelines for Bootstrap Hypothesis Testing*, Biometrics, Vol. 47, No. 2 (Jun., 1991), pp. 757-762.