An Essay on Linear Regression Bootstrapping

Pham Viet Hung

Content

Resampling

Introduction

- Extract all the info from the data
- Earlier: a statistic is a RV having a probability distribution, that is, a sampling distribution of the statistic. Sometimes, no need to approximate the properties ("relative accuracy") of an estimator with some strong prior hypothesis. E.x.: OLS is a BLUE (Gauss-Markov Theorem).
- However, most of the time, those hypothesis do not apply. An estimation of a statistic is needed $\Longrightarrow$ Jackknife, Bootstrap

Jackknife

Algorithm 1: The (Leave-One-Out) Jackknife

Input : $S_n = (X_1, X_2, \ldots, X_n)$ – observed samples of size n

Output: $J_{\theta_{jack}} = [\hat{\theta}^*_{(-1)}, \ldots, \hat{\theta}^*_{(-n)}]$ – jackknifed parameters

Init: $N :=$ sample size

$J_{\theta_{jack}} := [\ ]$ – an empty array of parameter estimates

T – A function of a statistic

for $i := 1$ to N **do**

$S^*_{(-i)} :=$ remove i^{th} element from S_n

$\hat{\theta}^*_{(-i)} := T\left(S^*_{(-i)}\right)$

$J_{\theta_{jack}}[i] := \hat{\theta}^*_{(-i)}$

end

Bootstrap

Algorithm 2: The Bootstrap

Input : $\hat{F}_n = (X_1, X_2, \ldots, X_n)$ – observed sample of size n

Output: $L_{\theta_{boot}} = [\hat{\theta}^*_1, \ldots, \hat{\theta}^*_n]$ – bootstrapped parameters

Init: B – # of bootstrap repetitions

$b := 1$

$L_{\theta_{boot}} := [\,]$ – an empty array of parameter estimates

T – A function of a statistic

while $b < B + 1$ **do**

$\hat{F}^{*(i)}_n$:= a sample size n from S_n by **sampling with replacement**

$\hat{\theta}^*_b := T\left(\hat{F}^{*(b)}_n\right)$

$L_{\theta_{boot}}[b] := \hat{\theta}^*_b$

$b{+}{+}$

end

Bias-Corrected and accelerated (BCa)

- bootstrap CDF: $\hat{G}(x) = \dfrac{\#\{\hat{\theta}_b^* \leq x\}}{B}$
- α^{th} percentile point $\hat{\theta}^{*(\alpha)}$ of the bootstrap distribution: $\hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha)$
- bias-correction factor: $\hat{z}_0 = \Phi^{-1}\left(\dfrac{\#\{\hat{\theta}_b^* < \hat{\theta}\}}{B}\right) \longrightarrow$ to measure the difference between the median of $\hat{\theta}^*$ and $\hat{\theta}$ in normal units.
- acceleration factor: $\hat{a} = \dfrac{1}{6} \dfrac{\sum_{i=1}^{n}\left(\hat{\theta}^*_{(\cdot)} - \hat{\theta}^*_{(-i)}\right)^3}{\left[\sum_{i=1}^{n}\left(\hat{\theta}^*_{(\cdot)} - \hat{\theta}^*_{(-i)}\right)^2\right]^{3/2}} \longrightarrow$ corrects for skewness in the bootstrap sampling distribution using **jackkinfe**.

Bias-Corrected and accelerated (BCa)

The $100\,(1-\alpha)\%$ **BCa confidence interval** is given by

$$\left[\hat{\theta}^{*(\alpha_1)},\, \hat{\theta}^{*(\alpha_2)}\right] = \left[\hat{G}^{-1}(\alpha_1),\, \hat{G}^{-1}(\alpha_2)\right] \tag{1}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{\alpha}\left(\hat{z}_0 + z_{\alpha/2}\right)}\right)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{\alpha}\left(\hat{z}_0 + z_{1-\alpha/2}\right)}\right), \tag{2}$$

and z_α is the 100αth percentile point of the standard normal distribution.

Speeding up BCa: Inner Group Jackknife, **Inner Random Subset Jackknife**

Linear Regression

Model Specification

The multivariable regression model with n observations and d independent variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_d X_{id} + \epsilon_i \tag{3}$$

Model assumptions:

- OLSA.1: Observations are Mutually Independent
- OLSA.2: Linear Model DGP and Strict Exogeneity
- OLSA.3: Condtional Homoskedasticity
- OLSA.4: Conditionally Uncorrelated Error Terms
- OLSA.5: No Exact Collinearity Between Regressors
- OLSA.6: Conditionally Normally Distributed Errors

Gauss-Markov Theorems

OLS Estimator: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}\,\mathbf{Y}$

Gauss-Markov Theorem states that If the conditions *OLSA.1 – OLSA.5* hold true, the OLS estimator $\hat{\boldsymbol{\beta}}$ is the **Best Linear Unbiased Estimator** (BLUE). It is also consistent with the true parameter values of the multivariate linear regression model.regression model.

Modern Gauss-Markov Theorem says that if $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ then $\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \leq \mathbb{V}_{\hat{\boldsymbol{\beta}}}$

However, what about relaxing the OLSA.3 (condtional homoskedasticity) assumption?

White Test for Heteroscedasticity

1. Estimate the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ using OLS.
2. Obtain the predicted $\hat{\boldsymbol{Y}}$ and the squared residual $\hat{\boldsymbol{\epsilon}}^2$.
3. Run the OLS for $\hat{\boldsymbol{\epsilon}} \odot \hat{\boldsymbol{\epsilon}} = \delta_0 + \delta_1 \hat{\boldsymbol{Y}} + \delta_2 \hat{\boldsymbol{Y}} \odot \hat{\boldsymbol{Y}} + u_i$. Retrieve the R-squared value $\boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2}$ from this regression.
4. For the following hypothesis test,

$$\begin{cases} H_0 : & \delta_1 = \delta_2 = 0 \\ H_1 : & \exists i \in \{1, 2\} : \delta_i \neq 0 \,, \end{cases}$$

   apply either the $F \dfrac{\boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2} \,/\, 2}{\left(1 - \boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2}\right) / \left(n-3\right)} \sim F_{2,\, n-3}$ or $\mathrm{LM} = n\boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2} \sim \chi^2_2$.
5. Calculate the p-value.

Heteroskedasticity-Consistent Covariance Matrix Estimator

The error covariance matrix is $\mathbb{V}_{\hat{\boldsymbol{\beta}}} := \mathbb{V}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

The estimated HCCME is $\mathbb{V}_{\hat{\boldsymbol{\beta}}}^{\mathsf{HCCME}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

Revolutionary idea by White: $\hat{\mathbb{V}}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2 \xrightarrow{a.s.} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\epsilon_i^2\right)$.

From which the asymptotic covariance matrix is $\left(\boldsymbol{X}^T\boldsymbol{X}/n\right)^{-1}\hat{\mathbb{V}}_n\left(\boldsymbol{X}^T\boldsymbol{X}/n\right)^{-1}$.

The finite version becomes the HC0: $\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{\mathsf{HC0}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

HC3: $\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{\mathsf{HC3}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$, where h_{ii} from $\boldsymbol{H}$.

Bootstrap Regression

Methods

Applying bootstrap principles on OLS. These techniques are nonparametric bootstrapping.

3 ways to do that:

- Pairs Bootstrap
- Residual Bootstrap
- Wild Bootstrap

However, from the modelling perspective, Wild Bootstrap is the desirable one.

Wild Bootstrap

The wild bootstrap DGP: $\boldsymbol{Y}_i^* = \boldsymbol{X}_{i.}^T \hat{\boldsymbol{\beta}} + \xi_i^* f(\hat{\epsilon}_i)$ where ξ_i^* is a random variable.

Mammen's two-point distribution:

$$\xi_i^* = \begin{cases} -\dfrac{\sqrt{5}-1}{2} & \text{with probability } \dfrac{\sqrt{5}+1}{2\sqrt{5}} \approx 0.7236 \\ \dfrac{\sqrt{5}+1}{2} & \text{with probability } \dfrac{\sqrt{5}-1}{2\sqrt{5}} \approx 0.2764. \end{cases} \tag{4}$$

Rademacher distribution:

$$\xi_i^* = \begin{cases} -1 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases} \tag{5}$$

Wild Bootstrap

- When conditional uncorrelatedness is assumed, for large sample size it's not worth using wild bootstrap to estimate the standard errors of the coefficients. A simple OLS with HCCME is a surprisingly good estimator in this case. However, bootstrap confidence interval may be worth using even for larger sample sizes.
- When conditional uncorrelatedness (OLSA.4) is **NOT** assumed, even for larger sample sizes it is worth using wild bootstrap. This is called **wild cluster bootstrap**.

Framework of the Simulations

- How bootstrap distributions and confidence intervals of regression coefficients change in wild bootstrap simulations when increasing the sample size and bootstrap replication numbers.
- CEU MicroData's Hungarian balance sheet data in the commerce sector from 2014.
- Python package ols_bootstrap
- The model: $ln\,(w_i/L_i) = \beta_0 + \beta_1\,ln(L_i) + \beta_2 D_i + \epsilon_i \qquad \forall i = 1, \ldots, n$ where
- $\beta_0 = 6.7833$ for the constant
- $\beta_1 = 0.2105$ for the log_employment ($ln(L_i)$)
- $\beta_2 = 0.3226$ for the is_exporter (D_i)

Confidence Interval Test

sample size	100	600	1100
BCa	0: **60** 1: **940**	0: **46** 1: **954**	0: **56** 1: **944**
BC	0: **59** 1: **941**	0: **44** 1: **956**	0: **55** 1: **945**
Percentile	0: **58** 1: **942**	0: **44** 1: **956**	0: **55** 1: **945**
Reverse Percentile	0: **58** 1: **942**	0: **45** 1: **955**	0: **56** 1: **944**

TABLE 4.1: The 95% Confidence Interval Test for $\hat{\beta}_1$

sample size	100	600	1100
BCa	0: **67** 1: **933**	0: **58** 1: **942**	0: **46** 1: **954**
BC	0: **59** 1: **941**	0: **51** 1: **949**	0: **42** 1: **958**
Percentile	0: **60** 1: **940**	0: **51** 1: **949**	0: **46** 1: **954**
Reverse Percentile	0: **57** 1: **943**	0: **50** 1: **950**	0: **47** 1: **953**

TABLE 4.2: The 95% Confidence Interval Test for $\hat{\beta}_2$

Bootstrap Distribution Test

https://github.com/pvh95/ols_bootstrap/blob/masterthesis2022/bca_application.ipynb

Thank you very much for your attention!

Reference 1

- Rao, C. R. (1989). *Statistics and Truth. Putting Chance to Work*. International Co-operative Publishing House, Burtonsville, Md.
- Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap*. New York, NY, USA: Springer Verlag.
- Tukey, J. W. (1958). *Bias and Confidence in not-quite Large Samples (Abstract)*. In: The Annals of Mathematical Statistics 29, p. 614
- Efron, B. (1979). *Bootstrap Methods: Another Look at the Jackknife: Annual Statistics*. In.
- Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Reference 2

- Efron, B. & Balasubramanian N. (2020). *The Automatic Construction of Bootstrap Confidence Intervals*. In: Journal of Computational and Graphical Statistics 29.3, pp. 608–619.
- Hansen, B. (2022b). *A Modern Gauss-Markov Theorem*. In: Econometrica 90.3, pp. 1283–1294
- White, H. (1980). *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. Econometrica: journal of the Econometric Society, 817-838.
- Davidson, R. & Flachaire, E. (2008). *The Wild Bootstrap, Tamed at Last*. In: Journal of Econometrics 146.1, pp. 162–169.

Reference 3

- MacKinnon, J.G. and White, H., 1985. *Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties*. Journal of econometrics, 29(3), pp.305-325.

- MacKinnon, J. G. (2013). *Thirty Years of Heteroskedasticity-Robust Inference*. In: Recent Advances and Future Firections in Causality, Prediction, and Specification Analysis. Springer, pp. 437–461.

- Liu, R.Y. (1988). *Bootstrap procedures under some non-I.I.D. models*. Annals of Statistics, 16, 1696–1708.

- Mammen, E. (1993). *Bootstrap and wild bootstrap for high dimensional linear models*. Annals of Statistics 21, 255–285.