Department of Probability and Mathematical Statistics

MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

Ondřej Mifek, Sophia Theodora Klímová

Bootstrap in Regression

23 March 2026

Definition. Regression model

Let $(Y_1, \mathbf{X_1})^T, \ldots, (Y_n, \mathbf{X_n})^T \overset{\text{iid}}{\sim} (Y, \mathbf{X})^T$. We say that $(Y, X)^T$ satisfy the regression model if

$$Y = m(\mathbf{X}) + \epsilon,$$

where $\mathbb{E}[\epsilon|\mathbf{X}] = 0$.

- Y - outcome/response/dependent variable
- $\mathbf{X}$ - regressors/predictors/independent variables
- ϵ - regression error, unobservable
- $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ - regression function, deterministic but unknown
- $\mathbb{E}[Y|\mathbf{X}] = \mathbb{E}[m(\mathbf{X})|\mathbf{X}] + \mathbb{E}[\epsilon|\mathbf{X}] = m(\mathbf{X})$

Parametric, nonparametric and semiparametric regression

Regression:

- parametric - $m \in \{m_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^p\}$
  - simple linear
    - $m(x) = \beta_0 + \beta_1 x$, kde $\theta = (\beta_0, \beta_1)^T$
  - multiple linear - m linear combination of regressors
  - nonlinear - m nonlinear function of regressors
- nonparametric - m satisfies a smoothness condition (usually also $\epsilon \sim N(0, \sigma^2)$)
- semiparametric

The aims of regression:

- estimate m or θ
- confidence sets for $m(x)$ or θ
- confidence bands for m
- lack-of-fit test, i.e. testing $H_0 : m \in \{m_\theta \mid \theta \in \Theta\}$
  - $H_0 : \exists \theta_0 \in \Theta;\ m = m_{\theta_0}$

Bootstrap in regression

Bootstrap in regression:

- residual bootstrap
  - $\mathbf{x}$ fixed, $\text{var}(\epsilon|\mathbf{X}) = \sigma^2 < \infty$
  - parametric regression
- wild (external) bootstrap
  - $\mathbf{x}$ nonrandom, $\text{var}(\epsilon|\mathbf{X} = \mathbf{x}) = \sigma^2(x) < \infty$
  - parametric/nonparametric regression
- paired (naive) bootstrap
  - $\mathbf{X}$ and ϵ are independent and $\text{var}(\epsilon|\mathbf{X}) = \sigma^2 < \infty$
- smooth residual bootstrap
  - $\mathbf{x}$ fixed, errors absolutely continuous
  - nonparametric regression
  - heteroscedastic, homoscedastic

Bootstrap in regression - general information

Parametric regression:

- θ (true parameter in real world) vs $\hat{\theta}_n$
- $\hat{\theta}_n$ (”true parameter” in bootstrap world) vs $\hat{\theta}_n^*$

Nonparametric regression:

- m vs $\hat{m}_n$
- $\hat{m}_n$ vs $\hat{m}_n^*$ $(=> \hat{m}_n(\mathbf{x}^*) = \mathbb{E}^*[Y^*|\mathbf{X}^* = \mathbf{x}^*])$
- $m^* = \hat{m}_n$

Real world: $Y = m(X) + \epsilon$, where $\mathbb{E}[\epsilon|\mathbf{X}] = 0$
Bootstrap world: $Y^* = m^*(X^*) + \epsilon^*$, where $\mathbb{E}^*[\epsilon^*|\mathbf{X}^*] = 0$

Residual bootstrap

Regression model, $\mathbf{x}$ fixed and $\text{var}(\epsilon|\mathbf{X}) = \sigma^2 < \infty$

- $Y^* = m^*(X^*) + \epsilon^*$, $\mathbb{E}^*[\epsilon^*|\mathbf{X}^*] = 0$
- $\forall i \in \{1, ..., n\}$ define $X_i^* = X_i$ and $Y_i^* = \hat{m}_n(X_i^*) + \epsilon_i^*$
  - $\hat{m}_n$ - consistent estimator of m
  - $\epsilon_1^*, ..., \epsilon_n^*$ drawn with replacement from:
    - residuals $\hat{\epsilon_1}, ..., \hat{\epsilon_n}$, where $\hat{\epsilon}_i = Y_i - \hat{m}_n(\mathbf{X_i})$
    - centered residuals $\hat{\epsilon_1} - \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i, ..., \hat{\epsilon_n} - \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i$

Residual bootstrap

Generating bootstrap regression errors ϵ_i^*:

- We want: $\mathbb{E}^*[\epsilon_i^* | \mathbf{X}_\mathbf{i}^*] = 0$
- $\mathbb{E}^*[\epsilon_i^* | \mathbf{X}_\mathbf{i}^*] = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i$
  - = 0 in linear regression models with intercept
  - = $o\left(\frac{1}{\sqrt{n}}\right)$ in many nonparametric regression models
    - in the asymptotics, it does not matter if we draw them from $\hat{\epsilon_1}, ..., \hat{\epsilon_n}$ or $\hat{\epsilon_1} - \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i, ..., \hat{\epsilon_n} - \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i$

Residual bootstrap for heteroscedastic regression models

Does not work.

Regression model, $\mathbf{x}$ fixed, $\mathrm{var}(\epsilon|\mathbf{X} = \mathbf{x}) = \sigma^2(x) < \infty$.

$$\mathrm{Var}^*(\varepsilon_i^* \mid X_i^*) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i - \frac{1}{n}\sum_{j=1}^{n}\hat{\varepsilon}_j\right)^2 \xrightarrow{P} \mathrm{Var}(\varepsilon) \neq \mathrm{Var}(\varepsilon \mid X).$$

Wild bootstrap

Regression model, $\mathbf{x}$ fixed, $\text{var}(\epsilon|\mathbf{X} = \mathbf{x}) = \sigma^2(x) < \infty$.

- $X_i^* = X_i$
- $Y_i^* = \hat{m}_n(\mathbf{X_i}) + \varepsilon_i^*$
  - $\hat{m}_n$ consistent estimator of m
  - $\epsilon_i^* = V_i \hat{\epsilon}_i$, V_i independent, centered with unit variance, and independent on the original sample.
- $\text{Var}^*(\epsilon_i^*|X_i^*) = \hat{\epsilon}_i^2$

$X_t = aX_{t-1} + \varepsilon_t, \quad X_0 = 0, a \in (0,1), t \in \{1, \dots, n\},$
$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, i.i.d.

X_t not i.i.d.
MS3:

$$\sqrt{n}\left(\begin{pmatrix} \hat{a}_n \\ \hat{\sigma}_n^2 \end{pmatrix} - \begin{pmatrix} a \\ \sigma^2 \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1-a^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix} \right)$$

Bootstrap:

$$\hat{\varepsilon}_t = X_t - \hat{a}_n X_{t-1}$$

Random sample with replacement

$$\varepsilon_1^\star, \ldots, \varepsilon_n^\star \quad \text{from} \quad \hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$$

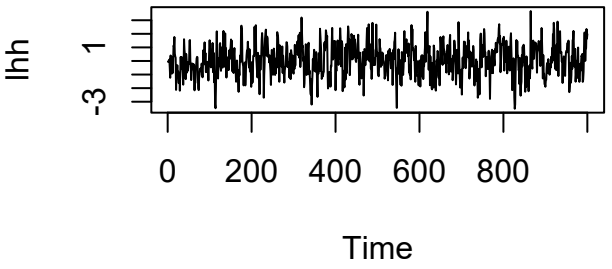
$X_0^\star = 0$,

$$X_t^\star = \hat{a}_n X_{t-1}^\star + \varepsilon_t^\star$$

Repeat B times - get B bootstrap estimates of a.

Ar(1)

$X_t = 0,5X_{t-1} + \varepsilon_t, \quad X_0 = 0, n = 1000, \quad \varepsilon_t \sim \mathcal{N}(0,1)$, i.i.d.



$X_t = 0,5X_{t-1} + \varepsilon_t, \quad X_0 = 0, n = 1000, \quad \varepsilon_t \sim \mathcal{N}(0,1)$, i.i.d.

$\hat{a}_n = 0,509$,
Asymptotic CI: (0,456, 0,563).
B = 500:

$$\frac{1}{500}\sum_{i=1}^{500} \hat{a}^{\star}_{(i)} = 0,507.$$

Bootstrap CI: (0,461, 0,567).

$X_t = 0,5X_{t-1} + \varepsilon_t, \quad X_0 = 0, n = 10\,000, \quad \varepsilon_t \sim \mathcal{N}(0,1)$, i.i.d.

$\hat{a}_n = 0,503$,
Asymptotic CI: (0,486, 0,520).
B = 2 000:

$$\frac{1}{2000}\sum_{i=1}^{2000} \hat{a}^{\star}_{(i)} = 0,503.$$

Bootstrap CI: (0,487, 0,520).

$X_t = 0,8X_{t-1} + \varepsilon_t, \quad X_0 = 0, n = 400, \quad \varepsilon_t \sim \mathcal{N}(0,1)$, i.i.d.

$\hat{a}_n = 0,816$,
Asymptotic CI: (0,760, 0,872).
B = 500:

$$\frac{1}{500}\sum_{i=1}^{500} \hat{a}^{\star}_{(i)} = 0,808.$$

Bootstrap CI: (0,770, 0,890).

Simple linear regression model

$$Y_i = \beta X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

$$\hat{\beta}_n = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

CLT + Slutsky + law of large numbers

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \tau^2),$$

where

$$\tau^2 = \frac{\mathrm{Var}(X_1 Y_1 - \beta X_1^2)}{(\mathbb{E}X_1^2)^2} = \frac{\mathrm{Var}(\varepsilon_1 X_1)}{(\mathbb{E}X_1^2)^2}.$$

Assuming homoskedasticity

$$\frac{\mathrm{Var}(\varepsilon_1 X_1)}{(\mathbb{E}X_1^2)^2} = \frac{\sigma^2 \mathbb{E}X_1^2}{(\mathbb{E}X_1^2)^2} = \frac{\sigma^2}{\mathbb{E}X_1^2} \approx \frac{\mathrm{SSE}/(n-1)}{(1/n)\sum_{i=1}^n x_i^2}$$

With normality

$$\frac{c^\top \hat{\beta}_n - c^\top \beta}{\sqrt{\hat{\sigma}_e^2 \, c^\top (X^\top X)^{-1} c}} \sim t_{n-p}.$$

Naive bootstrap?
Bootstrap, draw from (no intercept)

$$\hat{\varepsilon}_i - \frac{1}{n}\sum_{j=1}^{n} \hat{\varepsilon}_j.$$

Define

$$Y_i^\star = \hat{\beta}_n X_i + \varepsilon_i^\star,$$

and

$$\hat{\beta}_n^\star = \frac{\sum_{i=1}^n X_i Y_i^\star}{\sum_{i=1}^n X_i^2}.$$

Feller-Lindeberg

$$\sqrt{n}(\hat{\beta}_n^\star - \hat{\beta}_n) \xrightarrow{d} \mathcal{N}(0, \tau^2).$$

Simple linear regression model

$$Y_i = 2X_i + \varepsilon_i, \quad i = 1, \ldots, 100.$$

Where

$$\varepsilon_i \sim \mathcal{N}(0, 4), \quad X_i \sim U(0, 10).$$

$\hat{\beta} = 1,979$,
CI: (1,913, 2,046).
B = 2000

$$\frac{1}{2000} \sum_{i=1}^{2000} \hat{\beta}^{\star}_{(i)} = 1,980.$$

CI as before (1,916, 2,041)
Asymptotic bootstrap CI (1,913, 2,045)

Heteroscedastic linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, 100.$$

Where

$$\varepsilon_i | X_i \sim \mathcal{N}(0, \sigma^2(X_i)), \quad X_i \sim U(0, 2), \quad \sigma^2(X_i) = e^{2X_i},$$

and

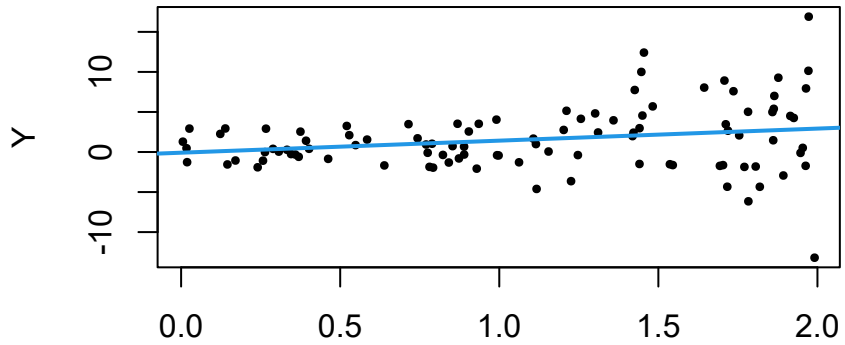
$$\beta_0 = 0, \quad \beta_1 = 1.$$

Summary

$$\hat{\beta}_0 = -0,086, \quad \hat{\beta}_1 = 1,505$$

B = 1000

Heteroscedastic linear regression model



Heteroscedastic linear regression model

- Nonparametric naive bootstrap
- Studentized nonparametric bootstrap
  Uses sandwich

$$\frac{\hat{\beta}_i^\star - \hat{\beta}_i}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}}$$

- Model based bootstrap - residuals not i.i.d.

We compute

$$\widehat{SE}_{\hat{\beta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\beta}^{\star}_{(i)} - \overline{\hat{\beta}^{\star}}\right)^2},$$

where

$$\overline{\hat{\beta}^{\star}} = \frac{1}{B}\sum_{i=1}^{B}\hat{\beta}^{\star}_{(i)}.$$

And classical

$$\sqrt{\hat{\sigma}^2(X^\top X)^{-1}}.$$

Heteroscedastic linear regression model

Method	beta0	beta1
Standard LM	0.825	0.656
Model boot.	0.846	0.670
Naive boot.	0.575	0.765
Studentized boot.	0.576	0.759

Heteroscedastic linear regression model

- Using t-distribution + sandwich SE (no bootstrap in this)
- Using t-distribution + naive bootstrap $\widehat{SE}_{\hat{\beta}}$

Heteroscedastic linear regression model

Method	**2.5%**	**97.5%**
Standard LM	0.2036	2.81
Model boot.	0.1341	2.79
Naive boot.	0.0292	2.98
Studentized boot.	-0.1921	2.99
Bootstrap SE	-0.0125	3.02
Sandwich	-0.0669	3.08

Heteroscedastic linear regression model

Method	Coverage	Cover Lower	Cover Upper	Avg. Length
Standard LM	0.903	0.954	0.949	2.50
Naive boot.	0.939	0.968	0.971	2.84
Studentized boot.	0.943	0.971	0.972	3.03
Bootstrap SE	0.947	0.972	0.975	2.89
Sandwich	0.940	0.966	0.974	2.88

Homoscedastic linear regression model

Method	Coverage	Cover Lower	Cover Upper	Avg. Length
Standard LM	0.945	0.969	0.976	2.20
Naive boot.	0.935	0.965	0.970	2.16
Studentized boot.	0.945	0.969	0.976	2.22
Bootstrap SE	0.942	0.967	0.975	2.19
Sandwich	0.938	0.967	0.971	2.21

Thank you for your attention