

Lecture Notes V – Bootstrap for Regression

Marina Meila
`mmp@stat.washington.edu`

Department of Statistics
University of Washington

February 2025

Linear Regression refresher

Empirical Bootstrap for linear regression

Reading: Ch.

Prediction

- ▶ **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \dots (x^n, y^n)\}$
- ▶ **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- ▶ **Outputs** $y^{1:n}$
- ▶ **Goal** Learn/estimate $f(x)$ **predictor** for y
- ▶ By type of output
 - ▶ **Classification** if $y \in S_Y$ discrete
 - ▶ $y \in \{0, 1\}$ (or $\{\pm 1\}$) **binary classification**
 - ▶ $y \in \{1, 2, \dots m\}$ **multiclass classification**
 - ▶ **Regression** if $y \in \mathbb{R}$ continuous

Prediction

- ▶ **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \dots (x^n, y^n)\}$
- ▶ **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- ▶ **Outputs** $y^{1:n}$

- ▶ **Goal** Learn/estimate $f(x)$ **predictor** for y

- ▶ By type of output
 - ▶ **Classification** if $y \in S_Y$ discrete
 - ▶ $y \in \{0, 1\}$ (or $\{\pm 1\}$) **binary classification**
 - ▶ $y \in \{1, 2, \dots m\}$ **multiclass classification**
 - ▶ **Regression** if $y \in \mathbb{R}$ continuous

- ▶ **Model family** $\mathcal{F} = \{f_\theta : \mathbb{R}^d \rightarrow S_Y, \theta \in \Theta\}$
 - ▶ $\mathcal{F} = \{ \text{linear functions} \}$ linear regression/classification
 - ▶ $\mathcal{F} = \{ \text{polynomials of degree 2, 3, } \dots \}$ polynomial regression/classification
 - ▶ $\mathcal{F} = \{ \frac{1}{1 + e^{-\beta^T x + \beta_0}} \}$ logistic regression
 - ▶ **neural network** regression/classification
 - ▶ **kernel** regression/classification
 - ▶ $\mathcal{F} = \{ \text{monotonic functions} \}$ **isotonic regression**
 - ▶ **support vector** regression/classification
 - ▶ regression/classification **trees** (and **random forests**)

Linear regression

Model

- ▶ $y^i = \beta_0 + \beta_1 x^i + \epsilon^i$ for $x^{1:n} \in \mathbb{R}$ (univariate regression)
- ▶ $y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \dots + \beta_d x_d^i + \epsilon^i$ for $x^{1:n} \in \mathbb{R}^d$ (multivariate regression)
 - $y \in \mathbb{R}$ is **output/response**
 - $x_j^{1:n}$ for $j = 1 : d$ are **input(s)/covariates/features/attributes/...**
 - $\beta_{1:d}$ are **regression coefficients**, β_0 is **intercept**
 - $\epsilon^{1:n} \in \mathbb{R}$ is **noise**, $\epsilon^{1:n} \sim N(0, \sigma^2)$ i.i.d.

Linear regression model in matrix form

For a single data point (x^i, y^i)

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_d \end{bmatrix} \in \mathbb{R}^d \quad x^i = \begin{bmatrix} x_1^i \\ x_2^i \\ \dots \\ x_d^i \end{bmatrix} \in \mathbb{R}^d \quad (1)$$

Then,

$$y^i = \beta_0 + (x^i)^T \beta + \epsilon^i \quad (2)$$

For all $\mathcal{D}$

$$y = \begin{bmatrix} y^1 \\ y^2 \\ \dots \\ y^n \end{bmatrix} \in \mathbb{R}^n \quad X = \begin{bmatrix} (x^1)^T \\ (x^2)^T \\ \dots \\ (x^n)^T \end{bmatrix} \in \mathbb{R}^{n \times d} \quad \epsilon = \begin{bmatrix} \epsilon^1 \\ \epsilon^2 \\ \dots \\ \epsilon^n \end{bmatrix} \quad (3)$$

$$y = \beta_0 \mathbf{1} + X\beta + \epsilon \quad \text{with } \text{Cov}(\epsilon) = \sigma^2 I_d \quad (4)$$

Solution by Maximum Likelihood

- ▶ **Likelihood** Probability $y | \mathbf{X}$, parameters
- ▶ **Parameters** $\beta_0, \beta_{1:d}, \sigma^2$
- ▶ $\beta_0^{ML}, \beta_{1:d}^{ML}, (\sigma^2)^{ML} = \operatorname{argmax}_{\beta_0, \beta_{1:d}, \sigma^2} l(y | \mathbf{X}, \beta_0, \beta_{1:d}, \sigma^2)$

The (log)-likelihood

- ▶ What is random? **the noise** $\epsilon^{1:n}$
- ▶ Express noise as function of $(x^{1:n}, y^{1:n})$

$$\epsilon^i = y^i - \beta_0 - \beta^T x^i \sim N(0, \sigma^2) \quad (5)$$

- ▶ Likelihood

- ▶ Let $p_{0,\sigma^2}(\epsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\epsilon^2}{2\sigma^2}} = N(\epsilon; 0, \sigma^2)$

- ▶ Then

$$L(\beta_0, \beta_{1:d}, \sigma^2) = \prod_{i=1}^n p_{0,\sigma^2}(\epsilon^i) \quad (6)$$

$$= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\epsilon^i)^2}{2\sigma^2}} \quad (7)$$

$$= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y^i - \beta_0 - \beta^T x^i)^2}{2\sigma^2}} \quad (8)$$

- ▶ **log-likelihood**

$$l(\beta_0, \beta_{1:d}, \sigma^2) = \quad (9)$$

$$= \sum_{i=1}^n \left\{ -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) - \frac{1}{2} (y^i - \beta_0 - \beta^T x^i)^2 \frac{1}{2\sigma^2} \right\} \quad (10)$$

$$= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) + \text{constant} - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \beta_0 - \beta^T x^i)^2 \quad (11)$$

Maximizing the log-likelihood w.r.t β

- ▶ For simplicity, let $\beta_0 = 0$; hence $y^i = \beta^T x^i + \epsilon^i$
- ▶ log-likelihood

$$l(\beta_{1:d}, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \beta_0 - \beta^T x^i)^2 + \text{constant} \quad (12)$$

- ▶ For any σ^2 ,

$$\operatorname{argmax}_{\beta} l(\sigma^2, \beta) = \operatorname{argmin}_{\beta} \sum_{i=1}^n (y^i - \beta_0 - \beta^T x^i)^2 \quad (13)$$

a Least Squares Problem

- ▶ In matrix form $\min_{\beta} \|y - X\beta\|^2$
- ▶ Solution

$$\beta^{ML} = (X^T X)^{-1} X^T y \quad (14)$$

with $(X^T X)^{-1} X^T \equiv X^\dagger$ the **pseudoinverse** of X

- ▶ β^{ML} is **linear** in y !

Statistical properties of β^{ML}

- ▶ Assume the true model is $y^i = \beta^T x^i + \epsilon^i$ with $\epsilon \sim N(0, \sigma^2)$.
- ▶ Here β, σ^2 are **true parameters** and we assume we know them.
- ▶ β^{ML} is a random variable. Let us calculate its mean and standard deviation.
- ▶ **Expectation** of β^{ML}

$$E[\beta^{ML}] = E[X^\dagger y] = E[X^\dagger (X\beta + \epsilon)] \quad (15)$$

$$= E[X^\dagger X]\beta + E[X^\dagger \epsilon] \quad (16)$$

$$= \underbrace{X^\dagger X}_{I_d} \beta + X^\dagger \underbrace{E[\epsilon]}_0 = \beta \quad (17)$$

Hence, $E[\beta^{ML}] = \beta$ and we say that β^{ML} is **unbiased**

- ▶ **Covariance** of β^{ML}
- ▶ By a similar (but longer) calculation, we obtain that

$$\text{Cov}(\beta^{ML}) = \sigma^2 (X^T X)^{-1} \in \mathbb{R}^{d \times d} \quad (18)$$

- ▶ In fact, β^{ML} has a Normal distribution. Remember from (15)

$$\beta^{ML} = \beta + X^\dagger \epsilon. \quad (19)$$

This is a linear transformation of ϵ , a Gaussian variable. Therefore,

$$\beta^{ML} \sim N(\beta, \sigma^2 (X^T X)^{-1}). \quad (20)$$

- ▶ This is a **multivariate Normal** distribution over $\mathbb{R}^d$,

Estimation of σ^2

- ▶ Let $\hat{\epsilon}^i = y^i - (x^i)^T \beta^{ML}$, for $i = 1 : n$
- ▶ $\hat{\epsilon}^i$ called the **residuals**
- ▶ If we plug in β^{ML} in the log-likelihood equation (12) we obtain

$$l(\sigma^2, \beta^{ML}) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (\hat{\epsilon}^i)^2 + \text{constant} \quad (21)$$

- ▶ Maximizing this expression w.r.t. σ^2 gives

$$(\sigma^2)^{ML} = \frac{1}{n} \sum_{i=1}^n (\hat{\epsilon}^i)^2 \quad (22)$$

- ▶ The predictor is $f(x) = x^T \beta^{ML}$
- ▶ Hence, for a new $x \in \mathbb{R}^d$, our guess of y is $\hat{y} = f(x)$

Empirical Bootstrap for linear regression

Residual Bootstrap for linear regression

Wild Bootstrap for linear regression

