MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Linear Regression Models

Segment 6 – Advanced Topics in Linear Regression

Topic 1 – Bootstrapping Regression Models

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

1. Statistic and Sampling Distribution

2. Bootstrap - Idea and Need

3. The Bootstrap Algorithm for Linear Regression

4. An Example Using Heteroskedastic Data

5. Bootstrap Applications and Limitations

Statistic and Sampling Distribution

Statistic and Sampling Distribution

- Recall the difference between a population parameter and sample statistic:

Statistic and Sampling Distribution

- Recall the difference between a population parameter and sample statistic: the average height of all biological females in a city is a population parameter;

Statistic and Sampling Distribution

- Recall the difference between a population parameter and sample statistic: the average height of all biological females in a city is a population parameter; the average height of n randomly chosen biological females from that city is a sample statistic.

Statistic and Sampling Distribution

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Recall the difference between a population parameter and sample statistic: the average height of all biological females in a city is a population parameter; the average height of n randomly chosen biological females from that city is a sample statistic.
- Sample statistic (or just statistic) is a *random variable* and has an associated distribution referred to as its *sampling distribution*.

Statistic and Sampling Distribution

- Recall the difference between a population parameter and sample statistic: the average height of all biological females in a city is a population parameter; the average height of n randomly chosen biological females from that city is a sample statistic.
- Sample statistic (or just statistic) is a *random variable* and has an associated distribution referred to as its *sampling distribution*.
- Recall that the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ of the corresponding population parameters $\beta_0, \beta_1, \ldots, \beta_p$ in a multiple linear regression model are also statistics.

Statistic and Sampling Distribution

- Recall the difference between a population parameter and sample statistic: the average height of all biological females in a city is a population parameter; the average height of n randomly chosen biological females from that city is a sample statistic.
- Sample statistic (or just statistic) is a *random variable* and has an associated distribution referred to as its *sampling distribution*.
- Recall that the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ of the corresponding population parameters $\beta_0, \beta_1, \ldots, \beta_p$ in a multiple linear regression model are also statistics.
- The OLS estimates have an associated variability depending on the dataset used to build them.

Bootstrap - Idea and Need

Bootstrap - Idea and Need

- A beautiful visual tool for understanding bootstrap: external web link

Bootstrap - Idea and Need

- A beautiful visual tool for understanding bootstrap: external web link
- In the bootstrap, we *resample* from the data which is already sampled from the population.

Bootstrap - Idea and Need

- A beautiful visual tool for understanding bootstrap: external web link
- In the bootstrap, we *resample* from the data which is already sampled from the population.
- We use the resampled data to calculate OLS estimates, associated standard errors, confidence intervals etc.

Bootstrap - Idea and Need

- A beautiful visual tool for understanding bootstrap: external web link
- In the bootstrap, we *resample* from the data which is already sampled from the population.
- We use the resampled data to calculate OLS estimates, associated standard errors, confidence intervals etc.
- Bootstrap is needed in situations when

Bootstrap - Idea and Need

- A beautiful visual tool for understanding bootstrap: external web link
- In the bootstrap, we *resample* from the data which is already sampled from the population.
- We use the resampled data to calculate OLS estimates, associated standard errors, confidence intervals etc.
- Bootstrap is needed in situations when (1) data collected is small w.r.t. sample size;

Bootstrap - Idea and Need

- A beautiful visual tool for understanding bootstrap: external web link
- In the bootstrap, we *resample* from the data which is already sampled from the population.
- We use the resampled data to calculate OLS estimates, associated standard errors, confidence intervals etc.
- Bootstrap is needed in situations when (1) data collected is small w.r.t. sample size; (2) assumptions we made in the population model might have been violated;

Bootstrap - Idea and Need

- A beautiful visual tool for understanding bootstrap: external web link
- In the bootstrap, we *resample* from the data which is already sampled from the population.
- We use the resampled data to calculate OLS estimates, associated standard errors, confidence intervals etc.
- Bootstrap is needed in situations when (1) data collected is small w.r.t. sample size; (2) assumptions we made in the population model might have been violated; (3) need sampling distributions for much more complex estimation procedures, where no closed form expressions exist.

The Bootstrap Algorithm for Linear Regression

The Bootstrap Algorithm for Linear Regression

From a dataset with n samples, draw n samples *with replacement*.

The Bootstrap Algorithm for Linear Regression

From a dataset with n samples, draw n samples *with replacement*.

Population

The Bootstrap Algorithm for Linear Regression

From a dataset with n samples, draw n samples *with replacement*.

$$\text{Population} \Rightarrow \underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix}}_{\text{Sample}}$$

The Bootstrap Algorithm for Linear Regression

From a dataset with n samples, draw n samples *with replacement*.

Population $\Rightarrow$ $\underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix}}_{\text{Sample}} \Rightarrow$

The Bootstrap Algorithm for Linear Regression

From a dataset with n samples, draw n samples *with replacement*.

Population $\Rightarrow \underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix}}_{\text{Sample}} \Rightarrow \left\{ \underbrace{\begin{bmatrix} x_1^{(3)} & x_2^{(3)} \\ x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{bmatrix}, \begin{bmatrix} y^{(3)} \\ y^{(1)} \\ y^{(2)} \end{bmatrix}}_{\text{Bootstrap sample-1}} \right.$

The Bootstrap Algorithm for Linear Regression

From a dataset with n samples, draw n samples *with replacement*.

Population $\Rightarrow$ $\underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix}}_{\text{Sample}} \Rightarrow$

$\underbrace{\begin{bmatrix} x_1^{(3)} & x_2^{(3)} \\ x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{bmatrix}, \begin{bmatrix} y^{(3)} \\ y^{(1)} \\ y^{(2)} \end{bmatrix}}_{\text{Bootstrap sample-1}}$ estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

The Bootstrap Algorithm for Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

From a dataset with n samples, draw n samples *with replacement*.

Population $\Rightarrow$ $\underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix}}_{\text{Sample}} \Rightarrow$

$\underbrace{\begin{bmatrix} x_1^{(3)} & x_2^{(3)} \\ x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{bmatrix}, \begin{bmatrix} y^{(3)} \\ y^{(1)} \\ y^{(2)} \end{bmatrix}}_{\text{Bootstrap sample-1}}$ estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

$\vdots$

The Bootstrap Algorithm for Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

From a dataset with n samples, draw n samples *with replacement*.

Population $\Rightarrow$ $\underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix}}_{\text{Sample}} \Rightarrow$

$\underbrace{\begin{bmatrix} x_1^{(3)} & x_2^{(3)} \\ x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{bmatrix}, \begin{bmatrix} y^{(3)} \\ y^{(1)} \\ y^{(2)} \end{bmatrix}}_{\text{Bootstrap sample-1}}$ estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

$\vdots$

$\underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(1)} \\ y^{(2)} \end{bmatrix}}_{\text{Bootstrap sample-B}}$

The Bootstrap Algorithm for Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

From a dataset with n samples, draw n samples *with replacement*.

Population $\Rightarrow$ $\underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \end{bmatrix}}_{\text{Sample}} \Rightarrow$

$\underbrace{\begin{bmatrix} x_1^{(3)} & x_2^{(3)} \\ x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{bmatrix}, \begin{bmatrix} y^{(3)} \\ y^{(1)} \\ y^{(2)} \end{bmatrix}}_{\text{Bootstrap sample-1}}$ estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

$\vdots$

$\underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{bmatrix}, \begin{bmatrix} y^{(1)} \\ y^{(1)} \\ y^{(2)} \end{bmatrix}}_{\text{Bootstrap sample-B}}$ estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

An Example Using Heteroskedastic Data

An Example Using Heteroskedastic Data

- Consider a dataset drawn from the population model $Y = X + \epsilon$,

An Example Using Heteroskedastic Data

- Consider a dataset drawn from the population model $Y = X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = X^4)$.

An Example Using Heteroskedastic Data

- Consider a dataset drawn from the population model $Y = X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = X^4)$.
- This dataset has heteroskedasticity;

An Example Using Heteroskedastic Data

- Consider a dataset drawn from the population model $Y = X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = X^4)$.
- This dataset has heteroskedasticity; the random error variance is not constant.

An Example Using Heteroskedastic Data

- Consider a dataset drawn from the population model $Y = X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = X^4)$.
- This dataset has heteroskedasticity; the random error variance is not constant.
- The assumptions for multiple linear regression to draw statistical inferences and for performing hypothesis tests for the coefficient estimates are *violated*.

An Example Using Heteroskedastic Data

- Consider a dataset drawn from the population model $Y = X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = X^4)$.
- This dataset has heteroskedasticity; the random error variance is not constant.
- The assumptions for multiple linear regression to draw statistical inferences and for performing hypothesis tests for the coefficient estimates are *violated*.
- We will use bootstrap to investigate the effect on the standard error of the coefficients reported by the `lm( )` function of R.

Bootstrap Applications and Limitations

Bootstrap Applications and Limitations

- Bootstrap can be used to measure variability in model estimates for modeling strategies such as ridge and lasso regression.

Bootstrap Applications and Limitations

- Bootstrap can be used to measure variability in model estimates for modeling strategies such as ridge and lasso regression.
- Bootstrap can be used to estimate test error in a prediction setting instead of cross validation.

Bootstrap Applications and Limitations

- Bootstrap can be used to measure variability in model estimates for modeling strategies such as ridge and lasso regression.
- Bootstrap can be used to estimate test error in a prediction setting instead of cross validation.
- The number of bootstrap samples should be *large* to dilute the effect of resampling in the estimated distribution, which is dictated by available computational power.

Bootstrap Applications and Limitations

- Bootstrap can be used to measure variability in model estimates for modeling strategies such as ridge and lasso regression.
- Bootstrap can be used to estimate test error in a prediction setting instead of cross validation.
- The number of bootstrap samples should be *large* to dilute the effect of resampling in the estimated distribution, which is dictated by available computational power.
- The dataset used for drawing bootstrap samples should be *representative* of the original population model.

Summary

Summary

- Core idea behind the bootstrap approach.

Summary

- Core idea behind the bootstrap approach.
- Bootstrap for linear regression and comparison with standard approach.

Summary

- Core idea behind the bootstrap approach.
- Bootstrap for linear regression and comparison with standard approach.
- Limitations of the bootstrap approach.