QTM 347 Machine Learning

Lecture 8: Bootstrap

Ruoxuan Xiong

Suggested reading: ISL Chapter 5

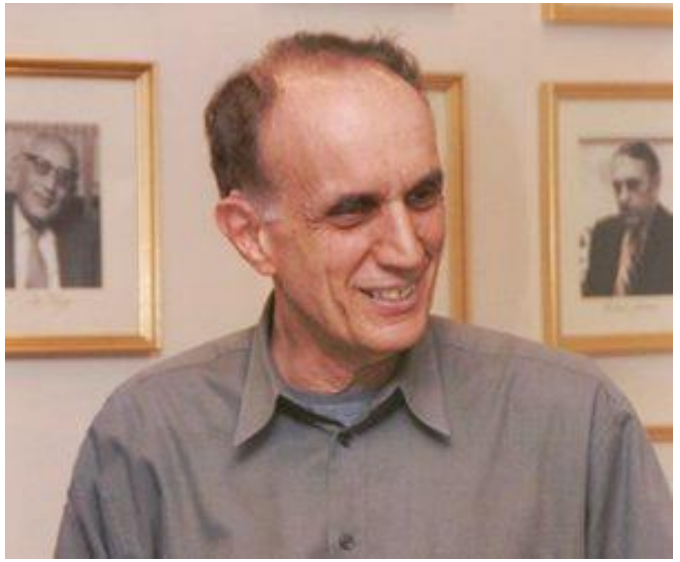
EMORY

Lecture plan

- Bootstrap

Cross-validation vs. Bootstrap

- Cross-validation: Provide the test error with an independent validation set
- Bootstrap: Provide the standard error of model estimates
  - One of the most important techniques in all of Statistics
  - Computationally intensive
  - Popularized by Brad Efron (Stanford)



Standard errors

- **Definition:** Standard error is the standard deviation of an estimate from a sample set of size n
  - Example: linear regression

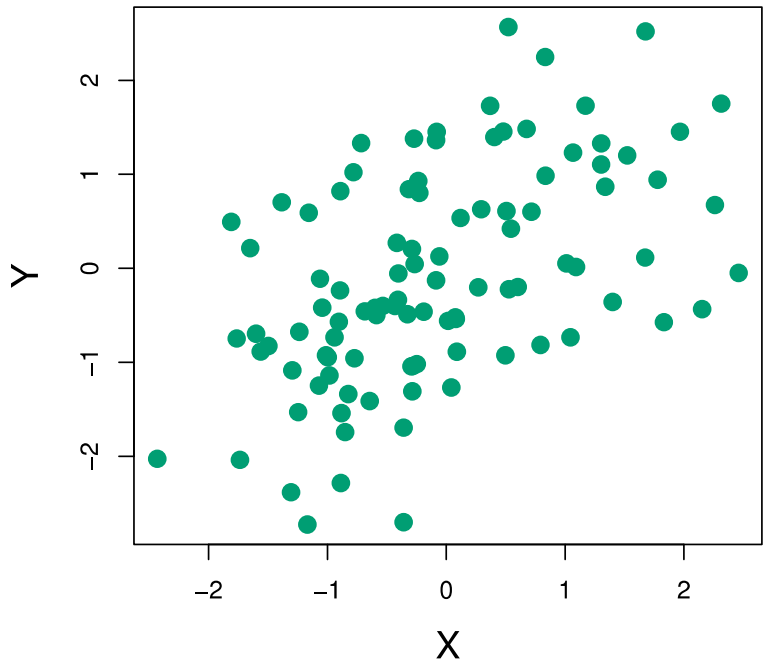
```
Residuals:
    Min      1Q  Median      3Q     Max
-15.594  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.761e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-Squared: 0.7406,     Adjusted R-squared: 0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

In many cases, we do not have a formula to calculate standard errors

- **Example**
  - Investing in two assets
  - Suppose that X and Y are the returns of two assets
  - These returns are observed every day: $(X_1, Y_1), \cdots, (X_n, Y_n)$



Example

- We have a fixed amount of money to invest: α fraction on X and $1-\alpha$ fraction on Y
  - Therefore, our return will be: $\alpha X + (1-\alpha)Y$
- We want to solve α that minimizes the variance of our return

$$\min_{\alpha} \mathrm{Var}(\alpha X + (1-\alpha)Y)$$

  - Solve α from the first order derivative $\frac{\mathrm{d}\,\mathrm{Var}(\alpha X+(1-\alpha)Y)}{\mathrm{d}\,\alpha} = 0$
  - The optimal α is: $\alpha = \frac{\sigma_Y^2 - \mathrm{Cov}(X,Y)}{\sigma_X^2+\sigma_Y^2-2\mathrm{Cov}(X,Y)}$ (a take-home exercise)
    - σ_X^2 is the variance of X; σ_Y^2 is the variance of Y
    - $\mathrm{Cov}(X,Y)$ is the covariance between X and Y

Example

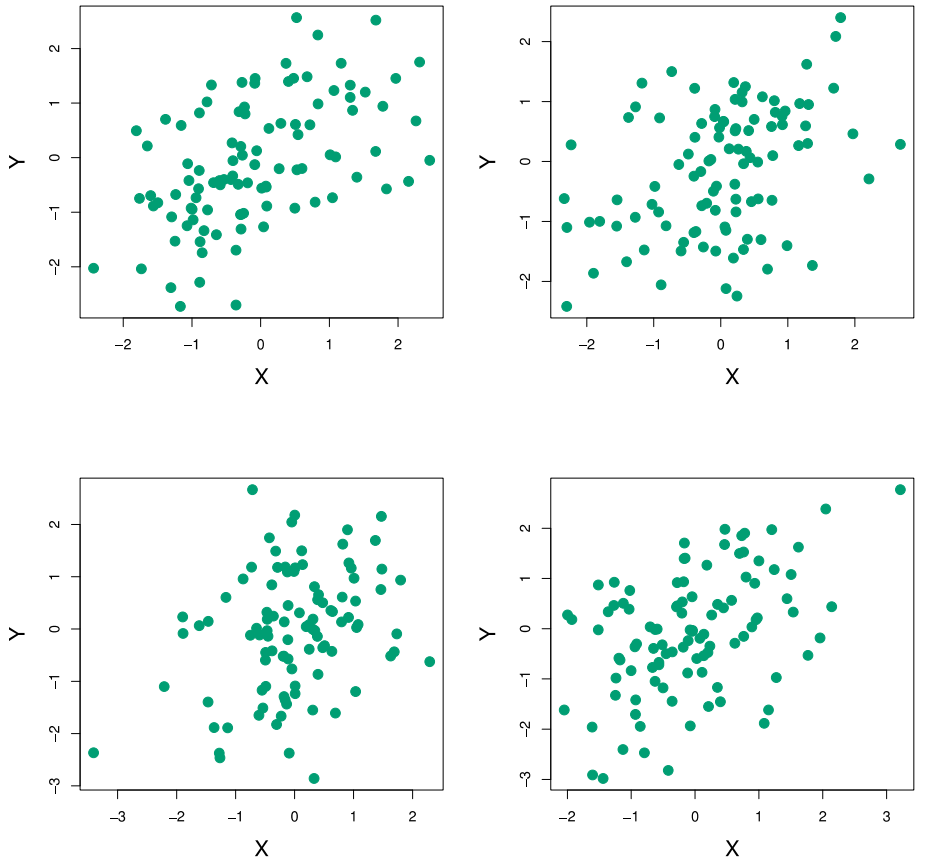
- We can approximate $\alpha = \frac{\sigma_Y^2 - \mathrm{Cov}(X,Y)}{\sigma_X^2 + \sigma_Y^2 - 2\mathrm{Cov}(X,Y)}$ with the observed data
  - $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, and $\widehat{\mathrm{Cov}}\,(X,Y)$ are from the observed data
  - Calculate $\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \widehat{\mathrm{Cov}}(X,Y)}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\widehat{\mathrm{Cov}}(X,Y)}$

Thought experiment

- Suppose we compute the estimate $\hat{\alpha} = 0.6$ using the observed data $(X_1, Y_1), \cdots, (X_n, Y_n)$
- How certain is this value?
- If we resample the observations, would we get a wildly different $\hat{\alpha}$ (say 0.1)?
- In this **thought experiment**, we know the actual joint distribution $P(X, Y)$, so we can resample the n observations

Thought experiment

- In this **thought experiment**, we know the actual joint distribution $P(X, Y)$, so we can resample the n observations

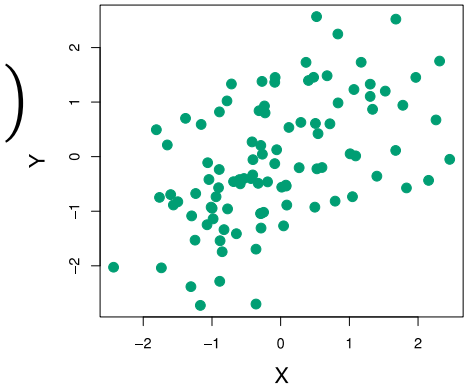


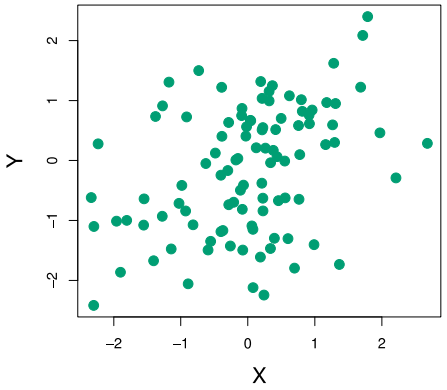
Thought experiment

- Estimate an $\hat{\alpha}$ from each sample

$\left(X_1^{(1)}, Y_1^{(1)}\right), \cdots, \left(X_n^{(1)}, Y_n^{(1)}\right)$

Get $\hat{\alpha}^{(1)}$



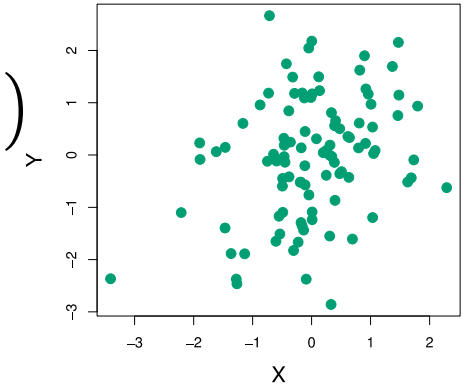


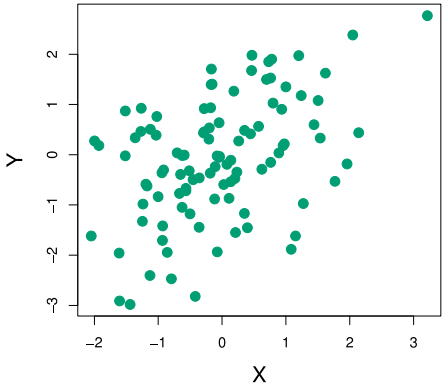
$\left(X_1^{(2)}, Y_1^{(2)}\right), \cdots, \left(X_n^{(2)}, Y_n^{(2)}\right)$

Get $\hat{\alpha}^{(2)}$

$\left(X_1^{(3)}, Y_1^{(3)}\right), \cdots, \left(X_n^{(3)}, Y_n^{(3)}\right)$

Get $\hat{\alpha}^{(3)}$





$\left(X_1^{(4)}, Y_1^{(4)}\right), \cdots, \left(X_n^{(4)}, Y_n^{(4)}\right)$

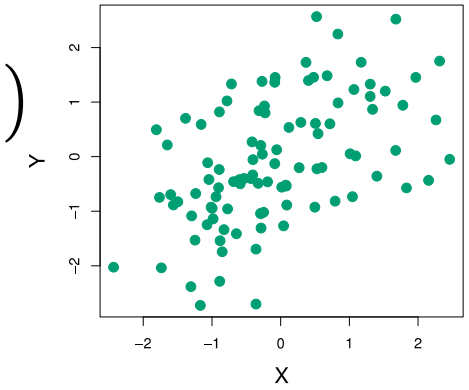
Get $\hat{\alpha}^{(4)}$

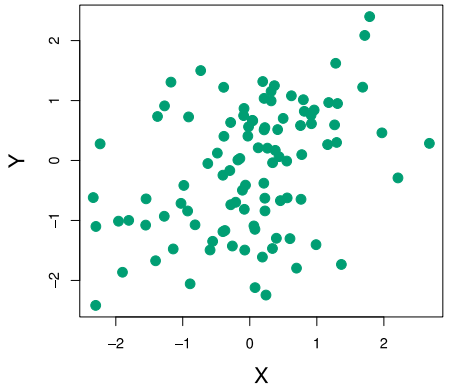
Thought experiment

- **Standard error** of $\hat{\alpha}$ is approximated by the **standard deviation** of $\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}, \hat{\alpha}^{(3)}, \hat{\alpha}^{(4)}, \ldots$

$\left(X_1^{(1)}, Y_1^{(1)}\right), \cdots, \left(X_n^{(1)}, Y_n^{(1)}\right)$

Get $\hat{\alpha}^{(1)}$



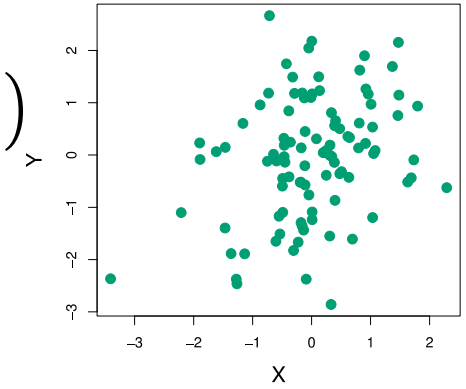


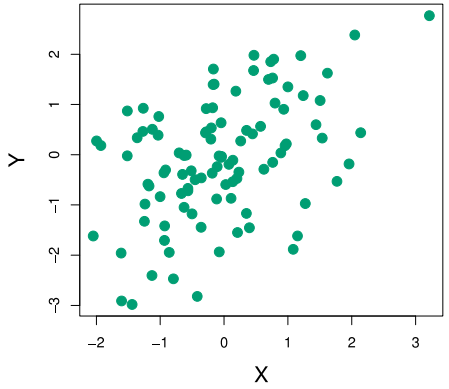
$\left(X_1^{(2)}, Y_1^{(2)}\right), \cdots, \left(X_n^{(2)}, Y_n^{(2)}\right)$

Get $\hat{\alpha}^{(2)}$

$\left(X_1^{(3)}, Y_1^{(3)}\right), \cdots, \left(X_n^{(3)}, Y_n^{(3)}\right)$

Get $\hat{\alpha}^{(3)}$



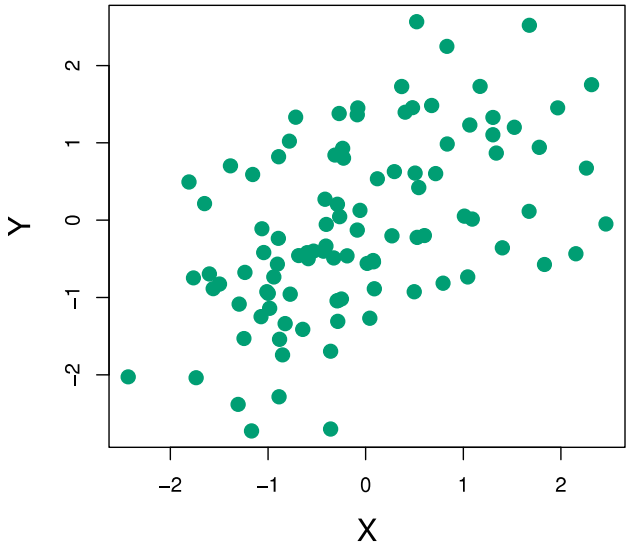


$\left(X_1^{(4)}, Y_1^{(4)}\right), \cdots, \left(X_n^{(4)}, Y_n^{(4)}\right)$

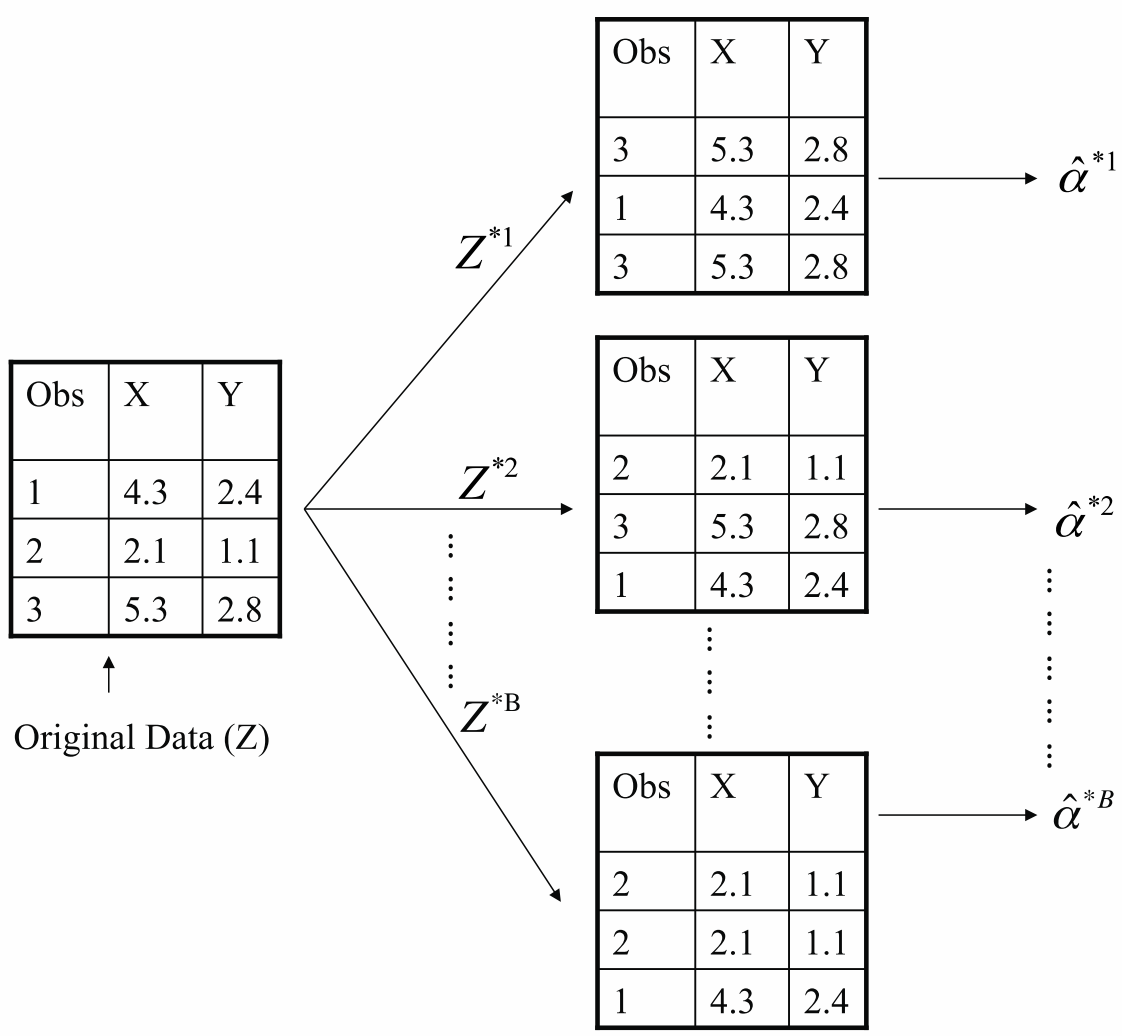
Get $\hat{\alpha}^{(4)}$

Bootstrap

- Back to reality: we cannot resample the data ☹
  - However, we can use the training data set to approximate the joint distribution of X and Y
- **Bootstrap**: Resample from the empirical distribution
  - Resample the data by drawing n samples **with replacement** from the actual observations
  - $\hat{P}(X = x, Y = y) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i = x, y_i = y)$



Bootstrap



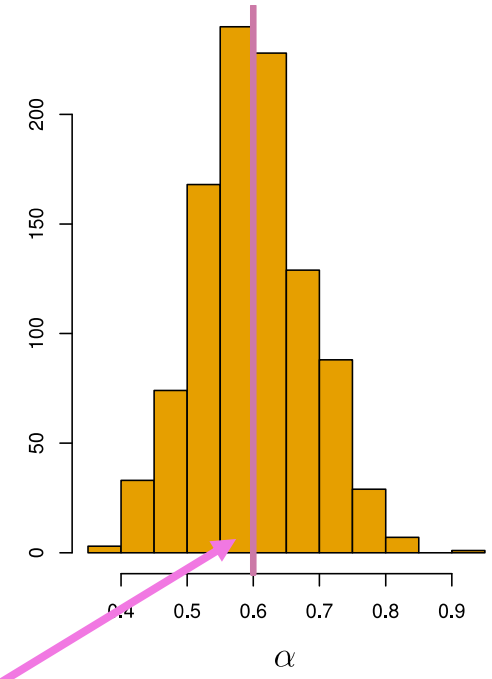
We have a fixed amount of money to invest: α fraction on X and $1-\alpha$ fraction on Y

Estimate the standard error of $\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \widehat{\mathrm{Cov}}(X,Y)}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\widehat{\mathrm{Cov}}(X,Y)}$

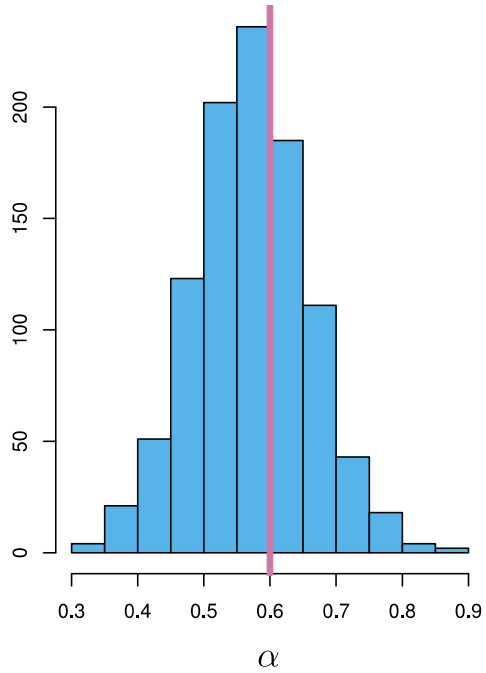
Use the standard error of $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \cdots, \hat{\alpha}^{*B}$ to approximate the standard error of $\hat{\alpha}$.

Bootstrap vs. Resampling from the true distribution

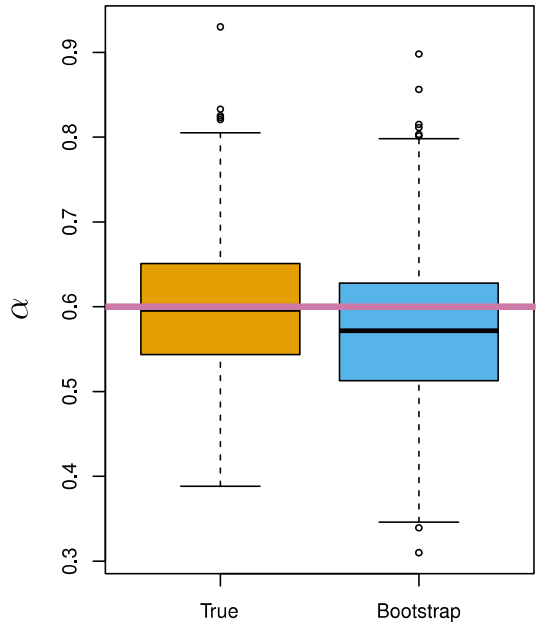
Histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population



True value of α



Histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set



Bootstrap can be used in broader settings

- Example 1: Estimate the standard error of the mean of medv
- Example 2: Estimate the standard error of the coefficient of lstat in the regression to predict medv
- Example 3: Estimate the standard error of the predicted medv for lstat = 22