Resampling Methods

Haixia Liu

1. Cross Validation

2. Bootstrap

About Resampling

- Pretending the data as population,
- repeatedly draw sample from the data,
- refitting a model of interest on each sample.
- Main task: assess the validity/accuracy of statistical methods and models.
  - **Cross-validation**: estimate the *test error* of models
  - **Bootstrap**: quantify the *uncertainty* of estimators

Training error is not sufficient enough

- training error easily computable with training data.
- because of possibility of over-fit, it cannot be used to properly assess test error.
- Test error would be also easily computable, if test data are well designated.

Training error is not sufficient enough

- training error easily computable with training data.
- because of possibility of over-fit, it cannot be used to properly assess test error.
- Test error would be also easily computable, if test data are well designated.
- Normally we are just given ... data.
- Shall have to create "test data" for the purpose of computing test error.

The validation set approach

Artificially separate data into “training data” and “test data” for validation purpose.

The validation set approach

Artificially separate data into "training data" and "test data" for validation purpose.

- The "test data" here should be more accurately called *validation data* or hold out data,
- meaning that they not used in training.
- Model fitting only uses the training data.

Ideal scenario for performance assessment

- In a "data-rich" scenario, we can afford to separate the data into three parts:
    - *training data*: used to train various models.
    - *validation data*: used to assess the models and identify the best.
    - *test data*: test the results of the best model.

Ideal scenario for performance assessment

- In a "data-rich" scenario, we can afford to separate the data into three parts:
    - *training data*: used to train various models.
    - *validation data*: used to assess the models and identify the best.
    - *test data*: test the results of the best model.
- Usually, people also call validation data or hold-out data as test data.

Validation and Cross validation

- Validation set approach.
- LOOCV (Leave-one-out cross valiation)
- K-fold cross validation.

Validation set approach

A set of n observations are *randomly* split into

- **a training set**
  – shown in blue, containing observations 7, 22, and 13, among others,
- **a validation set**
  – shown in beige, and containing observation 91, among others).

Figure: 5.1. A schematic display of the validation set approach. The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

A summary

- The validation estimate of the test error rate can be highly variable, depending on the random split.

A summary

- The validation estimate of the test error rate can be highly variable, depending on the random split.
- Only a subset of the observations–the training set are used to fit the model.

A summary

- The validation estimate of the test error rate can be highly variable, depending on the random split.
- Only a subset of the observations–the training set are used to fit the model.
- Statistical methods tend to perform worse when trained on fewer observations.

Drawback of validation set approach

- Our ultimate goal is to produce the best model with best prediction accuracy.
- Validation set approach has a drawback of using ONLY training data to fit model.
- The validation data do not participate in model building but only model assessment.
- A "waste" of data.
- We need more data to participate in model building.

Drawback of validation set approach

- Our ultimate goal is to produce the best model with best prediction accuracy.
- Validation set approach has a drawback of using ONLY training data to fit model.
- The validation data do not participate in model building but only model assessment.
- A "waste" of data.
- We need more data to participate in model building.
- We need to improve it.

Cross Validation.

The leave-one-out cross-validation

Suppose the data contain n data points: $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$.

- *Pick data point 1 as validation set, the rest as training set.*
  - Fit the model on the training set,
  - evaluate the test error on the validation set, $\text{MSE}_1 = (y_1 - \hat{y}_1)^2$.

The leave-one-out cross-validation

Suppose the data contain n data points: $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$.

- *Pick data point 1 as validation set, the rest as training set.*
  - Fit the model on the training set,
  - evaluate the test error on the validation set, $\mathrm{MSE}_1 = (y_1 - \hat{y}_1)^2$.
- *Pick data point 2 as validation set, the rest as training set.*
  - Fit the model on the training set,
  - evaluate the test error on the validation set, $\mathrm{MSE}_2 = (y_2 - \hat{y}_2)^2$.

The leave-one-out cross-validation

Suppose the data contain n data points: $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$.

- *Pick data point 1 as validation set, the rest as training set.*
  - Fit the model on the training set,
  - evaluate the test error on the validation set, $\text{MSE}_1 = (y_1 - \hat{y}_1)^2$.
- *Pick data point 2 as validation set, the rest as training set.*
  - Fit the model on the training set,
  - evaluate the test error on the validation set, $\text{MSE}_2 = (y_2 - \hat{y}_2)^2$.
- ..... (repeat the procedure for all data point.)
- Obtain an estimate of the test error by combining the MSE_i, $i =, ..., n$.

The LOOCV estimate for the test MSE is

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i.$$

LOOCV

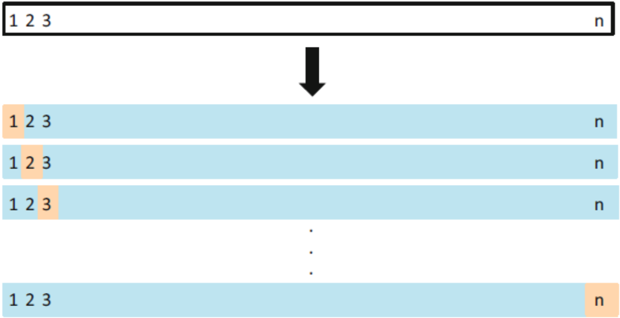


Figure: 5.3. A schematic display of LOOCV.

A set of n data points is repeatedly split into

- a *training* set (shown in blue): containing all but one observation,
- a *validation* set (shown in beige): contains only that observation.

Pros and cons of LOOCV

Advantages:

- Far less bias, since the training data size $(n-1)$ is close to the entire data size (n).
- One single test error estimate (thanks to the averaging), without the variablity validation set approach.

Disadvantage:

- could be computationally expensive since the model need to be fit n times.
- The MSE_i may be too much correlated.

Complexity of LOOCV in linear model?

- Consinder linear model:

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i, \qquad i = 1, .., n$$

  and the fitted values $\hat{y}_i = \mathbf{x}_i^T \hat{\beta}$, where $\hat{\beta}$ is the least squares estimate of β based on all data $(\mathbf{x}_i, y_i), i = 1, ..., n$.
- Using LOOCV, the

$$\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i^{(i)})^2$$

- Looks to be complicated to compute least squares estimate n times.

Recall: Leverage

- Recall the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad \text{as } \hat{y} = \mathbf{H}y.$$

  Let $h_{ij} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_j$ be the (i, j) elements of $\mathbf{H}$.
- The leverage of the i-th observation is just the i-th diagonal element of $\mathbf{H}$, denoted as h_{ii}.
- A high leverage implies that observation is quite influential. Note that the average of h_{ii} is $(p+1)/n$.
- E.g., if h_{ii} is greater than $2(p+1)/n$, twice of the average, is generally considered large.

Simple Formula of LOOCV in linear model

- Easy formula:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{y_i - \hat{y}_i}{1 - h_i}\Big)^2$$

where $\hat{y}_i$ is the fitted values of least squares method based on all data. h_i is the leverage.

Fast computation of cross-validation I

- The leave-one-out cross-validation statistic is given by

$$CV = \frac{1}{n}\sum_{i=1}^{n} e_{[i]}^2,$$

where $e_{[i]} = y_i - \hat{y}_{[i]}$, the observations are given by $y_1, \ldots, y_n$, and $\hat{y}_{[i]}$ is the predicted value obtained when the model is estimated with the ith case deleted.

- Suppose we have a linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. The $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix. It has this name because it is used to compute $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{HY}$. If the diagonal values of $\mathbf{H}$ are denoted by $h_1, \ldots, h_N$, then the leave-one-oout cross-validation statistic can be computed using

$$CV = \frac{1}{N}\sum_{i=1}^{N}[e_i/(1-h_i)]^2,$$

where $e_i = y_i - \hat{y}_i$ is predicted value obtained when the model is estimated with all data included.

Fast computation of cross-validation II

Proof

- Let $\mathbf{X}_{[i]}$ and $\mathbf{Y}_{[i]}$ be similar to $\mathbf{X}$ and $\mathbf{Y}$ but with the ith row deleted in each case. Let $\mathbf{x}_i^T$ be the ith row of $\mathbf{X}$ and let

$$\hat{\boldsymbol{\beta}}_{[i]} = (\mathbf{X}_{[i]}^T \mathbf{X}_{[i]})^{-1} \mathbf{X}_{[i]}^T \mathbf{Y}_{[i]}$$

  be the estimate of $\boldsymbol{\beta}$ without the ith case. Then $e_{[i]} = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{[i]}$.

- Now $\mathbf{X}_{[i]}^T \mathbf{X}_{[i]} = (\mathbf{X}^T \mathbf{X} - \mathbf{x}_i \mathbf{x}_i^T)$ and $\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = h_i$. So by the Sherman-Morrison-Woodbury formula,

$$(\mathbf{X}_{[i]}^T \mathbf{X}_{[i]})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_i}.$$

Fast computation of cross-validation III

Proof

- Also note that $\mathbf{X}_{[i]}^T \mathbf{Y}_{[i]} = \mathbf{X}^T \mathbf{Y} - \mathbf{x} y_i$. Therefore

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{[i]} &= \left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_i} \right] (\mathbf{X}^T\mathbf{Y} - \mathbf{x}_i y_i) \\
&= \hat{\boldsymbol{\beta}} - \left[\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_i} \right] [y_i(1-h_i) - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + h_i y_i] \\
&= \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i/(1-h_i)
\end{aligned}$$

- Thus

$$\begin{aligned}
e_{[i]} &= y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{[i]} \\
&= y_i - \mathbf{x}_i^T \left[\hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i/(1-h_i) \right] \\
&= e_i + h_i e_i/(1-h_i) = e_i/(1-h_i)
\end{aligned}$$

Simplicity of LOOCV in linear model

- One fit (with all data) does it all!
- The prediction error rate (in terms of MSE) is just weighted average of the least squares fit residuals.
- High leverage point gets more weight in prediction error estimation.

K-fold cross validation

- Divide the data into K subsets, sizes of each subset ($\approx n/K$).

K-fold cross validation

- Divide the data into K subsets, sizes of each subset ($\approx n/K$).
- Treat one subset as validation set, the rest as a training set.
  - Run the model fitting on training set.
  - Calculate the test error estimate on the validation set, denoted as MSE_i, say.

K-fold cross validation

- Divide the data into K subsets, sizes of each subset ($\approx n/K$).
- Treat one subset as validation set, the rest as a training set.
  - Run the model fitting on training set.
  - Calculate the test error estimate on the validation set, denoted as MSE_i, say.
- Repeat the procedures over every subset.
- Average over the above K estimates of the test errors,

$$\mathrm{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{MSE}_i.$$

K-fold cross validation

- Divide the data into K subsets, sizes of each subset ($\approx n/K$).
- Treat one subset as validation set, the rest as a training set.
  - Run the model fitting on training set.
  - Calculate the test error estimate on the validation set, denoted as MSE_i, say.
- Repeat the procedures over every subset.
- Average over the above K estimates of the test errors,

$$\text{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K}\text{MSE}_i.$$

- LOOCV is a special case of K-fold cross validation, actually n-fold cross validation.

K-fold cross validation

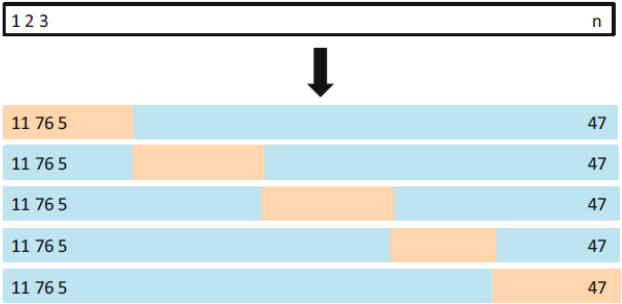


Figure: 5.5. A schematic display of 5-fold CV.

A set of n observations is randomly split into five non-overlapping groups.

- Each of these fifths acts as a validation set (shown in beige),
- the remainder as a training set (shown in blue).

K-fold cross validation

- Common choices of K: $K = 5$ or $K = 10$.
- Advantage over LOOCV:
  1. computationally lighter, espeically for complex model with large data.
  2. Likely less variance (to be addressed later).
- Advantage over validation set approach:
  Less variability resulting from the data-split, thanks to the averaging.

Bias variance trade-off

- In terms of bias of estimation of test error:
  - Validation set approach has more bias due to smaller size of training data;
  - LOOCV is nearly unbiased;
  - K-fold (e.g, $K = 5$ or 10) has intermediate bias.
- K-fold cross validation has smaller variance than that of LOOCV.
  - The n traing sets LOOCV are too similar to each other. As a result, the trained models are too postively correlated.
  - The K training sets of K-fold cross validation are much less similar to each other.
- The K-fold cross validation generally has less variance than LOOCV.

Cross validation for classification

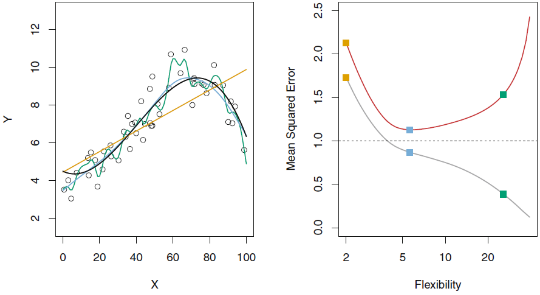
- MSE is a popular criterion to measure predition/estimation accuracy for regression. There are other criteria.

Cross validation for classification

- MSE is a popular criterion to measure predition/estimation accuracy for regression. There are other criteria.
- For classification with qualitative response, a natural choice is:
  - 1 for incorrect classification,
  - 0 for correct classification.

Cross validation for classification

- MSE is a popular criterion to measure predition/estimation accuracy for regression. There are other criteria.
- For classification with qualitative response, a natural choice is:
  - 1 for incorrect classification,
  - 0 for correct classification.
- For LOOCV,
  - Let $y_i^{(i)}$ is the classification of i-th observation based on model fitted not using i-th observation,
    $$\mathrm{Err}_i = I(y_i \neq \hat{y}_i^{(i)}).$$
  - The average number of incorrect classification
    $$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i.$$



- Training error declines in general when model complexity increases. Some times even reaches 0.
- Test error general declines first and then increases.
- 10-fold cross validation provides reasonable estimate of the test error, with slight under-estimation.

Bootstrap as a resampling procedure.

- Suppose we have data $x_1, ..., x_n$, representing the ages of n randomly selected people in Wuhan.
- Use sample mean $\bar{x}$ to estimate the population mean μ, the avearge age of all residents of Wuhan.

Bootstrap as a resampling procedure.

- Suppose we have data $x_1, ..., x_n$, representing the ages of n randomly selected people in Wuhan.
- Use sample mean $\bar{x}$ to estimate the population mean μ, the avearge age of all residents of Wuhan.
- How to justify the estimation error $\bar{x} - \mu$?
  - They rely on normality assumption or central limit theorm.
  - Usually by t-confidence interval, test of hypothesis.

Bootstrap as a resampling procedure.

- Suppose we have data $x_1, ..., x_n$, representing the ages of n randomly selected people in Wuhan.
- Use sample mean $\bar{x}$ to estimate the population mean μ, the avearge age of all residents of Wuhan.
- How to justify the estimation error $\bar{x} - \mu$?
  - They rely on normality assumption or central limit theorm.
  - Usually by t-confidence interval, test of hypothesis.
- Is there another reliable way?

Bootstrap as a resampling procedure.

- Suppose we have data $x_1, ..., x_n$, representing the ages of n randomly selected people in Wuhan.
- Use sample mean $\bar{x}$ to estimate the population mean μ, the avearge age of all residents of Wuhan.
- How to justify the estimation error $\bar{x} - \mu$?
  - They rely on normality assumption or central limit theorm.
  - Usually by t-confidence interval, test of hypothesis.
- Is there another reliable way?
- Just bootstrap.

Boostrap as a resampling procedure.

- Take n random sample (with replacement) from $x_1, ..., x_n$.

Boostrap as a resampling procedure.

- Take n random sample (with replacement) from $x_1, ..., x_n$.
- calculate the sample mean of the "re-sample", denoted as $\bar{x}_1^*$.

Boostrap as a resampling procedure.

- Take n random sample (with replacement) from $x_1, ..., x_n$.
- calculate the sample mean of the "re-sample", denoted as $\bar{x}_1^*$.
- Repeat the above a large number M times. We have $\bar{x}_1^*, \bar{x}_2^*, ..., \bar{x}_M^*$.

Boostrap as a resampling procedure.

- Take n random sample (with replacement) from $x_1, ..., x_n$.
- calculate the sample mean of the "re-sample", denoted as $\bar{x}_1^*$.
- Repeat the above a large number M times. We have $\bar{x}_1^*, \bar{x}_2^*, ..., \bar{x}_M^*$.
- Use the distribution of $\bar{x}_1^* - \bar{x}, ..., \bar{x}_M^* - \bar{x}$ to approximate that of $\bar{x} - \mu$.

Example

- X and Y are two random variables. Then minimizer of $\mathrm{var}(\alpha X + (1-\alpha)Y)$ is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}.$$

Example

- X and Y are two random variables. Then minimizer of $\text{var}(\alpha X + (1-\alpha)Y)$ is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}.$$

- Data: $(X_1, Y_1), ..., (X_n, Y_n)$.
- We can compute sample variances and covariances.
- Estimate α by

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

Example

- X and Y are two random variables. Then minimizer of $\text{var}(\alpha X + (1-\alpha)Y)$ is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}.$$

- Data: $(X_1, Y_1), ..., (X_n, Y_n)$.
- We can compute sample variances and covariances.
- Estimate α by

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

- How to evalute $\hat{\alpha} - \alpha$? (remember $\hat{\alpha}$ is random and α is unknown).

Example

- X and Y are two random variables. Then minimizer of $\text{var}(\alpha X + (1-\alpha)Y)$ is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}.$$

- Data: $(X_1, Y_1), ..., (X_n, Y_n)$.
- We can compute sample variances and covariances.
- Estimate α by

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

- How to evalute $\hat{\alpha} - \alpha$? (remember $\hat{\alpha}$ is random and α is unknown).
- Use Bootstrap.

Example

- Sample n resamples from $(X_1, Y_1), ..., (X_n, Y_n)$ with replacement.

Example

- Sample n resamples from $(X_1, Y_1), ..., (X_n, Y_n)$ with replacement.
- Compute the sample the sample variance and covariances for this resample.

Example

- Sample n resamples from $(X_1, Y_1), ..., (X_n, Y_n)$ with replacement.
- Compute the sample the sample variance and covariances for this resample.
- Compute

$$\hat{\alpha}^* = \frac{(\hat{\sigma}_Y^*)^2 - \hat{\sigma}_{XY}^*}{(\hat{\sigma}_X^*)^2 + (\hat{\sigma}_Y^*)^2 - 2\hat{\sigma}_{XY}^*}.$$

Example

- Sample n resamples from $(X_1, Y_1), ..., (X_n, Y_n)$ with replacement.
- Compute the sample the sample variance and covariances for this resample.
- Compute

$$\hat{\alpha}^* = \frac{(\hat{\sigma}^*_Y)^2 - \hat{\sigma}^*_{XY}}{(\hat{\sigma}^*_X)^2 + (\hat{\sigma}^*_Y)^2 - 2\hat{\sigma}^*_{XY}}.$$

- Repeat this procedure, and we have $\hat{\alpha}^*_1, ..., \hat{\alpha}^*_M$ for a large M.
- Use the distribution of $\hat{\alpha}^*_1 - \hat{\alpha}, ..., \hat{\alpha}^*_M - \hat{\alpha}$ to approximate the distribution of $\hat{\alpha} - \alpha$.

Example

- Sample n resamples from $(X_1, Y_1), ..., (X_n, Y_n)$ with replacement.
- Compute the sample the sample variance and covariances for this resample.
- Compute
$$\hat{\alpha}^* = \frac{(\hat{\sigma}_Y^*)^2 - \hat{\sigma}_{XY}^*}{(\hat{\sigma}_X^*)^2 + (\hat{\sigma}_Y^*)^2 - 2\hat{\sigma}_{XY}^*}.$$
- Repeat this procedure, and we have $\hat{\alpha}_1^*, ..., \hat{\alpha}_M^*$ for a large M.
- Use the distribution of $\hat{\alpha}_1^* - \hat{\alpha}, ..., \hat{\alpha}_M^* - \hat{\alpha}$ to approximate the distribution of $\hat{\alpha} - \alpha$.
- For example, we can use
$$\frac{1}{M}\sum_{j=1}^{M}(\hat{\alpha}_j^* - \hat{\alpha})^2$$
to estimate $E(\hat{\alpha} - \alpha)^2$.

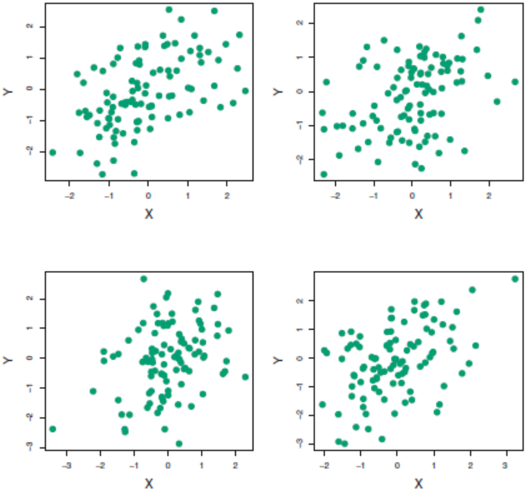
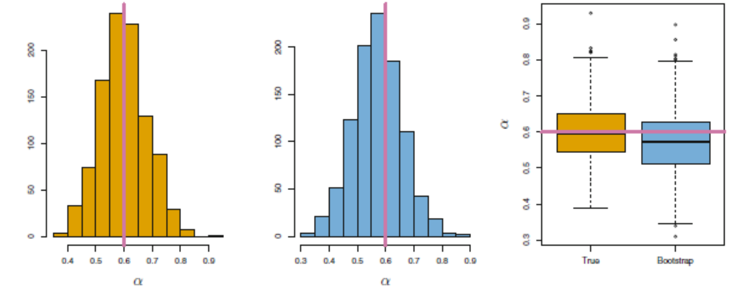


Figure: 5.9. Each panel displays 100 simulated returns for investments X and Y. From left to right and top to bottom, the resulting estimates for α are 0.576, 0.532, 0.657, and 0.651.



- **Left**: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population.
- **Center**: A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set.
- **Right**: The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.