Data driven sampling (Lecture 8)

Resampling Method: Cross-validation, Bootstrap

Hongwei YUAN

University of Macau

Contents

Resampling methods

- Resampling methods involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.
- Resampling approaches can be computationally expensive, because they involve fitting the same statistical method multiple times using different subsets of the training data.
- Due to recent advances in computing power, the computational requirements of resampling methods generally are not prohibitive.
- We study two of the most commonly used resampling methods:
  - Cross-validation,
  - Bootstrap.

Resampling methods

- Cross-validation can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility, i.e.
  - *model assessment*,
  - *model selection*.
- Bootstrap is most commonly used to provide a measure of accuracy of a parameter estimate or of a given statistical method.

Statistical Inferences

A statistical inference model can be generally represented as:

$$Y = f(\mathbf{X}) + \epsilon.$$

$$E[Y] = E[f(X)]$$

where

- Y is a quantitative response; $\mathbf{X} = (X_1, \ldots, X_p)$ is p covariates.
- f is some fixed but unknown function of $\mathbf{X}$.
- ϵ is a random error term, which is independent of X and has mean zero.
- The goal is to estimate f from the observed data:

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n).$$

The obtained estimation of f is denoted by $\hat{f}$.

Statistical Inferences: Sale v.s. Advertisements

The statistical inference model is

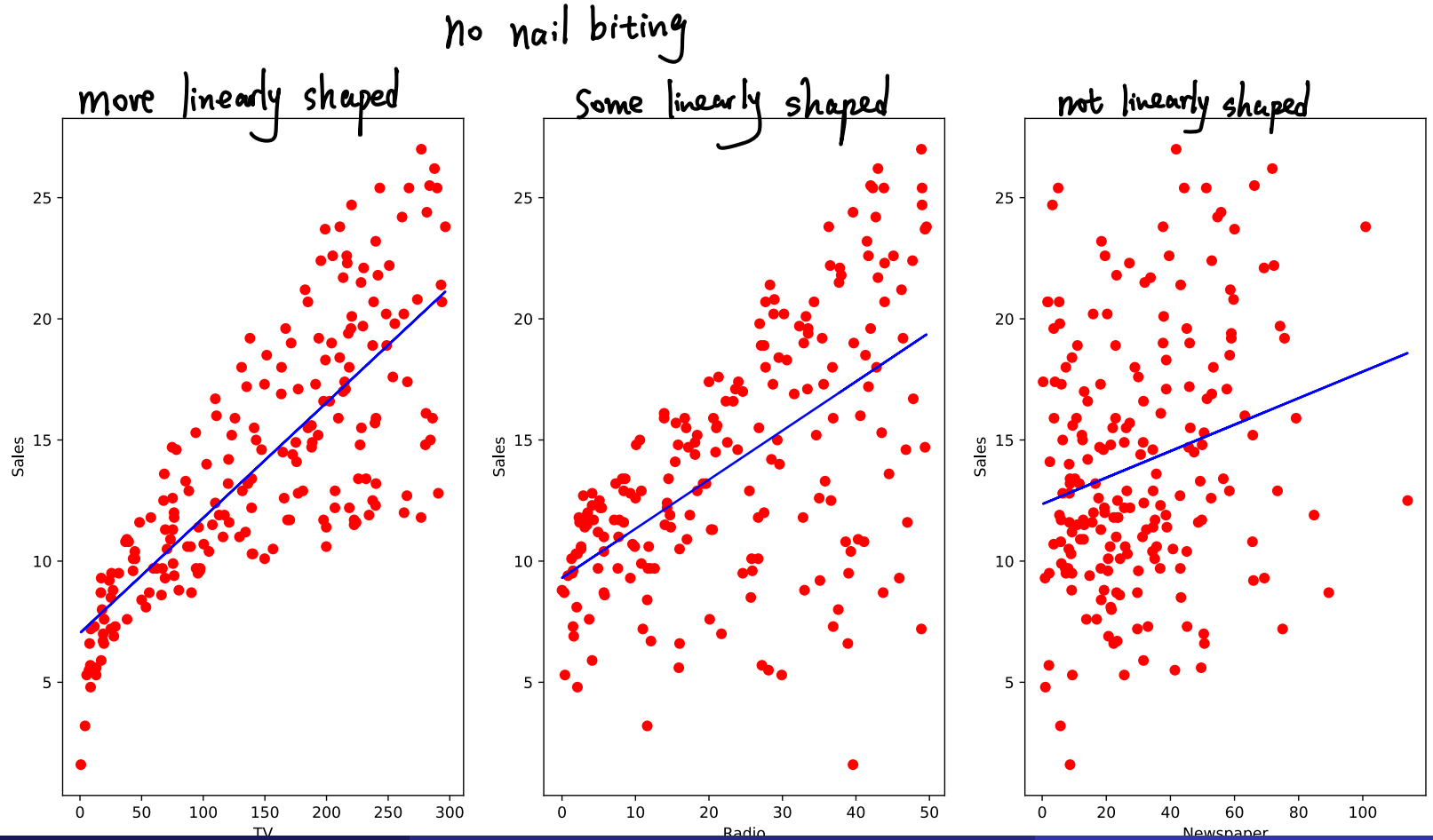
$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_0 \epsilon,$$

- Y: Sales
- X_1: TV advertisement budget
- X_2: Radio advertisement budget
- X_3: Newspaper advertisement budget
- $\beta_0, \beta_1, \beta_2, \beta_3 \in \mathbb{R}$ are the ~~three~~ Four parameters to be estimated,
- the estimation of $(\beta_0, \beta_1, \beta_2, \beta_3)$ based on the data is denoted by $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$.

Statistical Inferences: Sale v.s. Advertisements

The data are from 200 markets.

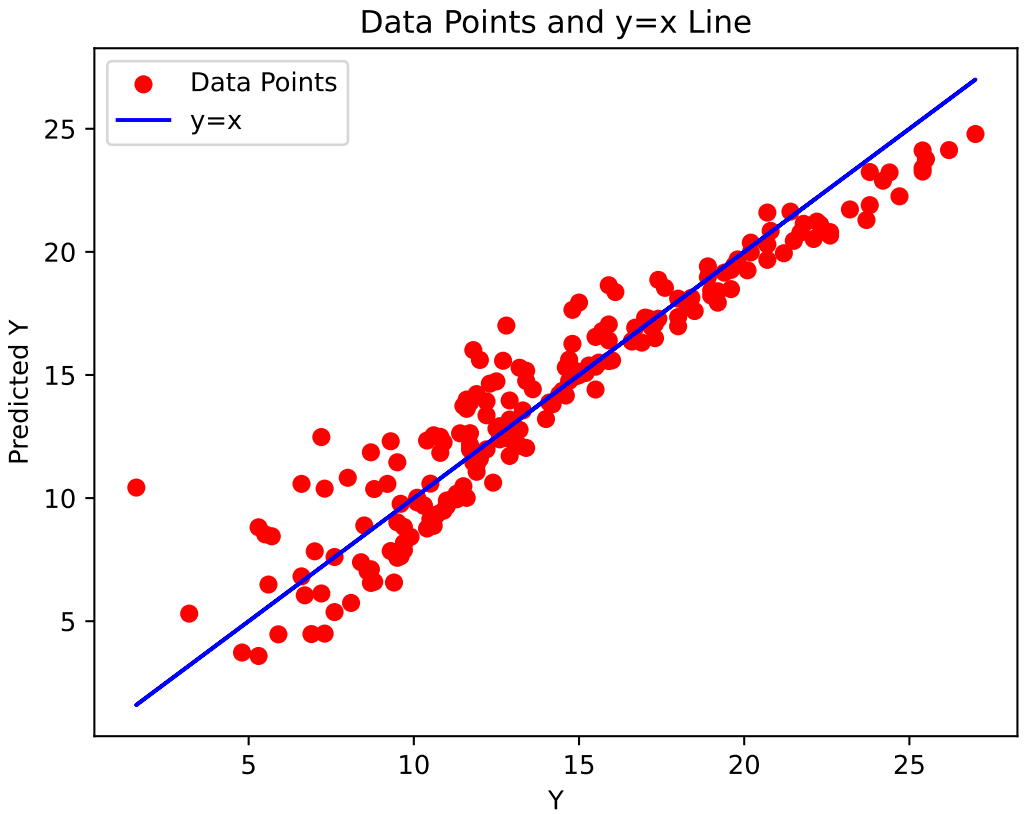
Assignment 1: Use Python to plot the following pictures: (The data is from 'Advertising.csv' and recall the python codes in Lecture 6 Linear Regression)



Statistical Inferences: Sale v.s. Advertisements

Assignment 1: Use Python to obtain the coefficients $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ and plot the following pictures:

$$Sales = 0.046 * TV + 0.189 * Radio - 0.001 * Newspaper + 2.939$$



linear regression behaves relatively well

Assessing Model Accuracy: Prediction Error

- If the population distribution is known, then we naturally use

$$\mathbb{E}[(Y - \hat{Y})^2]$$

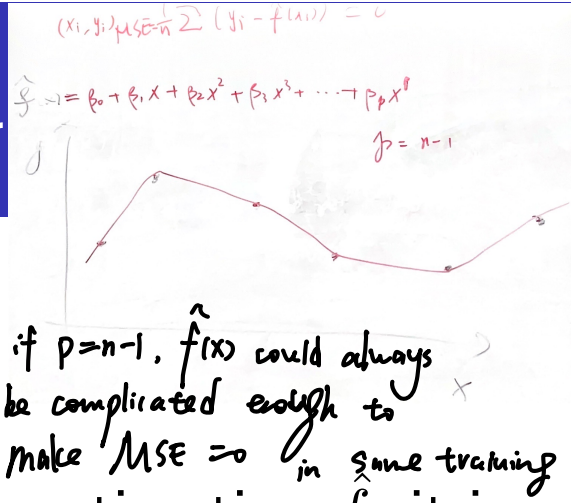
to give the prediction error, where $\hat{Y} = \hat{f}(X)$ and $\hat{f}$ is the estimation of f based on the observed data.

- However, the population distribution is usually unknown, we replace $\mathbb{E}[(Y - \hat{Y})^2]$ by its estimator based on the data, a popular candidate of such an estimator is the mean squared error (MSE), given by

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(\mathbf{x}_i\right)\right)^2.$$

The MSE will be small if the predicted responses are very close to the true responses.

Assessing Model Accuracy: Prediction Error



- If we use the data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ to obtain an estimation $\hat{f}$, it is not reasonable to use the same data to define MSE:

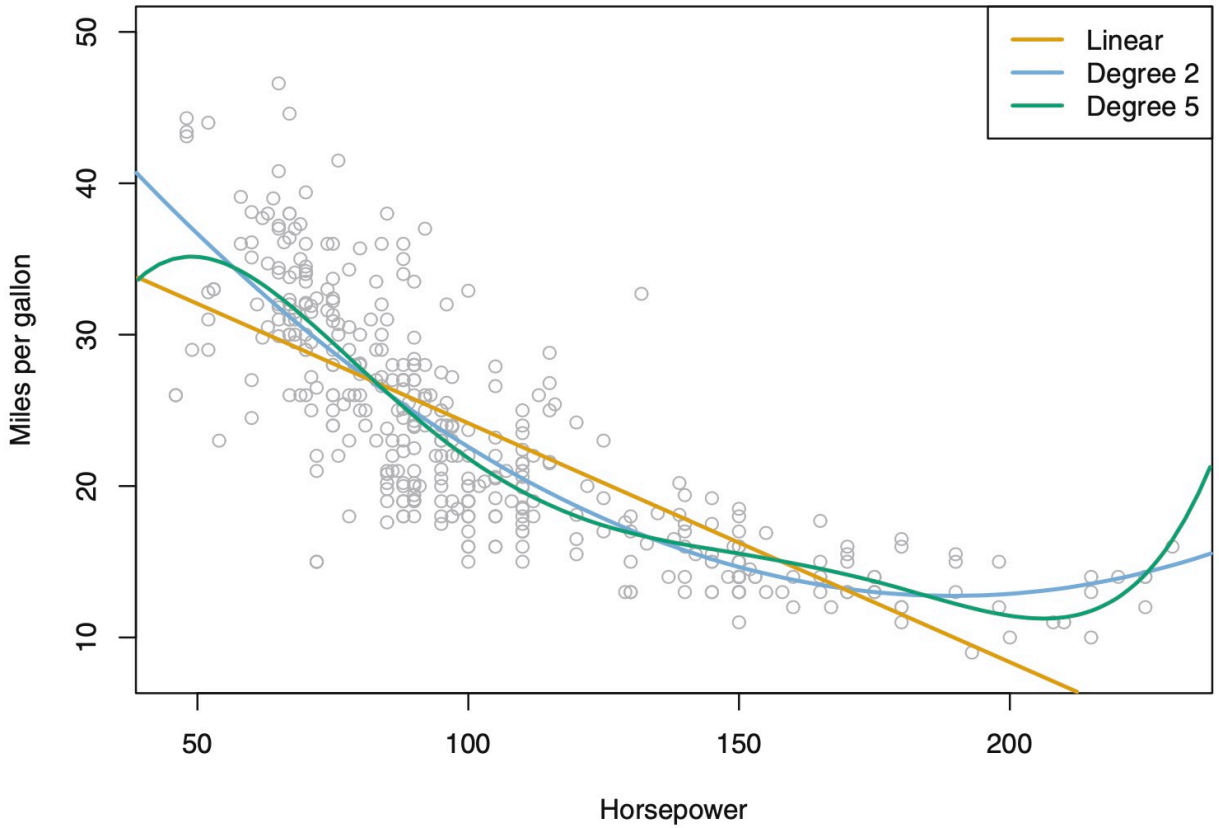
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(\mathbf{x}_i\right)\right)^2.$$

- In practice, we split the data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ into two groups.
  - training data: these data is used to get the estimation $\hat{f}$
  - testing data: these data is used to assess the model accuracy by MSE.

Contents

An example about model selection: Auto Data

The Auto data set with 392 observations which shows the mpg (gas mileage in miles per gallon) versus horsepower in cars:



An example about model selection

- It is obvious that linear model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where Y is the mpg and X is the horsepower, is NOT fit for this data.

- Alternatively, we would like to choose polynomial regression:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_k X^k + \epsilon. \quad (*)$$

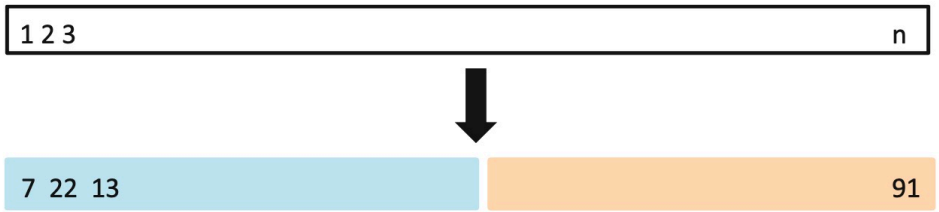
Which k shall we choose?

- Let us learn how to use the cross validation method to select the best k.

A validation method

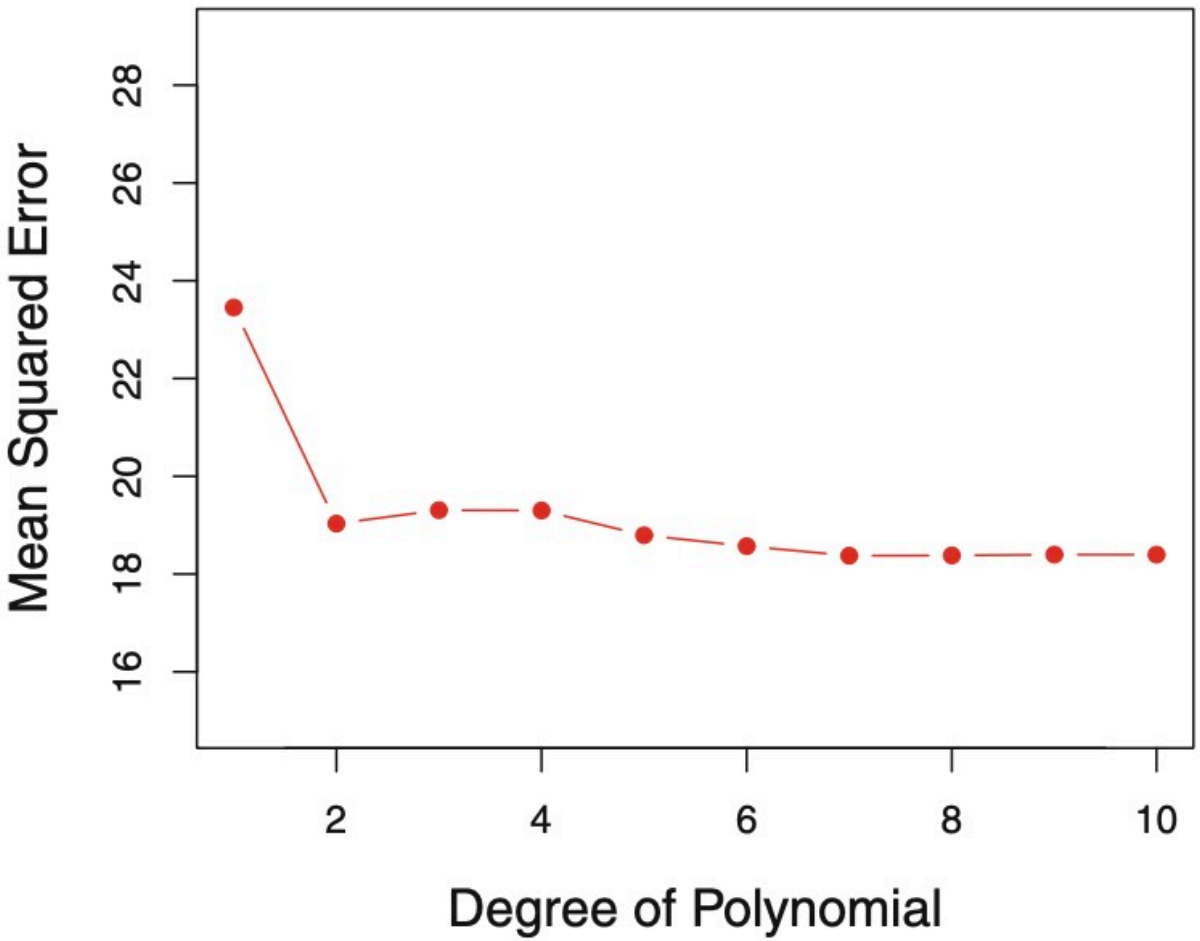
In order to determine the best k, we split the 392 observations into two sets: training set, validation set.

- training set with 196 observations, validation set with 196 observations



- For a given k, we use the training data to estimate the model (*) and calculate the MSE by the validation data.

A validation method



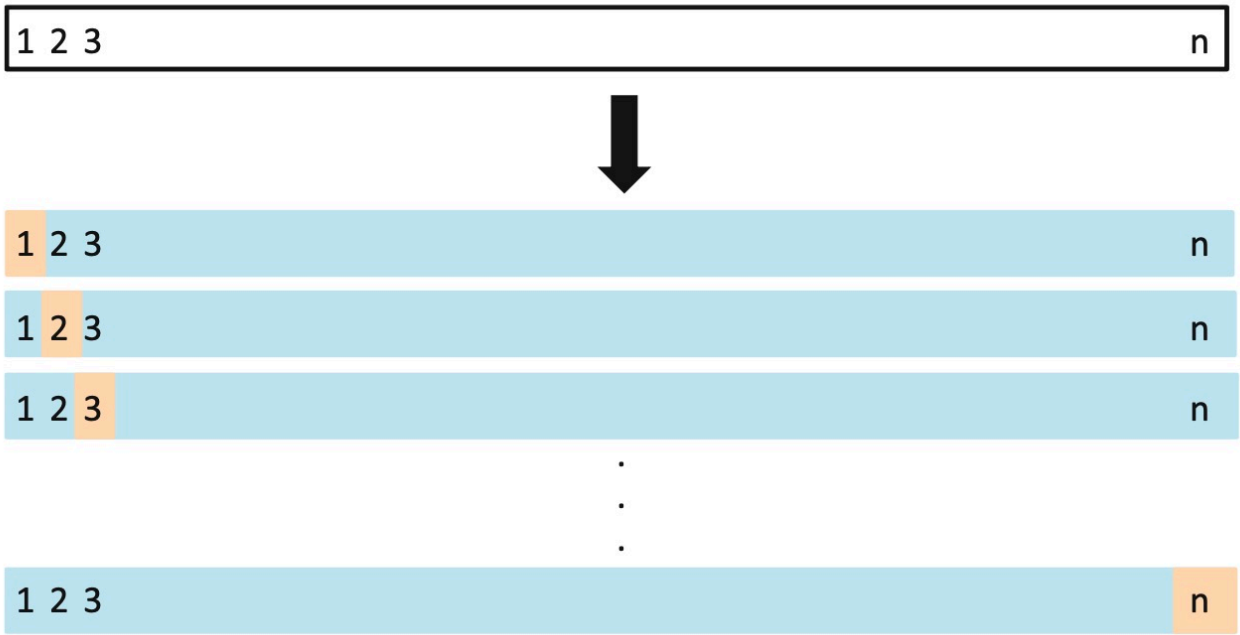
We can see that $k = 2$ is the best choice.

Leave one out cross validation (LOOCV)

Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the observed data.

- Let (x_1, y_1) be the validation set, let the remaining $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ be the training set:
  - train the model by $\{(x_2, y_2), \ldots, (x_n, y_n)\}$,
  - get a prediction $\hat{y}_1$ for y_1,
  - $\mathrm{MSE}_1 = (y_1 - \hat{y}_1)^2$.
- Repeat the procedure by selecting (x_2, y_2) for the validation data, training the model by $\{(x_1, y_1), (x_3, y_3), \ldots, (x_n, y_n)\}$, and computing $\mathrm{MSE}_2 = (y_2 - \hat{y}_2)^2$.
- Repeating this approach n times produces: $\mathrm{MSE}_1, \ldots, \mathrm{MSE}_n$. Define the average test error estimates as $\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i$

Leave one out cross validation (LOOCV)



The example of Auto Data

Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the observed data:

- x represents horsepower, y represents mpg (mileage per galon), $n = 392$.
- For each i,
  - let (x_i, y_i) be the validation set
  - train the model by $\{(x_1, y_1), \ldots, (x_n, y_n)\} \setminus \{(x_i, y_i)\}$,
  - get a prediction $\hat{y}_i$ for y_i,
  - $\mathrm{MSE}_i = (y_i - \hat{y}_i)^2$.
- $\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i$.

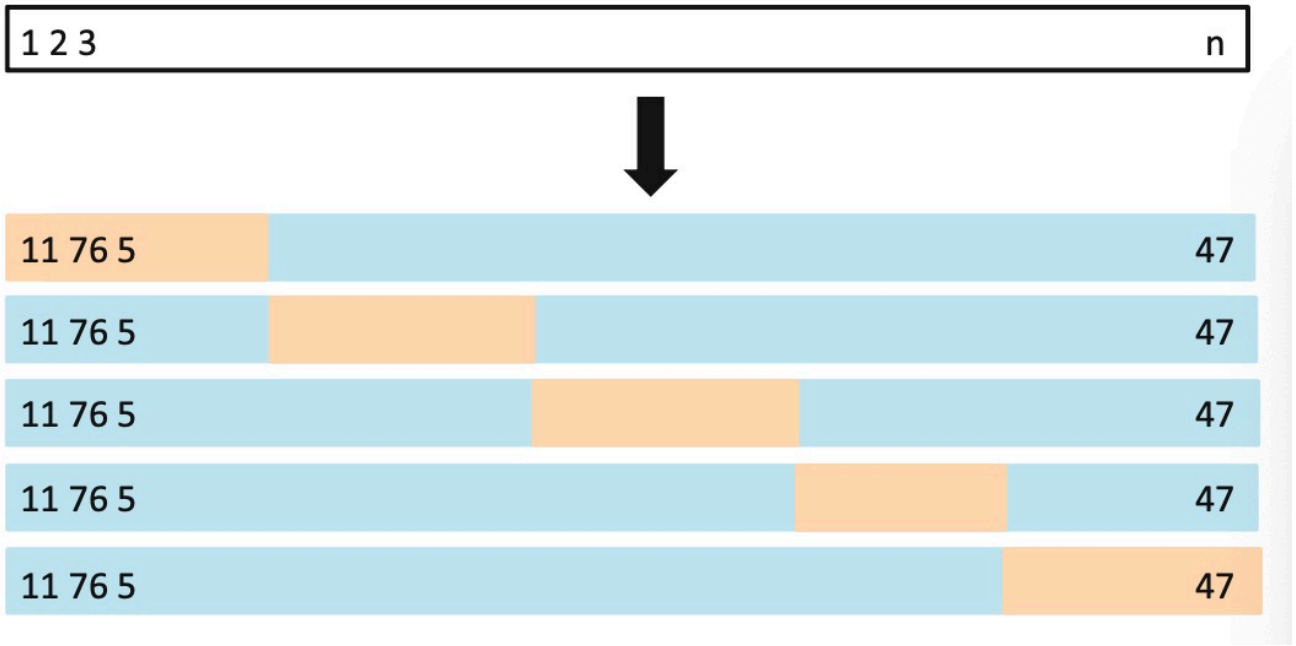
k-fold cross validation (CV)

An alternative to LOOCV is k-fold CV. This approach involves

- randomly dividing the set of observations into k groups, or folds, of approximately equal size.
- calculate the error MSE_1:
  - the first fold is treated as a validation set,
  - the model is trained by the remaining $k-1$ folds,
  - MSE_1, is computed by the first fold,
- this procedure is repeated k times, and results in k test errors: MSE_1, $\mathrm{MSE}_2, \ldots, \mathrm{MSE}_k$.
- the k-fold CV estimate is computed by averaging these values,

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

A demonstration of 5-fold CV



Contents

Bootstrap

- The bootstrap is a widely applicable and extremely powerful statistical tool to quantify the uncertainty associated with an estimator or a statistical learning method.
- The power of the bootstrap lies in the fact that it can be easily applied to a wide range of statistical learning methods, including some problems in which a measure of variability is difficult to obtain.
- We illustrate the bootstrap on a toy example.

The toy model

We wish to invest 1 dollar in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities:

- invest a fraction α of our money in X,
- invest the remaining $1-\alpha$ in Y,
- choose α to minimize the total risk, or variance, of our investment:

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

One can show that the value that minimizes the risk is given by

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

where $\sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y)$, and $\sigma_{XY} = \mathrm{Cov}(X, Y)$.

$\alpha^2 \underbrace{\mathrm{Var}(X)}_{=\sigma_X^2} + 2\alpha(1-\alpha)\underbrace{\mathrm{Cov}(X,Y)}_{=\sigma_{XY}} + (1-\alpha)^2 \underbrace{\mathrm{Var}(Y)}_{=\sigma_Y^2}$

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

$$0 = f'(\alpha) = 2\cdot\alpha\cdot\sigma_X^2 + 2(1-\alpha-\alpha)\sigma_{XY} + 2(1-\alpha)\cdot(-1)\sigma_Y^2$$

$$= 2\sigma_X^2\cdot\alpha + 2\sigma_{XY} - 4\sigma_{XY}\alpha + 2\alpha\sigma_Y^2 - 2\sigma_Y^2$$

$$\Rightarrow 2(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\cdot\alpha = 2(\sigma_Y^2 - \sigma_{XY})$$

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

Observed $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

$$\hat{\sigma}_X^2 := \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu}_X)^2, \quad \hat{\mu}_X := \frac{1}{n}\sum_{i=1}^n x_i$$

$\mathrm{Var}(X) = \mathbb{E}((X - \mu_X)^2)$, $\mu_X = \dots$

$$\hat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu}_X)\cdot(y_i - \hat{\mu}_Y), \quad \hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^n y_i$$

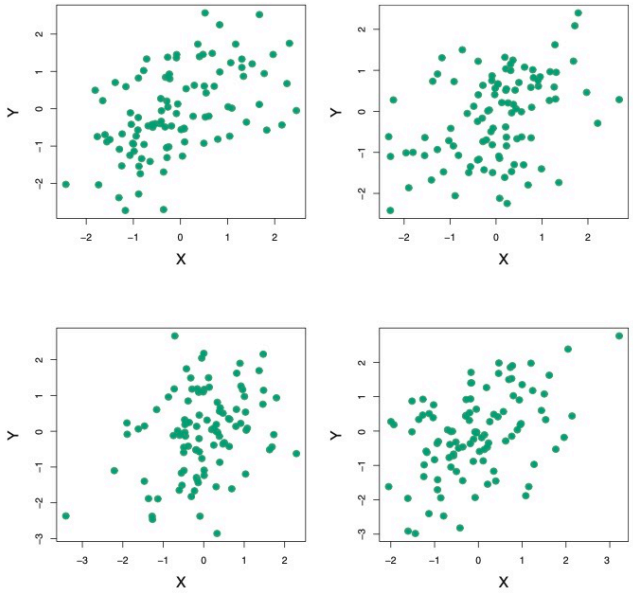
$$\hat{\sigma}_Y^2 := \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\mu}_Y)^2$$

The toy model

- In reality, the quantities σ_X^2, σ_Y^2, and σ_{XY} are unknown.
- It is natural for us to compute estimates for these quantities, $\hat{\sigma}_X^2, \hat{\sigma}_Y^2$, and $\hat{\sigma}_{XY}$, using a data set that contains past measurements for X and Y.
- We can then estimate the value of α that minimizes the variance of our investment using

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

The toy model: simulation data



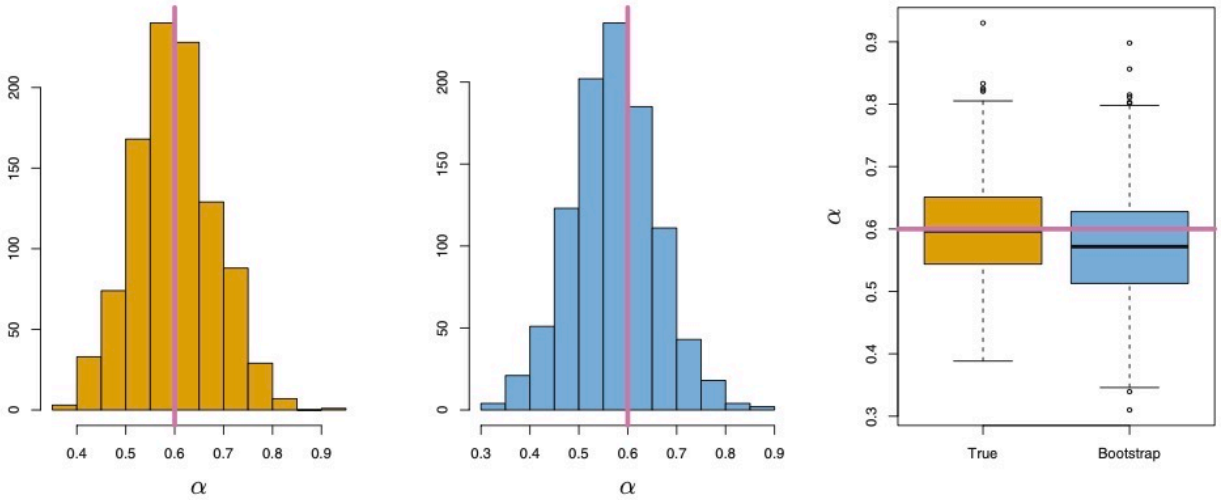
- The figure illustrates this approach for estimating α on a simulated data set.
- In each panel, we simulated 100 pairs of returns for the investments X and Y. We used these returns to estimate σ_X^2, σ_Y^2, and σ_{XY}, which we then substituted into (*) in order to obtain $\hat{\alpha}$.
- $\hat{\alpha} = 0.576, 0.532, 0.657, 0.651.$

The toy model: simulation data

- One usually uses the variance to evaluate the estimation $\hat{\alpha}$, i.e. $\mathbb{E}[(\hat{\alpha} - \alpha)^2]$. But this is not practical due to:
  - α is not known,
  - the distribution of $\hat{\alpha}$ is not known.
- We usually use sample variance to replace $\mathbb{E}[(\hat{\alpha} - \alpha)^2]$, i.e.
  - We repeated the above estimation of α 1000 times, obtaining $\hat{\alpha}_1, \ldots, \hat{\alpha}_{1000}$.
  - The mean over all 1,000 estimates for α is $\bar{\alpha} = \frac{1}{1{,}000} \sum_{r=1}^{1{,}000} \hat{\alpha}_r = 0.5996$,
  - The standard deviation of the estimates is $\sqrt{\frac{1}{1{,}000-1} \sum_{r=1}^{1{,}000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$.
  - We would expect $\hat{\alpha}$ to differ from α by approximately 0.08, on average.

The toy model: simulation data

The histogram and boxplot for the simulation data:



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.

The toy model: bootstrap

- In the real world practice, it is often not have sufficient observed data for estimating σ_X^2, σ_Y^2, and σ_{XY}.
- For given finite number of observed data, we often apply bootstrap method to reuse the data.
- Bootstrap: Based on the given observed data, we do:
  - NOT repeatedly obtain independent data sets from the population,
  - use a computer to emulate the process of obtaining new sample from the given observed data,
  - obtain distinct data sets and do statistical learning.

The toy model: bootstrap

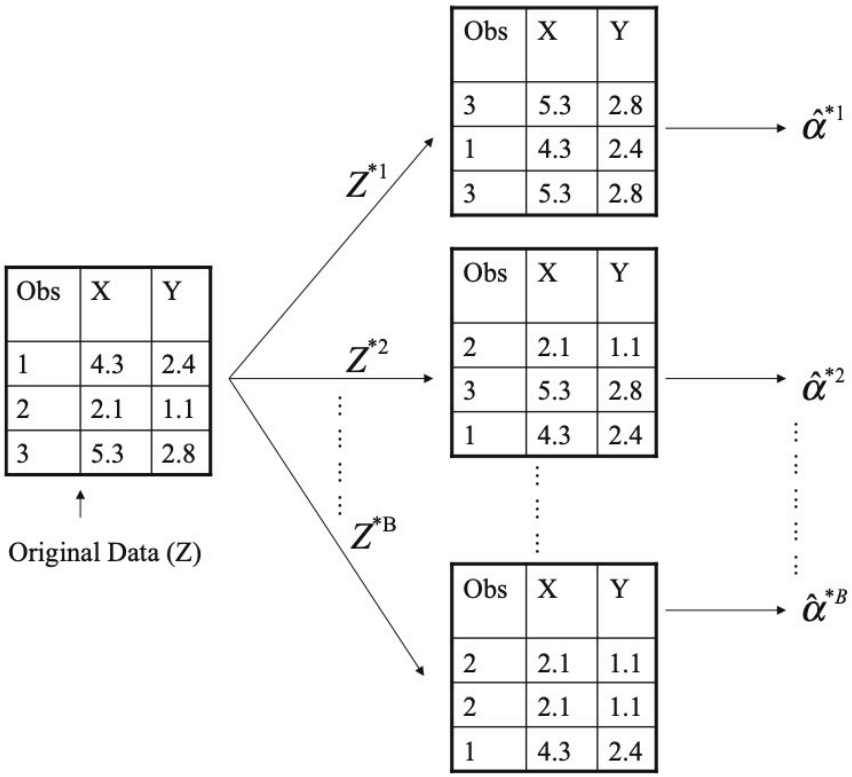
In a real world practice, we only observed three data:

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data (Z)

The toy model: bootstrap

In a real world practice, we only observed three data, we take these three data as the population and practise drawing samples from population:



The toy model: bootstrap

- Randomly select the data in Z for B times with some large value of B, and produce B different bootstrap data sets, $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$.
- Conduct B corresponding α estimates, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$.
- Compute the standard error of these bootstrap estimates using the formula[1]

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}.$$

- $\mathrm{SE}_B(\hat{\alpha})$ serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set.

[1] The unbiased estimate of the standard deviation is obtained by using the sample standard deviation and applying the correction factor $(B-1)$, which is derived based on the unbiased property of the sample variance. This correction factor is derived from adjusting the degrees of freedom to eliminate the bias caused by the known information of the sample mean.

Advantages of Bootstrap Resampling

- Non-parametric: The bootstrap method does not rely on assumptions about the underlying distribution of the data, making it applicable to a wide range of data types and situations.
- Flexibility: It can be applied to various statistical estimators and models, allowing for the estimation of parameters, confidence intervals, and hypothesis testing.
- Reliability: The bootstrap method provides robust estimates and measures of uncertainty, even with small sample sizes or when the underlying assumptions are violated.
- Versatility: It can handle complex study designs, such as stratified sampling or clustered data, by resampling within these structures.

Disadvantages of Bootstrap Resampling

- Sampling Variability: The bootstrap estimates may still be subject to sampling variability, especially when the sample size is small or the original sample is biased or skewed.
- Reliance on the Original Sample: The accuracy of bootstrap estimates heavily depends on the representativeness and quality of the original sample. If the original sample is not representative of the population, the bootstrap estimates may be biased or unreliable.

Overall, the bootstrap resampling method is a powerful tool for statistical inference and estimation. Its advantages of flexibility, reliability, and non-parametric nature make it a valuable technique for a wide range of applications, but it is essential to consider its limitations and potential sources of bias or variability.