MONASH University

ETC3250

Business Analytics

Week 6

Resampling methods

11 April 2018

Outline

Week	Topic	Chapter	Lecturer
1	Introduction	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression	3	Souhaib
4	Classification	4	Souhaib
5	Clustering	10	Souhaib
	Semester break		
6	Model selection and resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	Advanced regression	6	Souhaib
9	Advanced regression	6	Souhaib
10	Advanced classification	9	Souhaib
11	Tree-based methods	8	Souhaib
12	Project presentation		Souhaib

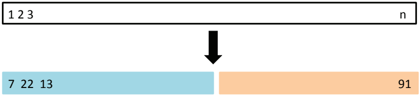
Resampling methods

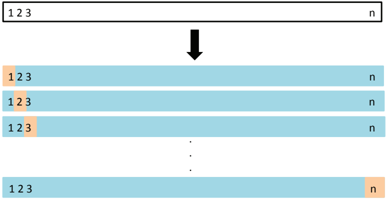
Resampling methods are used in

1. **validating models** by using (random) subsets of the data (e.g cross validation and bootstrapping),
2. **estimating uncertainty** in sample statistics by drawing randomly with replacement from the data set (e.g. bootstrapping),
3. performing **(non-parametric) significance tests** (permutation tests).

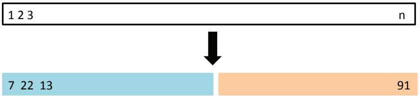
Outline

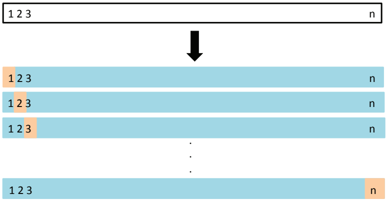
Validation set and Leave-one-out

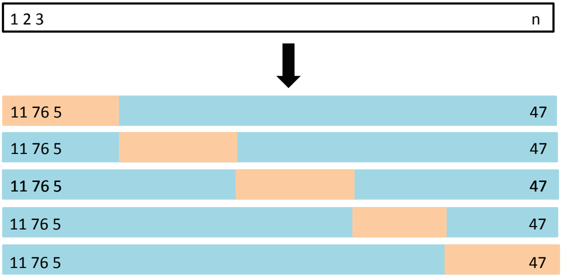




Cross-validation







k-fold Cross-validation

- Divide the data set into k different parts.
- Remove one part, fit the model on the remaining $k-1$ parts, and compute the MSE on the omitted part.
- Repeat k times taking out a different part each time

By averaging the k MSEs we get an estimated validation (test) error rate for new observations.

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

LOOCV is a special case where $k = n$.

k-fold Cross-validation

- Each training set is only $(k-1)/k$ as big as the original data set. So the estimates of prediction error will be biased upwards.
- Bias minimized when $k = n$ (LOOCV).
- But variance increases with k (as there are overlapping observations in each part).
- $k = 5$ or $k = 10$ provide a good compromise for this bias-variance tradeoff.

The wrong way to do cross validation

Consider a simple regression procedure applied to a dataset with 500 predictors and 50 samples:

1. Find the 5 predictors having the largest correlation with the response
2. Apply linear regression using only these 5 predictors

How to use cross-validation to estimate the test error of this procedure?

1. Find the 5 predictors having the largest correlation with the response
2. Estimate the test error of linear regression with these 5 predictors via 10-fold cross validation.

$\rightarrow$ Wrong!

The wrong way to do cross validation

Consider a simple regression procedure applied to a dataset with 500 predictors and 50 samples:

1. Find the 5 predictors having the largest correlation with the response
2. Apply linear regression using only these 5 predictors

How to use cross-validation to estimate the test error of this procedure?

1. Find the 5 predictors having the largest correlation with the response
2. Estimate the test error of linear regression with these 5 predictors via 10-fold cross validation.

$\rightarrow$ Wrong!

The wrong way to do cross validation

Consider a simple regression procedure applied to a dataset with 500 predictors and 50 samples:

1. Find the 5 predictors having the largest correlation with the response
2. Apply linear regression using only these 5 predictors

How to use cross-validation to estimate the test error of this procedure?

1. Find the 5 predictors having the largest correlation with the response
2. Estimate the test error of linear regression with these 5 predictors via 10-fold cross validation.

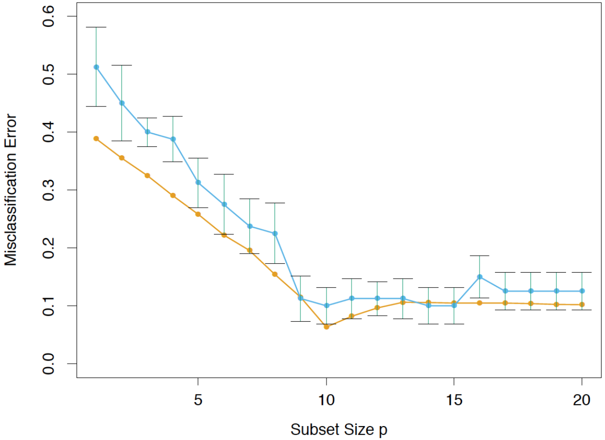
$\rightarrow$ Wrong!

The right way to do cross-validation

1. Divide the data into 10 folds
2. For $i = 1, \ldots, 10$
    1. Using every fold except i, find the 5 predictors having the largest correlation with the response, and run linear regression with these 5 predictors
    2. Compute the error on fold i
3. Average the 10 test errors obtained

Every aspect of the procedure that involves using the data — variable selection, scaling, etc — must be cross-validated

The one standard error rule



Choose the simplest model whose CV error is no more than one standard error above the model with the lowest CV error

Outline

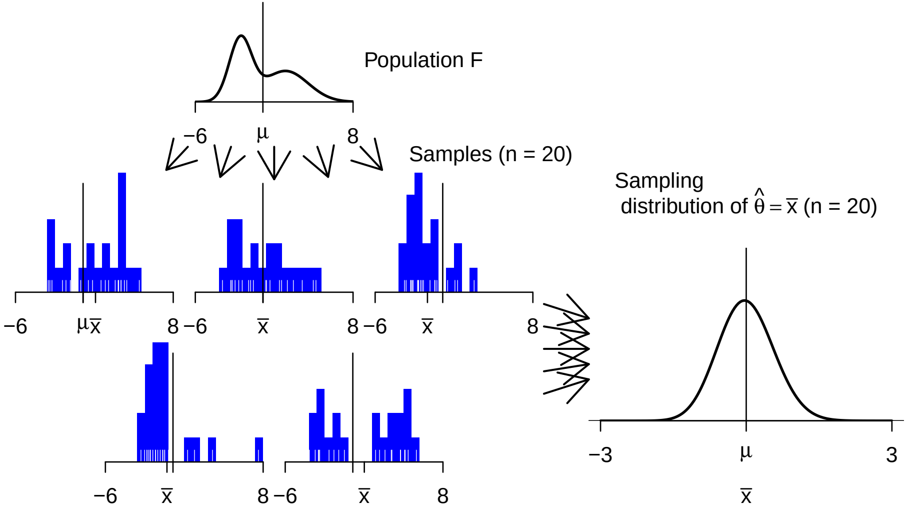
Pull yourself up by your bootstraps

What is the bootstrap?

The bootstrap is a flexible statistical tool to **quantify the uncertainty** associated with a *given estimator* or *statistical learning method*.

- The bootstrap allows us to use a computer to **mimic the process of obtaining new data sets**, so that we can estimate the **variability of our estimate** without generating additional samples
- We obtain distinct data sets (with the same size as our original dataset) by repeatedly sampling observations **from the original data set with replacement** (nonparametric) or **from an estimated model** (parametric).

Bootstrapping: Ideal world



Bootstrapping: Bootstrap world

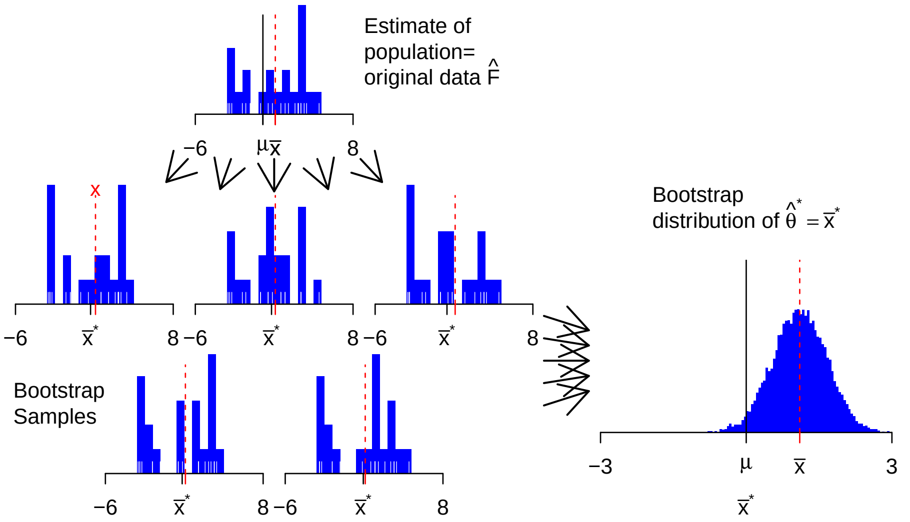
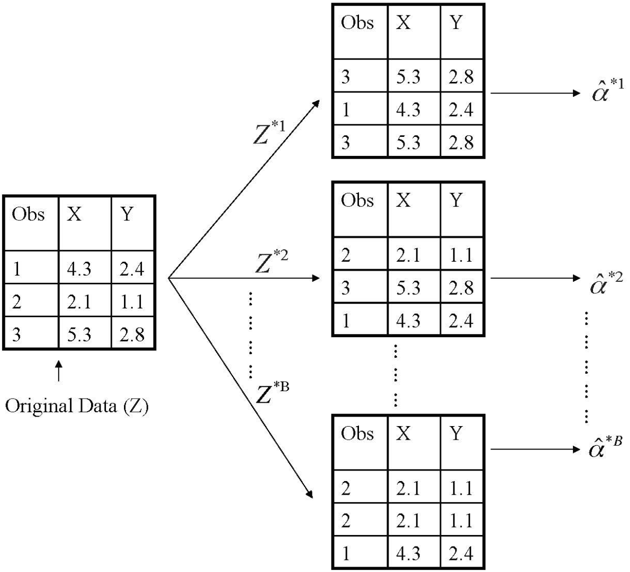


Illustration of the bootstrap



The bootstrap procedure

- Find a good estimate $\hat{P}$ of P
  - Parametric bootstrap
  - Nonparametric bootstrap
- Draw B independent bootstrap samples $X^{*(1)}, \ldots, X^{*(B)}$ from $\hat{P}$:

$$X_1^{*(b)}, \ldots, X_n^{*(b)} \sim \hat{P} \quad b = 1, \ldots, B.$$

- Evaluate the bootstrap replications:

$$\hat{\theta}^{*(b)} = s(X^{*(b)}) \quad b = 1, \ldots, B.$$

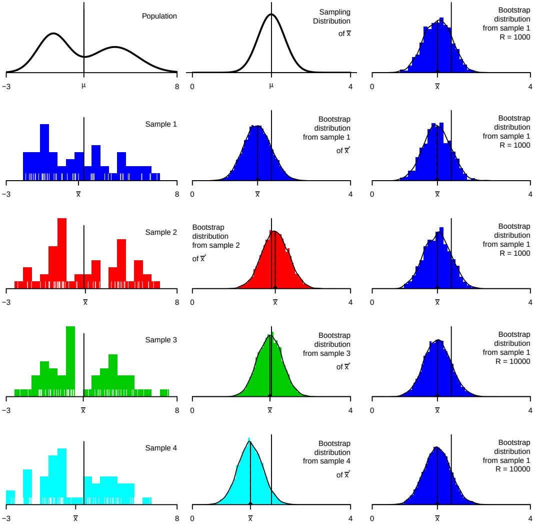
- Estimate the quantity of interest from the distribution of the $\hat{\theta}^{*(b)}$

Examples

What is the standard error of $\hat{\theta}$ (i.e., the standard deviation of the sampling distribution of $\hat{\theta}$)?

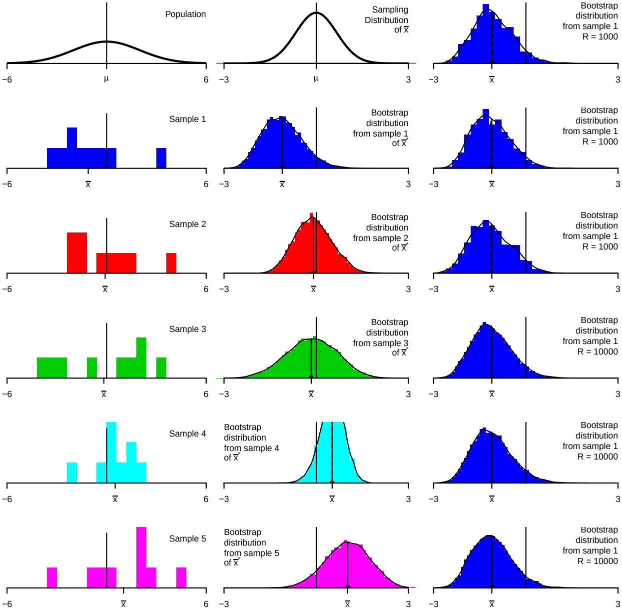
1. $\hat{\theta} =$ sample mean
2. $\hat{\theta} =$ sample median
3. $\hat{\theta} =$ expected shortfall at 5%
4. $\hat{\theta} =$ lag 1 autocorrelation.

Sample mean: $n = 50$



- Two types of random variation

Sample mean: $n = 9$



Prediction error estimation

- Fit the model on a set of bootstrap samples, and then keep track of how well it predicts the original dataset

$$\mathrm{Err}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

- Each of these bootstrap data sets is created by sampling with replacement, and is the same size as our original dataset. As a result **some observations may appear more than once in a given bootstrap data set and some not at all**.

Prediction error estimation

- Fit the model on a set of bootstrap samples, and then keep track of how well it predicts the original dataset

$$\mathrm{Err}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

- Each of these bootstrap data sets is created by sampling with replacement, and is the same size as our original dataset. As a result **some observations may appear more than once in a given bootstrap data set and some not at all**.

Prediction error estimation

- Training and validation sets have observations in common! Overfit predictions will look very good.

$$P(\text{observation } i \in \text{ bootstrap sample } b) = ??$$

Prediction error estimation

- Training and validation sets have observations in common! Overfit predictions will look very good.

$$
\begin{aligned}
&\mathrm{P}(\text{observation } i \in \text{ bootstrap sample } b) = ??\\
&= 1 - \left(1 - \frac{1}{n}\right)^n\\
&\approx 1 - \frac{1}{e}\\
&= 0.632
\end{aligned}
$$

- Remember that cross-validation uses **non-overlapping** data for the training and validation samples

Prediction error estimation

Better bootstrap version: we only keep track of predictions from bootstrap samples not containing that observation. The leave-one-out bootstrap estimate of prediction error can be defined as

$$\mathrm{Err}_{\text{loo-boot}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

where C^{-i} is the set of indices of the bootstrap samples b that do not contain observation i. Problem of overfitting with $\mathrm{Err}_{\text{boot}}$ solved but **training-set-size bias as with cross-validation**.

Many applications

- Computing standard errors for complex statistics
- Prediction error estimation
- Bagging (Bootstrap aggregating)
- ...

Variations

There are several types of bootstrap based on different assumptions:

- block bootstrap
- sieve bootstrap
- smooth bootstrap
- residual bootstrap
- wild bootstrap