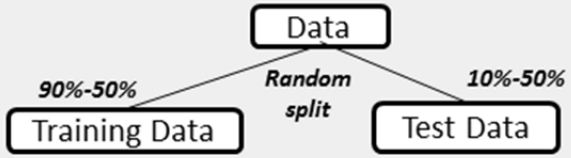
Appendix 1. Strategies for Assessing Model Performance and Avoiding Resubstitution Bias.

Data-Splitting



Many models explored. Favored model(s) selected.

Evaluate performance of favored model(s).

***Key: Only "one look" at the test data**

Advantages

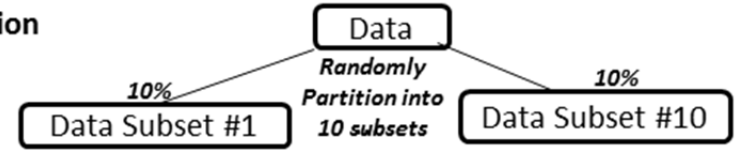
Model performance on test data is unbiased estimate of how the model will perform on new data from the same population.

Disadvantages

When the size of the test dataset is small, estimates of model performance are highly variable.

Statistically inefficient because the test data are only used for validation and not model selection or fitting[7]

Cross Validation (ex. 10-fold)



Procedure for 1st iteration

- Data Subsets 2 to 10 are combined and model selection and fitting is performed
- Fitted model from Subsets 2-10 predict risk for individuals in Subset 1
- Performance measure is calculated on predicted risks for Subset 1

Procedure is repeated 10 times with each subset left out in turn.

In the end, 10 performance measures for 10 subsets are appropriately averaged.

***Key: Cross Validation can only be performed once**

Advantages

Within each iteration, the model performance is unbiased.

Variation of data-splitting that avoids "wasting data"

Disadvantages

Model selection procedure must be prescribed and automated. No flexibility to explore data, exercise judgment and refine procedures.

Each iteration typically produces a different model - unclear what model to report.

Computationally Intensive.

Bootstrapping

Advantages

Avoids "wasting data"- full dataset is used to fit model

Disadvantages

Model selection procedure must be prescribed and automated. No flexibility to explore data, exercise judgment and refine procedures.

Computationally intensive

Step 1. Apply model selection procedure to data and compute performance measure S

Step 2. For each bootstrap dataset, apply model selection procedure from step 1 and compute performance measure. Take the final, fitted model from the bootstrap dataset and apply it to the original dataset and compute a performance statistic S. The difference in performance statistics for the fitted model from the bootstrap dataset and the fitted model from the original dataset is the optimism. Repeat the procedure many times, say n=1000, so that there are 1000 estimates of the optimism.

Step 3. Subtract the average optimism from the apparent performance statistic S to get an optimism-corrected S value.