Resampling method[1]

[1] Based on materials in ISLR Ch 5

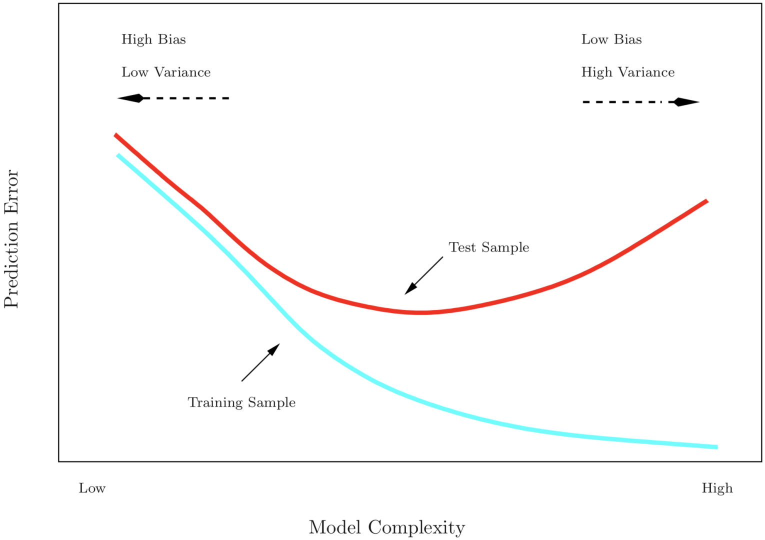
Cross-validation and the Bootstrap

- In the section we discuss two **resampling** methods: cross-validation and the bootstrap.
- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- Cross-validation provides estimates of testing prediction error
- Bootstrap estimates standard error of parameter estimates.

Training Error versus Test error

- The **test error** is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the **training error** can be easily calculated by applying the statistical learning method to the observations used in its training.
- But the training error rate often is quite different from the test error rate, and in particular the former can **dramatically underestimate** the latter.
- This problem is due to **over-fitting**.

Training- versus Test-Set Performance



More on prediction-error estimates:

- Best solution: a large designated test set. Often not available
- We consider a class of methods that estimate the test error by **holding out** a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

SBP Thursday 2/16/22 (week 4, lecture 8)

Validation-set approach

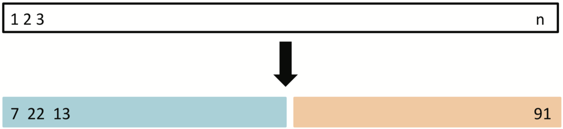
START Tuesday (Week 5, lecture 8)
2/15/22

- Here we randomly divide the available set of samples into two parts: a **training set** and a **validation** or **hold-out** set.
- The model is fit on the **training set**, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. This is typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

- can calculate test error (best model)
- select tuning parameters

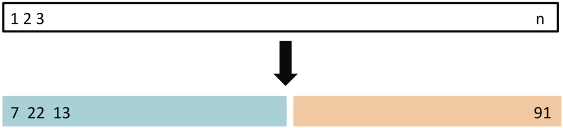
} can't do both simultaneously.

The Validation process



A random splitting into two halves: left part is training set, right part is validation set.

The Validation process



A random splitting into two halves: left part is training set, right part is validation set.

Goals:

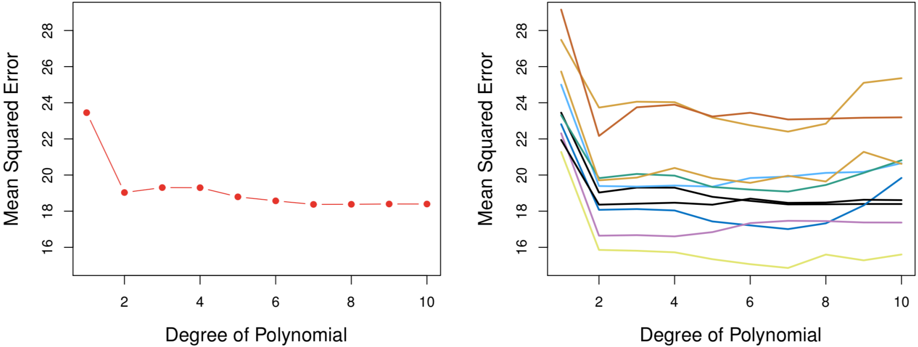
(1) Pick a good tuning parameter

(2) Estimate the test error

> Cant do both simultaneously.

Example: automobile data

- Want to compare linear vs higher-order polynomial terms in a linear regression.
- We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.



Left panel shows single split; right panel shows multiple splits

Drawbacks of validation set approach

- The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.
- This suggests that the validation set error may tend to **overestimate** the test error for the model fit on the entire data set.

Drawbacks of validation set approach

- The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.
- This suggests that the validation set error may tend to **overestimate** the test error for the model fit on the entire data set. **Why**?

b/c sample size is small

1/2 of the total dataset

K-fold Cross-validation

- **Widely used approach** for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1, 2, \cdots, K$, and then the results are combined.

K-fold Cross-validation in detail.

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train

K-fold Cross-validation in detail.

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train
Train	Validation	Train	Train	Train

K-fold Cross-validation in detail.

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train
Train	Validation	Train	Train	Train
Train	Train	Validation	Train	Train

K-fold Cross-validation in detail.

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train
Train	Validation	Train	Train	Train
Train	Train	Validation	Train	Train
Train	Train	Train	Validation	Train

K-fold Cross-validation in detail.

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train
Train	Validation	Train	Train	Train
Train	Train	Validation	Train	Train
Train	Train	Train	Validation	Train
Train	Train	Train	Train	Validation

The details

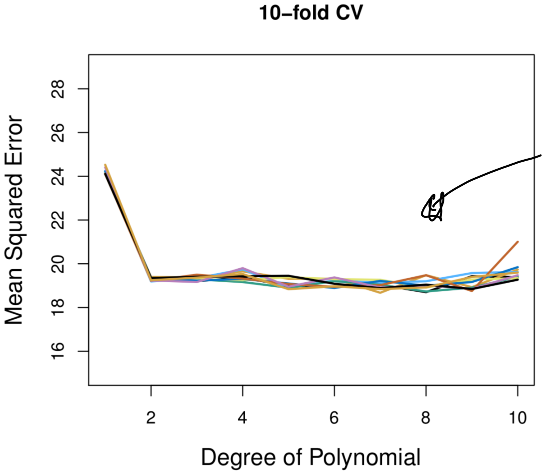
- Let the K parts be $C_1, C_2, \cdots, C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/K$.
- Compute

$$\text{CV}_{(K)} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in C_k}(y_i - \hat{y}_i)^2 = \sum_{k=1}^{K}\frac{n_k}{n}\text{MSE}_k,$$

where $\text{MSE}_k = \sum_{i \in C_k}(y_i - \hat{y}_i)^2/n_k$ and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.

- Typically, $K = 5$ or 10.
- Setting $K = n$ yields n-fold or leave-one out cross-validation (LOOCV).

Auto data revisited



Stabilize
our estimator

Cross-Validation for Classification Problems

- We divide the data into K roughly equal-sized parts $C_1, C_2, \cdots, C_K$. C_k denotes the indices of the observations in part k. There are n_k observations in part k: if n is a multiple of K, then $n_k = n/K$.
- Compute

$$\mathrm{CV}_K = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} I(y_i \neq \hat{y}_i) = \sum_{k=1}^{K} \frac{n_K}{n} \mathrm{Err}_k$$

where $\mathrm{Err}_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.

Cross-validation: right and wrong

- Consider a simple classifier applied to some two-class data:
    - Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
    - We then apply a classifier such as logistic regression, using only these 100 predictors.

How do we estimate the test set performance of this classifier?

Cross-validation: right and wrong

- Consider a simple classifier applied to some two-class data:
  - Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
  - We then apply a classifier such as logistic regression, using only these 100 predictors.

How do we estimate the test set performance of this classifier?
Can we apply cross-validation in step 2, forgetting about step 1?

↳ we need to apply cross validation to the whole data set.

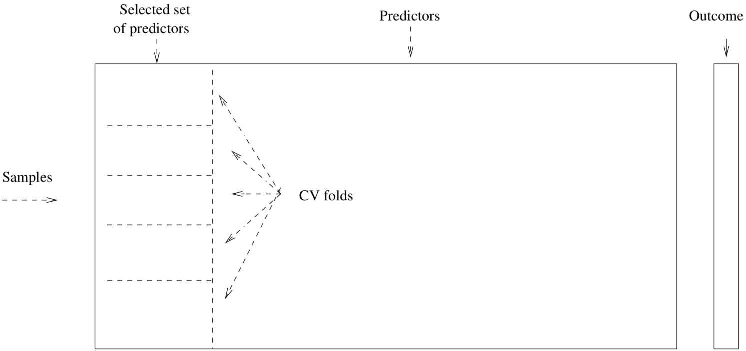
NO!

- This would ignore the fact that in Step 1, the procedure **has already seen the labels of the training data**, and made use of them. This is a form of training and must be included in the validation process.
- It is easy to simulate realistic data with the class labels independent of the outcome, so that true test error =50%, but the CV error estimate that ignores Step 1 is zero! **Try to do this yourself!**
- This error has made in many high profile genomics papers.

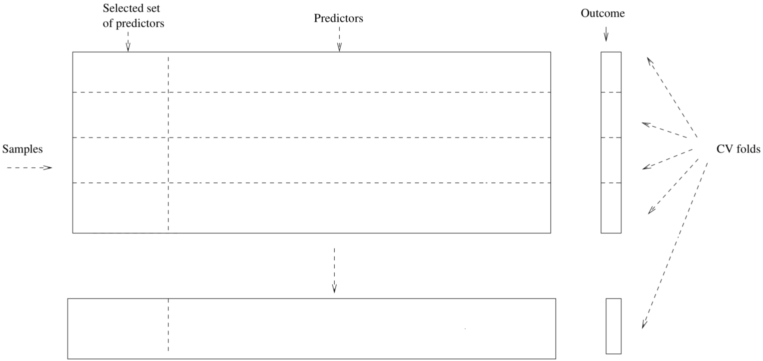
The Wrong and Right Way

- **Wrong**: Apply cross-validation in step 2.
- **Right**: Apply cross-validation to steps 1 and 2.

Wrong way


Selected set of predictors
Predictors
Outcome
Samples
CV folds

Right way


Selected set
of predictors
Predictors
Outcome
Samples
CV folds

The Bootstrap

— not used for test error estimation.

- The **bootstrap** is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

A simple example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities.
- We will invest a fraction α of our money in X, and will invest the remaining $1 - \alpha$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $\mathrm{Var}(\alpha X + (1 - \alpha)Y)$.
- One can show that the value that minimizes the risk is given by

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

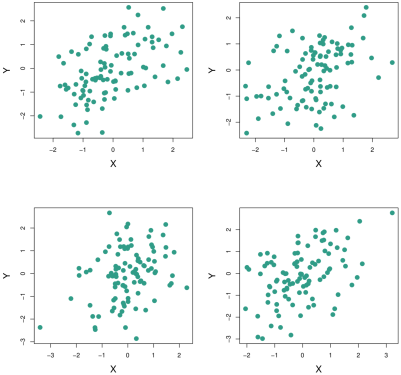
where $\sigma_X^2 = \mathrm{Var}(X)$, $\sigma_Y^2 = \mathrm{Var}(Y)$, and $\sigma_{XY} = \mathrm{Cov}(X, Y)$.

Example continued

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

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

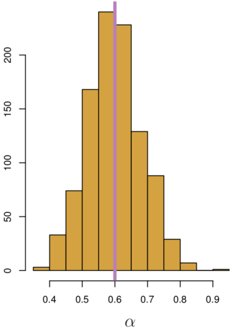
Example continued



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.

Example continued

- To estimate the standard deviation of $\hat{\alpha}$, we repeated the process of simulating 100 paired observations of X and Y , and estimating α 1,000 times.
- We thereby obtained 1,000 estimates for α, which we can call $\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_{1000}$.



- For these simulations the parameters were set to $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$, and $\sigma_{XY} = 0.5$, and so we know that the true value of α is 0.6 (indicated by the red line).

Example continued

- The mean over all 1,000 estimates for α is

$$\bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\alpha}_r = 0.5996,$$

  very close to $\alpha = 0.6$, and the standard deviation of the estimates is

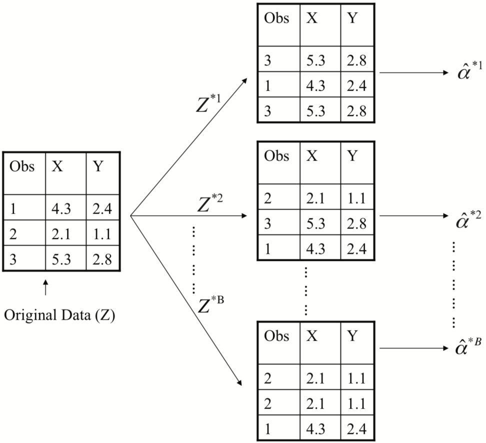
$$\sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$

- This gives us a very good idea of the accuracy of $\hat{\alpha}$: $\mathrm{SE}(\hat{\alpha}) = 0.083$.

Now back to the real world

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach 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.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set **with replacement**.
- 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.

Example with just 3 observations



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

Original Data (Z)

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

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

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

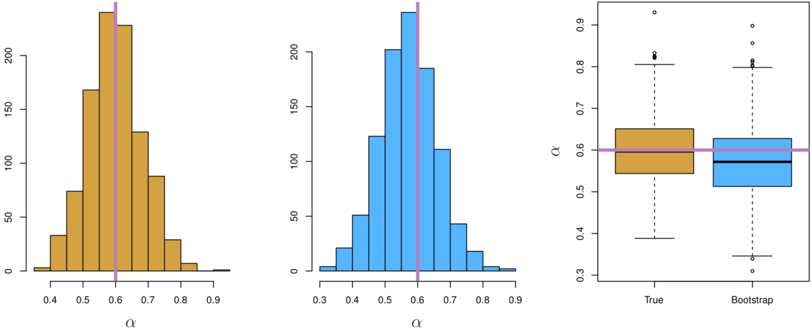
A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.

- Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α, which we call $\hat{\alpha}^{*1}$.
- This procedure is repeated B times for some large value of B (say 100 or 1000), in order to produce B different bootstrap data sets, $Z^{*1}, Z^{*2}, \cdots, Z^{*B}$, and B corresponding α estimates, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \cdots, \hat{\alpha}^{*B}$.
- We estimate the standard error of these bootstrap estimates using the formula

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

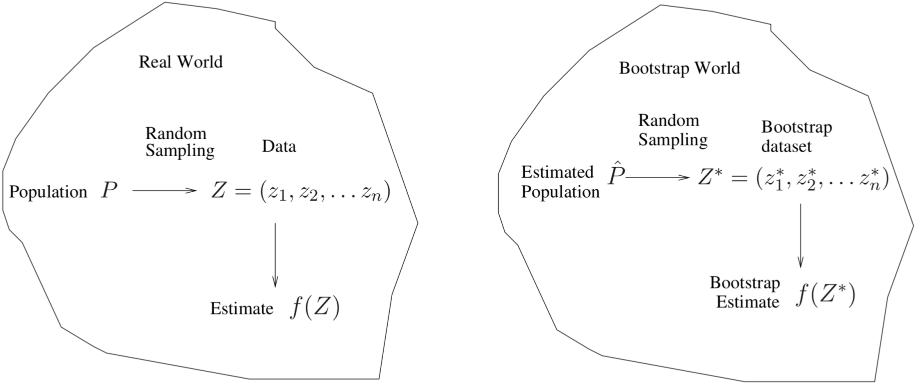
- This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set.

Results



Left: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population. $\mathrm{SE}(\hat{\alpha}) = 0.083$. **Center**: A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set. $\mathrm{SE}_B(\hat{\alpha}) = 0.087$. **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.

A general picture for the bootstrap



Real World

Population P Random Sampling $\longrightarrow$ Data $Z = (z_1, z_2, \ldots z_n)$

$\downarrow$

Estimate $f(Z)$

Bootstrap World

Estimated Population $\hat{P}$ Random Sampling $\longrightarrow$ Bootstrap dataset $Z^* = (z_1^*, z_2^*, \ldots z_n^*)$

$\downarrow$

Bootstrap Estimate $f(Z^*)$

STOP: Tuesday 2/15/22 (Week 5, Lecture 8)