Chapter 5: Resampling Methods

Yonghyun Kwon

Department of Mathematics, Korea Military Academy

Cross-Validation and the Bootstrap

- In this chapter 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.

- For example, they provide estimates of
  - *test-set prediction error*,
  - the standard deviation and bias of our parameter estimates.

Training Error versus Test Error

- Recall the distinction between the *test error* and the *training 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 not used in training the method.
- The *training error* can be easily calculated by applying the 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.

More on Prediction-Error Estimates

- **Best solution**: a large designated test set. Often not available.

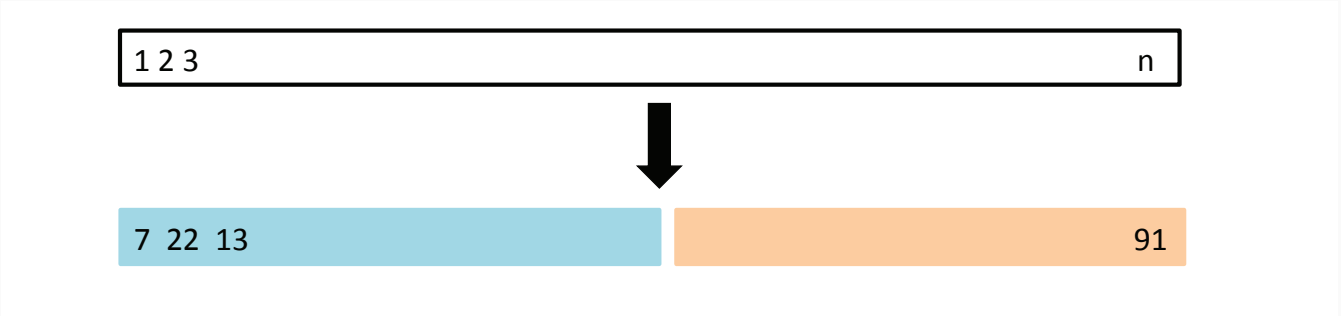
- Some methods make a *mathematical adjustment* to the training error rate to estimate the test error. These include the C_p *statistic*, *AIC*, and *BIC*. They are discussed elsewhere in this course.

- Here we instead 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.

Validation-Set Approach

- We randomly divide the available 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** for a quantitative response,
  - **misclassification rate** for a qualitative response.

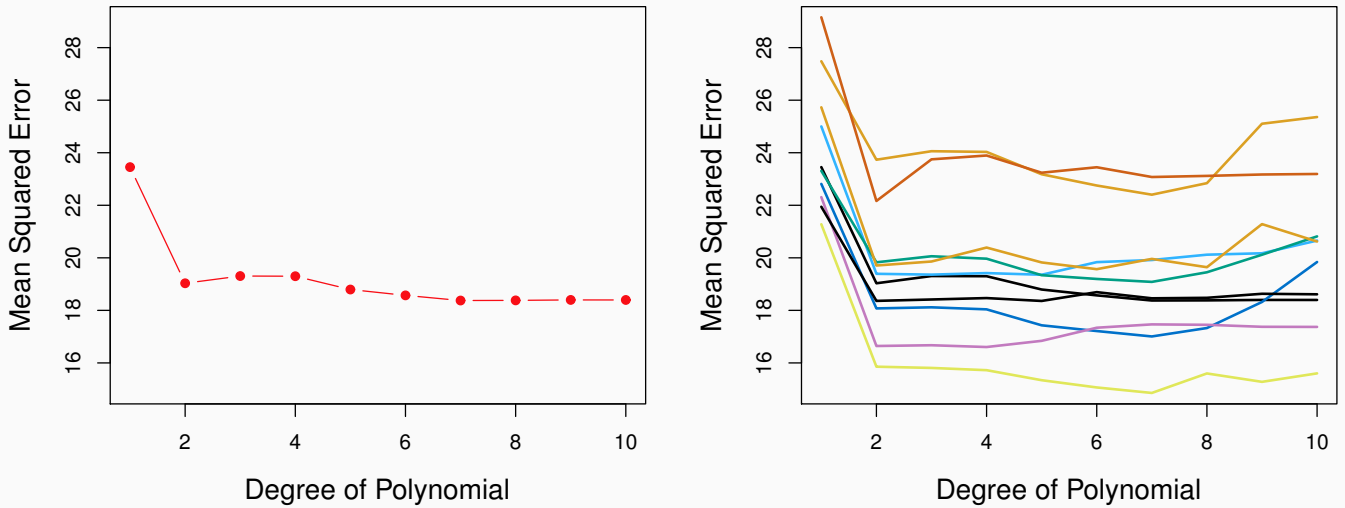
The Validation Process



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

Example: Automobile Data

- Goal: compare linear vs. higher-order polynomial terms in a regression of `mpg` on `horsepower`.
- We randomly split the 392 observations into a training set (196) and a validation set (196).



Left: single split. Right: ten different random splits — note the high variability in the estimated test error.

Drawbacks of the Validation Set Approach

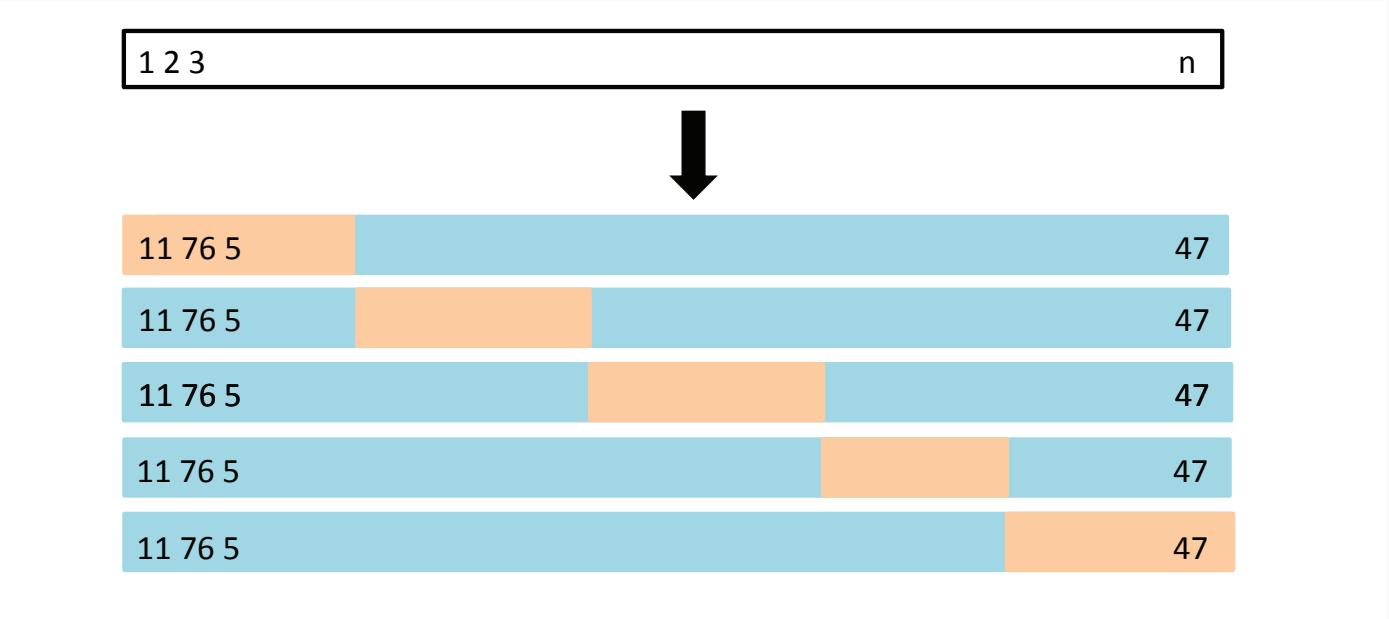
- The validation estimate of the test error can be *highly variable*, depending on which observations are included in the training set vs. the validation set.
- Only a subset of the observations — those in the training set — are used to fit the model.
- This suggests that the validation set error may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

K-Fold Cross-Validation

- *Widely used approach* for estimating test error.
- Estimates can be used to **select the best model** and to give an idea of the test error of the final chosen model.

- **Idea**: 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 $k = 1, 2, \ldots, K$, and then the results are combined.

K-Fold Cross-Validation in Detail



Divide data into K roughly equal-sized parts ($K = 5$ shown). At each step, one fold is the *validation* set; the rest form the *training* set.

The Details

- Let the K parts be $C_1, C_2, \ldots, C_K$, where C_k denotes the indices in part k, with n_k observations. If n is a multiple of K, then $n_k = n/K$.

K-Fold CV Error

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

where $\hat{y}_i$ is the fit for observation i obtained from the data with part k removed.

- Setting $K = n$ yields n-fold or *leave-one-out cross-validation (LOOCV)*.

A Nice Special Case: LOOCV Shortcut

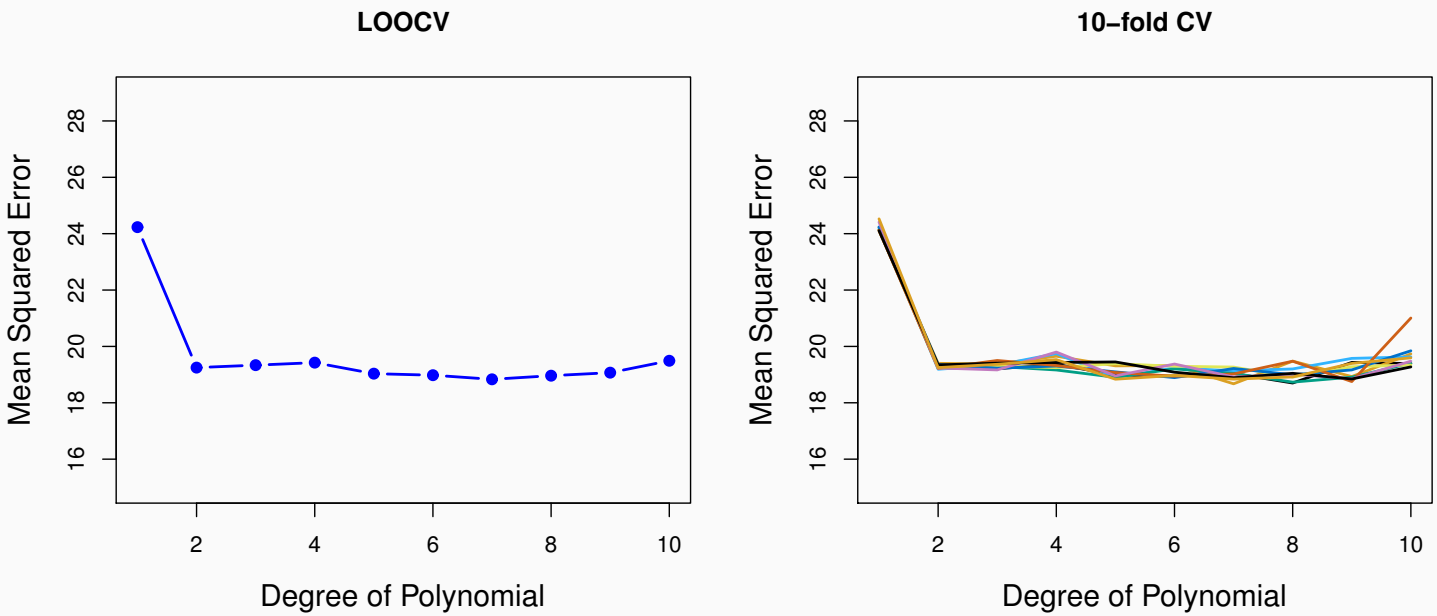
LOOCV Shortcut for Least Squares

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

where $\hat{y}_i$ is the ith fitted value from the original least squares fit, and h_i is the *leverage* (diagonal of the hat matrix). This is like the ordinary MSE, except the ith residual is divided by $1 - h_i$.

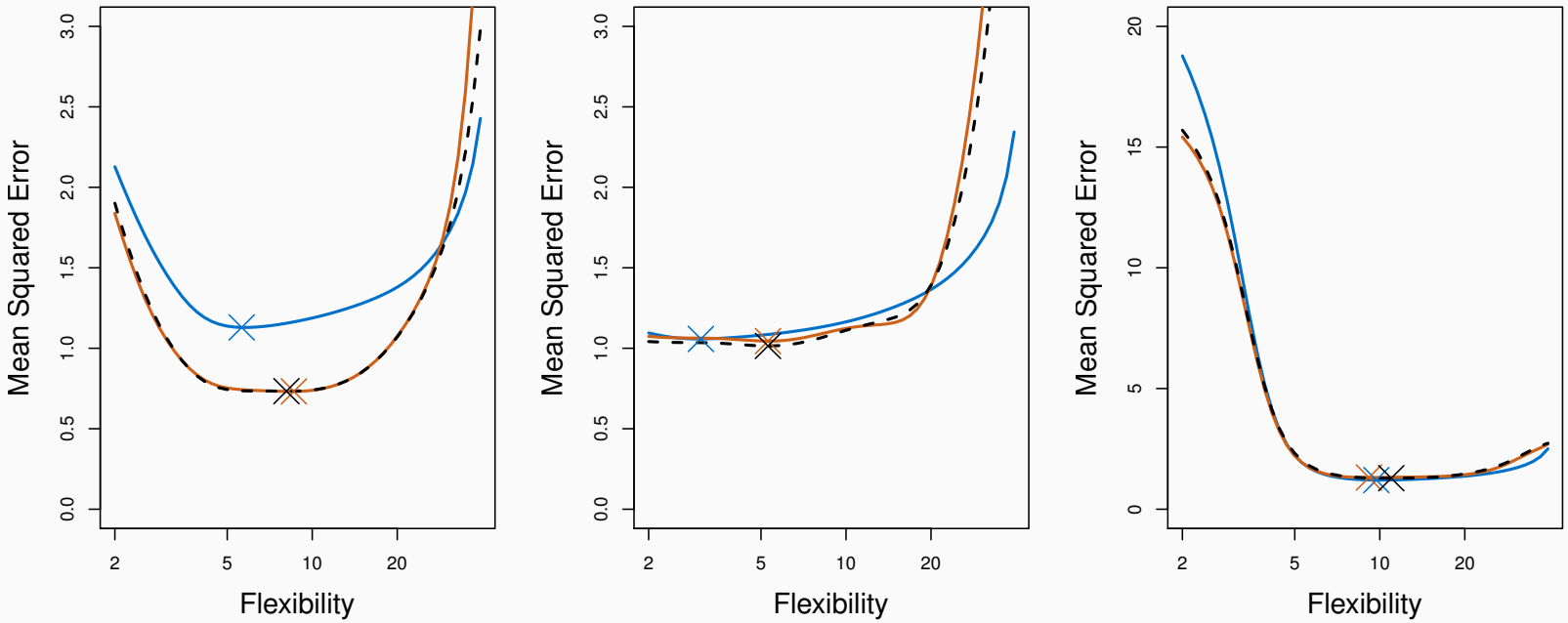
- An *amazing shortcut*: LOOCV costs no more than fitting the model once!
- LOOCV sometimes useful, but typically doesn't *shake up* the data enough — high fold correlation leads to *high variance*.
- A better choice is $K = 5$ or 10.

Auto Data Revisited: LOOCV vs. 10-Fold CV



Left: LOOCV (single deterministic curve). Right: 10-fold CV (multiple independent runs). Both correctly identify a quadratic fit as optimal; 10-fold CV shows more run-to-run variability.

True and Estimated Test MSE for Simulated Data



Three simulated scenarios. Blue = true test MSE; black dashed = 10-fold CV estimate; orange = LOOCV. The × marks the model selected by CV — it correctly identifies the optimal flexibility in each case.

Other Issues with Cross-Validation

- Since each training set is only $(K-1)/K$ as large as the original training set, the estimates of prediction error will typically be *biased upward*.
- This bias is minimized when $K = n$ (LOOCV), but this estimate has *high variance*, as noted earlier.
- $K = 5$ or 10 provides a good compromise for this *bias–variance tradeoff*.

Cross-Validation for Classification

K-Fold CV for Classification

$$\mathrm{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{Err}_k, \qquad \mathrm{Err}_k = \frac{1}{n_k} \sum_{i \in C_k} \mathbf{1}(y_i \neq \hat{y}_i).$$

The estimated standard deviation of CV_K is

$$\widehat{\mathrm{SE}}(\mathrm{CV}_K) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \frac{(\mathrm{Err}_k - \overline{\mathrm{Err}}_k)^2}{K-1}}.$$

- This is a useful estimate, but strictly speaking *not quite valid*.

Cross-Validation: Right and Wrong

Consider a simple classifier applied to some two-class data:

1. Starting with 5000 predictors and 50 samples, find the **100 predictors** having the largest correlation with the class labels.
2. 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?

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 the true test error $= 50\%$, but the CV error that ignores Step 1 is essentially *zero*!
- We have seen this error made in many high-profile **genomics papers**.

Wrong: Apply CV in step 2 only.

Right: Apply CV to steps 1 and 2.

The Bootstrap

- 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.
- The name derives from the phrase *to pull oneself up by one's bootstraps* — the idea of generating information from only what you already have.

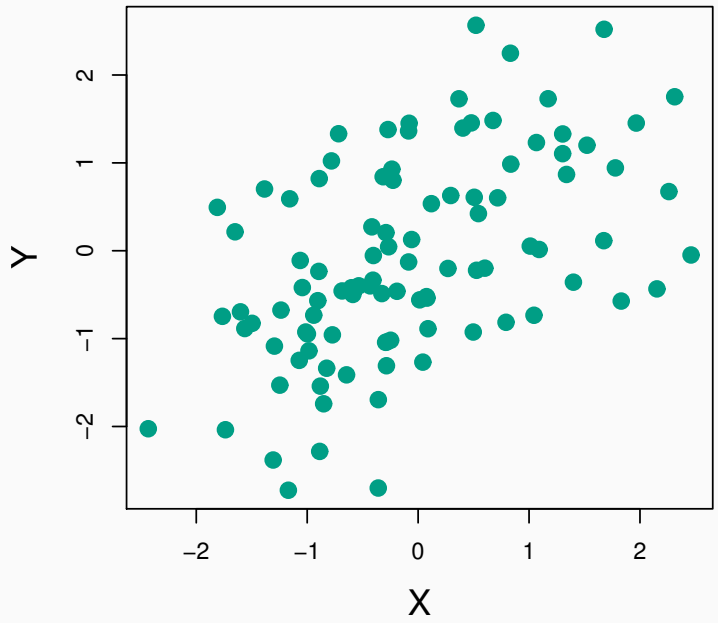
A Simple Example

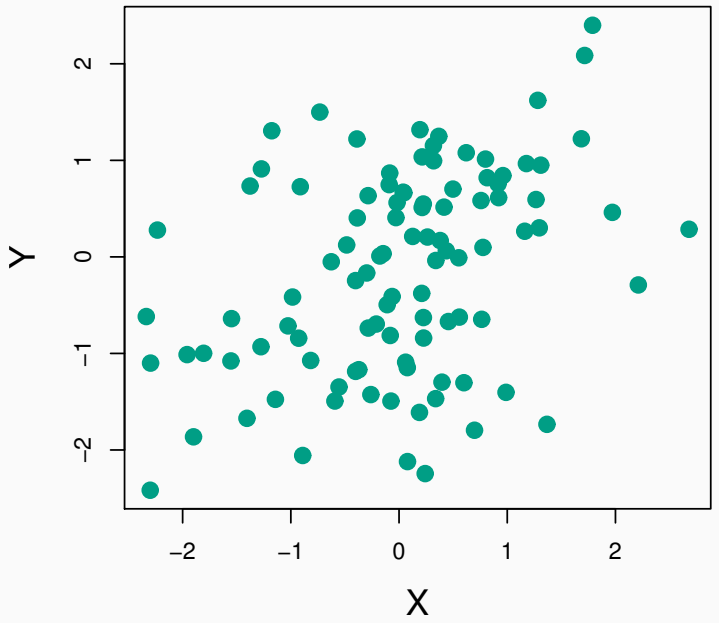
- Invest a fixed sum in two financial assets with random returns X and Y.
- Invest fraction α in X, remaining $1-\alpha$ in Y.
- Minimize $\mathrm{Var}(\alpha X + (1-\alpha)Y)$.

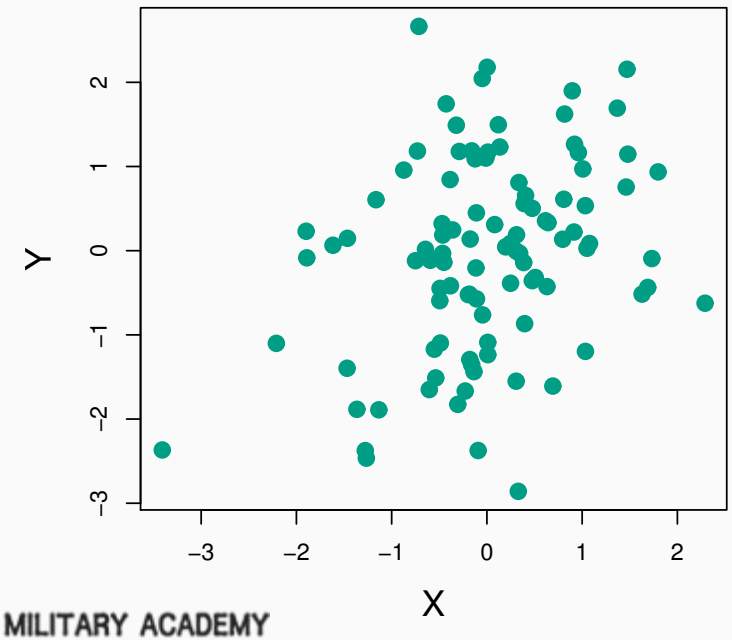
Optimal Allocation

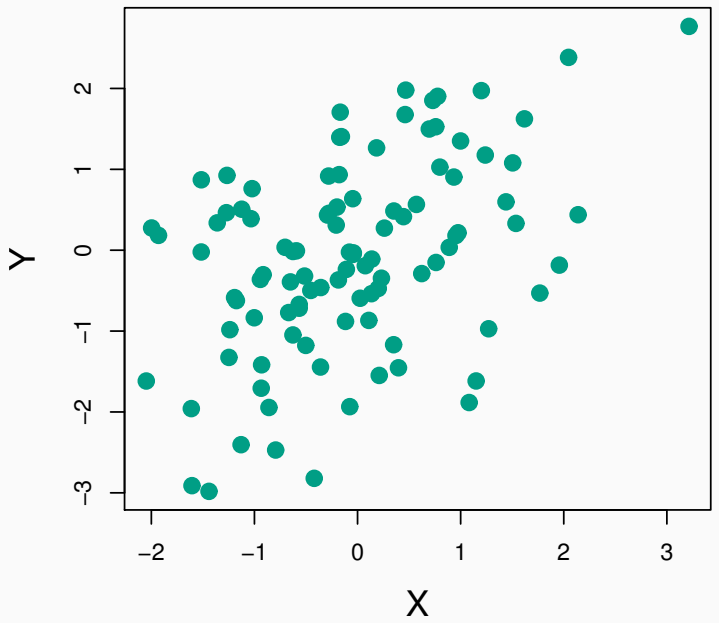
$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}, \qquad \hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

- **Question:** What is the standard error of $\hat{\alpha}$?


Y
X
−2
−1
0
1
2

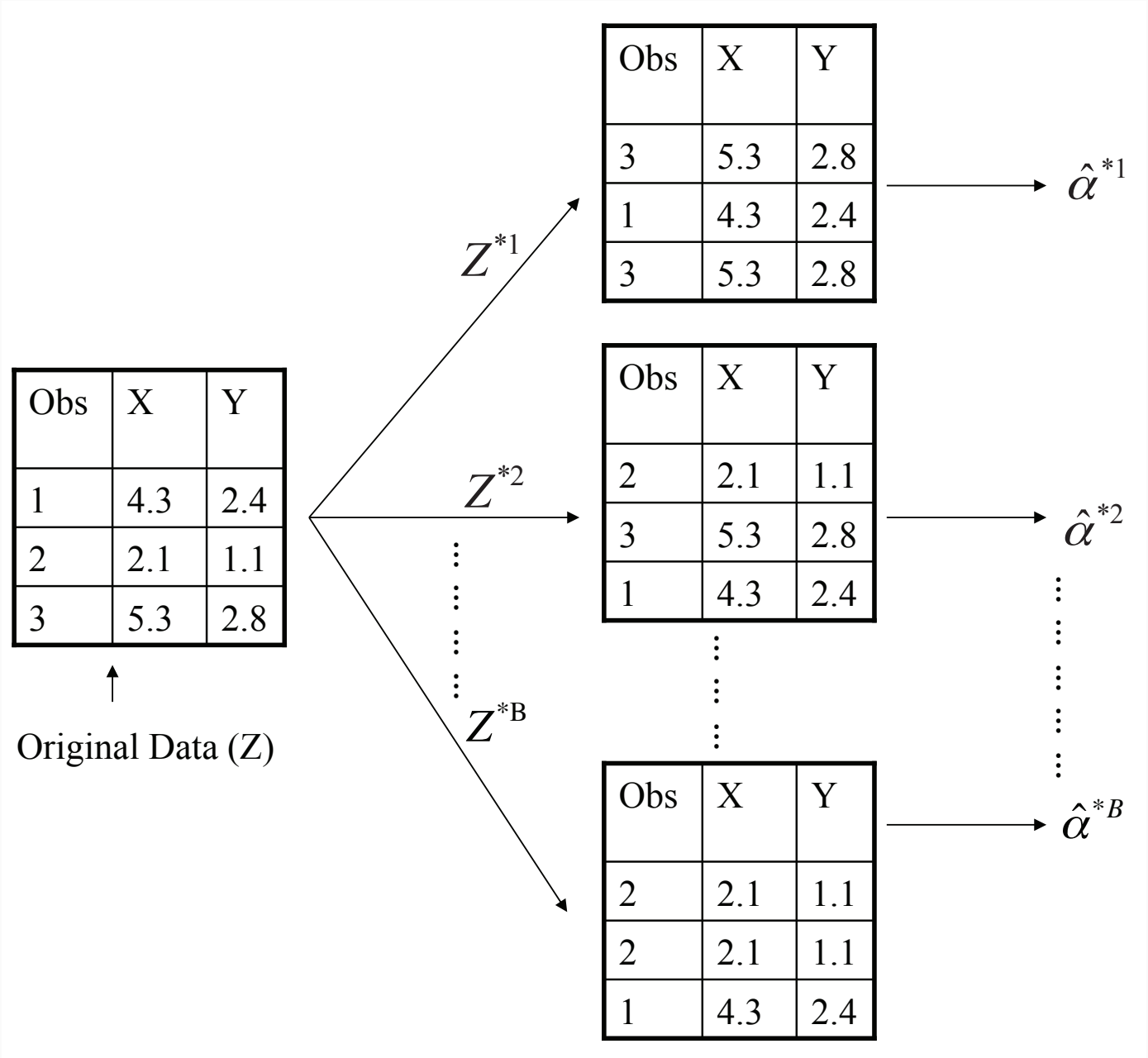

Y
X
−2
−1
0
1
2


Y
X
−3
−2
−1
0
1
2


Y
X
−3
−2
−1
0
1
2
3

Now Back to the Real World

- For real data we *cannot* generate new samples from the original population.
- 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 bootstrap dataset is the **same size** as the original; some observations may appear *more than once* and some *not at all*.

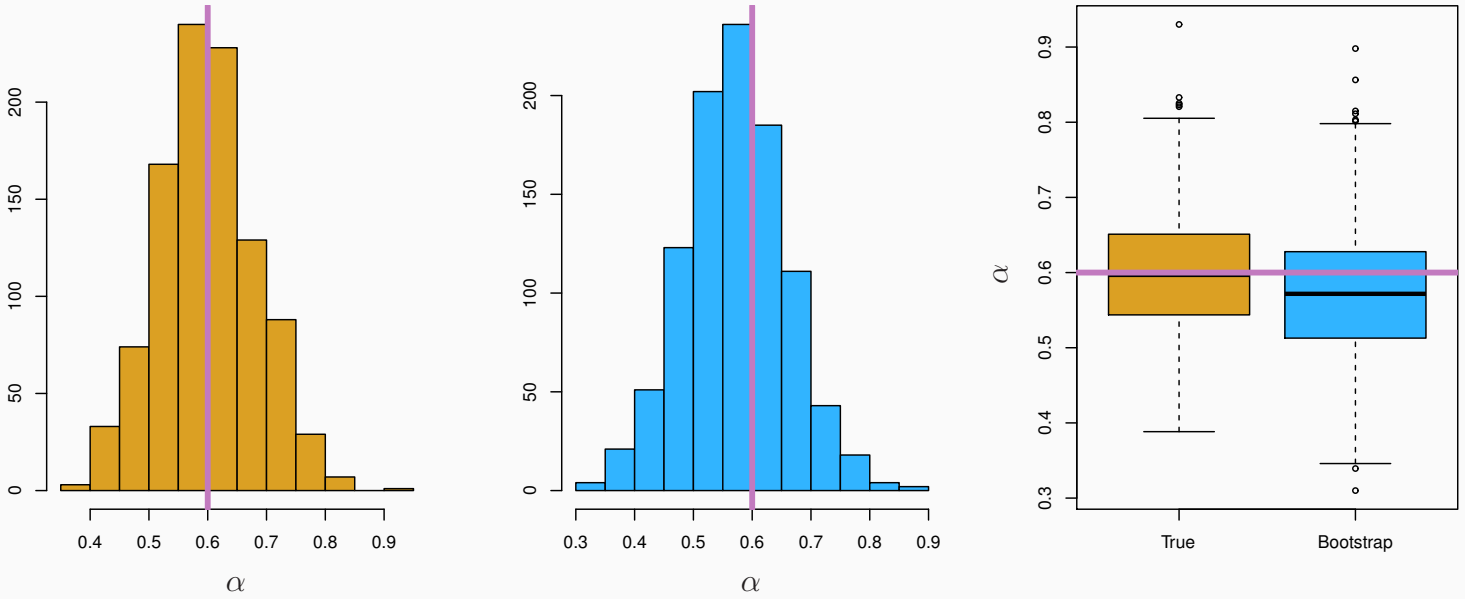

Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$

Bootstrap Standard Error

Bootstrap Standard Error Estimate

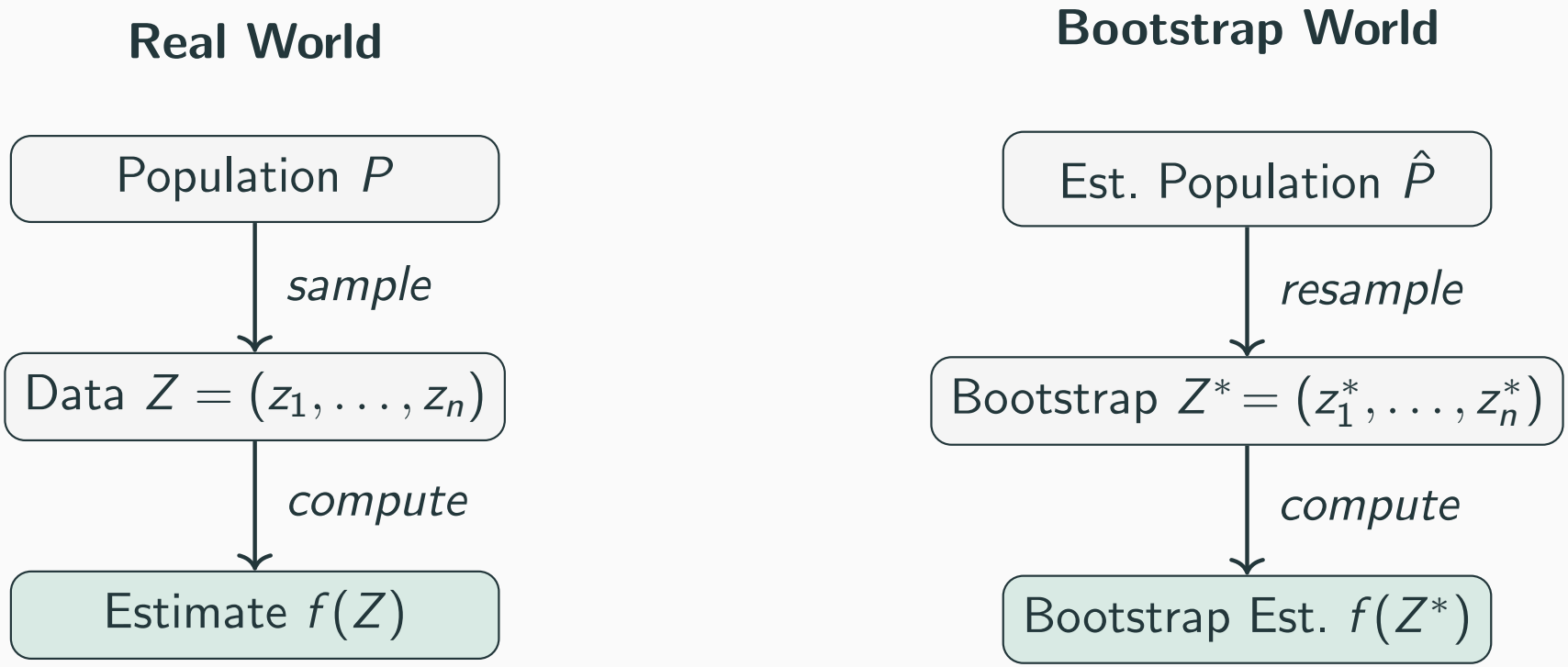
$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^*\right)^2}, \qquad \bar{\hat{\alpha}}^* = \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}^{*r}.$$

- Repeat for B (say 100 or 1,000) bootstrap samples to get $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$.
- This serves as an estimate of $\mathrm{SE}(\hat{\alpha})$ estimated from the original data.
- For this example, $\mathrm{SE}_B(\hat{\alpha}) = 0.087 \approx 0.083$ (true value).



Left: 1,000 estimates from true repeated sampling. *Center*: 1,000 bootstrap estimates from a **single** dataset. **Right**: Boxplot comparison. The bootstrap closely approximates the true sampling distribution. Pink line $=$ true $\alpha = 0.6$.

A General Picture for the Bootstrap



The bootstrap replaces the unknown P with the *empirical distribution* $\hat{P}$ and mimics the sampling process computationally.

Other Uses of the Bootstrap

- Primarily used to obtain *standard errors* of an estimate.
- Also provides approximate *confidence intervals* for a population parameter:
  - The *Bootstrap Percentile CI* uses quantiles of $\{\hat{\alpha}^{*b}\}$.
  - *Example*: the 5% and 95% quantiles of 1,000 bootstrap estimates give an approximate **90% CI** of $(0.43,\ 0.72)$.
- **Time series**: observations are serially correlated, so naive resampling destroys dependence. Instead, sample *blocks of consecutive observations* with replacement, then paste the sampled blocks to form a bootstrap dataset.

Can the Bootstrap Estimate Prediction Error?

- In cross-validation, each validation fold is *distinct* from the training folds — *no overlap*. This is crucial.
- Each bootstrap sample has *significant overlap* with the original data. About two-thirds of original observations appear in each bootstrap sample.
- This overlap causes the bootstrap to *seriously underestimate* the true prediction error.
- In the end, *cross-validation* provides a simpler, more attractive approach for estimating prediction error.

The Bootstrap versus Permutation Tests

- The *bootstrap* samples from the estimated population $\hat{P}$, and uses the results to estimate *standard errors* and *confidence intervals*.
- *Permutation methods* sample from an estimated *null distribution* for the data, and use this to estimate *p-values* and *False Discovery Rates* for hypothesis tests.
- The bootstrap can be used to test a null hypothesis in simple situations. E.g., if $\theta = 0$ is the null hypothesis, check whether the confidence interval for θ contains zero.
- Can also adapt the bootstrap to sample from a null distribution (see Efron & Tibshirani, 1993, Ch. 16), but there is no real advantage over permutations.

Exercises

1. Show that about two-thirds of the original observations appear in each bootstrap sample. That is, show

$$P(\text{obs } i \in Z^*) = 1 - \left(1 - \frac{1}{n}\right)^n \xrightarrow{n\to\infty} 1 - e^{-1} \approx 0.632.$$

2. Consider LOOCV applied to a least-squares linear regression.

(A) Explain why the shortcut formula $\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$ requires fitting the model only once.

(B) For simple linear regression, show that $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$.

(C) Explain intuitively why an observation with *large leverage* h_i receives a larger penalty in the LOOCV formula.

Exercises (continued)

3. Consider the following two-stage classifier:

1. Select the top 100 predictors (out of 5,000) by correlation with the class labels.
2. Fit logistic regression on those 100 predictors.

(A) Explain why applying 10-fold CV only to step 2 gives an *overly optimistic* estimate of test error.

(B) Describe the *correct* cross-validation procedure.

(C) In a simulation with truly independent labels, a researcher finds CV error $\approx 0\%$ using the wrong procedure. What would the correct procedure give, and why?

4. Explain the *bias–variance tradeoff* in K-fold CV as a function of K. Which value of K minimizes bias? Which minimizes variance? What is a practical recommendation?