Ch. 5 **Resampling**

web.stanford.edu/class/stats202

Sergio Bacallado, Jonathan Taylor

Autumn 2022

- Test/train split
- Cross-validation
- Bootstrap

Not covered: Permutation tests ...

Validation

Thinking about the *true* loss function is important

- Most of the **regression** methods we've studied aim to minimize the RSS, while **classification** methods aim to minimize the 0-1 loss.
- In classification, we often care about certain kinds of error more than others; i.e. the natural loss function is not the 0-1 loss.
- Even if we use a method which minimizes a certain kind of training error, we can *tune* it to optimize our true loss function.
- Example: in the `default` study we could find the threshold that brings the False negative rate below an acceptable level.

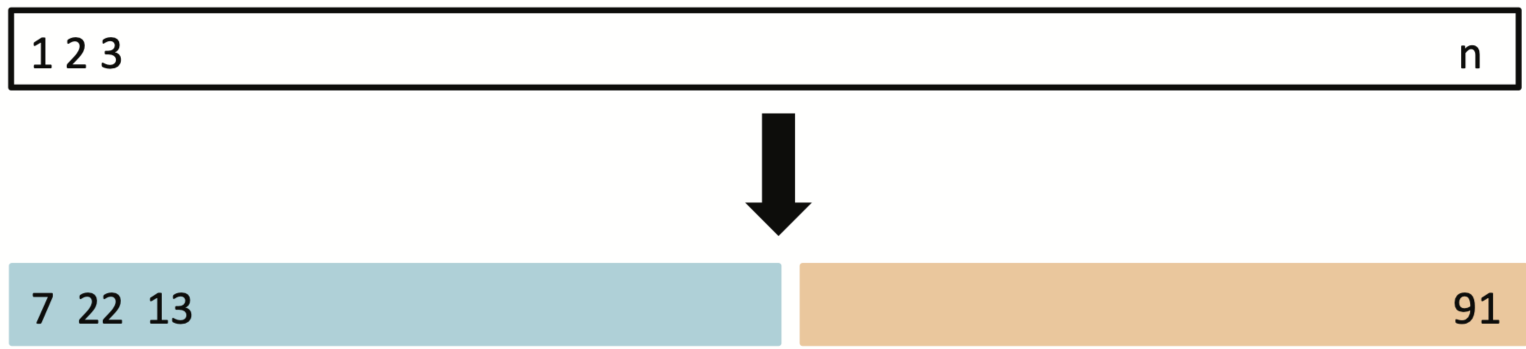
How to choose a supervised method that minimizes the test error

- In addition, *tune* the parameters of each method: maybe
  - *k in k-nearest neighbors.*
  - *The number of variables to include in forward or backward selection.*
  - *The order of a polynomial in polynomial regression.*

Validation set approach

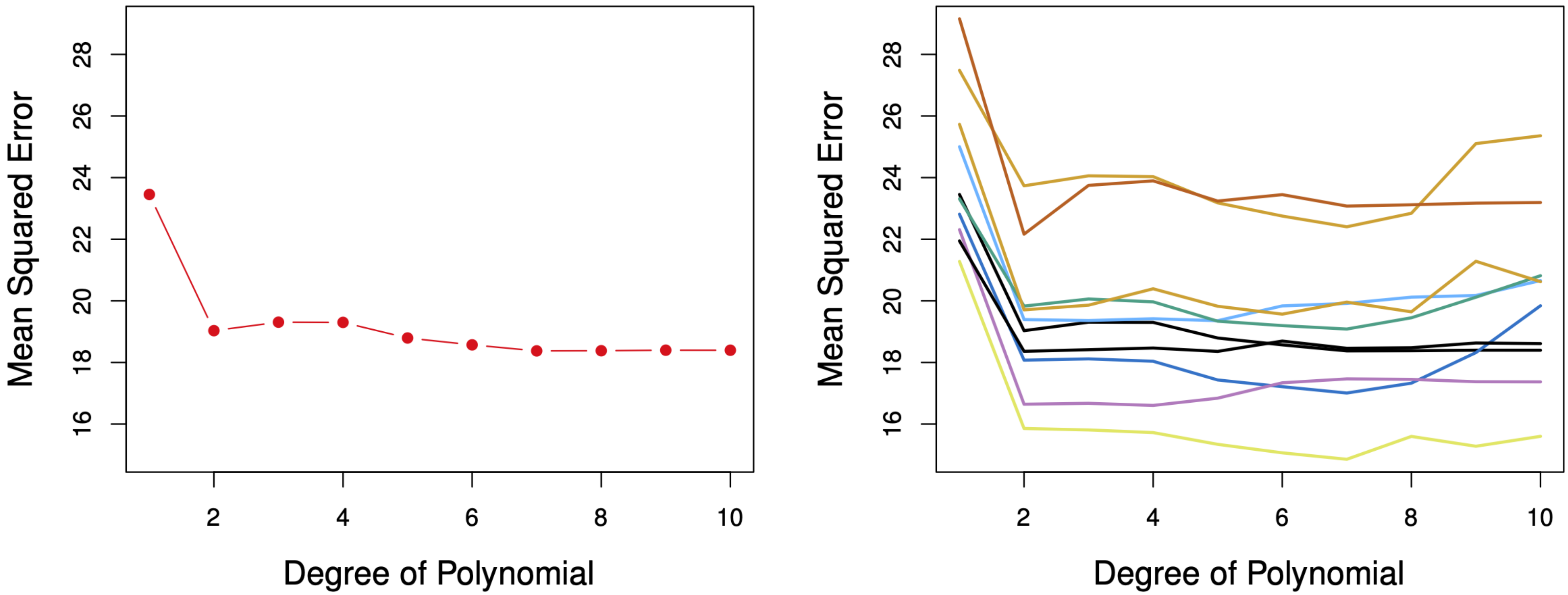
Use of a **validation set** is one way to approximate the test error:

- Divide the data into two parts.
- Train each model with one part.
- Compute the error on the remaining *validation* data.



Schematic of validation set approach.

Example: choosing order of polynomial



Left: validation error as a function of degree. Right: multiple splits into validation and training.

- Polynomial regression to estimate `mpg` from `horsepower` in the Auto data.
- **Problem:** Every split yields a different estimate of the error.

Leave one out cross-validation (LOOCV)

- For every $i = 1, \ldots, n$:
  - *train the model on every point except i,*
  - *compute the test error on the held out point.*
- Average the test errors.

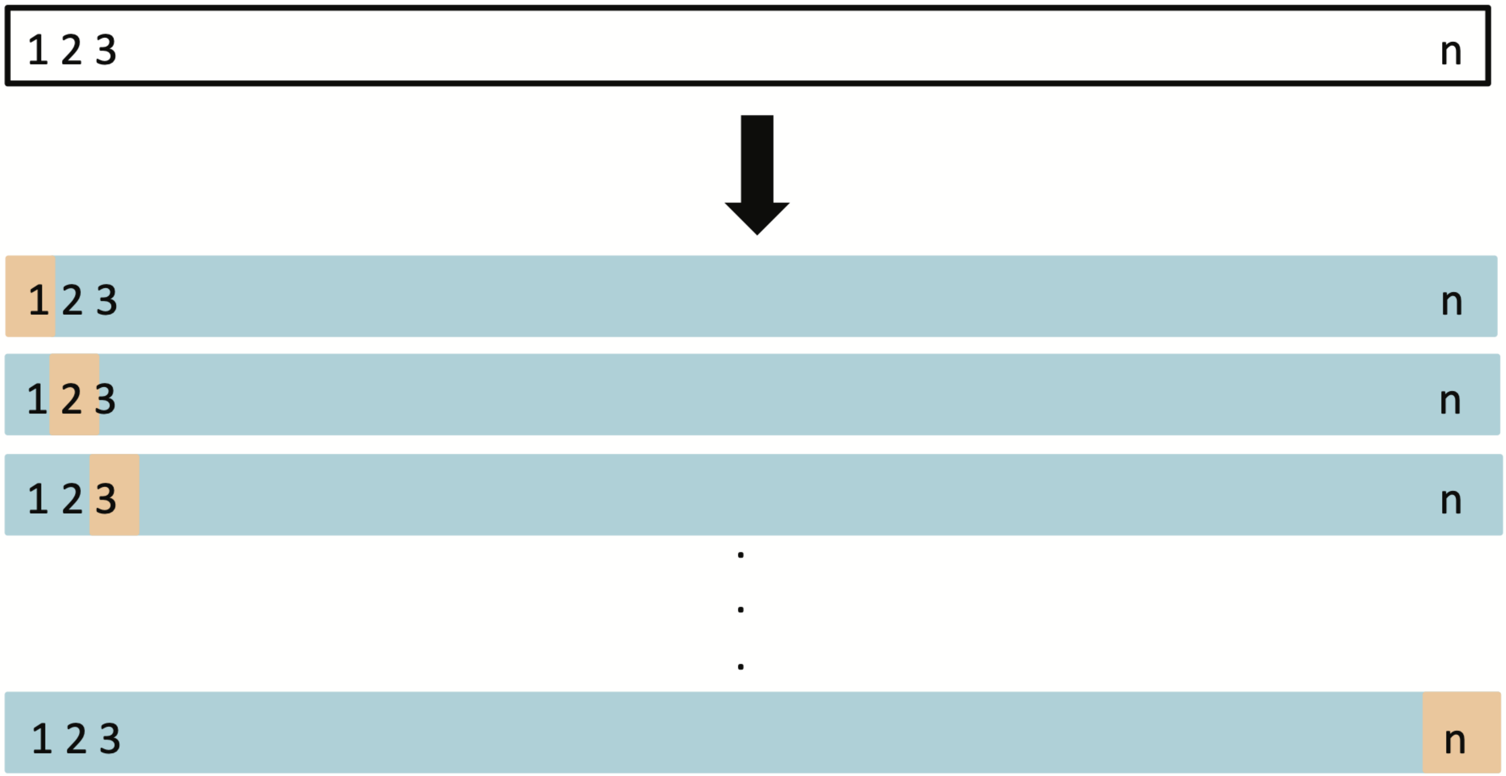
Regression

- Overall error:

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

- Notation $\hat{y}_i^{(-i)}$: prediction for the i sample when learning without using the ith sample.

Schematic for LOOCV



Schematic of leave-one-out cross-validation (LOOCV) set approach.

Requires fitting method n times...

Classification

- Overall error:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i \neq \hat{y}_i^{(-i)})$$

- Here, $\hat{y}_i^{(-i)}$ is predicted label for the i sample when learning without using the ith sample.

Shortcut for linear regression

- Computing $\mathrm{CV}_{(n)}$ can be computationally expensive, since it involves fitting the model n times.

- For linear regression, there is a shortcut:

$Y_i - \hat{Y}_i^{(-i)}$

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

- Above, h_{ii} is the leverage statistic.

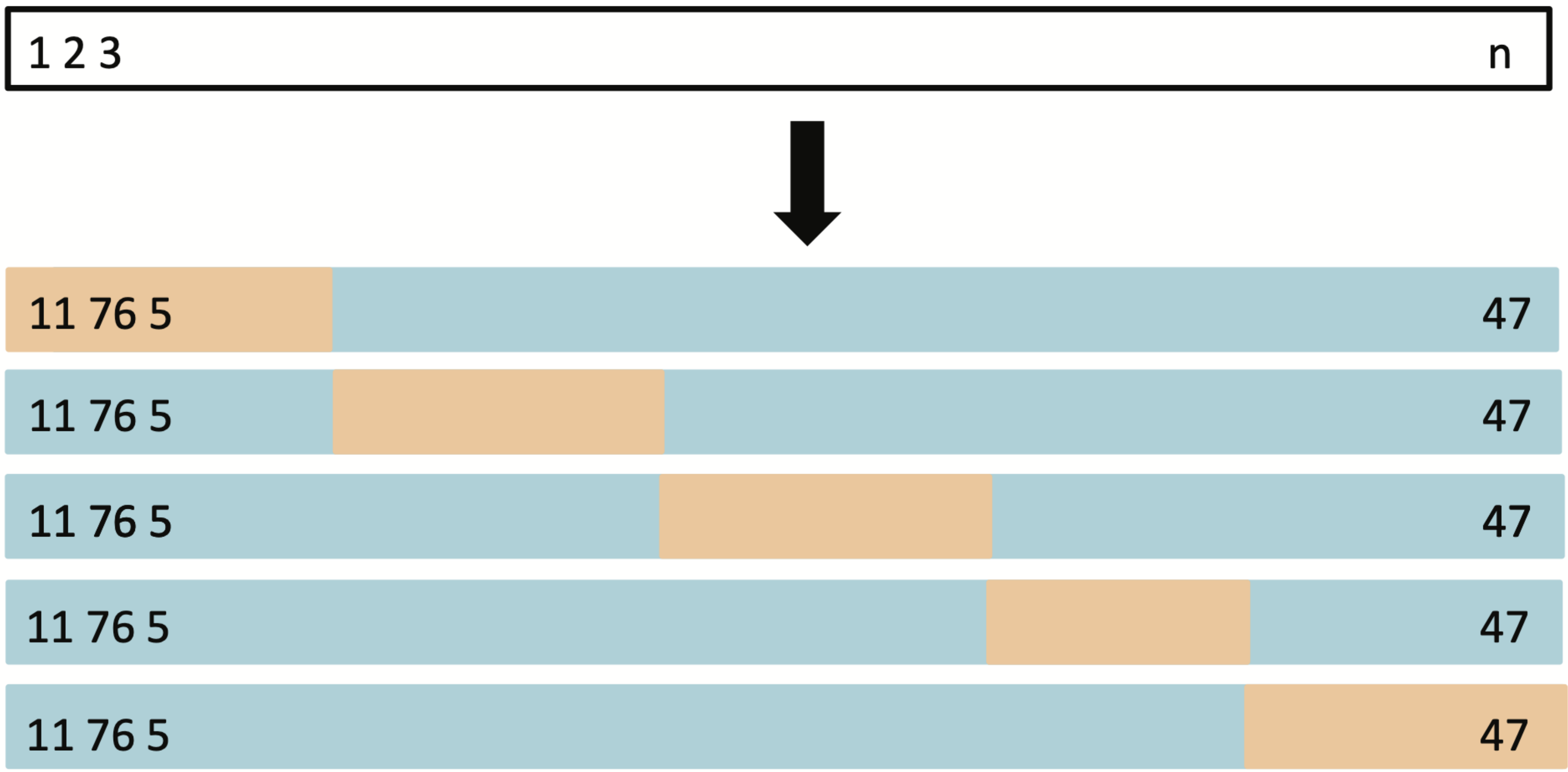
- Approximate versions sometimes used for logistic regression…

K-fold cross-validation

Algorithm 5.3? K-fold CV

- Split the data into K subsets or *folds*.
- For every $i = 1, \ldots, K$:
  - *train the model on every fold except the ith fold,*
  - *compute the test error on the ith fold.*
- Average the test errors.

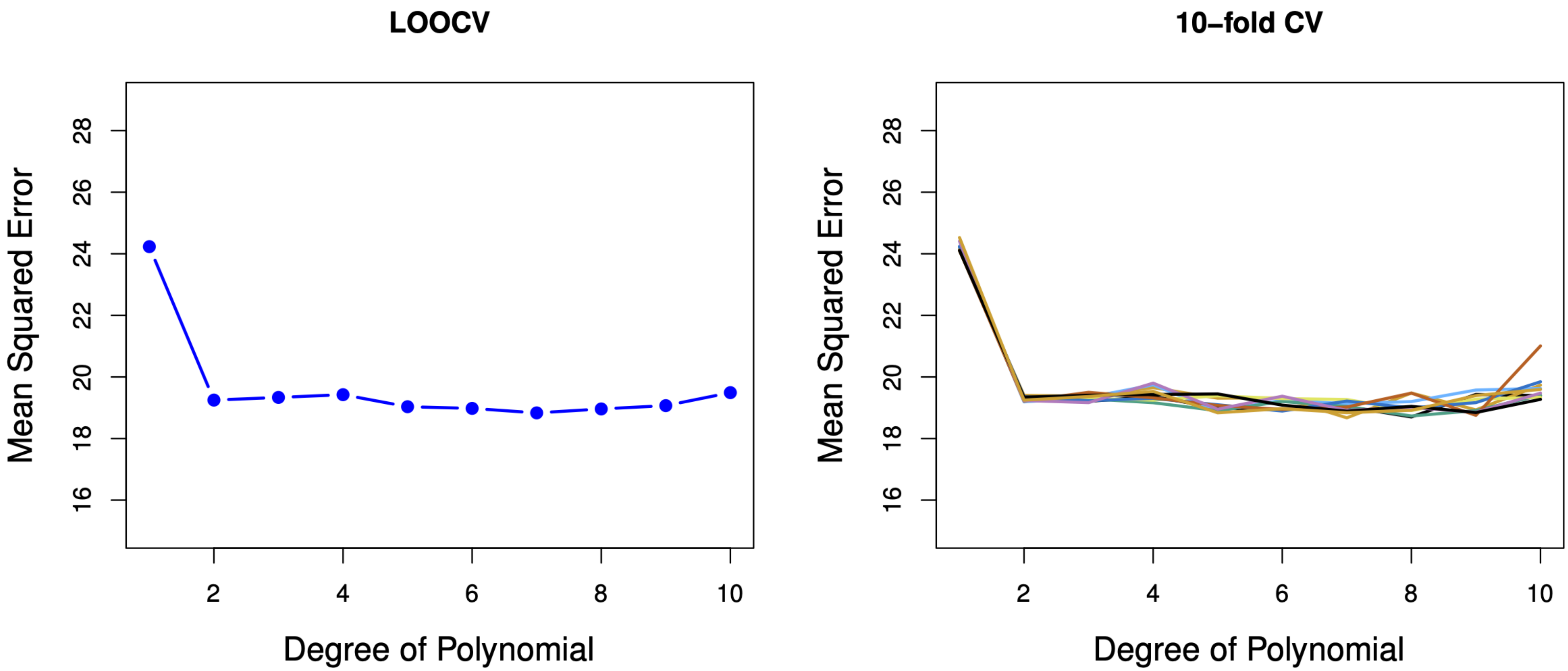
Schematic for K-fold CV



Schematic of K-fold CV fold approach.

Unlike LOOCV,
K-fold has
some randomness

LOOCV vs. K-fold cross-validation



Comparison of LOOCV and K-fold CV.

Much less randomness than single test/train split...

Common choice for k:

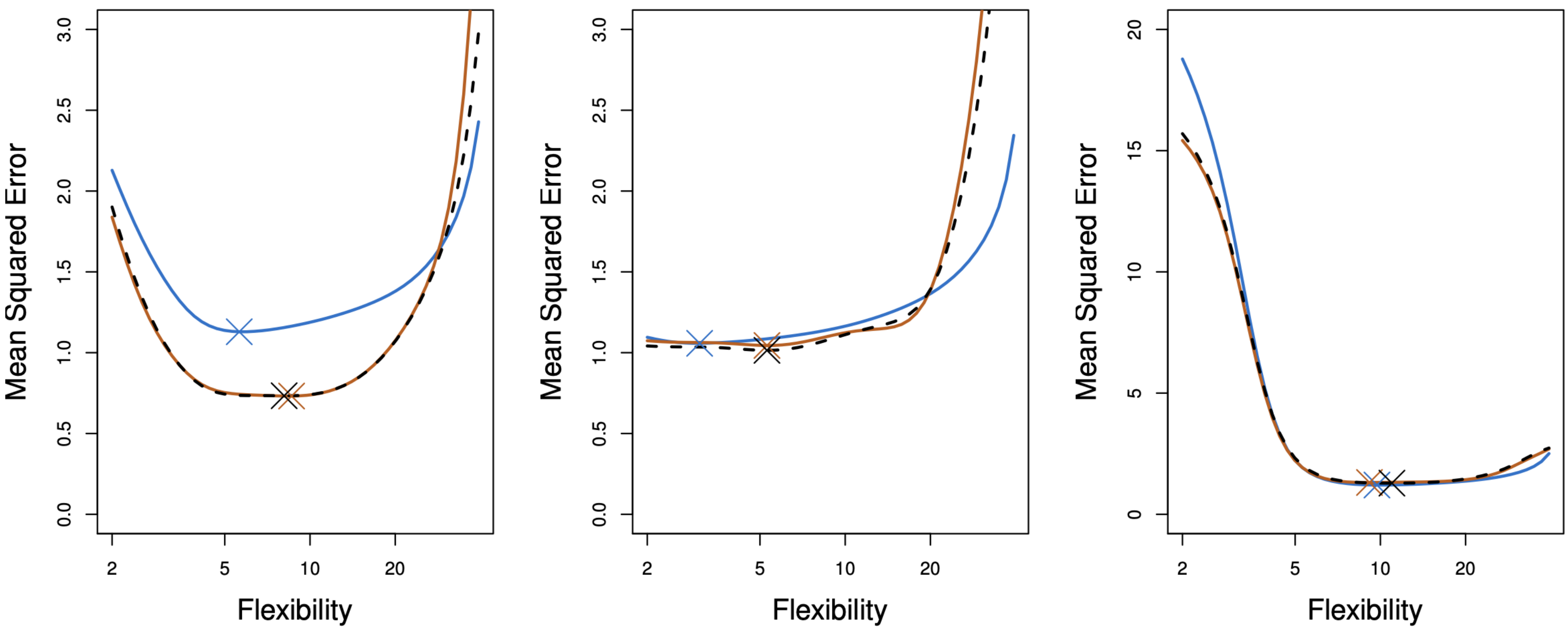
5, 10, n

LOOCV

Comments

- K-fold CV depends on the chosen split (somewhat).
- In K-fold CV, we train the model on less data than what is available to LOOCV. This introduces *some* bias into the estimates of test error.
- In LOOCV, the training samples highly resemble each other. This increases the *some* variance of the test error estimate.
- n-fold CV is equivalent LOOCV.

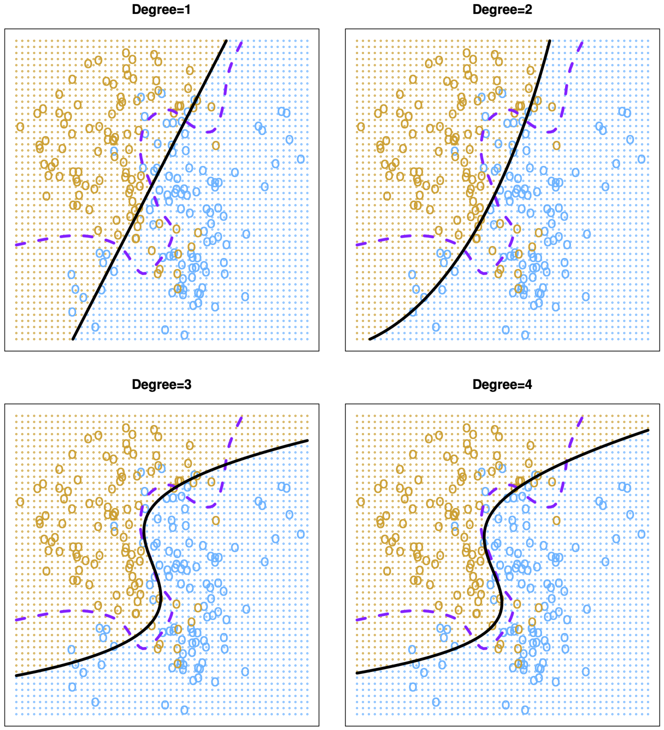
Choosing an optimal model



Comparison of LOOCV and K-fold CV to test MSE.

Even if the error estimates are off, choosing the model with the minimum cross validation error (10 fold in orange) often leads to a method with near minimum ~~test~~ error.

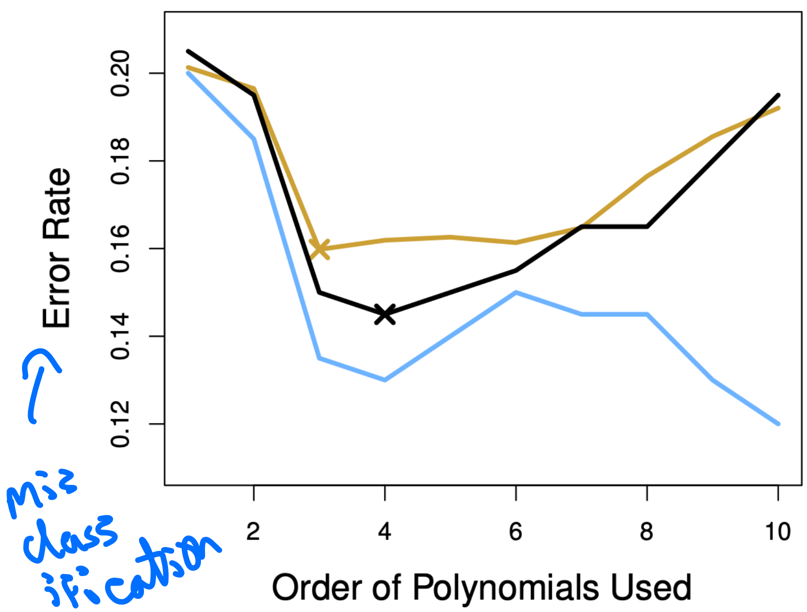
population

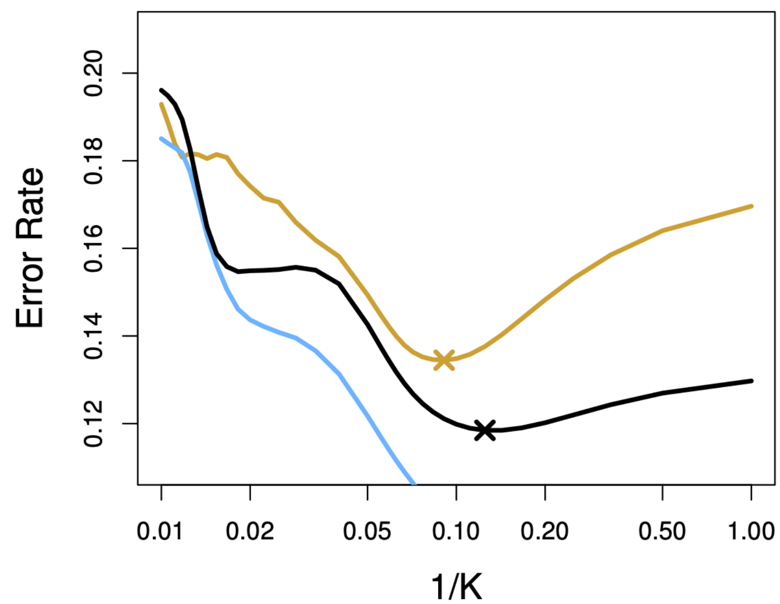


In a classification problem, things look similar.

- Logistic regression with polynomial predictors of increasing degree. (– – – – ––)
- – – – – –– Bayes boundary

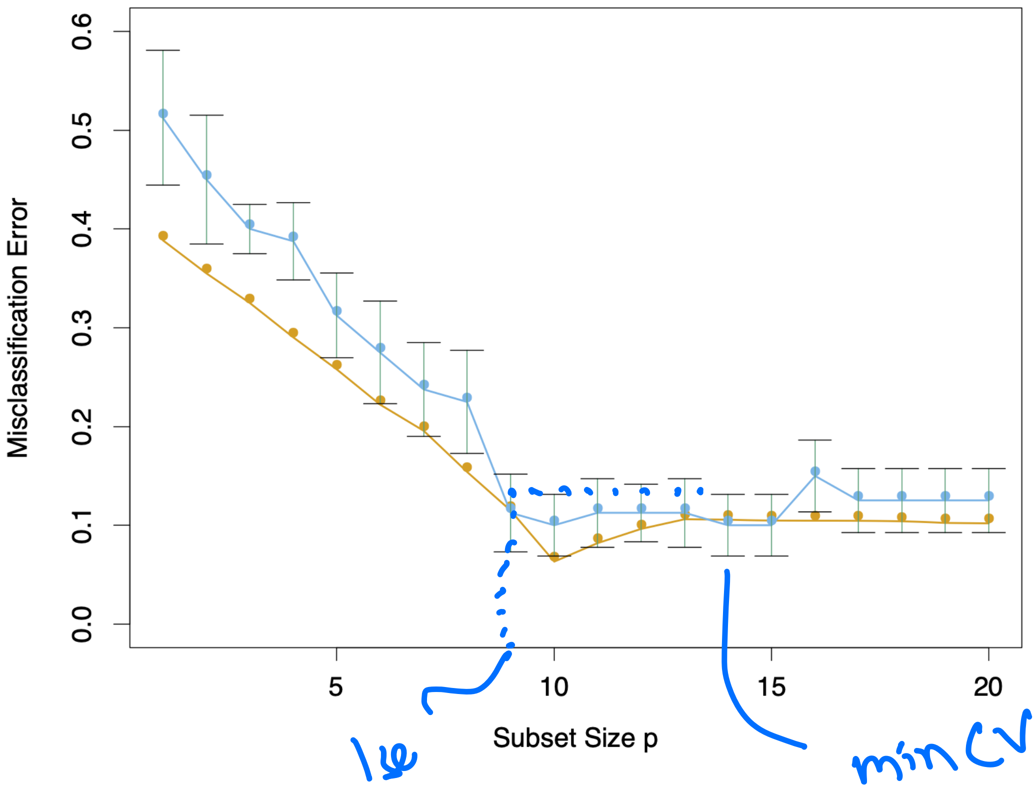
Choosing an optimal model





- Cubic model has best test error.
- Quartic has best CV.
- Curves look similar.
- *Q: Why doesn't training error keep decreasing?*

The one standard error (1SE) rule of thumb



Goal:
"simpler" model w/ similar CV to min. CV...

- Forward stepwise selection (we'll see in more detail shortly)
- 10-fold cross validation, True test error
- **1-SE rule of thumb:**
  - *A number of models with $10 \leq p \leq 15$ have almost the same CV error.*
  - *The vertical bars represent 1 standard error in the test error from the 10 folds.*
  - *Choose the simplest model whose CV error is no more than one standard error above the model with the lowest CV error.*

The wrong way to do cross validation

- *Reading:* Section 7.10.2 of The Elements of Statistical Learning.
- We want to classify 200 individuals according to whether they have cancer or not.
- We use logistic regression onto 1000 measurements of gene expression.
- **Proposed strategy:**

  1. *Using all the data, select the 20 most significant genes using z-tests.*
  2. *Estimate the test error of logistic regression with these 20 predictors via 10-fold cross validation.*

- To see how that works, let's use the following simulated data:

    1. *Each gene expression is standard normal and independent of all others.*

    2. *The response (cancer or not) is sampled from a coin flip — no correlation to any of the "genes".*

- Q: What should the misclassification rate be for any classification method using these predictors?

- A: Roughly 50%.

- We run this simulation, and obtain a CV error rate of 3%!
- Why?
  - *Since we only have 200 individuals in total, among 1000 variables, at least some will appear correlated with the response.*
  - *We had run variable selection using all the data, so the variables we select have some correlation with the response in every subset or fold in the cross validation.*

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, perform the variable selection and fit the model with the selected variables.*

    2. *Compute the error on fold i.*

    3. *Average the 10 test errors obtained.*

- In our simulation, this produces an error estimate of close to 50%.

- **Moral of the story:** Every aspect of the learning method that involves using the data — variable selection, for example — must be cross-validated.

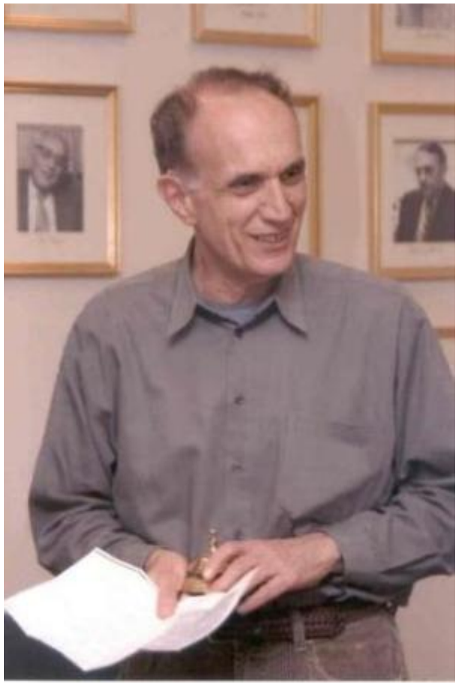
Bootstrap

- Another resampling technique often seen in practice.

Cross-validation vs. the Bootstrap

- **Cross-validation:** provides estimates of the (test) error
- **The Bootstrap:** provides the (standard) error of estimates

Bootstrap



Brad Efron

- One of the most important techniques in all of Statistics.
- Computer intensive method.
- Popularized by Brad Efron ← Stanford pride!

Standard errors in linear regression from a sample of size n

```r
Advertising = read.csv('https://www.statlearning.com/s/Advertising.csv')
M.sales = lm(sales ~ TV, data=Advertising)
summary(M.sales)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

uses the model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

Classical way to compute Standard Errors

- **Example:** Estimate the variance of a sample $x_1, x_2, \ldots, x_n$:

- Unbiased estimate of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

CI: $\hat{\sigma}^2 \pm 2 \cdot SE(\hat{\sigma}^2)$

- What is the Standard Error of $\hat{\sigma}^2$?

  - *Assume that $x_1, \ldots, x_n$ are normally distributed with common mean μ and variance σ^2.*
  - *Then $\hat{\sigma}^2(n-1)$ has a χ-squared distribution with $n-1$ degrees of freedom.*
  - *For large n, $\hat{\sigma}^2$ is normally distributed around σ^2.*
  - *The SD of this sampling distribution is the Standard Error.*

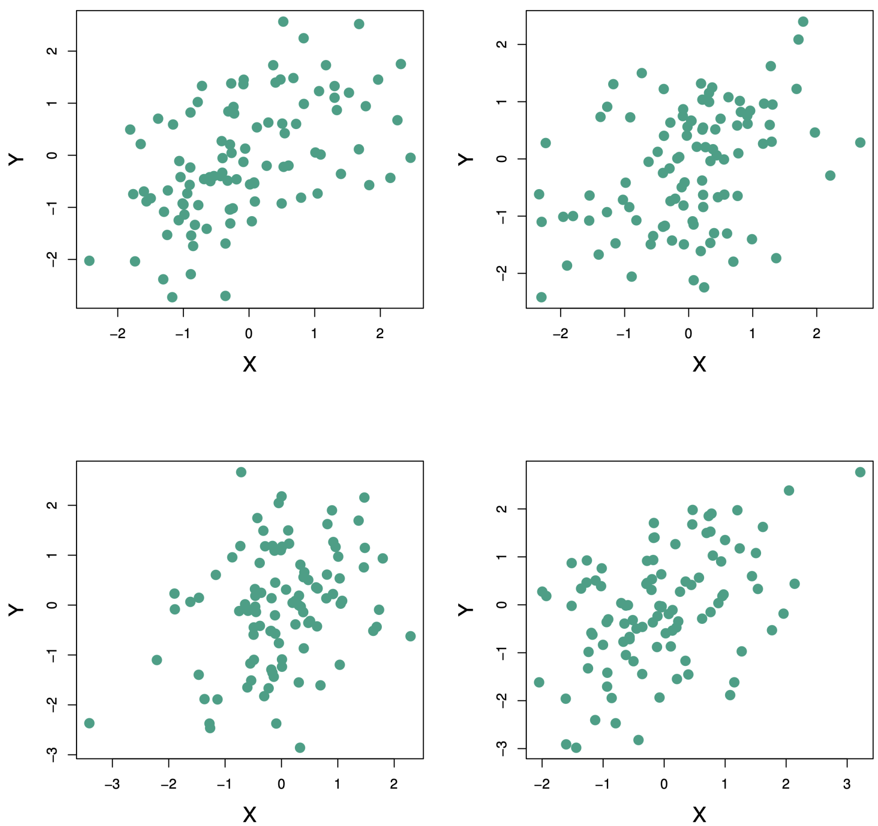
Assumptions

↳ can compute $SE(\hat{\sigma}^2)$

Limitations of the classical approach

- This approach has served statisticians well for many years; however, what happens if:
  - *The distributional assumption — for example,* $x_1, \ldots, x_n$ *being normal — breaks down?*
  - *The estimator does not have a simple form and its sampling distribution cannot be derived analytically?*
- Bootstrap can handle (at least some of) these departures from the usual assumptions!

Example: Investing in two assets



- Suppose that X and Y are the returns of two assets.
- These returns are observed every day: $(x_1, y_1), \ldots, (x_n, y_n)$.

- We have a fixed amount of money to invest and we will invest a fraction α on X and a fraction $(1-\alpha)$ on Y.
- Therefore, our return will be

$$\alpha X + (1-\alpha)Y.$$

- Our goal will be to minimize the variance of our return as a function of α.
- One can show that the optimal α is:

Truth

$$\alpha = \frac{\sigma_Y^2 - \mathrm{Cov}(X,Y)}{\sigma_X^2 + \sigma_Y^2 - 2\mathrm{Cov}(X,Y)}.$$

- **Proposal:** Use an estimate:

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \widehat{\mathrm{Cov}}(X,Y)}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\widehat{\mathrm{Cov}}(X,Y)}.$$

Estimate

CI: $\hat{\alpha} \pm 2 \cdot SE(\hat{\alpha})$

How do we get this?

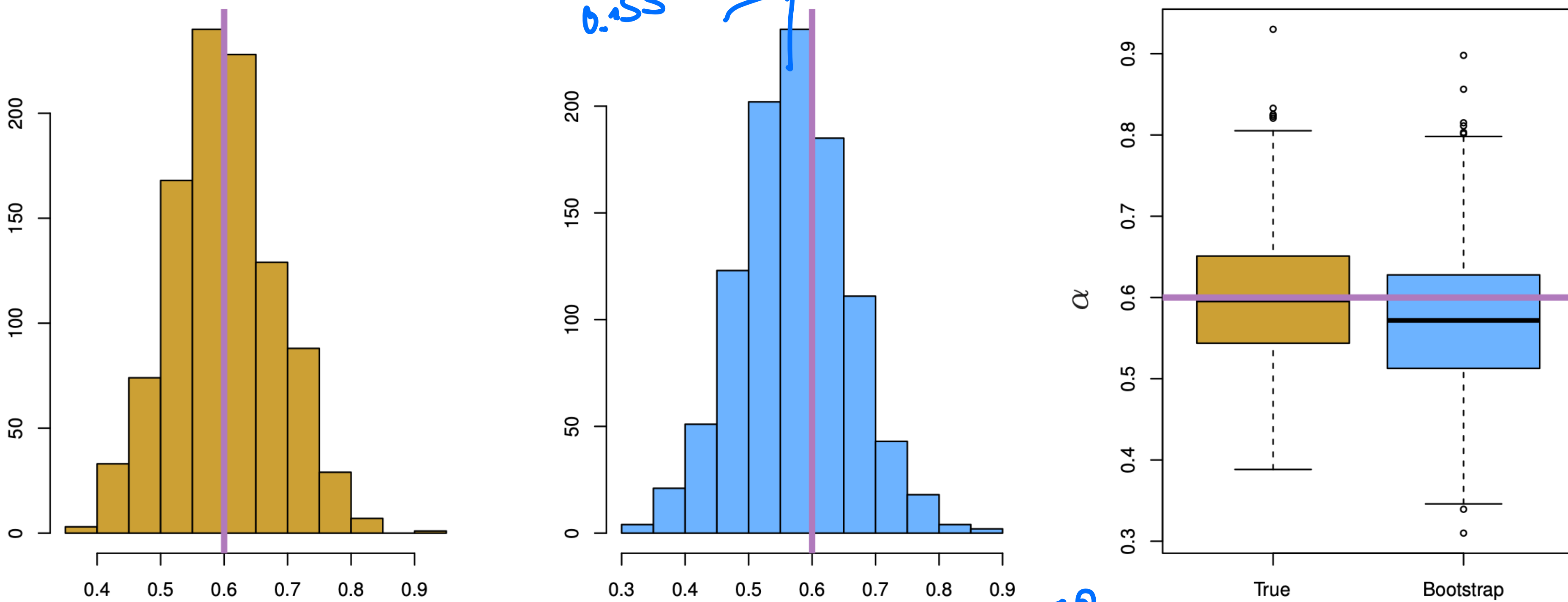
- Suppose we compute the estimate $\widehat{\alpha} = \cancel{0.6}$ using the samples $(x_1, y_1), \ldots, (x_n, y_n)$.
  0.55

- How sure can we be of this value? (*A little vague of a question.*)

- If we had sampled the observations in a different 100 days, would we get a wildly different $\widehat{\alpha}$? (*A more precise question.*)

Resampling the data from the true distribution

0.55



$O\left(\frac{1}{\sqrt{n}}\right)$

True dbn of $\hat{\alpha}$

Bootstrap

- In this thought experiment, we know the actual joint distribution $P(X, Y)$, so we can resample the n observations to our hearts' content.

- True distribution of $\widehat{\alpha}$

$SE(\text{Bootstrap}) \simeq SE(\text{Truth})$

Computing the standard error of $\widehat{\alpha}$

- We will use S samples to estimate the standard error of $\widehat{\alpha}$.

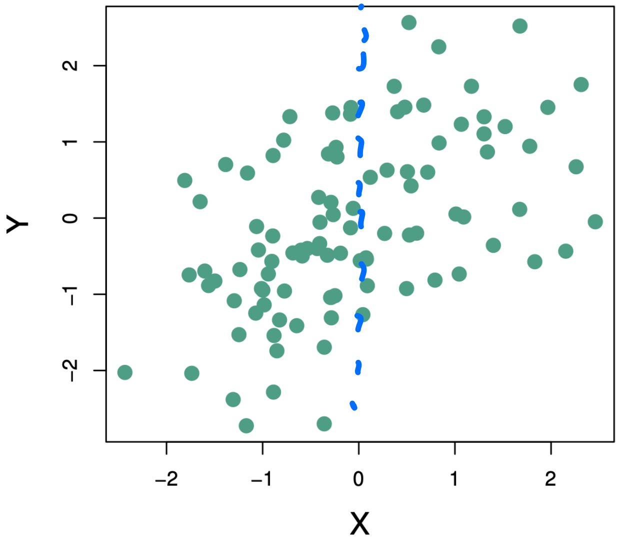
- For each sampling of the data, for $1 \leq s \leq S$

$$(x_1^{(s)}, \ldots, x_n^{(s)})$$

we can compute a value of the estimate $\widehat{\alpha}^{(1)}, \widehat{\alpha}^{(2)}, \ldots$.

- The Standard Error of $\widehat{\alpha}$ is approximated by the standard deviation of these values.

In reality, we only have n samples



$\hat{P}(X>0) \approx 40\%$

an estimate of $P(X>0)$

$\hat{E}[f(X,Y)]$

A single panel of Fig 5.9

- However, these samples can be used to approximate the joint distribution of X and Y.

Joint Dbn of X & Y.

$Var(\hat{\alpha}) \simeq$ function of Joint Dbn of X & Y.

- **The Bootstrap:** Sample from the *empirical distribution*:

$$\widehat{P}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}.$$

- Equivalently, resample the data by drawing n samples *with replacement* from the actual observations.

- *Why it works:* variances computed under the empirical distribution are good approximations of variances computed under the true distribution (in many cases).

$$\mathrm{Var}_{P}(h) \simeq \mathrm{Var}_{\hat{P}_B}(h)$$

A schematic of the Bootstrap

n=100

Z^{*1}

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

$\rightarrow \hat{\alpha}^{*1}$

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

Original Data (Z)

n=100

Z^{*2}

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

$\rightarrow \hat{\alpha}^{*2}$

⋮

Z^{*B}

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

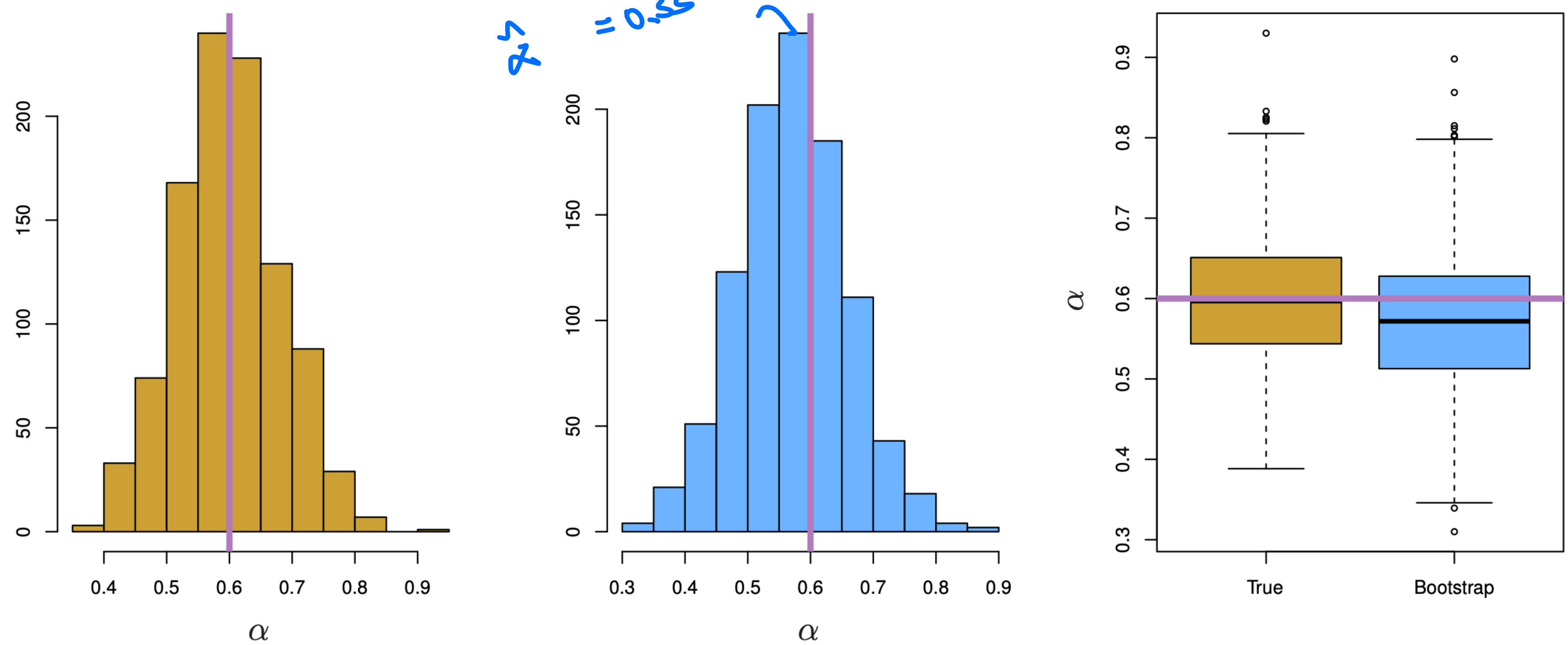
$\rightarrow \hat{\alpha}^{*B}$

A single dataset

$B \simeq 1000$

Comparing Bootstrap sampling to sampling from the true distribution

$\hat{\alpha} = 0.55$



- Left panel is population distribution of $\hat{\alpha}$ – centered (approximately) around the true α.
- Middle panel is bootstrap distribution of $\hat{\alpha}$ – centered (approximately) around observed $\hat{\alpha}$.