Resampling Methods

Slides on *Introduction to Statistical Learning*, Chapter 5

Edward Thompson

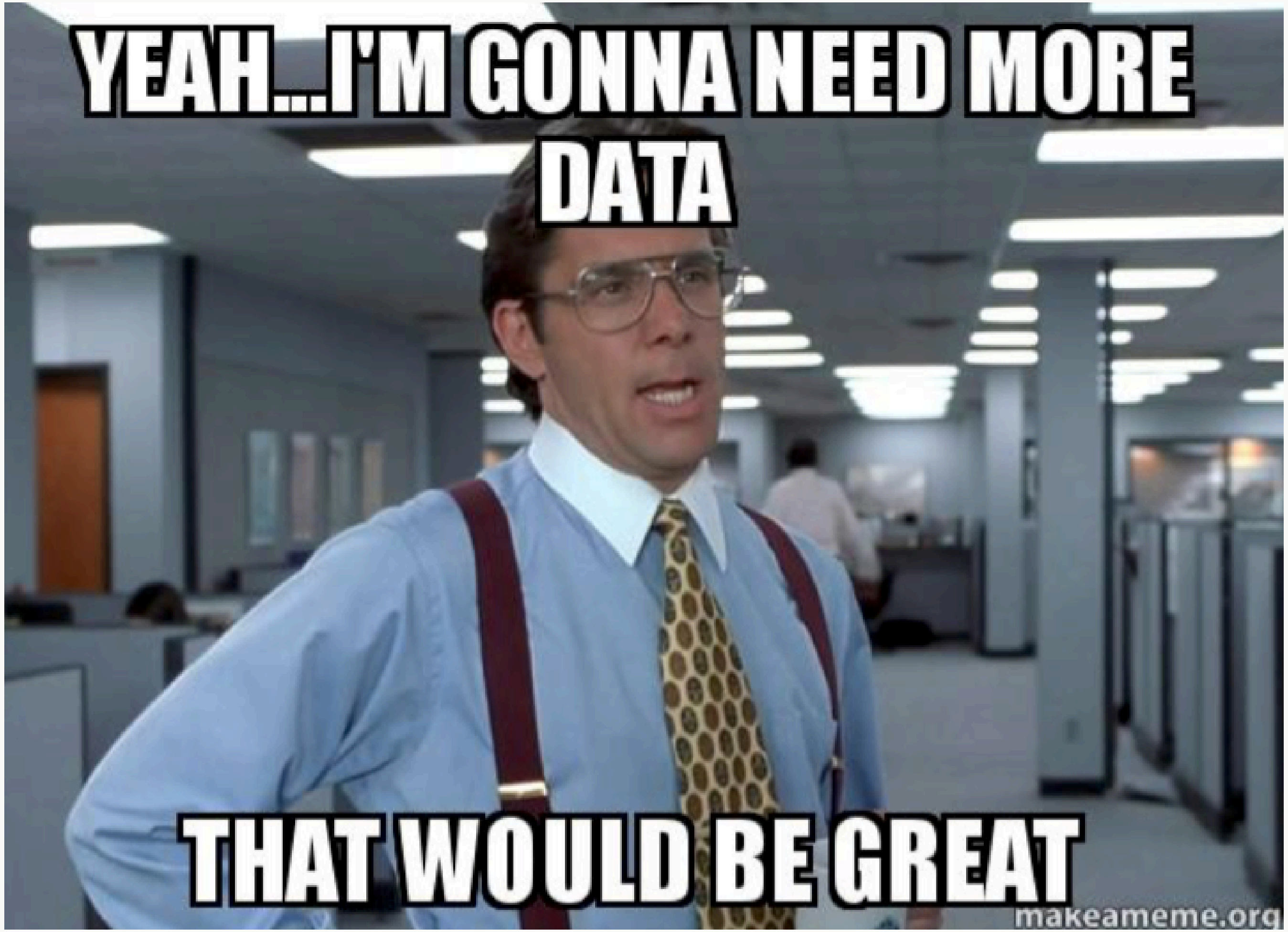
March 2024

See the Typst source: https://typst.app/project/peoR92wvjRSDgQBOtYFqgk

Table of contents

Introduction

Motivation


YEAH...I'M GONNA NEED MORE DATA
THAT WOULD BE GREAT
makeameme.org

In a Nutshell

- Repeatedly draw samples of data from a **training** set and refit models on each sample in order to obtain additional information.

Google Scholar Search Results

- Linear Regression: $644,000$
- Logistic Regression: $311,000$
- Cross Validation: $130,000$

High Level Steps

1. Draw a sample of data S, called the **training** data, from your **available** data D.
2. Fit the model on the training data sample S.
3. Check model fit on the remaining available data $D \setminus S$, called the **validation** data.
4. (Optional) Repeat 1-3 with other samples.
5. Use the model fits to learn about your model.

What do we learn?

- **Model assessment**
  - Estimating test errors
  - Checking robustness

- **Model selection**
  - Choose between multiple models
  - Choose hyper-parameter values

- Works for either categorical or quantitative variables.

Recap: Mean Squared Error

- Test MSE $= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$
- We can write the expected value of this quantity for a given test value x_0 as

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = E\left[\left(\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right] + \left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2 + \sigma_\varepsilon^2$$

$$\text{Test MSE} \quad = \quad \text{Method Variance} \quad + \quad \text{Method Bias}^2 \quad + \text{Irred. Error}^2$$

- **We can use the same formula and decomposition when checking validation data.**

High Level Steps

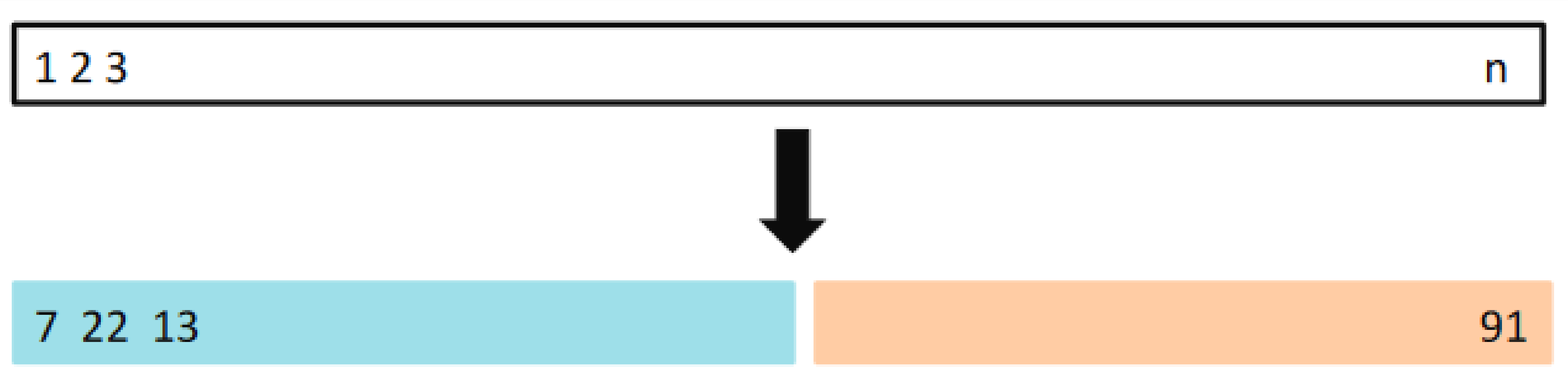
1. Draw a sample of data S, called the **training** data, from your **available** data D.
2. Fit the model on the training data sample S.
3. Check model fit on the remaining available data $D \setminus S$, called the **validation** data.
4. (Optional) Repeat 1-3 with other samples.
5. Use the model fits to learn about your model.

- How you draw samples and refit defines the resampling method: we look at *cross-validation* and *bootstrapping* approaches.

Cross-Validation

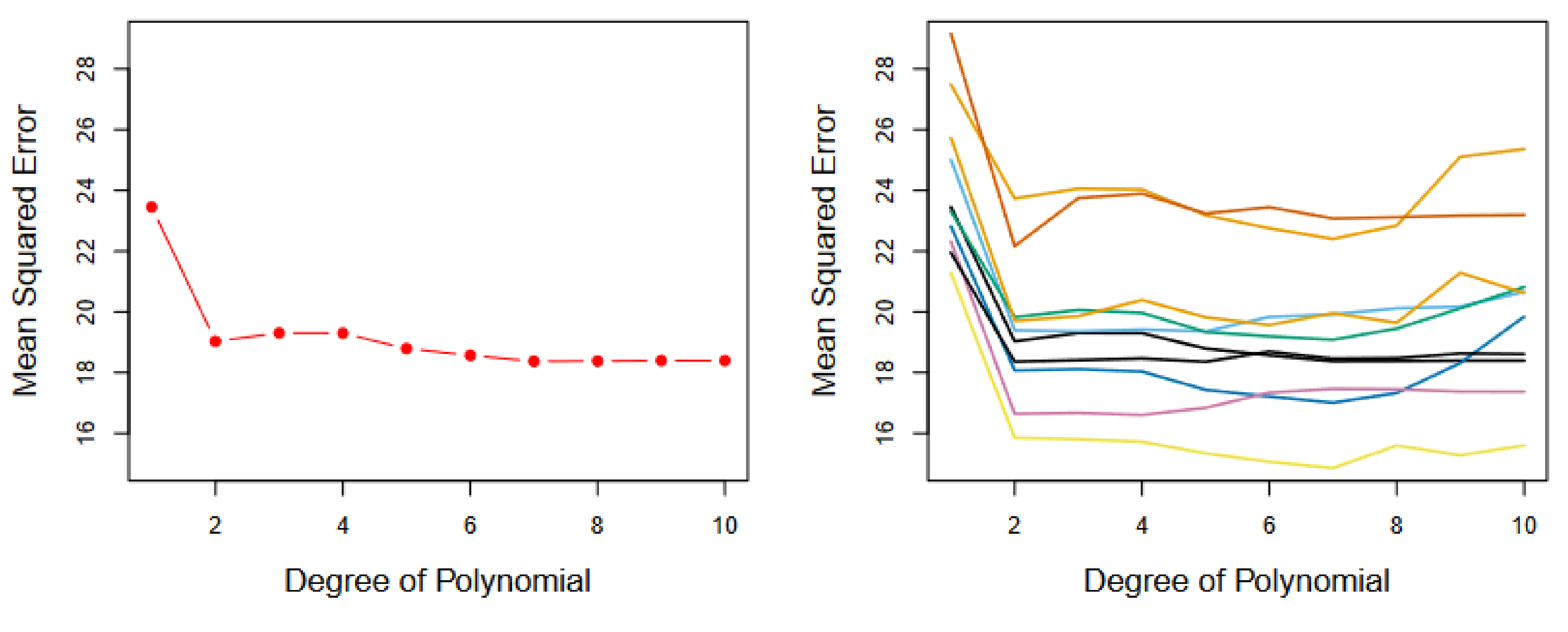
Validation Set - Definition

- **Randomly hold some of your training data back to use as the validation set data.**



- **Discussion**: what fraction of data should you use as training data?

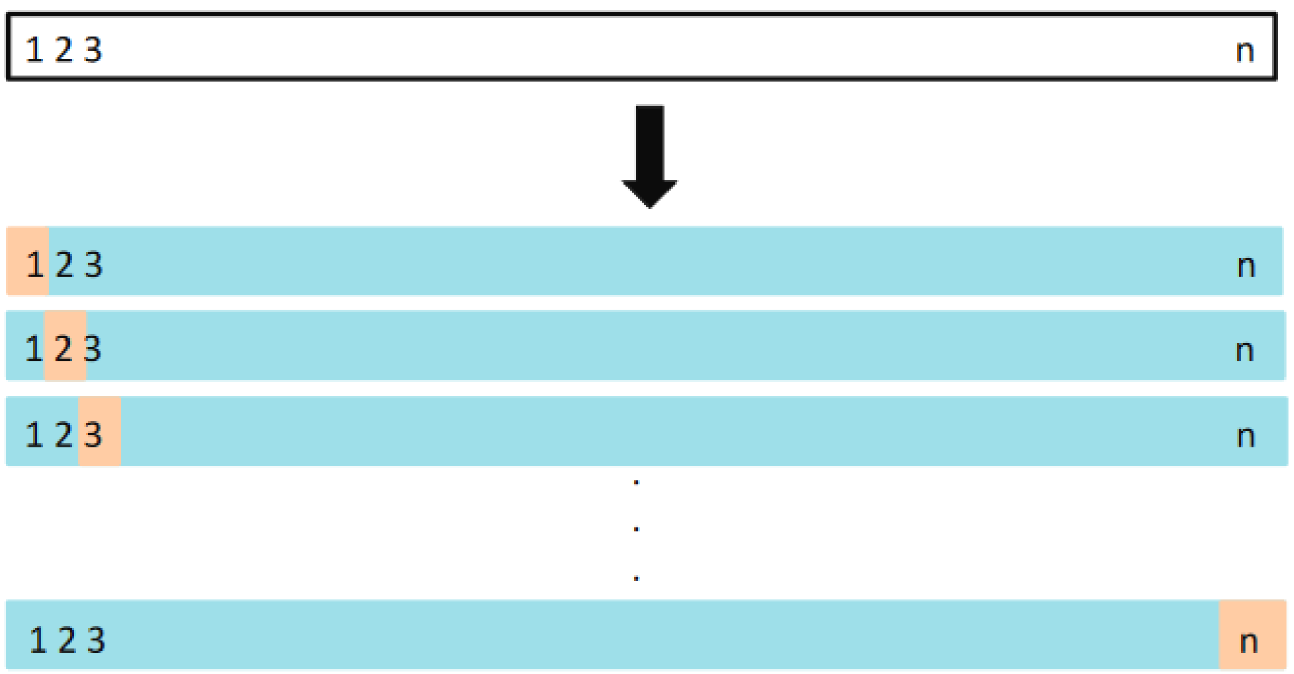
Validation Set - Example



- Test error can vary depending on what subset you use.
- Test error over-estimated due to smaller sample size.

Leave-One-Out Cross-Validation - Definition

- **Hold back one data point (x_i, y_i) to use as a validation set, and repeat for all i.**



- Define $\mathrm{MSE}_i = (y_i - \hat{y}_i)^2$. Then estimate overall MSE as:

$$\mathrm{CV}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i$$

Leave-One-Out Cross-Validation - MSE

- Define $\mathrm{MSE}_i = (y_i - \hat{y}_i)^2$. Then estimate overall MSE as:

$$\mathrm{CV}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i$$

- This may be slow!
- Recall from chapter 3. on linear regression we defined the *leverage statistic* of a data point x_i as

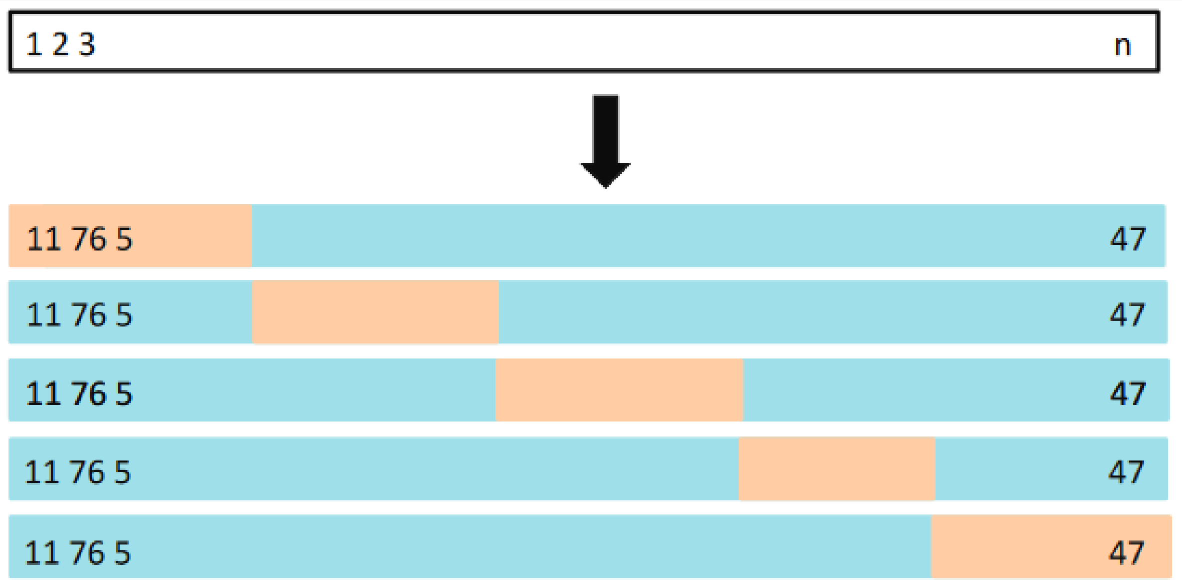
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}.$$

- Then *for linear regression*

$$\mathrm{CV}_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{y}_i}{1 - h_i} \right)^2.$$

k-Fold Cross-Validation - Definition

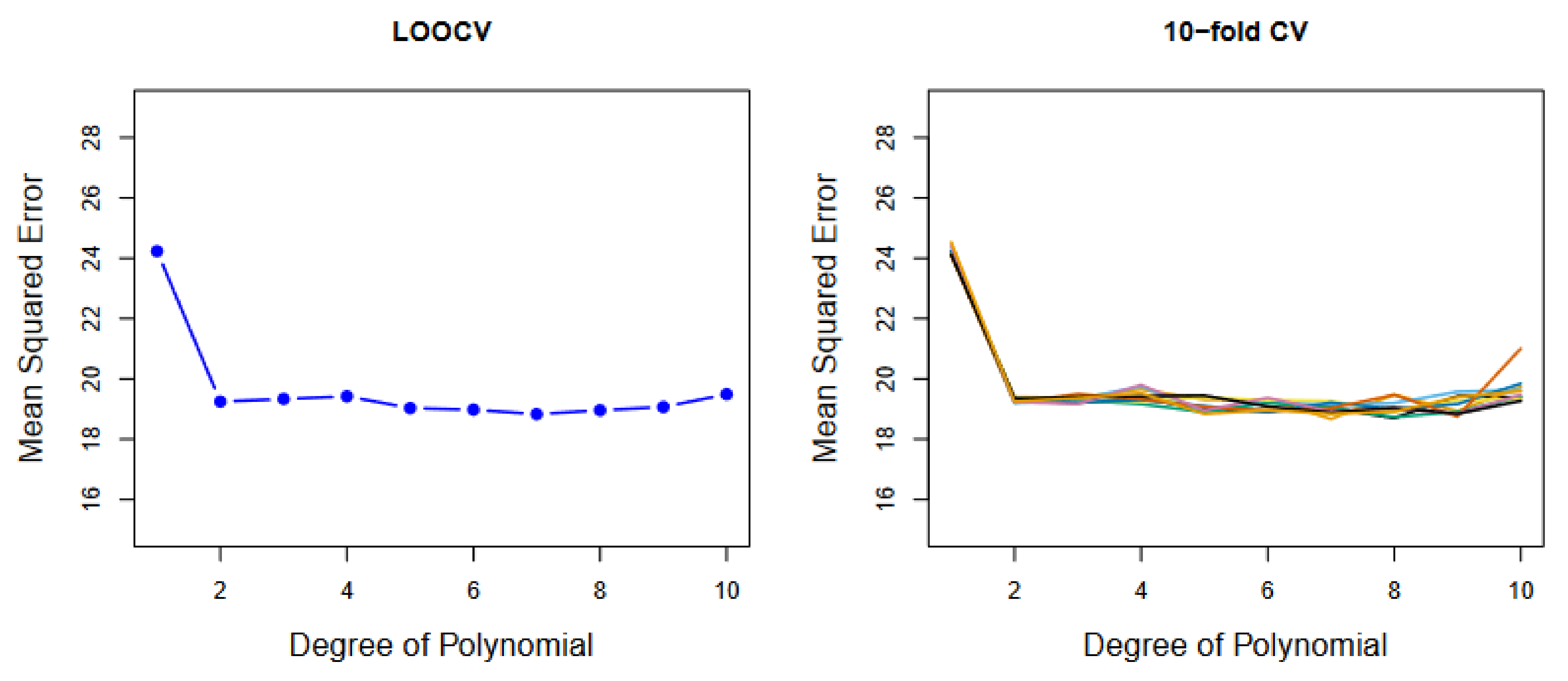
- **Randomly divide your data into k roughly equal "folds"; treat each as a validation set and fit on the remaining $k-1$ folds.**



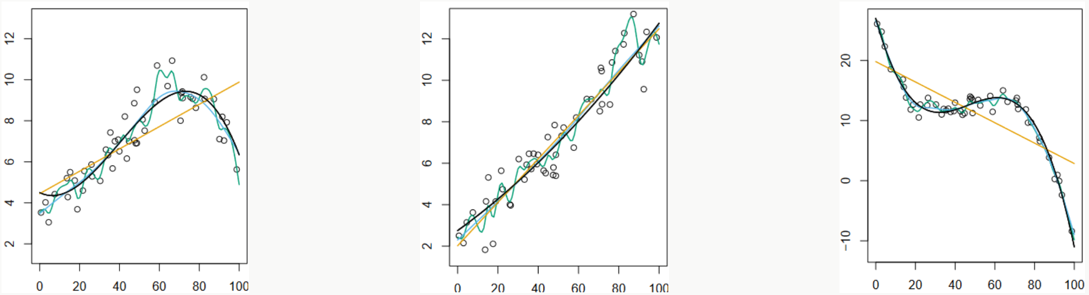
- Denoting the mean squared error from each fold as MSE_i, then

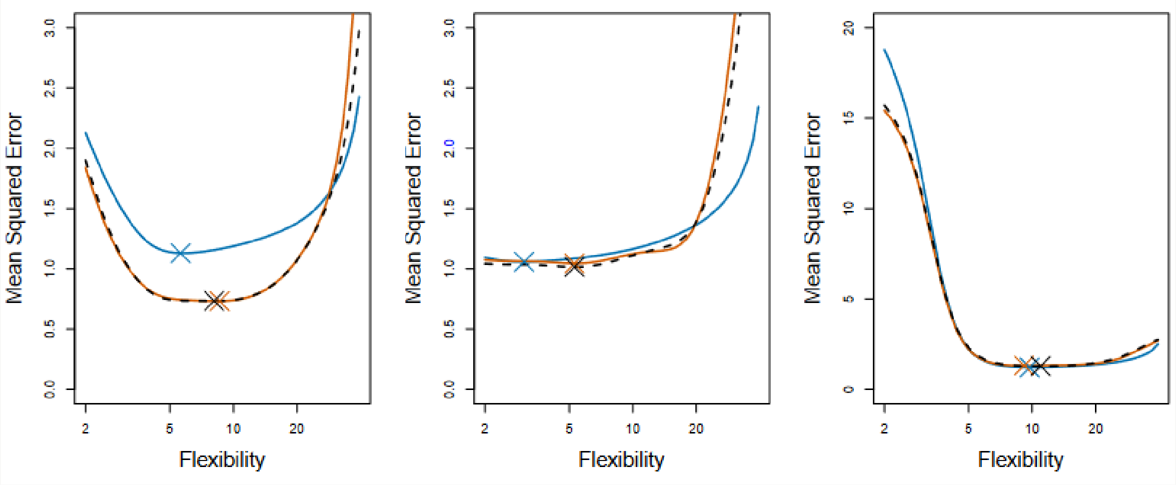
$$\mathrm{CV}_n = \frac{1}{n}\sum_{i=1}^{k} \mathrm{MSE}_i$$

Cross-Validation On Auto Data



Cross-Validation On Simulated Data





True = blue, LOOCV = black, 10-CV = orange

Cross-Validation: Bias Variance Trade Off

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = E\left[\left(\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right] + \left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2 + \sigma_\varepsilon^2$$

$$\text{Test MSE} \quad = \quad \text{Method Variance} \quad + \quad \text{Method Bias}^2 \quad + \text{Irred. Error}^2$$

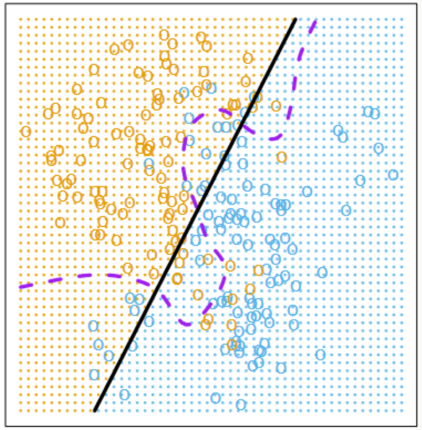
- LOOCV has less bias but higher variance (than k-fold CV).
- **Discussion**:
  - Why?
  - What is $\hat{f}$ here?
  - Do we agree with the books suggestion of using $k = 5$ or $k = 10$?

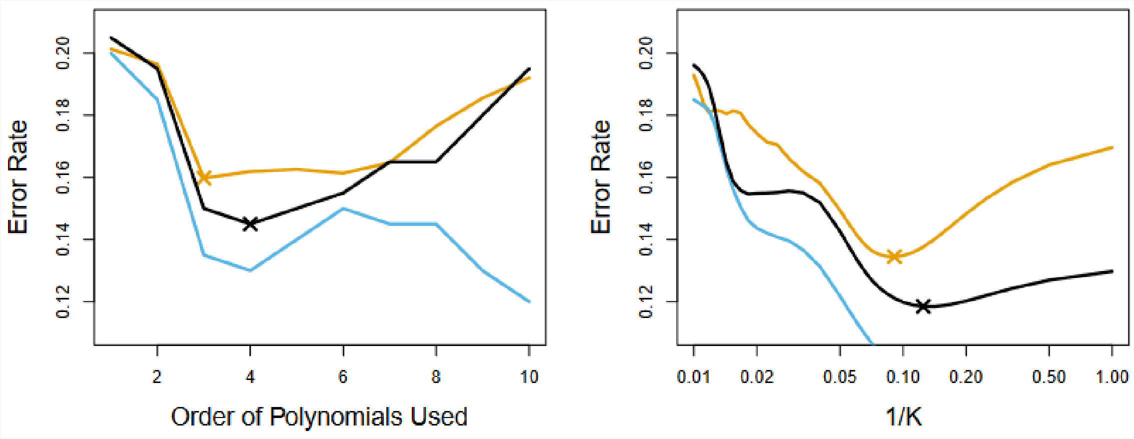
- We can perform cross-validation on qualitative problems too:

$$\mathrm{CV}_n = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i$$

where

$$\mathrm{Err}_i = I\left(y_i \neq \hat{y}_i\right)$$





Test = brown, training = blue, 10-fold CV error = black. LHS = Logistic Regression, RHS = KNN

Bootstrapping

Bootstrapping - Approach

- **Repeatedly randomly sample your data.**

- On a data set Z:
  1. From Z randomly select n observations **with replacement**, Z^{*1}
  2. For a quantity of interest α get an estimate $\hat{\alpha}^{*1}$ using Z^{*1}
  3. Repeat steps 1-2 B times to get multiple estimates of α: $\hat{\alpha}^{*1}$, ..., $\hat{\alpha}^{*B}$
  4. Use these to estimate the standard error of estimates of α:

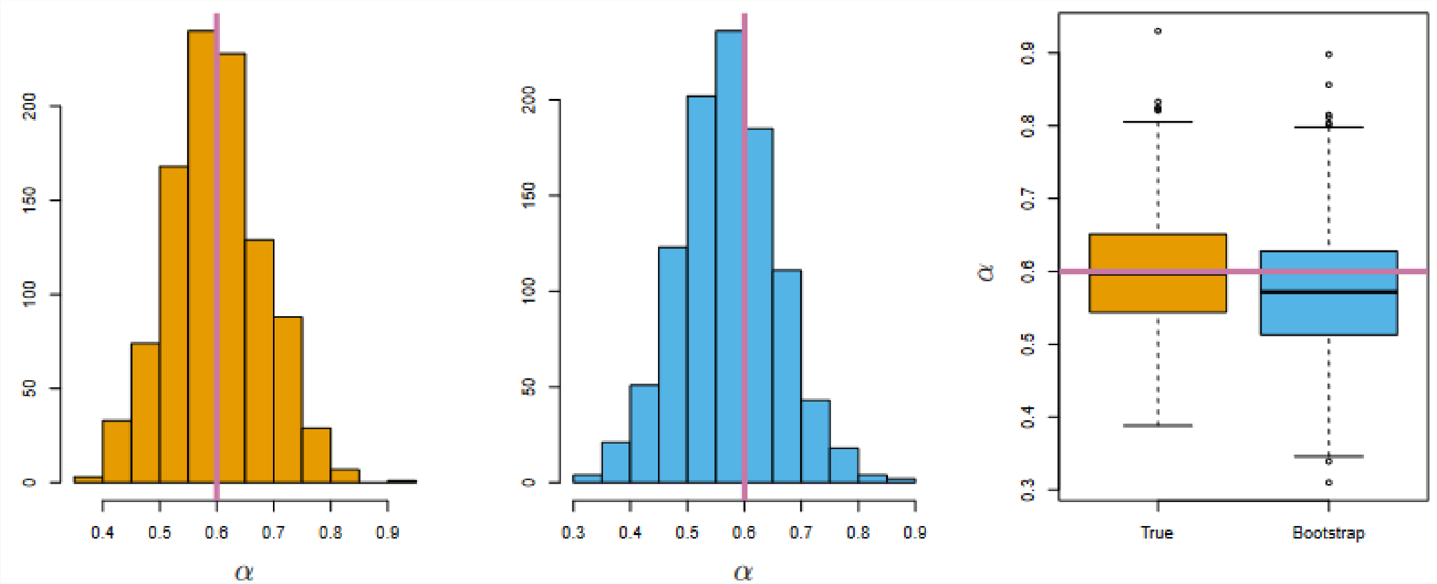
$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^{B} \left(\hat{\alpha}^{*r} - \frac{1}{B} \sum_{s=1}^{B} (\hat{\alpha}^{*s})\right)}$$

Bootstrapping - Example

- For an investment portfolio of two assets with returns X and Y, calculate the allocations α and $1-\alpha$ that minimises the portfolio variance.
- *True* and *estimate* values of alpha can be calculated as:

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

- For given example σ_Y, σ_{XY} and σ_X by repeatedly simulating $n = 100$ data points $B = 1000$ times we get $\mathrm{SE}_B(\hat{\alpha}) = 0.08$.



Summary

Comparison

Method	**Validation Set CV**	**LOOCV**	**K-fold CV**	**Bootstrap**
Description	Randomly split	Leave one out	Leave fraction out	Randomly sample
Quick?	☑	✗	☑	✗
Non-Random?	✗	☑	✗	✗
Low-Bias?	✗	☑	✗	☑
Low-Variance?	✗	✗	☑	☑

- **Discussion**: What do people recommend?