15.2) Cross-Validation

Vitor Kamada

January 2020

Reference

Tables, Graphics, and Figures from:

1) An Introduction to Statistical Learning

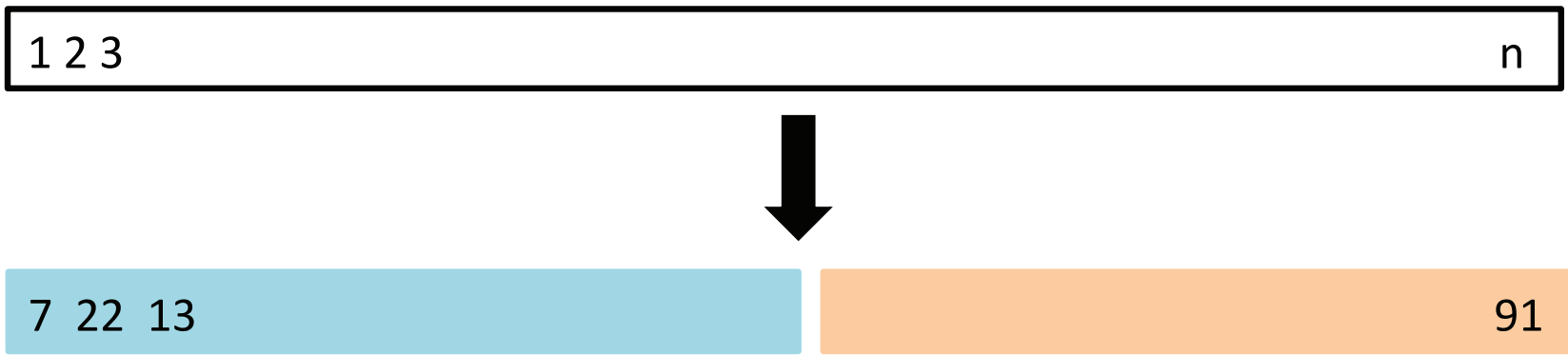
James et al. (2017): Chapters: 5.1

2) The Elements of Statistical Learning

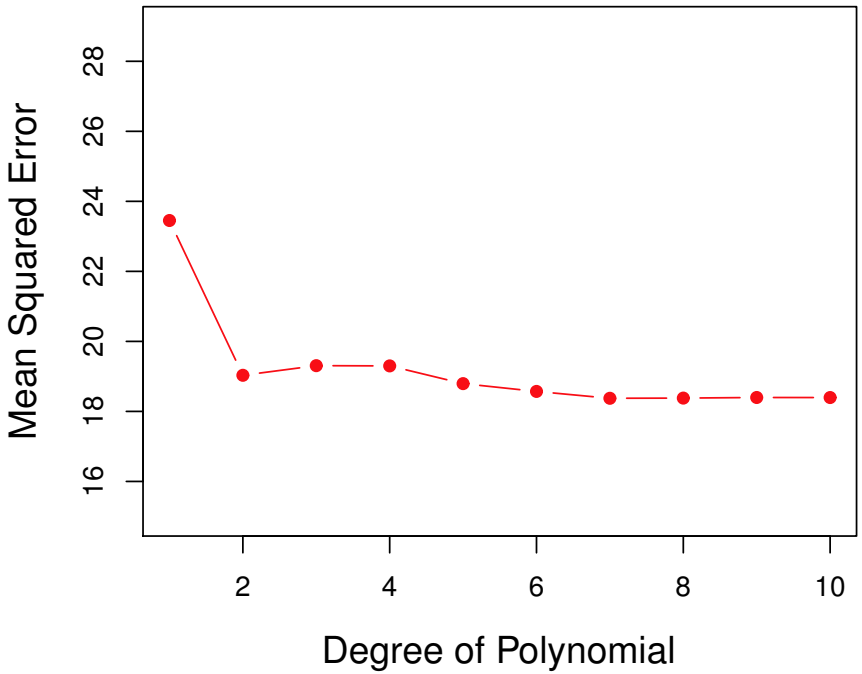
Hastie et al. (2017): Chapter: 7.10

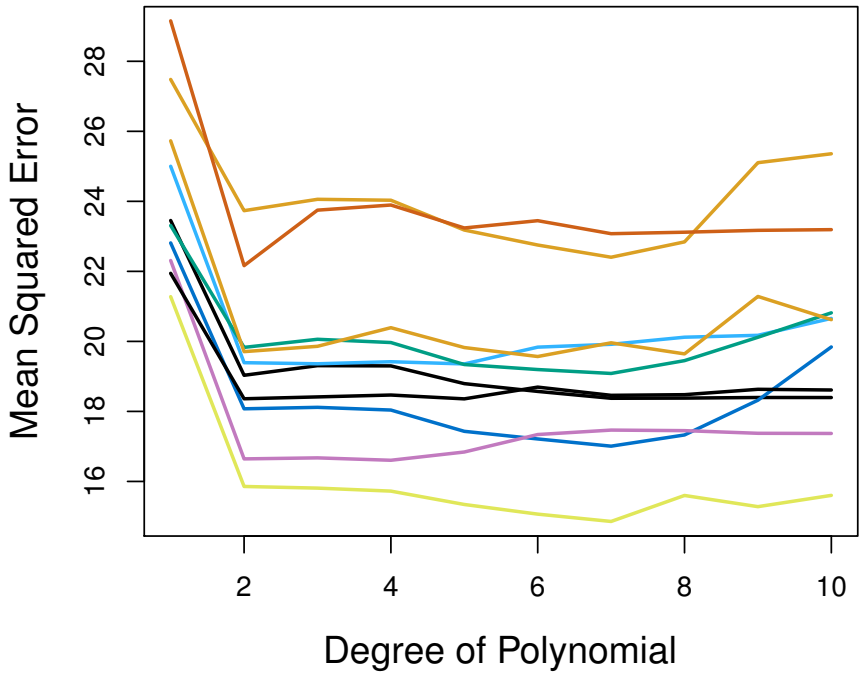
Training Set vs Validation or Hold-out Set

Randomly division in two part:

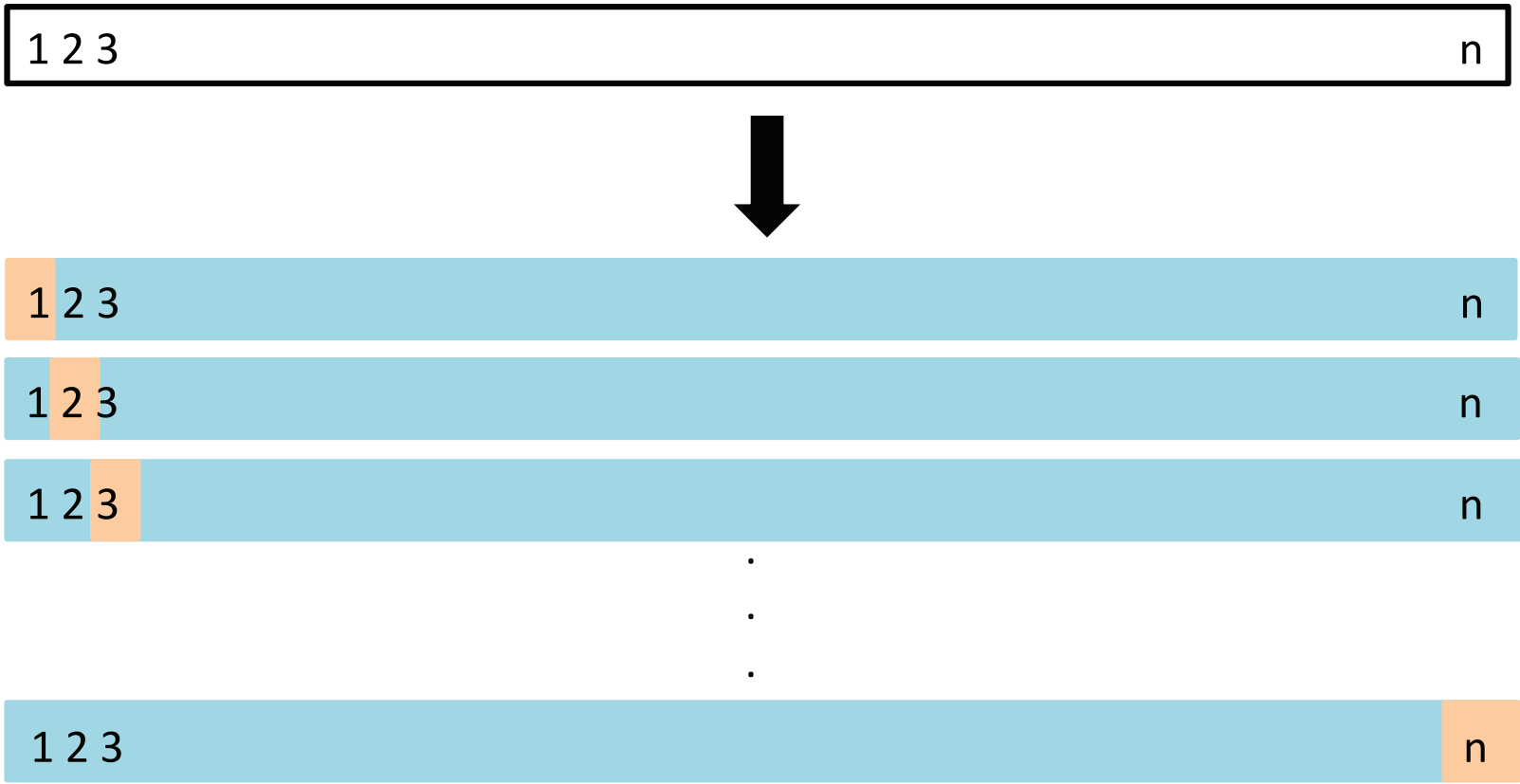


Random Split (10x)





Leave-One-Out Cross-Validation (LOOCV)



LOOCV

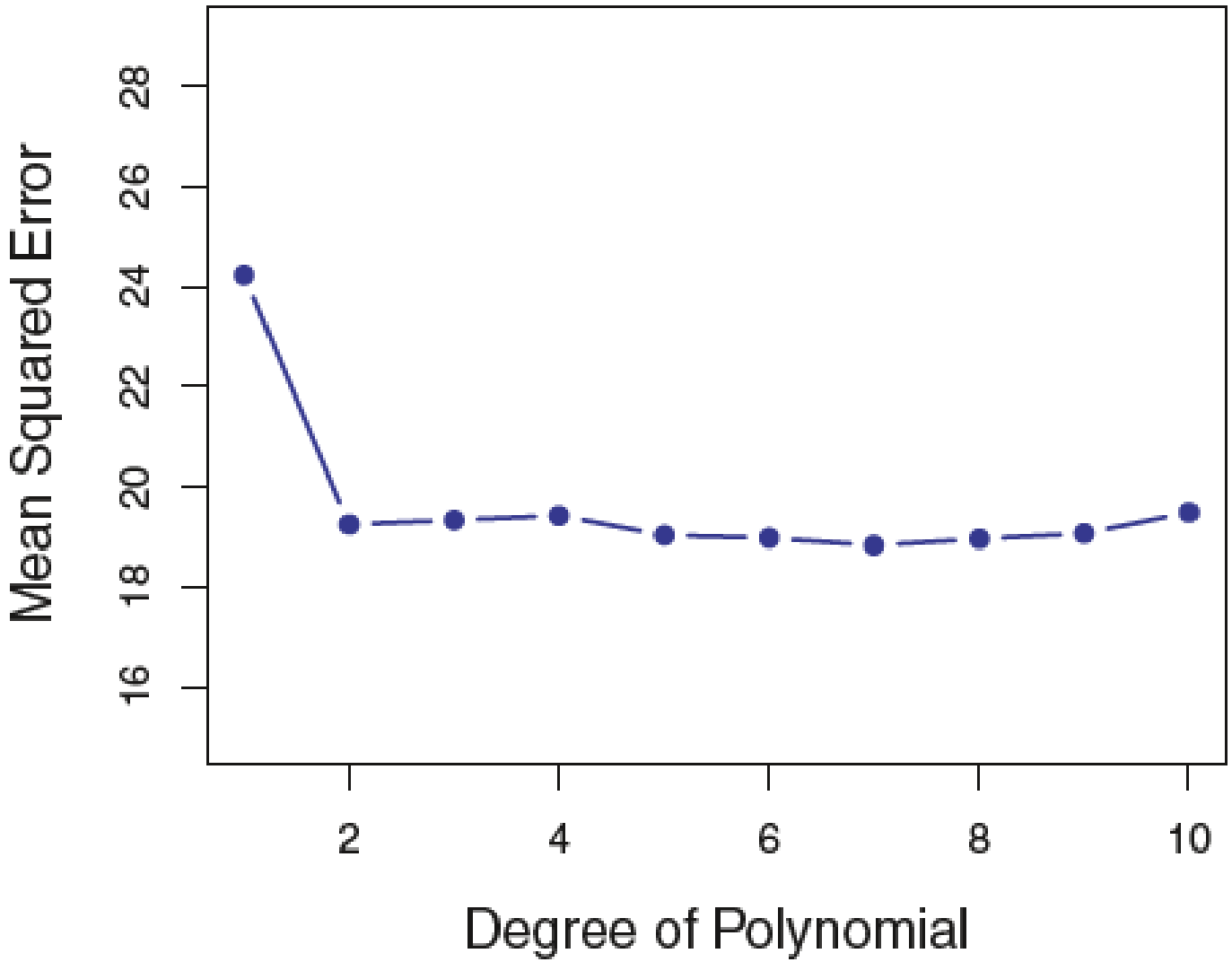
$$\{(x_2, y_2), \ldots, (x_n, y_n)\} \rightarrow \text{Training Set}$$

$$(x_1, y_1) \rightarrow \text{Validation Set}$$

$$MSE_1 = (y_1 - \hat{y}_1)^2$$

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

LOOCV: mpg on hp ...



Generalized Cross-Validation (GCV)

$$\hat{y} = Sy$$

$$\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}^{-i}(x_i)]^2 = \frac{1}{N}\sum_{i=1}^{N}[\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}]^2$$

$$GCV(\hat{f}) \cong \frac{1}{N}\sum_{i=1}^{N}[\frac{y_i - \hat{f}(x_i)}{1 - \frac{trace(S)}{N}}]^2$$

$trace(S)$: Effective # of Parameters

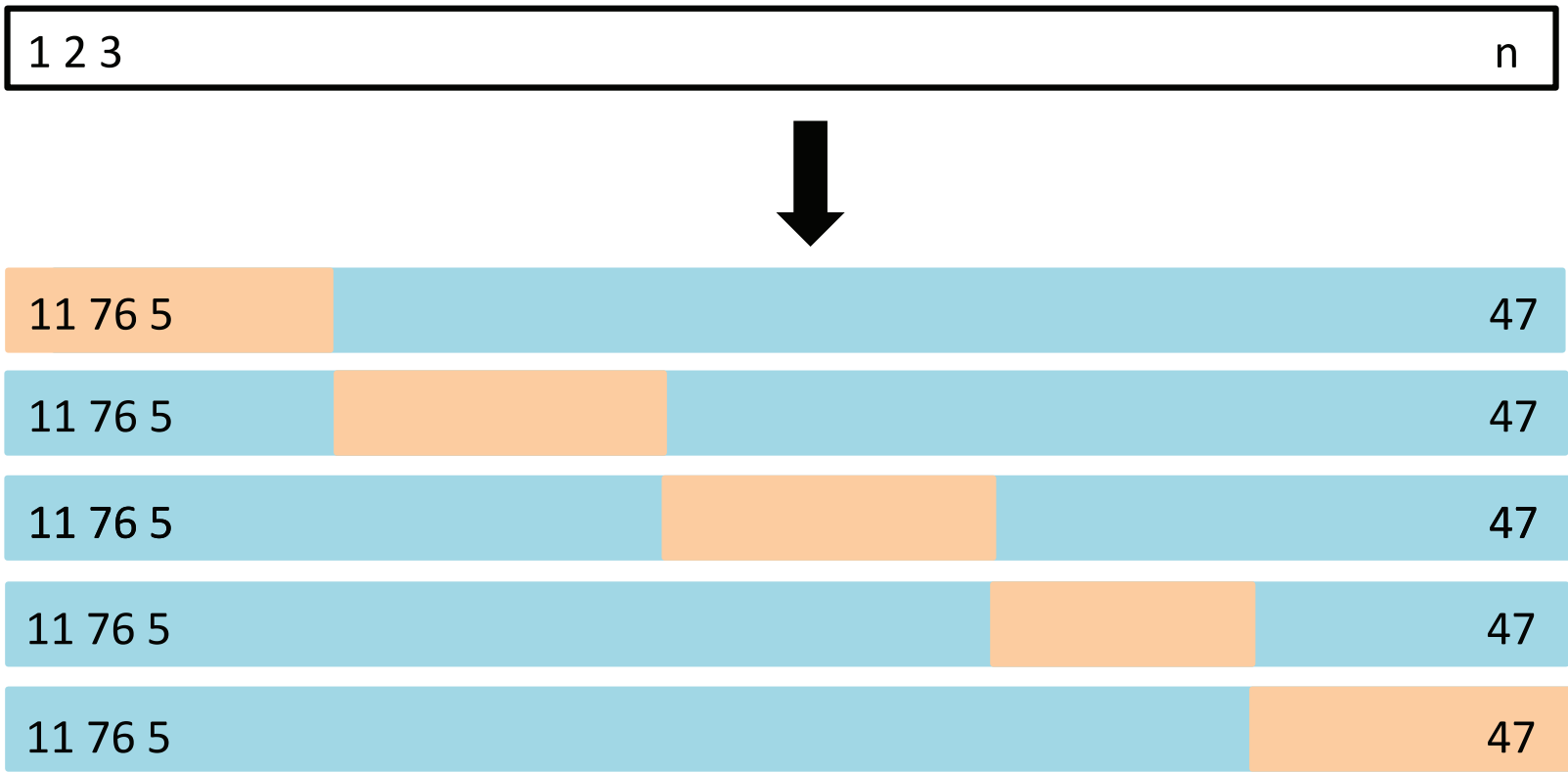
LOOCV for OLS

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

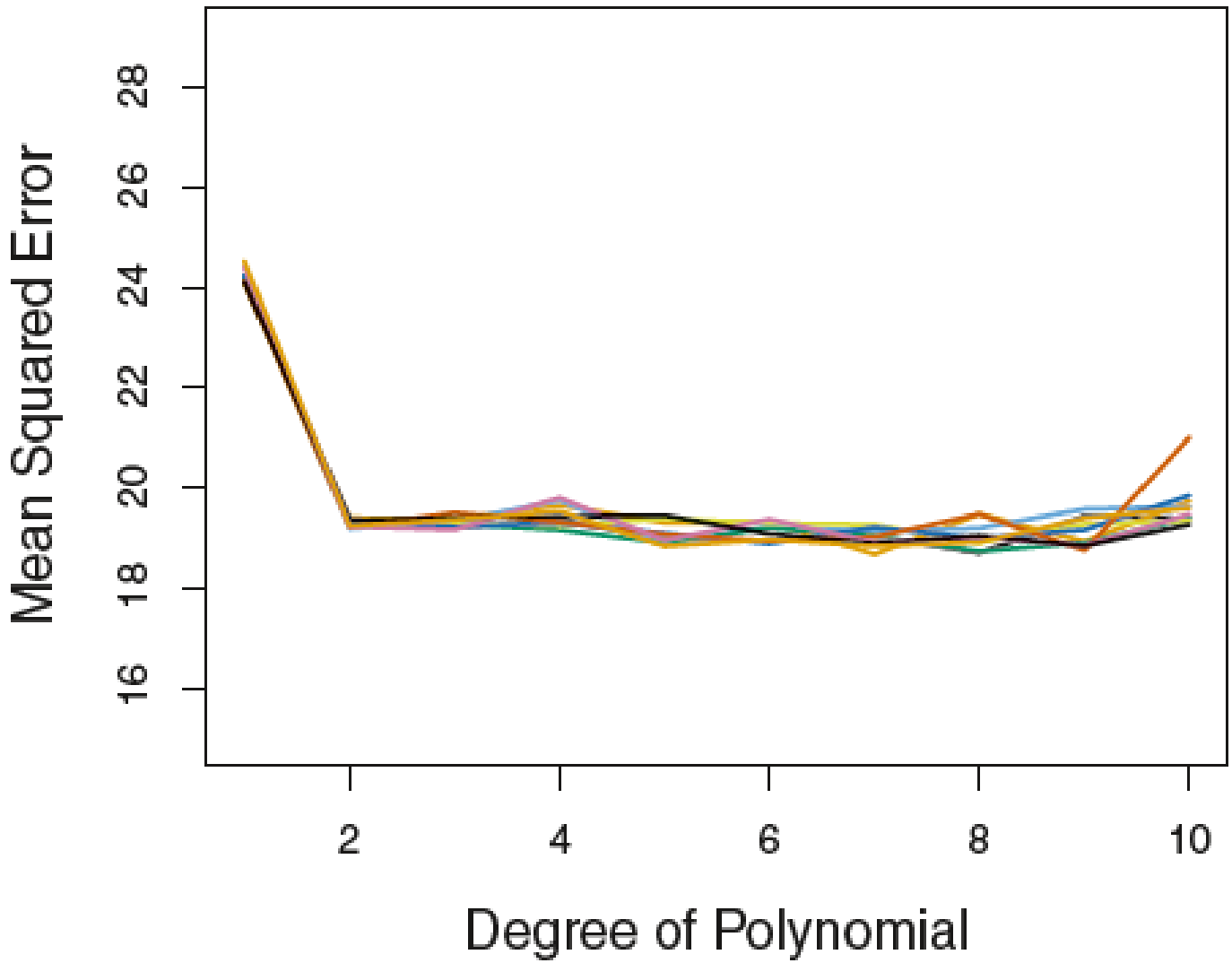
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

k-Fold Cross-Validation

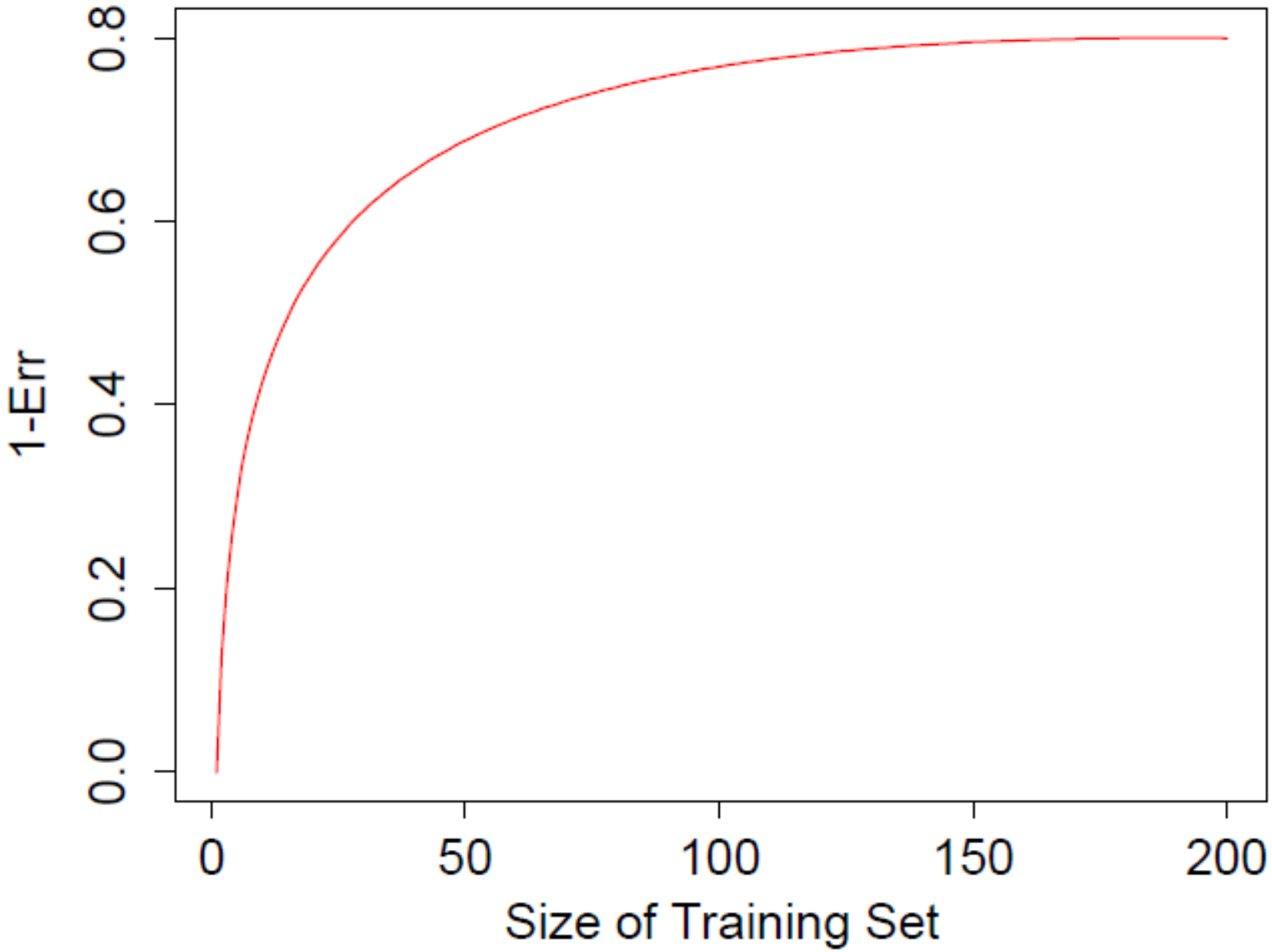
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$



10-fold CV: mpg on hp ...



Hypothetical Learning Curve for a Classifier



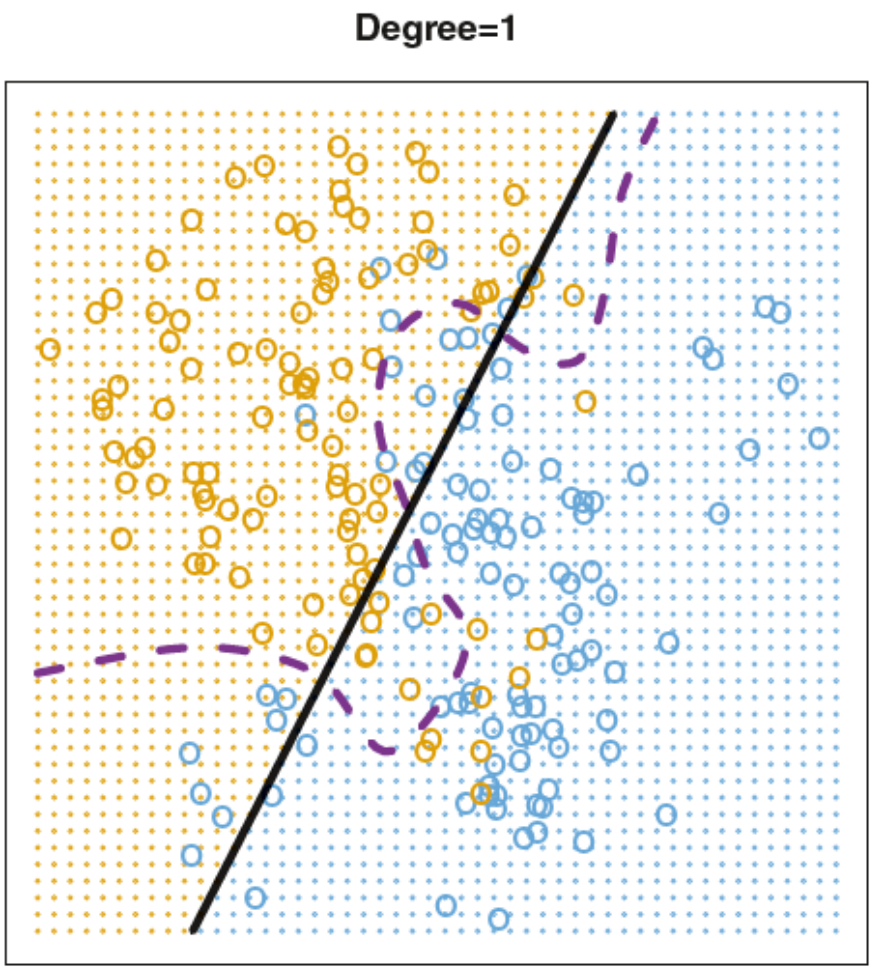
Cross-Validation on Classification Problems

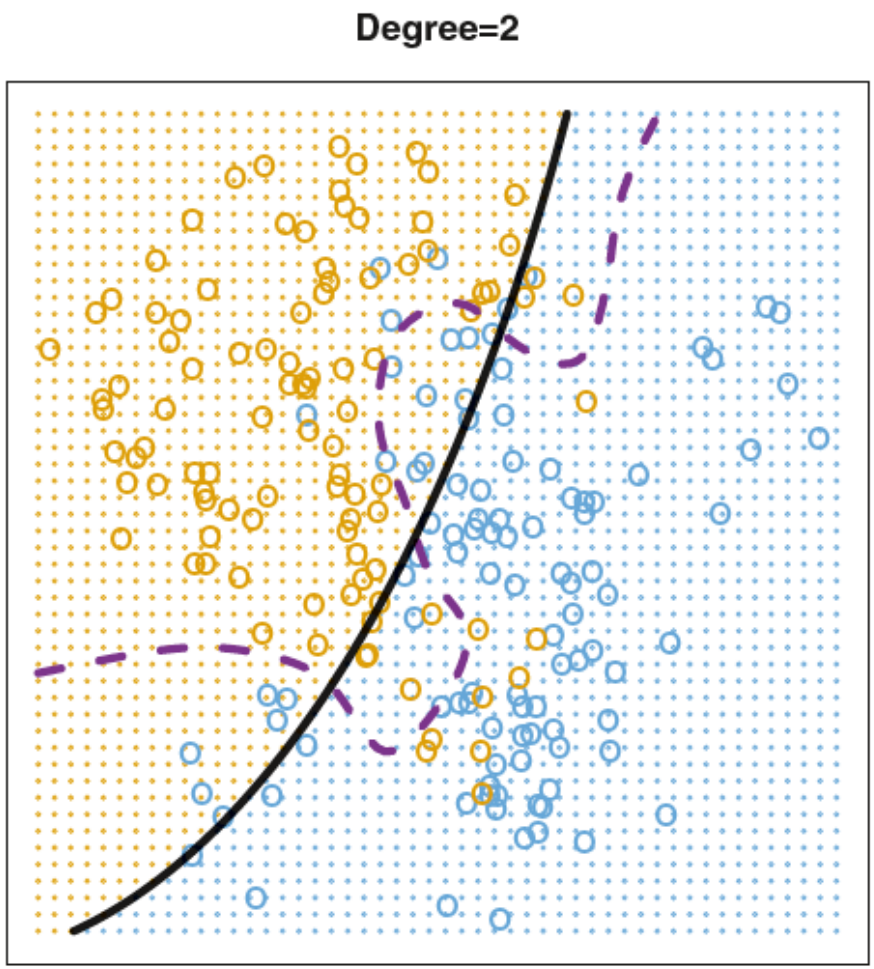
$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

$$log(\frac{p}{1-p}) =$$
$$\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

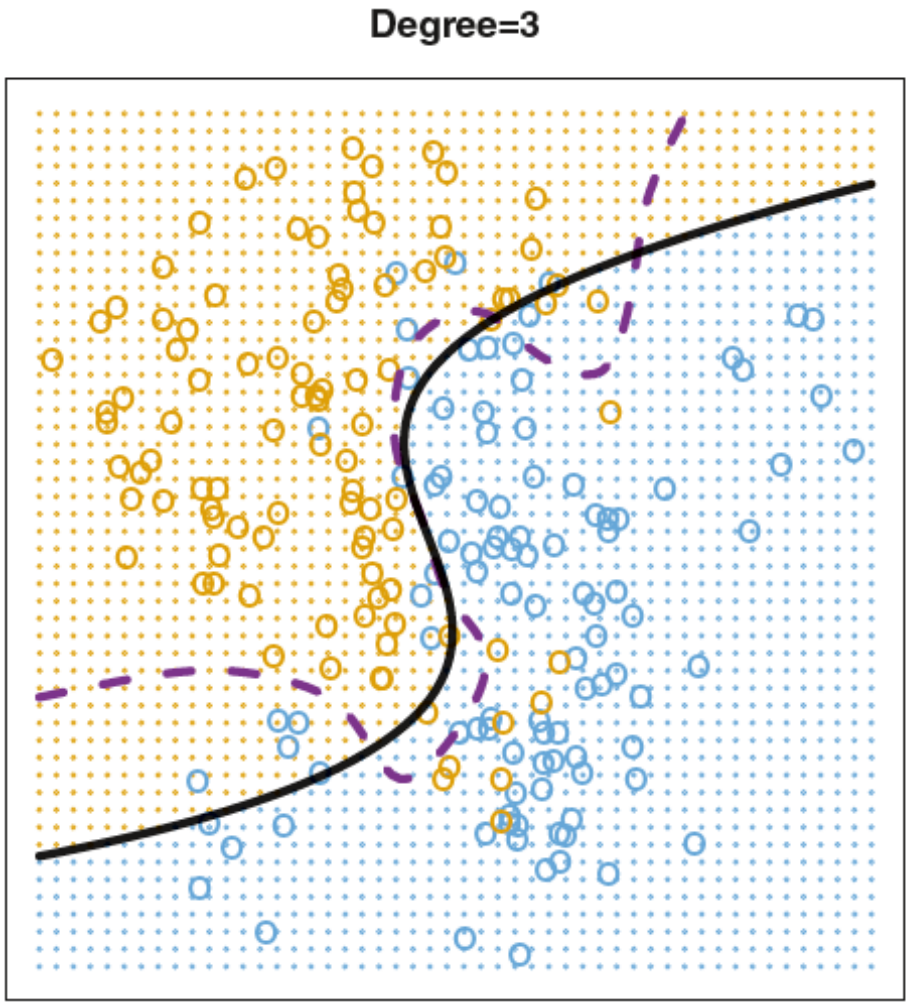
Test Error Rates: 20.1% and 19.7%

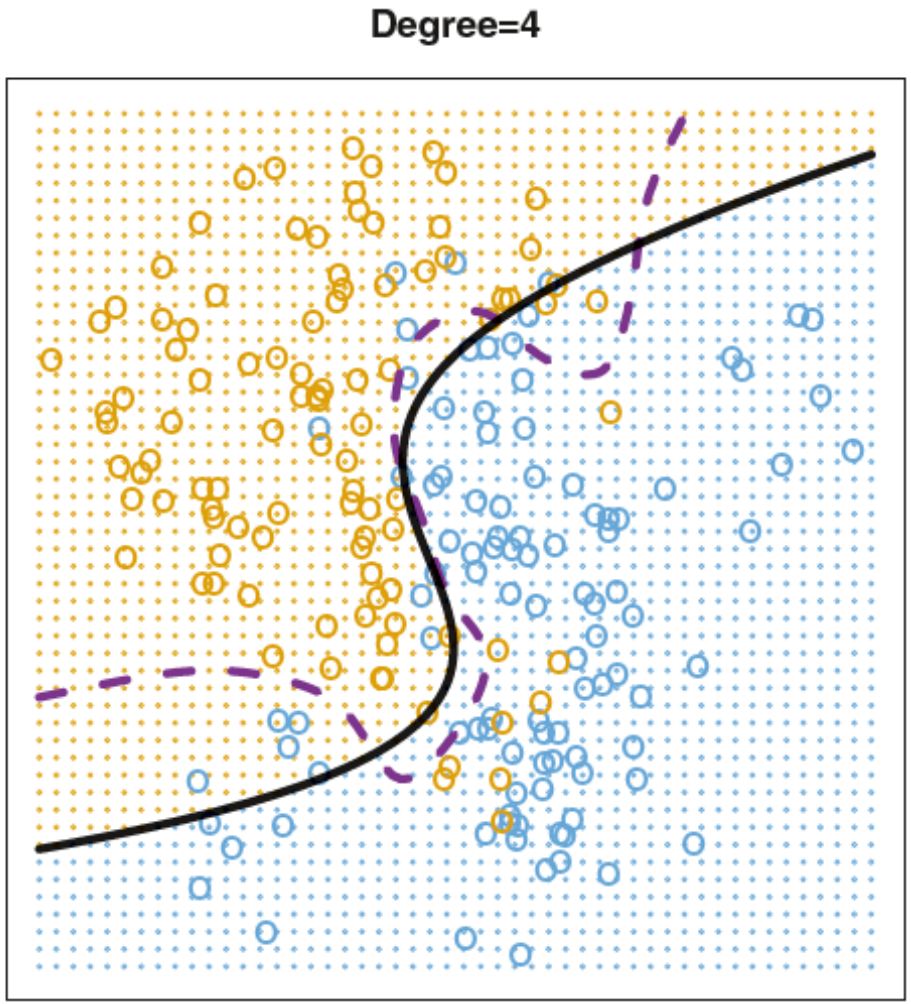
Bayes Error Rate: 13.3%





Test Error Rates: 16% and 16.2%





Test (brown), Training (blue), and 10-fold CV Error (black)