

# Methodenwerkstatt Statistik

## Introduction to Python



Hochschule für  
Wirtschaft und Recht Berlin  
Berlin School of Economics and Law

Prof. Dr. Natalie Packham  
Berlin School of Economics and Law  
Summer Term 2023

### Table of Contents

- ▼ ▼ [5 Model validation and measures of fit](#)
  - [5.1 Mean-square error and overfitting](#)
  - [5.2 The classification setting](#)
  - [5.3 Choosing the optimal model](#)
  - [5.4 Cross-validation](#)

## 5 Model validation and measures of fit

- In Data Science, there is no one single statistical method that performs *best* across all data sets.
- It is an important -- and at times difficult -- task to select the appropriate method or model for a given data set.
- We therefore study a number of measures to assess the quality of fit, which in turn allows to compare methods and models.
- For a more in-depth treatment, see Chapters 2.2, 5.1 and 6.1.3 of

James, Witten, Hastie, Tibshirani: An Introduction to Statistical Learning. Springer, 2013.

### 5.1 Mean-square error and overfitting

- A commonly used measure for assessing how well predictions match observed data is the **mean squared error (MSE)**, which you know e.g. from Ordinary Least Squares (OLS) in linear regression:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2,$$

where  $\hat{f}(x_i)$  is the prediction that the fitted method  $\hat{f}$  gives for the  $i$ -th observation.

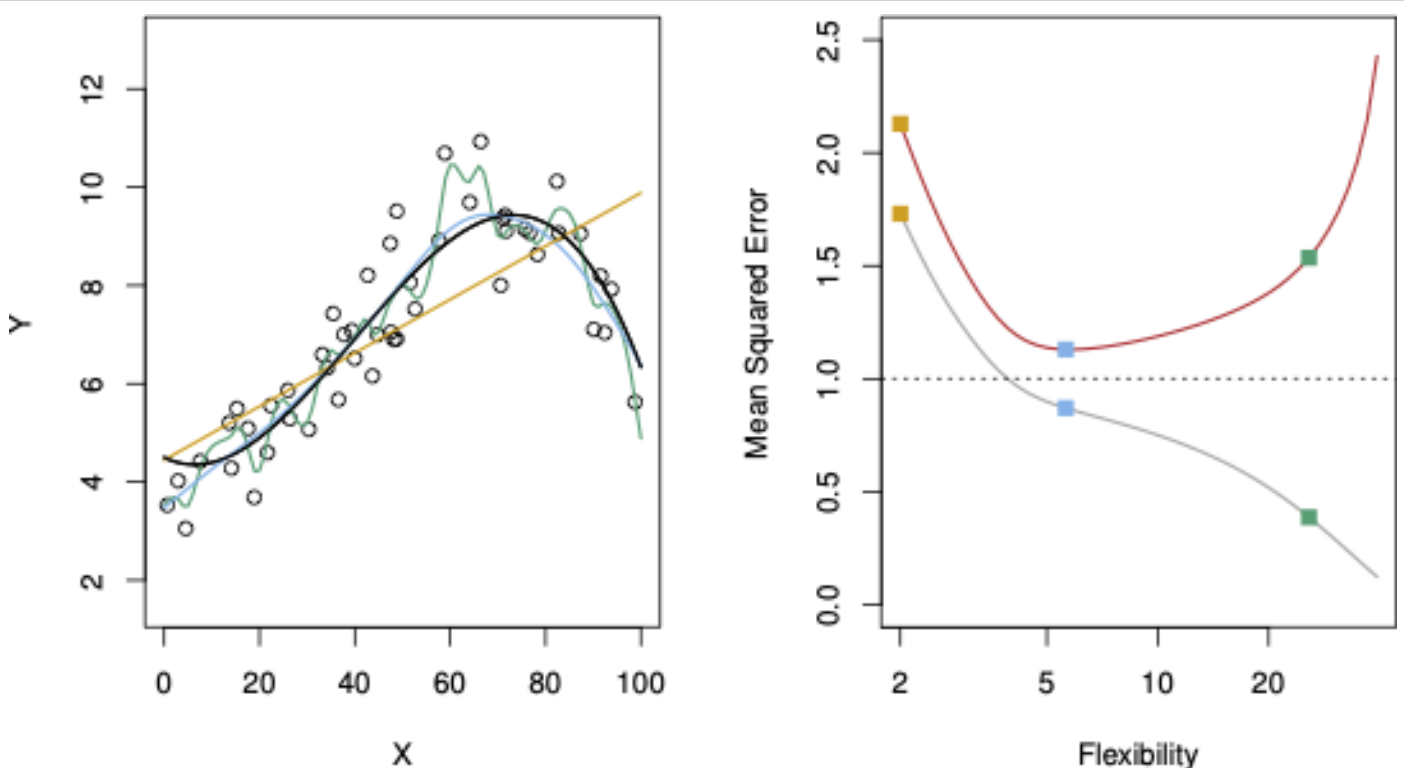
## Training and tests MSE

- In Linear Regression (a statistics method), the whole data set is used for finding a linear function  $\hat{f}$  that minimises the MSE.
- In Data Science (i.e., statistical learning, machine learning, artificial intelligence), it is common to split the data set into a **training data set** and a **test data set**.
- This reflects that we do not really care how well a method works on the training data, but rather we are interested in the accuracy of the prediction when applying the method to previously unseen data (the test data).
- In other words, first we fit the training data  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  to obtain the estimate  $\hat{f}$ , e.g. by minimising the MSE on the training data.
- Then we calculate the MSE on the test data, which are data points  $\{(x_0, y_0)\}$  that were not used to train the method.
- We want to choose the method that gives the lowest *test MSE*.

## Remarks

- Sometimes an additional **validation data set** is added to allow for tuning parameters such as the number of hidden units in a neural network.
- In econometrics, instead of the terminology training and test data set, we speak of **in-sample** and **out-of-sample testing**.

## Training and test MSE



Source: James et al.: An Introduction to Statistical Learning. Springer, 2013.

Left:

- **Test data**, simulated from  $f$  (black smooth line) are shown as black dots.

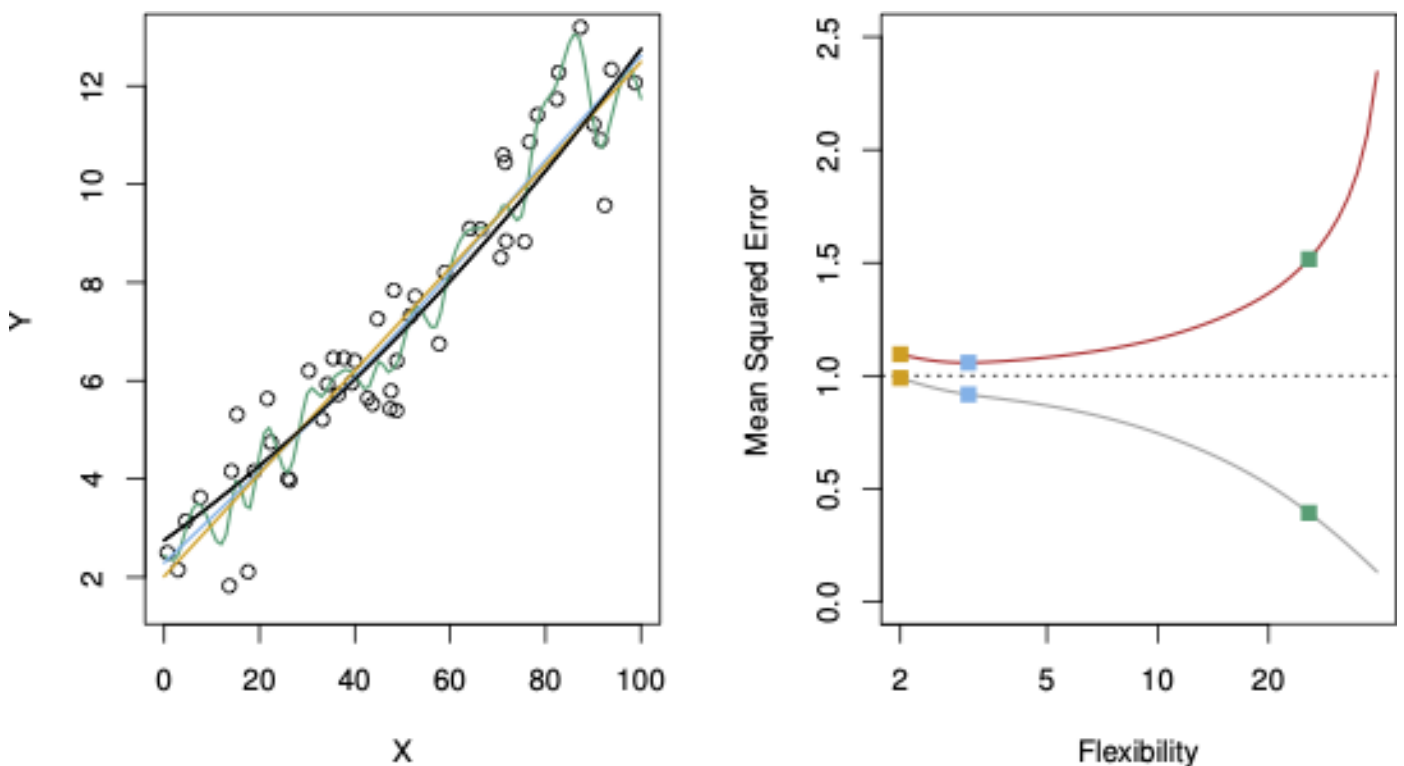
- Estimates of  $f$ :
  - linear regression line (orange)
  - smoothing spline I (blue)
  - smoothing spline II (green)

Right:

- Training MSE (grey)
- Test MSE (red)
- Flexibility denotes the complexity of the models (e.g. number of parameters)

## Training and test MSE

- The green smoothing spline will have the smallest MSE on the training data set.
- It does not perform well, however, when **extrapolating** to the test data set.
- This effect is called **overfitting**.
- **Overfitting** refers to choosing a model that fails to capture the *general* effects due to a too many *degrees of freedom*.
- In other words, a less flexible model performs better on the test data than a more flexible model.
- A second example is shown on the next slide.



Source: James et al.: An Introduction to Statistical Learning. Springer, 2013.

## 5.2 The classification setting

- The measure MSE applies in a regression setting.
- In a classification setting, we seek to estimate  $f$  on the basis of training observations  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ , where  $y_1, \dots, y_n$  are qualitative.
- Here, the training **error rate**, which denotes the proportion of mistakes when applying the  $\hat{f}$  to the training observations is a measure of **accuracy**:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{y_i \neq \hat{y}_i},$$

where  $\mathbf{1}_A$  is an *indicator function* taking value 1 if  $A$  is true and 0 otherwise.

- The **test error rate** is giving as the error rate from applying  $\hat{f}$  to the test data set.

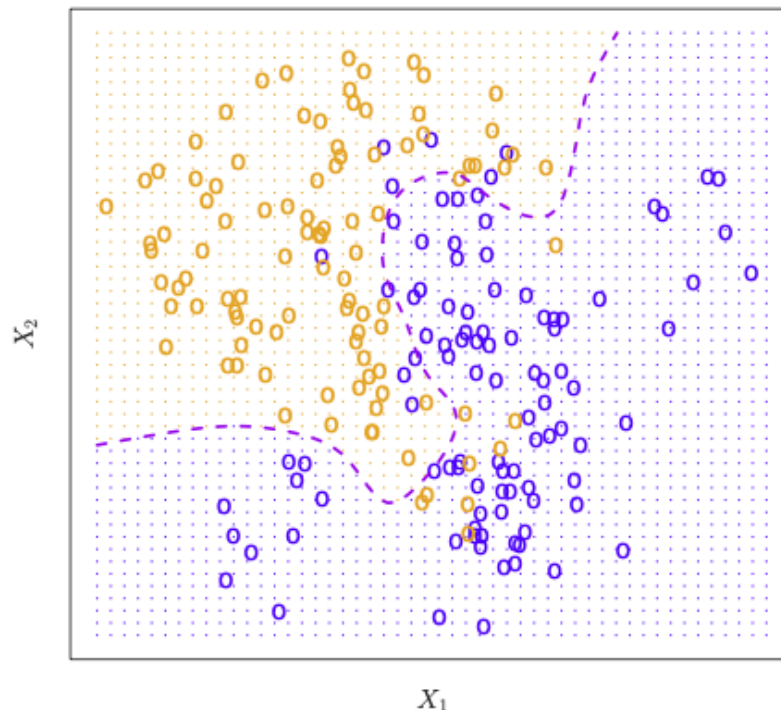
## The Bayes classifier

- The test error rate is minimised, on average, by the **Bayes classifier** which assigns each observation to the most likely class given its predictor value, i.e., for an observation  $x_k$ , and classes  $1, \dots, j$ , it determines the conditional probabilities

$$\mathbb{P}(Y = 1|X = x_k), \dots, \mathbb{P}(Y = J|X = x_k)$$

and assigns the class, for which the conditional probability is highest.

## The Bayes classifier



Source: James et al.: An Introduction to Statistical Learning. Springer, 2013.

- Two-dimensional predictors  $X_1$  and  $X_2$  and two classes (blue, orange).
- Dashed line is the Bayes decision boundary.

## 5.3 Choosing the optimal model

### AIC, BIC, Adjusted $R^2$

- If the optimal model is chosen based on the training data set, then measures such as MSE or  $R^2$  might be misleading.

- This is the case for example in a standard regression setting, where we do not differentiate between training and test data sets.
- In a regression setting, the  $R^2$  will always improve if more independent variables are added.
- There are a number of techniques for *adjusting* the training error according to model size (e.g. different number of independent variables).
- These can be used to select amongst models of different size.
- Typical measures are:
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - adjusted  $R^2$

## AIC, BIC, Adjusted $R^2$

- Consider the problem of finding the appropriate predictors (independent variables) in a regression model.
- Should you include all variables? Or just a subset of the variables?
- Define the *residual sum of squares (RSS)* as

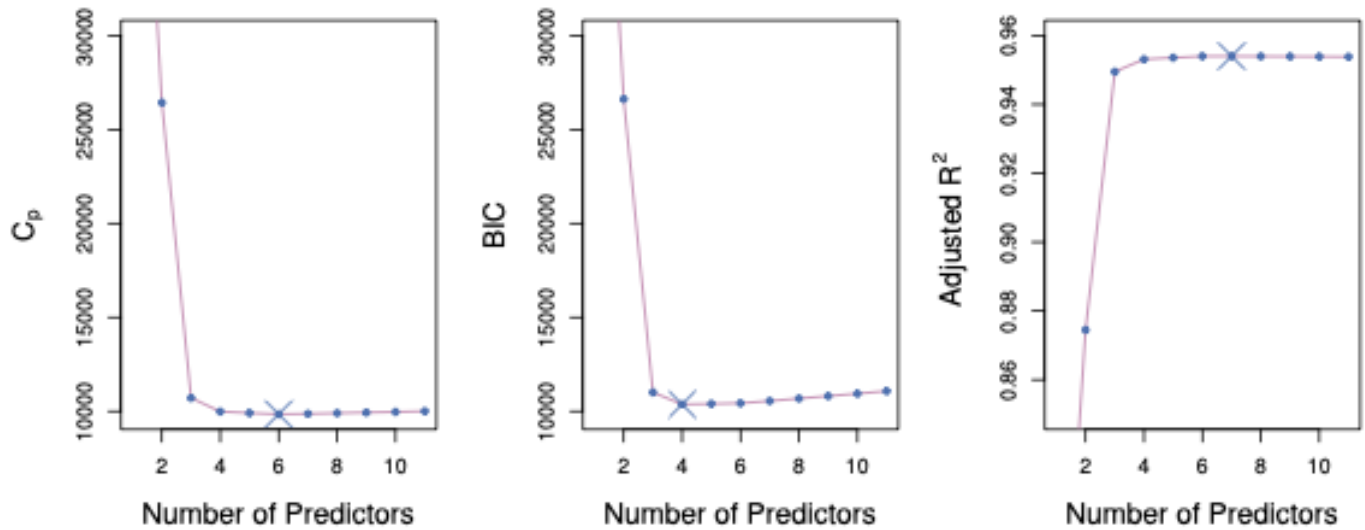
$$\text{RSS} = \epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_n^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

where  $\hat{\beta}_i$  and  $x_k$  may be vectors.

## AIC, BIC, Adjusted $R^2$

- Then, the measures above are defined as
  - $\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2)$ ;
  - $\text{BIC} = \frac{1}{n}(\text{RSS} + \ln(n)d\hat{\sigma}^2)$ ;
  - $\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n - d - 1)}{\text{TSS}/(n - 1)}$ , with  $\text{TSS} = \sum (y_i - \bar{y})^2$  the total sum of squares of the response.
- A further measure, which in OLS regression yields an equivalent model choice as AIC is:
  - $C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$ , with  $\hat{\sigma}^2$  an estimate of the error variance and  $d$  the dimension of  $x_k$ .
- All measures have in common that they place a penalty on a more complex model, measured by the number of explanatory variables  $d$ .
- Each measure has a theoretical justification; this is beyond the scope of the course, however.

## AIC, BIC, Adjusted $R^2$



Source: James et al.: An Introduction to Statistical Learning. Springer, 2013.

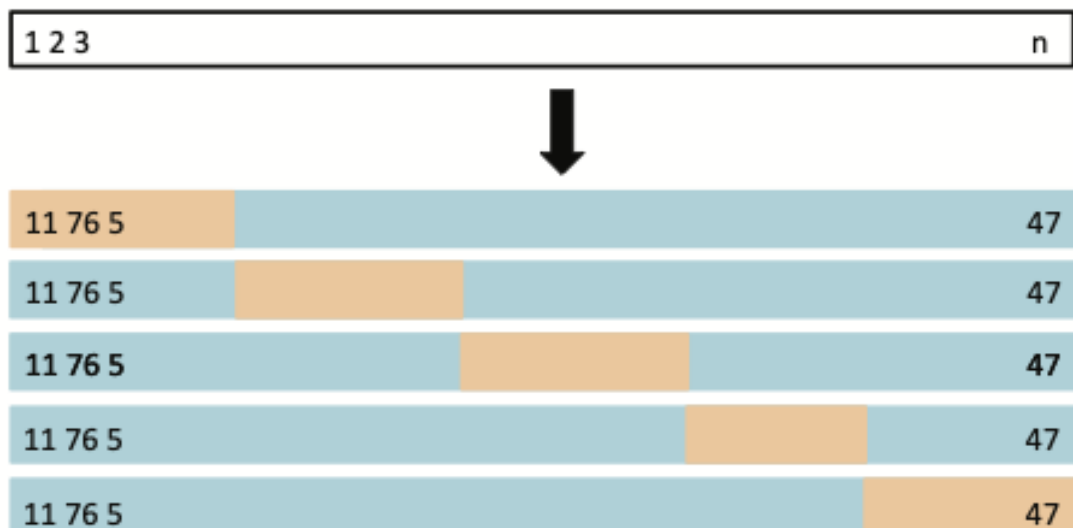
- Estimates of  $C_p$  (proportional to AIC), BIC and Adjusted  $R^2$  for a data set of credit card defaults with predictors such as age, income, marital status, etc.
- A lower  $C_p$  and BIC indicate a superior model; likewise a higher Adjusted  $R^2$ .

## 5.4 Cross-validation

- **Cross validation (CV)** refers to several methods of building the test and training data sets.
- In  $k$ -fold CV, the data set is randomly divided in  $k$  groups or *folds* of approximately equal size.
- In  $k$  iterations, each first fold is treated as the test or validation data set, while the  $k - 1$  other folds are taken as the training data.
- In this way,  $k$  MSE's of the test error are estimated and the  $k$ -fold CV estimate is given by

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

### Cross-validation



Source: James et al.: An Introduction to Statistical Learning. Springer, 2013.

- A schematic display of 5-fold CV.
- A set of  $n$  observations is randomly split into five non-overlapping groups.
- Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). \* The test error is estimated by averaging the five resulting MSE estimates.