Programa de Especialización en Econometría Aplicada
Centro de Formación Continua -UNI
Machine Learning (Basic)
Clase 3

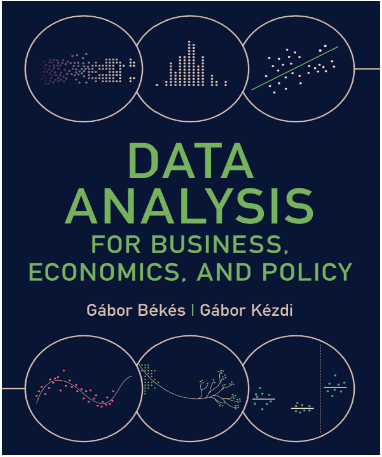
Edinson Tolentino
MSc Economics
email: edinson.tolentino@gmail.com

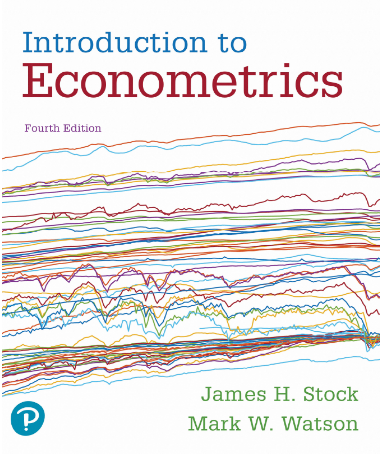
Twitter: @edutoleraymondi

Universidad Nacional de Ingeneria

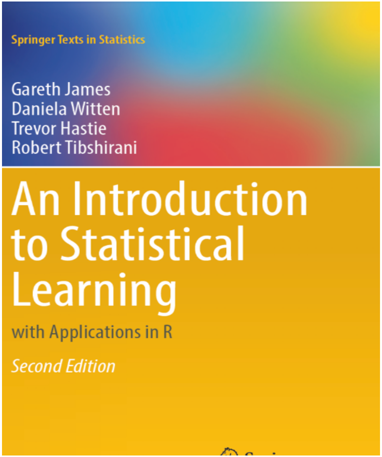
Content

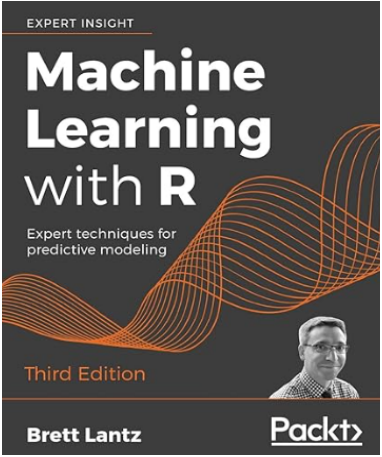
Books


DATA ANALYSIS
FOR BUSINESS, ECONOMICS, AND POLICY
Gábor Békés | Gábor Kézdi


Introduction to Econometrics
Fourth Edition
James H. Stock
Mark W. Watson

Books


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition


EXPERT INSIGHT
Machine Learning with R
Expert techniques for predictive modeling
Third Edition
Brett Lantz
Packt>

Supervised Versus Unsupervised

- Supervised methods
  - Each observations of predictor (x_i) is an associated of the measurement y_i
  - Methods
    - Lineal regression
    - logistic regression
    - GAM
    - Boosting Support vector machine
- Unsupervised methods
  - Each observations of predictor (x_i) but no associated of the measurement y_i
  - We can seek to understand the relationships between the variables or between the observations
  - Methods:
    - cluster analysis

Estimate f

- `Parametric Methods`
  - Involve a two-step model-based approach
    - First, we make an assumption about the functional form

$$f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_p X_p \tag{1}$$

    - After a model has been selected, we need a procedure that uses the **training data** to fit or **train the model**

$$Y \equiv \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_p X_p$$

- The potential disadvantage of a parametric approach is that the model we choose **will usually not match the true unknown form of f.**
  - If the chosen model is too far from the true f, then our estimate will be poor.
  - We can choose flexible model that can fit different possible functional form of f
  - However, more flexible model requires estimating a greater number of parameters, these can lead to phenomenon known as **overfitting**

Assessing model accuracy

- It is an important task to decide for any given set of data which method produces the best results.
- Selecting the best approach can be one of the most challenging parts
- In order to evaluate the performance of our methods on a given data set, we need a measure how well its predictions actually match the observed data.
- In the regression setting, the most commonly-used measure is the mean squared error (MSE),

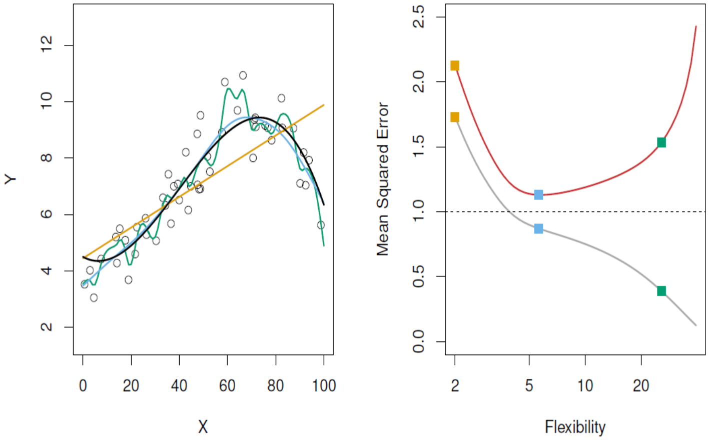
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2 \tag{2}$$

  Where $\widehat{f}(x_i)$ is the prediction that $\widehat{f}$ gives for the ith observations.
- The MSE **will be small if the predicted responses are very close to the true responses**

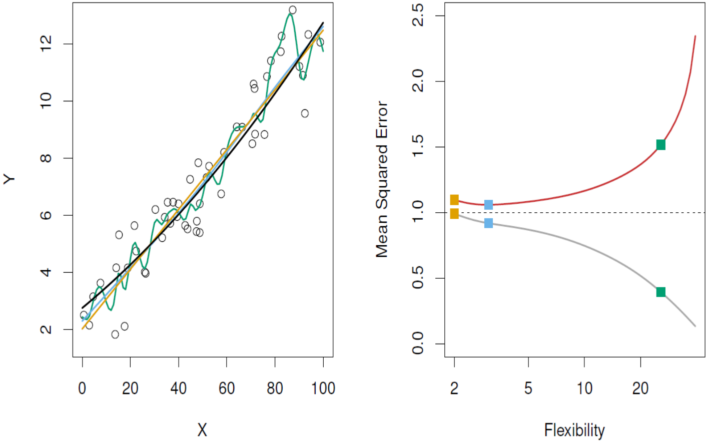
Assessing model accuracy

- In general, we do not really care how well the method works training on the training data. Rather, we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.
- How can we go about trying to select a method that minimizes the test MSE? In some settings, we may have a test data set available
- But what if no test observations are available?, you can select methods or models to minimizes the training MSE. Therefore, **there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.**

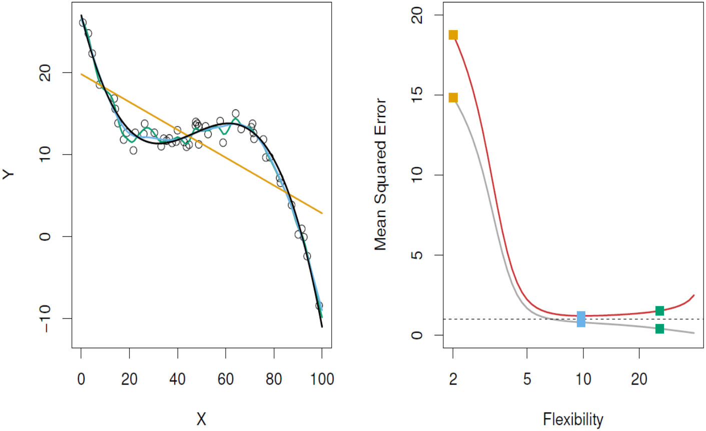
Assessing model accuracy



Assessing model accuracy

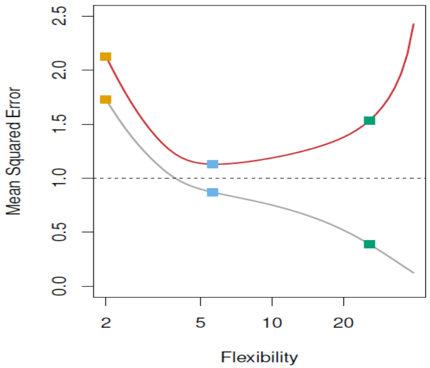


Assessing model accuracy



Assessing model accuracy

- The method picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f
- Due to overfit training data, the MSE is large. This is because the supposed patterns that the method found in the training data simply don't exist in the test data.
- Possible solutions: **cross-validation**



The Bias-Variance Trade-Off

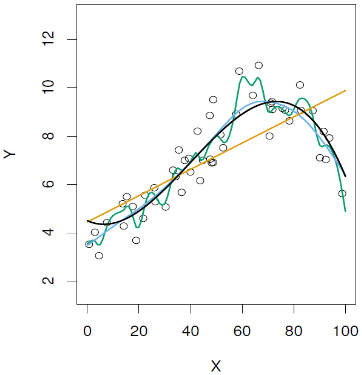
- The expected test MSE, for a given value x_0, can always be decomposed into the sum of three fundamental quantities:

$$E\left(y_0 - \widehat{f}(x_0)\right)^2 = Var(\widehat{f}(x_0)) + \left[Bias(\widehat{f}(x_0)) + Var(\varepsilon)\right]^2 \tag{3}$$

- We need to select simultaneously achieves low variance and low bias.
- What do we mean by the variance and bias of a method?
  - More flexible methods have higher **variance**
  - **Bias** refers to the error that is introduced by approximating a real-life problem. more flexible means less bias.

The Bias-Variance Trade-Off

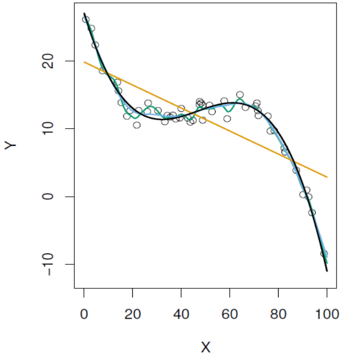
- The flexible green curve is following the observations very closely. It has high variance because changing any one of these data points may cause the estimate $\hat{f}$ to change considerably
- The orange least squares line is relatively inflexible and has low variance, because moving any single observation will likely cause only a small shift in the position of the line.



The Bias-Variance Trade-Off

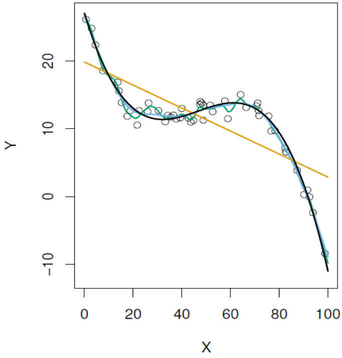
The Bias-Variance Trade-Off

(**Right side**): According to least square (orange line), means more bias

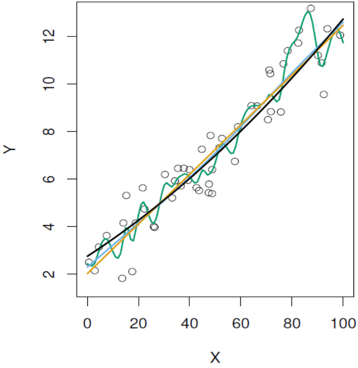


The Bias-Variance Trade-Off

(**Right side**): According to least square (orange line), means more bias



(**Left side**): less bias (more accurate)



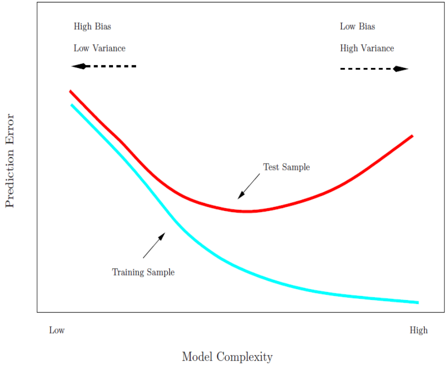
Cross validation

- There are two re-sampling methods:
  - Cross-validation
  - Bootstrap
- These methods refit a model of interest to samples formed from the training set
- In order to obtain additional information about the model

Training 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 one that was not used in training the method
- The **training error** can be easily calculated by applying the statistical learning method to the observations used in its training

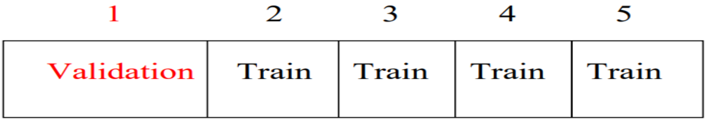


Validation-set approach

- Here we randomly divide the available set of 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 in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

K-fold Cross-validation

- In the validation approach, only a subset of the observations ? those that are included in the training set rather than in the validation 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
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model
- Idea is to 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.
- Divide data into K roughly equal-sized parts (K = 5 here)



K-fold Cross-validation

- Let the K parts be $C_1, C_2, \cdots, C_K$ where C_K denotes the indices of the observations in part k . There are n_K observations in part k: if N is a multiple of K, then $n_k = \frac{n}{K}$
- Compute :

$$CV_K = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

- A better choice is $K = 5$ or 10. Moreover, these provides a good compromise for this bias-variance tradeoff.

Edinson Tolentino

Summary Cross-validation