STATISTICAL LINEAR REGRESSION - PART II:

ASSESSING MODEL ACCURACY

(Rafael Alcalá)

Bibliography:

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
An Introduction to Statistical Learning with Applications in R
Springer, 2013

Chapter 02

(Some of the figures in this presentation are taken from this book and some slides are based on Abbass Al Sharif‘s slides for his course DSO 530)

Outline

- Assessing Model Accuracy
  - Measuring the Quality of Fit
  - The Bias-Variance Trade-off
  - How do we do in practice?
    - K-fold Crossvalidation Assessment

Measuring Quality of Fit

- Suppose we have a regression problem.
- One common measure of accuracy is the mean squared error (MSE) i.e.

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

- Where $\hat{y}_i$ is the prediction our method gives for the observation in our training data.

A Problem

- In either case our method has generally been designed to make MSE small on the training data we are looking at e.g. with linear regression we choose the line such that MSE is minimized.

- What we really care about is how well the method works on new data. We call this new data “**Test Data**”.

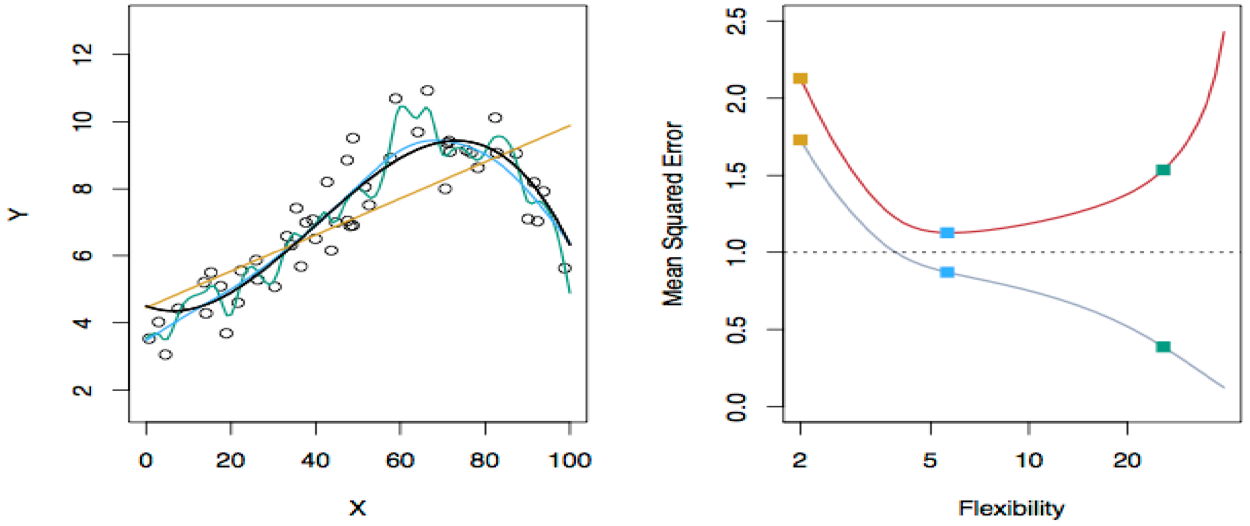
- There is no guarantee that the method with the smallest training MSE will have the smallest **test** (i.e. new data) MSE (**generalization** ability).

Training vs. Test MSE's

- In general the more flexible a method is the lower its training MSE will be i.e. it will "fit" or explain the training data very well.
  - Side Note: More Flexible methods (such as splines) can generate a wider range of possible shapes to estimate f as compared to less flexible and more restrictive methods (such as linear regression). The less flexible the method, the easier to interpret the model. Thus, there is a trade-off between flexibility and model interpretability.

- However, the **test** MSE may in fact be higher for a more flexible method than for a simple approach like linear regression.

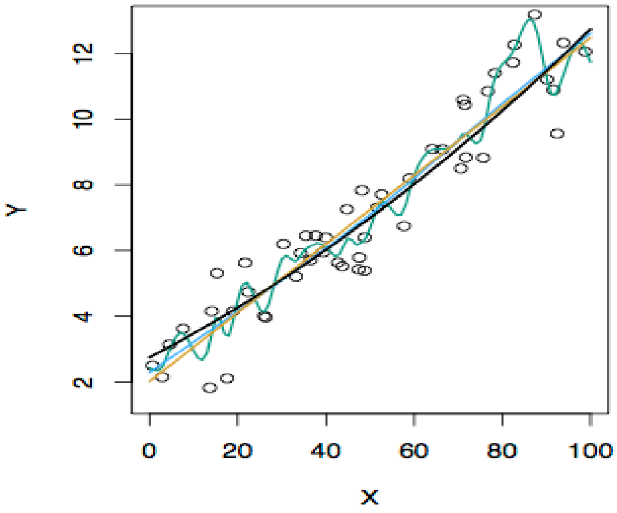
Examples with Different Levels of Flexibility: Example 1

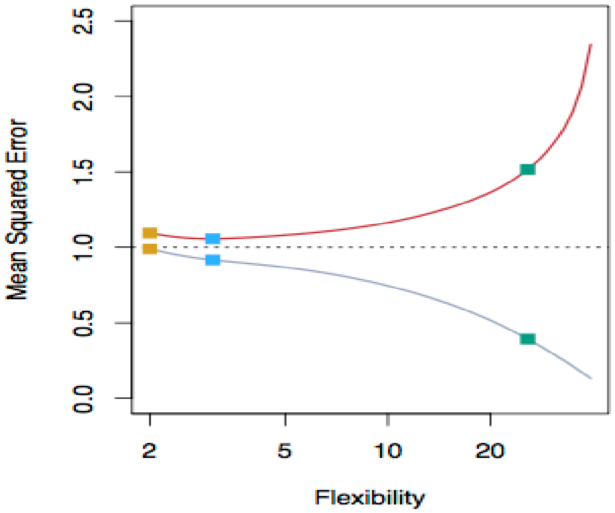


LEFT
Black: Truth
Orange: Linear Estimate
Blue: smoothing spline
Green: smoothing spline (more flexible)

RIGHT
RED: Test MES
Grey: Training MSE
Dashed: Minimum possible test MSE (irreducible error)

Examples with Different Levels of Flexibility: Example 2

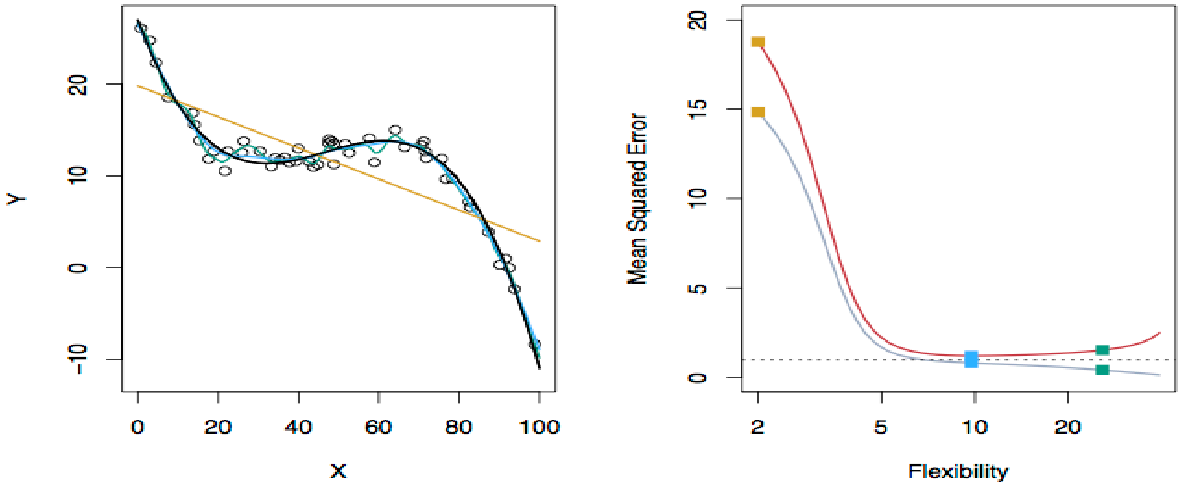




LEFT
Black: Truth
Orange: Linear Estimate
Blue: smoothing spline
Green: smoothing spline (more flexible)

RIGHT
RED: Test MES
Grey: Training MSE
Dashed: Minimum possible test MSE (irreducible error)

Examples with Different Levels of Flexibility: Example 3



LEFT
Black: Truth
Orange: Linear Estimate
Blue: smoothing spline
Green: smoothing spline (more flexible)

RIGHT
RED: Test MES
Grey: Training MSE
Dashed: Minimum possible test MSE (irreducible error)

Bias/ Variance Tradeoff

- The previous graphs of test versus training MSE's illustrates a very important tradeoff that governs the choice of statistical learning methods.

- There are always two competing forces that govern the choice of learning method i.e. bias and variance.

Bias of Learning Methods

- Bias refers to the error that is introduced by modeling a real life problem (that is usually extremely complicated) by a much simpler problem.

- For example, linear regression assumes that there is a linear relationship between Y and X. It is unlikely that, in real life, the relationship is exactly linear so some bias will be present.

- The more flexible/complex a method is the less bias it will generally have.

Variance of Learning Methods

- Variance refers to how much your estimate for f would change by if you had a different training data set.

- Generally, the more flexible a method is the more variance it has.

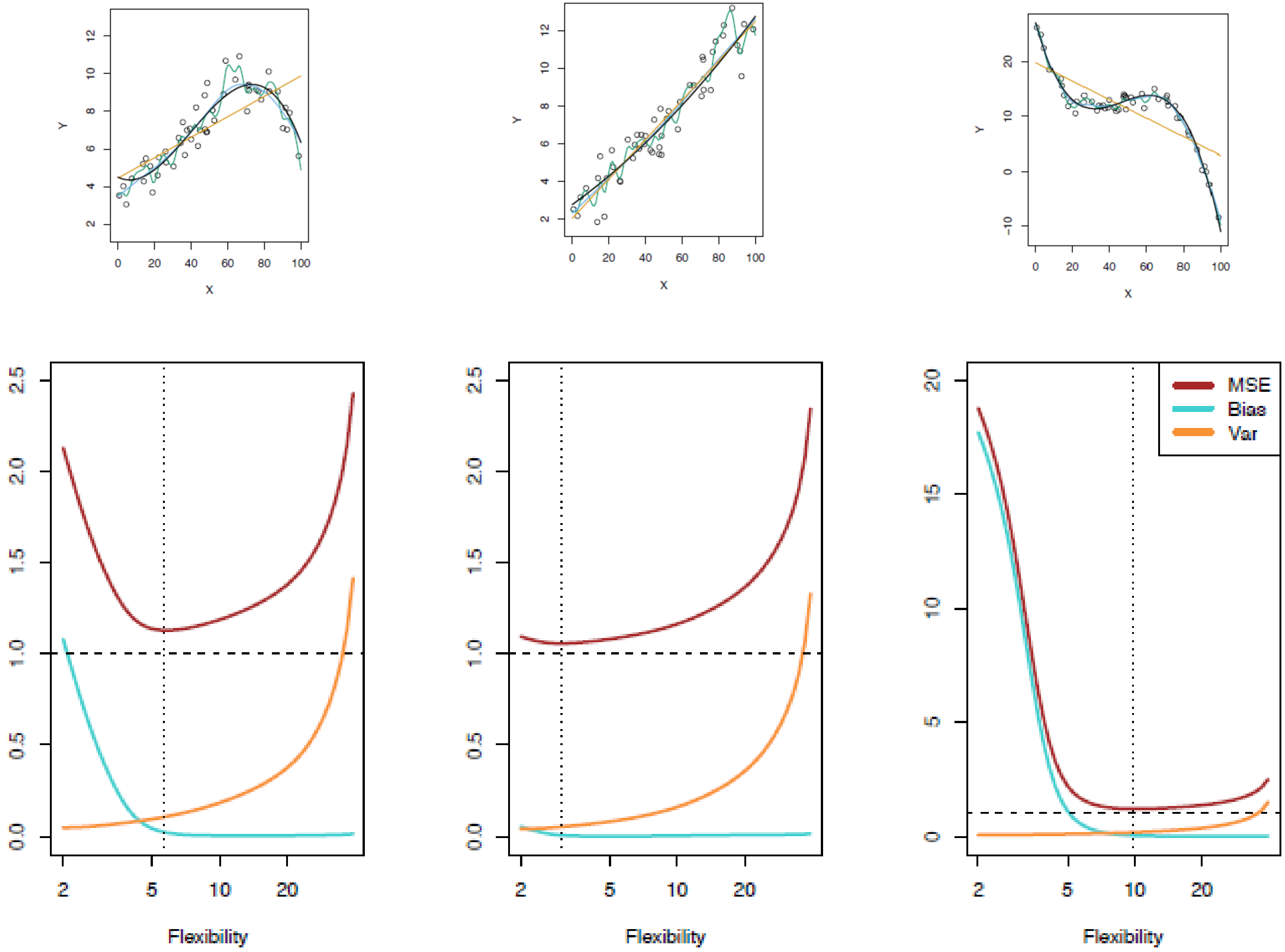
The Trade-off

- It can be shown that for any given, X=x_0, the expected test MSE for a new Y at x_0 will be equal to

$$ExpectedTestMSE = E\left(Y - f(x_0)\right)^2 = Bias^2 + Var + \underbrace{\sigma^2}_{\text{Irreducible Error}}$$

- What this means is that as a method gets more complex the bias will decrease and the variance will increase but expected test MSE may go up or down!

Test MSE, Bias and Variance



How do we do in Practice?

- Two issues arise at this point
  - Model Selection: How do we select the “optimal” model for a given problem?
  - Validation: Once we have chosen a model, how do we estimate its true error?

  (true error: Model error when tested on the entire population, i.e., *expectedTest MSE*)

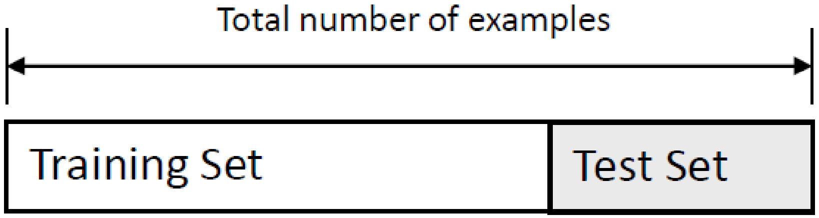
- Validation addresses the problem of over-fitting.

- If we have access to an unlimited number of examples, we choose the model that provides the lowest error rate on the entire population

- However, in real-world applications only a finite set of examples is available

The holdout method

- Split dataset randomly into two groups
  - Training set: Used to learn the model
  - Test set: Used to estimate the model true error



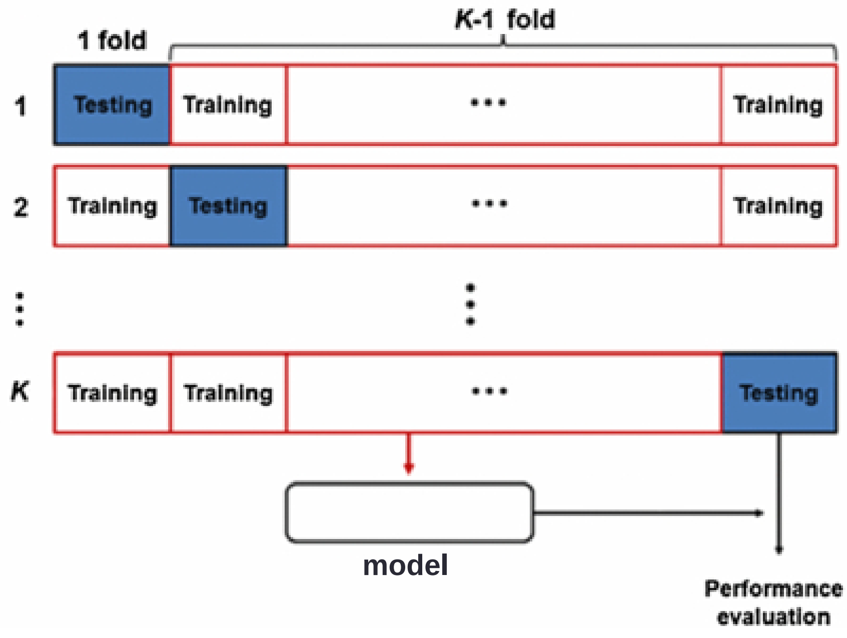
Usual data distributions
Regression: 80%-20%, 70%-30%
Classification: 90%-10%

- If over-fitting is present, the model will perform well in your training dataset but poorly in your test dataset.

- Two important drawbacks
  - In sparse datasets -> setting aside a portion becomes a “luxury”
  - Single train-and-test experiment -> misleading on “unfortunate” split

K-fold Cross-Validation

- These limitations of the holdout can be solved with a family of resampling methods.
- Among others, K-fold cross-validation is the one most used
- Create and use a K-fold partition of the dataset:
  - Randomly divide your data into *K* pieces
  - For each of K experiments, use a different fold for testing and the remaining *K−1* folds for training



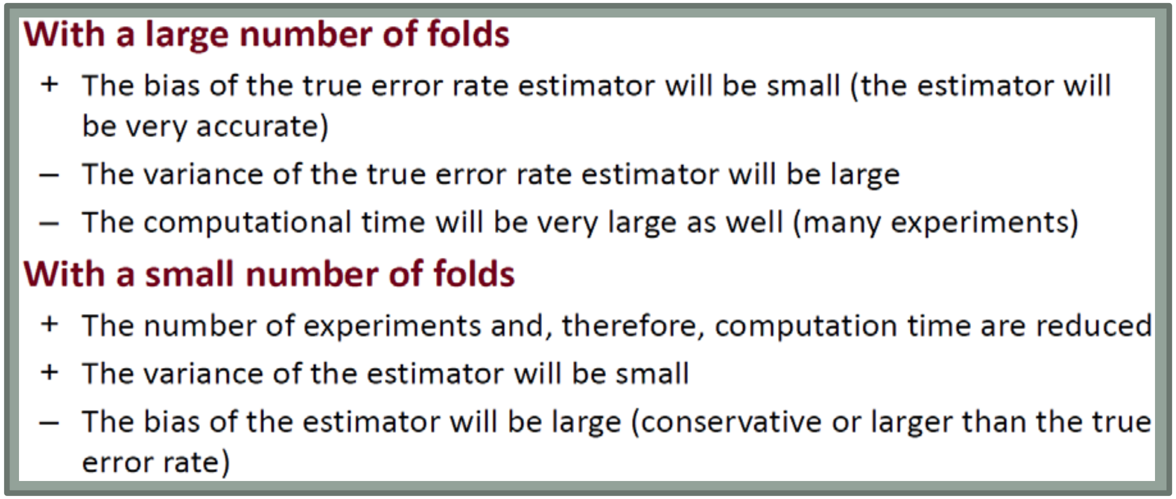
For **non-deterministic** algorithms several runs per experiment with **different seeds** fixed in a systematic way for all the algorithms:
123456, 612345, 561234, … or
0.123456, 0.612345, 0.561234, …

K-fold Cross-Validation

- All the examples in the dataset are eventually used for both training and testing.

- The true error is estimated as the average error rate on test examples

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i\,, whereE_i = MSE\ or\ equivalent$$



With a large number of folds

+ The bias of the true error rate estimator will be small (the estimator will be very accurate)
– The variance of the true error rate estimator will be large
– The computational time will be very large as well (many experiments)

With a small number of folds

+ The number of experiments and, therefore, computation time are reduced
+ The variance of the estimator will be small
– The bias of the estimator will be large (conservative or larger than the true error rate)

Usual *K* Values
Regression: 5-fold
Classification: 10-fold