Bias/Variance

This section is based on James et al. (2021), Chapter 2.

1 What is our objective?

We want to model the relationship between a response variable Y, which may be quantitative, qualitative, or of a different nature, and a set of p explanatory variables $X = (X_1, \dots, X_p)$, also of (potentially) different types. The central idea is that there is a relationship between Y and the explanatory variables X. Generally speaking, we model this relationship using the following model:

$$Y = f(X) + \varepsilon. \tag{1}$$

Here, f is a deterministic (non-random) function representing the systematic information that the explanatory variables $X_1, \dots, X_p$ provide about Y, and ε is a random error term, modeling the variations in Y not explained by X. For the purposes of this course, we will make the following assumptions: the random variable ε is independent of the explanatory variables X, $\mathbb{E}[\varepsilon] = 0$, and $\text{Var}(\varepsilon) = \sigma^2$. The Equation 1 is general. It serves as a framework for all the methods we will study, even when the explicit form of f is not known.

The Figure 1 illustrates the different elements of the model: the observed data (X_i, Y_i), the function f (in blue), and the random deviations ε_i represented by dotted lines.

In the rest of the course, we will examine different methods for estimating the function f from data. However, before examining how to construct a $\widehat{f}$ estimator of f, we will examine the quality of such an estimator: what does it mean to "properly estimate" f? And how can we evaluate the quality of the estimate?

> Example: Simple Linear Regression
>
> In this very simple framework, we assume that the function f is of the form: f(x) = a x + b. In this case, estimating the function f is reduced to estimating the coefficients a and

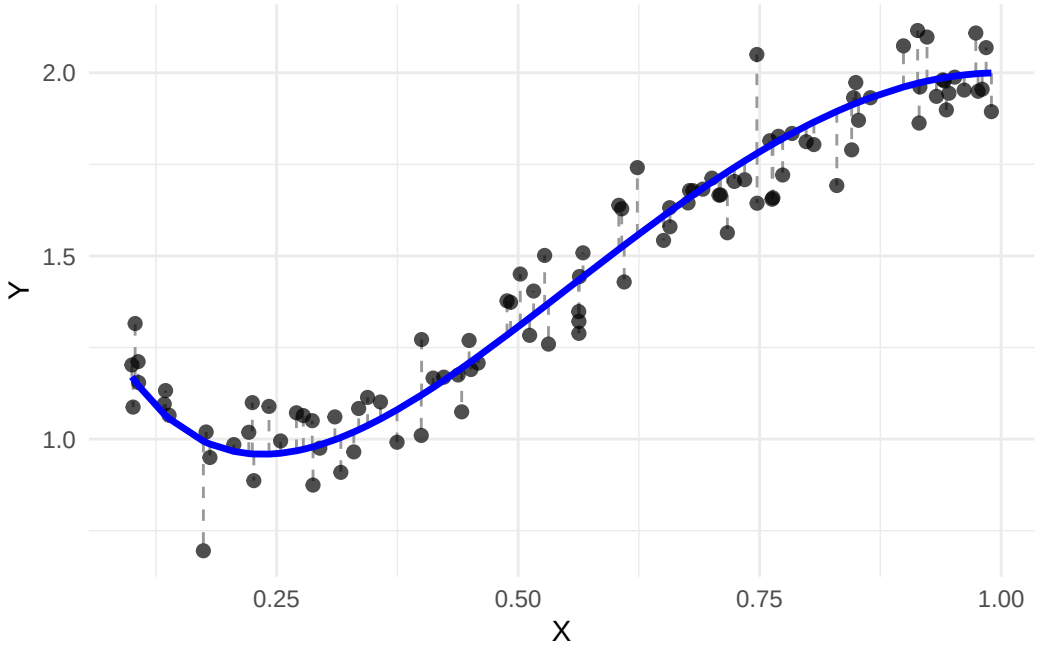

Figure 1: The different elements of the model. The points represent the observed data (X_i, Y_i). The blue curve represents the function f and the dotted lines represent the error associated with each observation.

b.

Note: Tradeoff between accuracy and interpretability

Depending on the objective of the study, we generally have to make a choice between the accuracy of our predictions and the interpretability of our model. A simple model, such as linear regression, will be easy to interpret but will poorly capture complex relationships. Conversely, a more flexible model, such as a random forest, will have better predictions but will be more difficult to interpret. The choice therefore depends on the objective of the analysis: understanding or predictive performance?

Note: *No free lunch in statistics*

Why not simply use the "ultimate" model, the one that would always be optimal regardless of the dataset? Because such a model doesn't exist! There is no universally best method for all datasets and all objectives. A method that performs well in a given context may fail elsewhere. It is therefore always necessary to adapt the approach to the problem (explanation, prediction, classification, etc.).

2 How to measure the quality of an estimator?

Once we have an estimator $\hat{f}$ of the function f, obtained from n observations $(y_1, x_1), \dots, (y_n, x_n)$, we seek to evaluate the accuracy of the predictions $\widehat{Y} = \hat{f}(X)$. The idea is to verify how close $\widehat{Y}$ is to the true value of Y.

Definition: Mean Square Error

When Y is a quantitative variable, a classic measure of the quality of $\hat{f}$ is the **mean square error** (*MSE*):

$$MSE(Y, \widehat{Y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2,$$

where $\hat{y}_i = \hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for observation x_i.

A low MSE indicates that the predictions are close to the observations. We can also interpret it as the average distance between the observed values and the predicted values. We therefore aim for a low average distance.

When Y is a qualitative variable, e.g., a class or a label, we use another measure: the error rate.

Definition: Error Rate

When Y is a qualitative variable, a classic measure of the quality of $\hat{f}$ is the **error rate** (*ER):

$$ER(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{f}(x_i)).$$

where $\hat{y}_i = \hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for observation x_i.

The error rate measures the proportion of incorrect predictions. It is, again, a measure of the average distance between Y and $\widehat{Y}$, suitable for qualitative variables.

3 The bias/variance trade-off

Our goal is often to minimize the prediction error, not only on observed data, but especially on new data (recovered after estimating the model). To do this, we focus on the prediction error:

$$\mathbb{E}\left[\left(Y - \widehat{Y}\right)^2\right] = \mathbb{E}\left[\left(Y - \hat{f}(X)\right)^2\right].$$

This error can be broken down into three components:

- Bias: the error due to a systematic approximation, e.g., if we impose a linear model when the relationship is nonlinear.
- Variance: the sensitivity of the estimator to fluctuations in the training sample.
- Irreducible error: the intrinsic variance of the noise ε, denoted σ^2.

Décomposition bias/variance

On to:

$$\mathbb{E}\left[(Y - \widehat{Y})^2\right] = \mathbb{E}\left[(Y - \hat{f}(X))^2\right] = \text{Biais}(\hat{f}(X))^2 + \text{Var}(\hat{f}(X)) + \sigma^2.$$

Preuve

On the other hand, we point out that the hope of the estimator's error is broken down into a reductible part and an irreductible part.

$$\begin{aligned}
\mathbb{E}\left[\left(Y-\widehat{Y}\right)^2\right] &= \mathbb{E}\left[\left(Y-\hat{f}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(X)+\varepsilon-\hat{f}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] + 2\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)\varepsilon\right] + \mathbb{E}[\varepsilon^2] \\
&= \mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] + 2\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)\right]\underbrace{\mathbb{E}[\varepsilon]}_{=0} + \sigma^2 \\
&= \underbrace{\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right]}_{\text{réductible}} + \underbrace{\sigma^2}_{\text{irréductible}}.
\end{aligned}$$

Use the linearity of hope and make sure that X and ε are independent. He is interested in maintaining the "reductible" party. The trick is to make apparatus $\mathbb{E}\left[\hat{f}(X)\right]$.

$$\begin{aligned}
\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] &= \mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]+\mathbb{E}\left[\hat{f}(X)\right]-\hat{f}(X)\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right]}_{\text{A}} \\
&\quad - 2\underbrace{\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)\left(\hat{f}(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)\right]}_{\text{B}} \\
&\quad + \underbrace{\mathbb{E}\left[\left(\hat{f}(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right]}_{\text{C}}.
\end{aligned}$$

A. The $f(X)$ function is not aléatoire, on a $\mathbb{E}[f(X)] = f(X)$ et donc

$$\begin{aligned}
\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[f(X)-\hat{f}(X)\right]\right)^2\right] \\
&= \mathbb{E}\left[f(X)-\hat{f}(X)\right]^2 \\
&= \text{Biais}(\hat{f}(X))^2.
\end{aligned}$$

B. To develop the expression and use the independence of the variables, on finding that $B = 0$.
C. Using the definition of variance,

$$\mathbb{E}\left[\left(\hat{f}(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right] = \text{Var}(\hat{f}).$$

Finally, hon a

$$\mathbb{E}\left[\left(f(X) - \hat{f}(X)\right)^2\right] = \text{Biais}(\hat{f}(X))^2 + \text{Var}(\hat{f}(X)).$$

Hence the results.

This breakdown before a fundamental compromise in analyzing women:

- If you choose a more flexible model, the bias will be higher, but the variance will be easier.
- If you choose a flexible model, the bias will be easy, but the variance will be too high.

Our objective is to find a just balance between bias and variance, i.e. a model here predicted correctly, tout en étant généralisable à de new women. The Figure 2 presents a number of women and different estimateurs $\hat{f}$. Simply vary the λ parameters, making the models more or less flexible (for example $\lambda = 0.15$, the model is flexible and for example $\lambda = 1$, the models are rigid). The Figure 3 shows the value of the bias, the variance and the MSE for the models estimated for the Figure 2. On the other hand, it should be noted that the more λ is smaller, the more the variance is larger, but the bias is smaller (the model is flexible). Inversely, plus λ is large, plus the bias is large and the variance is small (the models are rigid). The MSE panel in function of the parameters is a U panel. How to minimize the MSE, i.e. To make a compromise between the bias and the variance, you can take $\lambda = 0.5$.

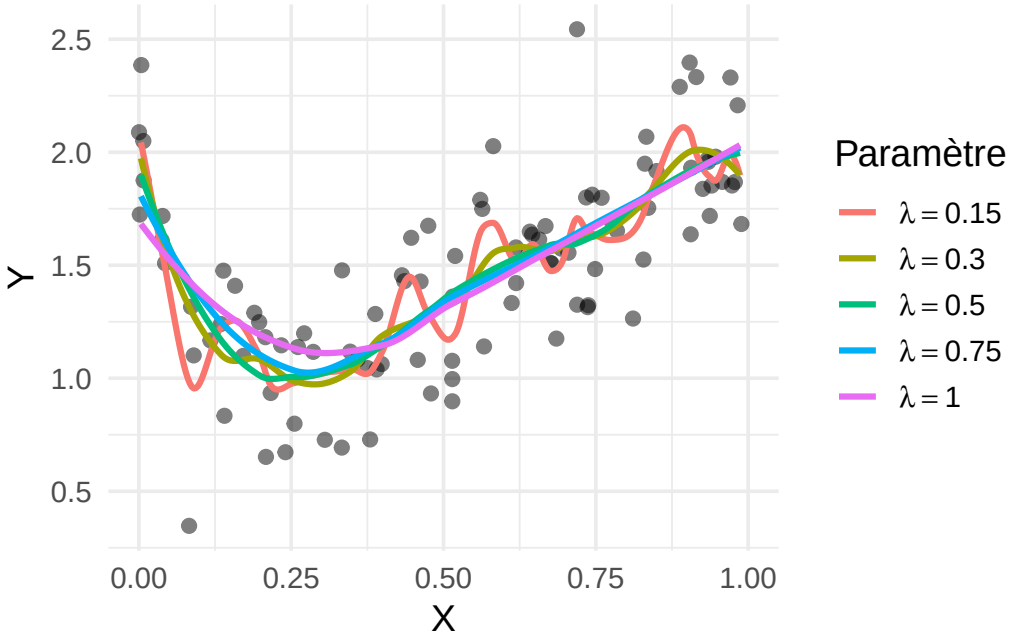

Figure 2: Différents estimateurs de la fonction f.

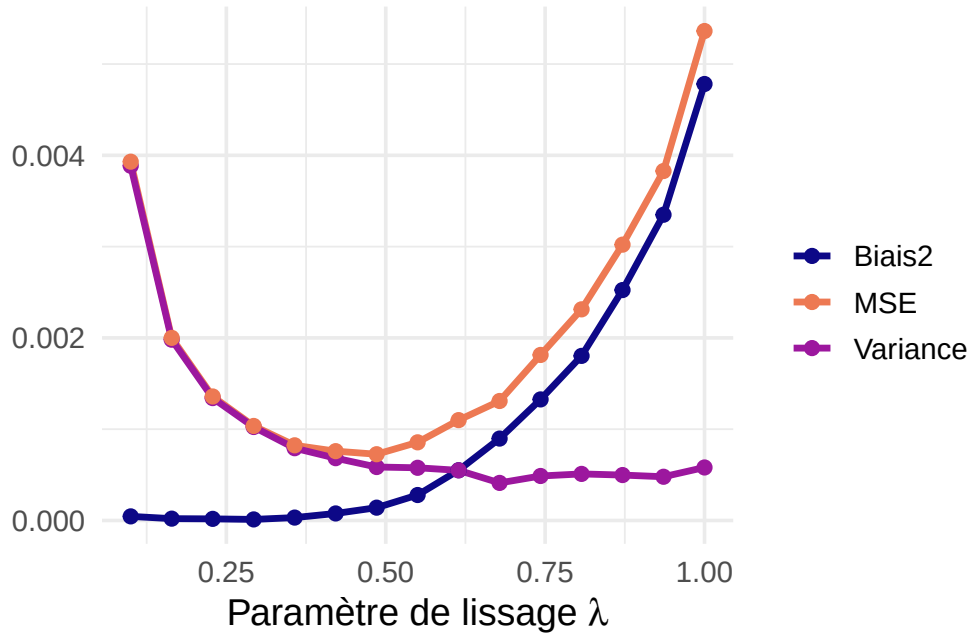

Figure 3: Compromis biais/variance.

Generally speaking, as flexibility increases, the decrease in bias is greater than the increase in variance, resulting in a decrease in prediction error. However, beyond a certain level of flexibility, the bias becomes negligible, and any further decrease is offset by the rapid increase in variance. Prediction error therefore begins to increase. This results in a U-shaped curve of prediction error as a function of model flexibility: a model that is too rigid generates a high bias, while a model that is too flexible leads to too much variance.

Note: Why compromise?

It is always possible to build a very flexible model with zero bias, e.g., a model that passes through all observation points, but which will have enormous variance. Conversely, a model that is too rigid, e.g., a constant, will have a very large bias but almost zero variance. The bias/variance compromise consists of choosing a model that controls both of these quantities.

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. New York,

NY: Springer US. https://doi.org/10.1007/978-1-0716-1418-1.