École chercheurs MEXICO, La Rochelle, Mars 2018

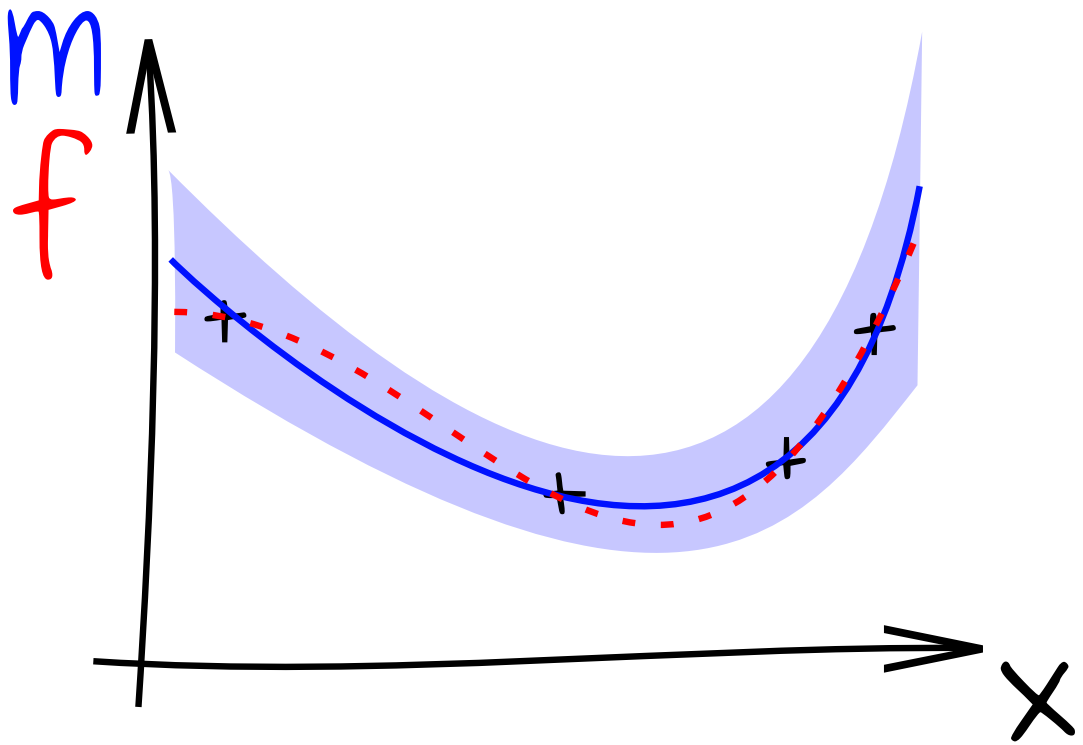
Introduction to statistical modelling

Nicolas Durrande, nicolas@prowler.io

PROWLER.io, Cambridge (UK) – Mines St-Étienne (France)

Introduction

Why **statistical models**?
We want to be able to quantify the model error:



The confidence intervals can be used to obtain a **measure of uncertainty on the value of interest**.

In the sequel, we will use the following notations :

- The set of observation points will be represented by a $n \times d$ matrix $X = (X_1, ..., X_n)^t$
- The vector of observations will be denoted by F : $F_i = f(X_i)$ (or $F = f(X)$).

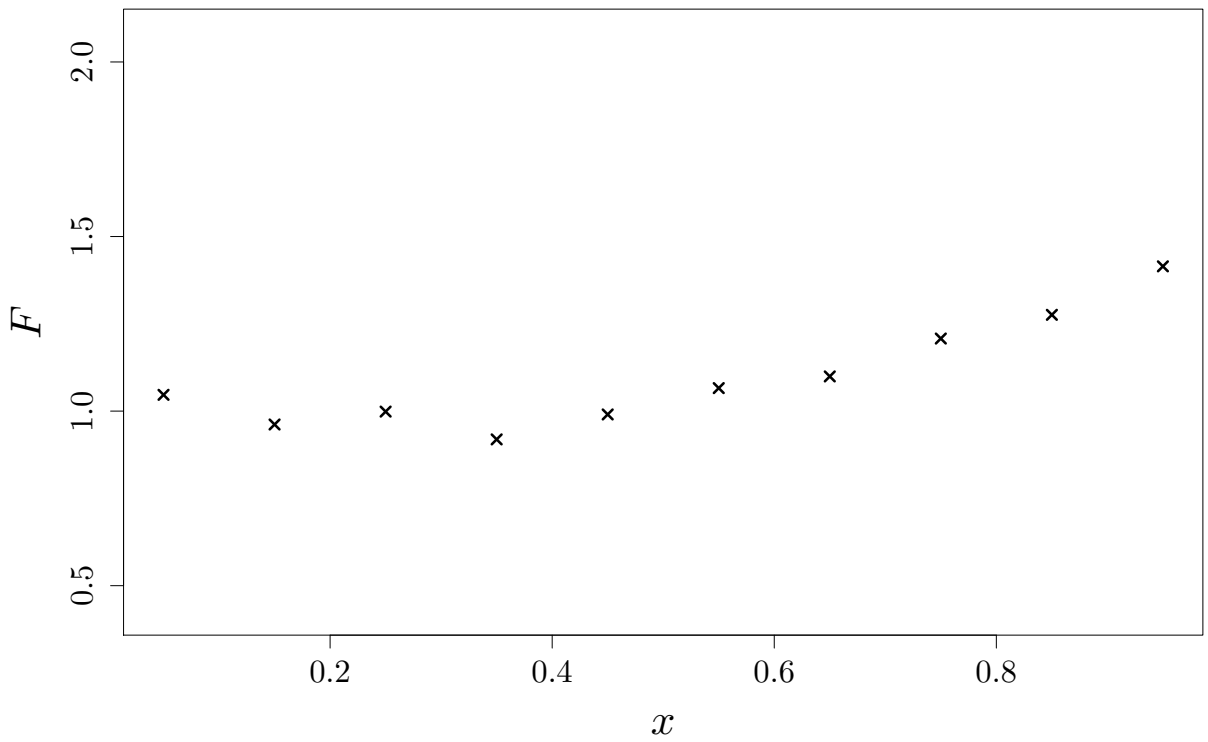
We will now discuss two types of statistical models:

- Linear regression
- Gaussian process regression

Linear Regression

Example

If we consider the following observations:

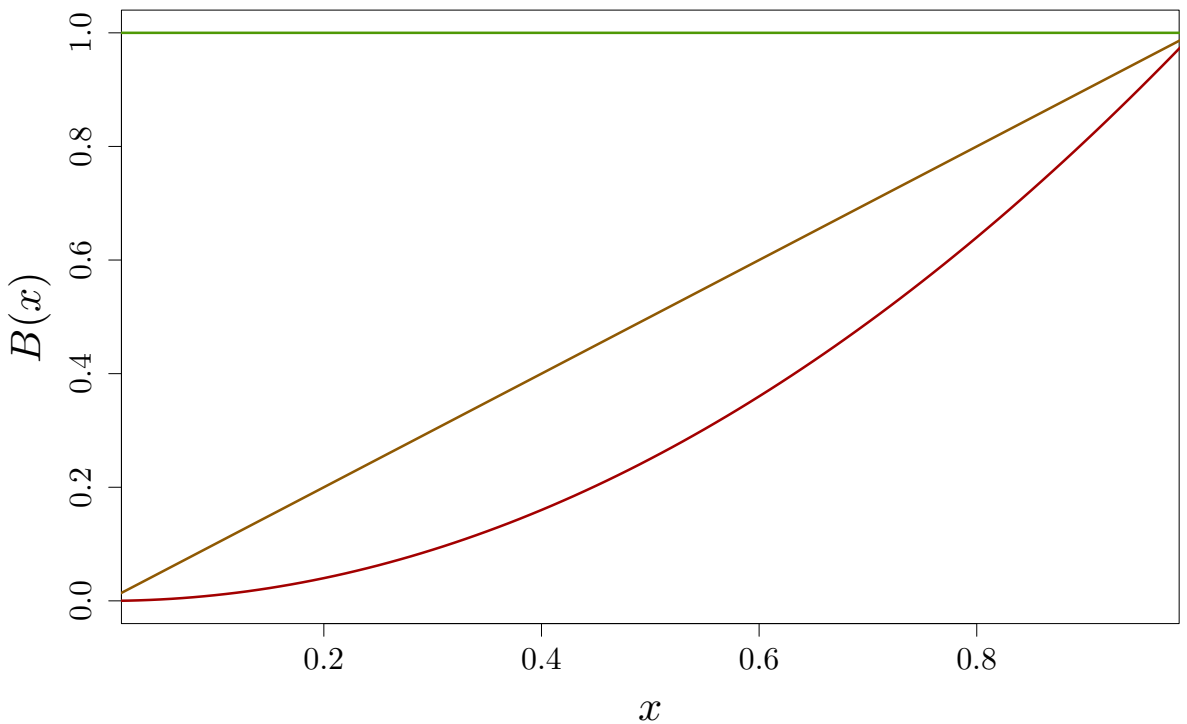


Example

We assume the observations are drawn from

$$F_i = \sum_{k=0}^{2} \beta_k b_k(X_i) + \varepsilon_i \qquad (= B(X_i)\beta + \varepsilon_i)$$

with $b_0(x) = 1,\ b_1(x) = x,\ b_2(x) = x^2$, unknown β_i and i.i.d ε_i.

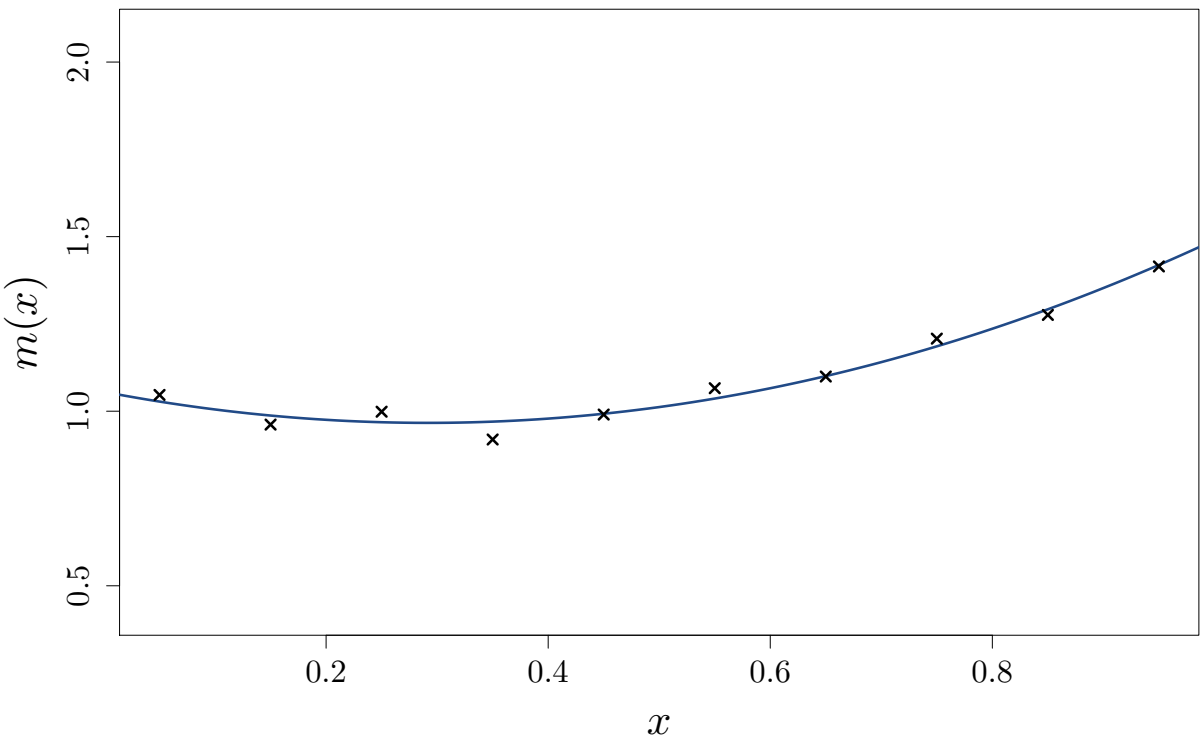


Example

The best linear unbiased estimator of β is

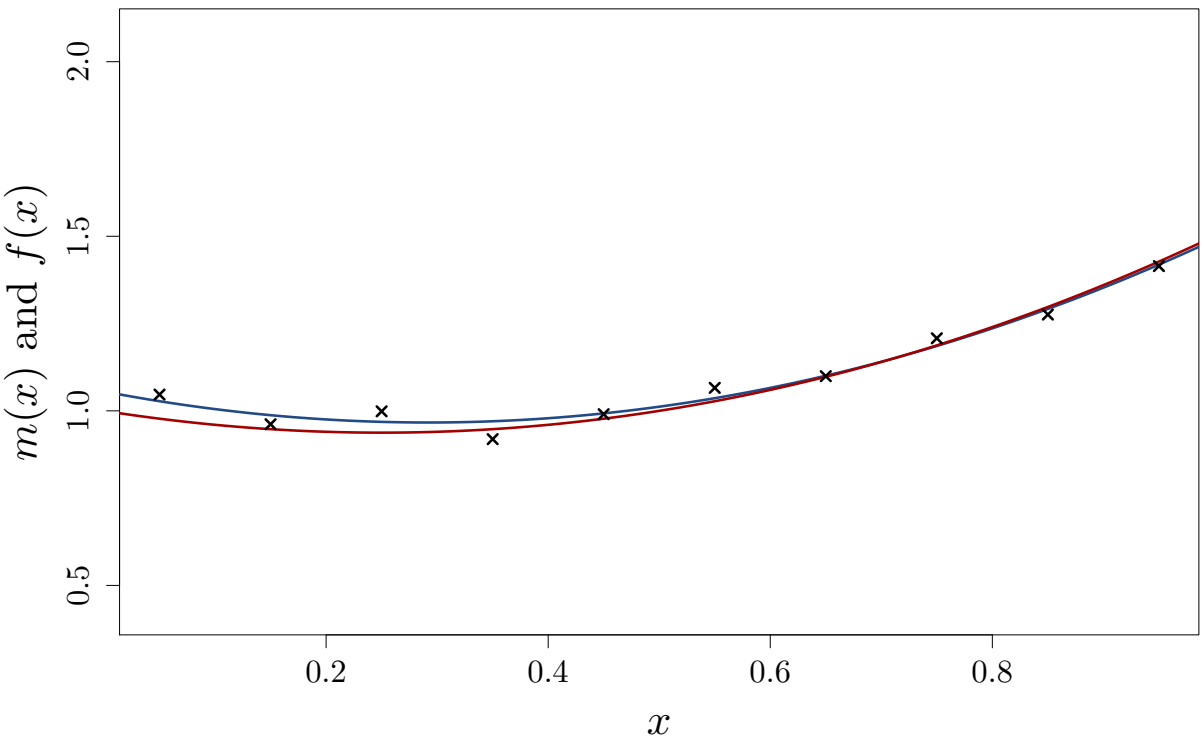
$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t F.$$

We obtain $\hat{\beta} = (1.06, -0.61, 1.04)^T$ and the model is:



Example

There is of course an error between the true generative function and the model



Can this error be quantified?

The initial assumption is $F = B(X)\beta + \varepsilon$ and we have computed an estimator of β:

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t F.$$

$\hat{\beta}$ can thus be seen as a sample from the random variable:

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t (B(X)\beta + \varepsilon).$$

What about the distribution of $\hat{\beta}$?

The initial assumption is $F = B(X)\beta + \varepsilon$ and we have computed an estimator of β:

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t F.$$

$\hat{\beta}$ can thus be seen as a sample from the random variable:

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t (B(X)\beta + \varepsilon).$$

What about the distribution of $\hat{\beta}$?

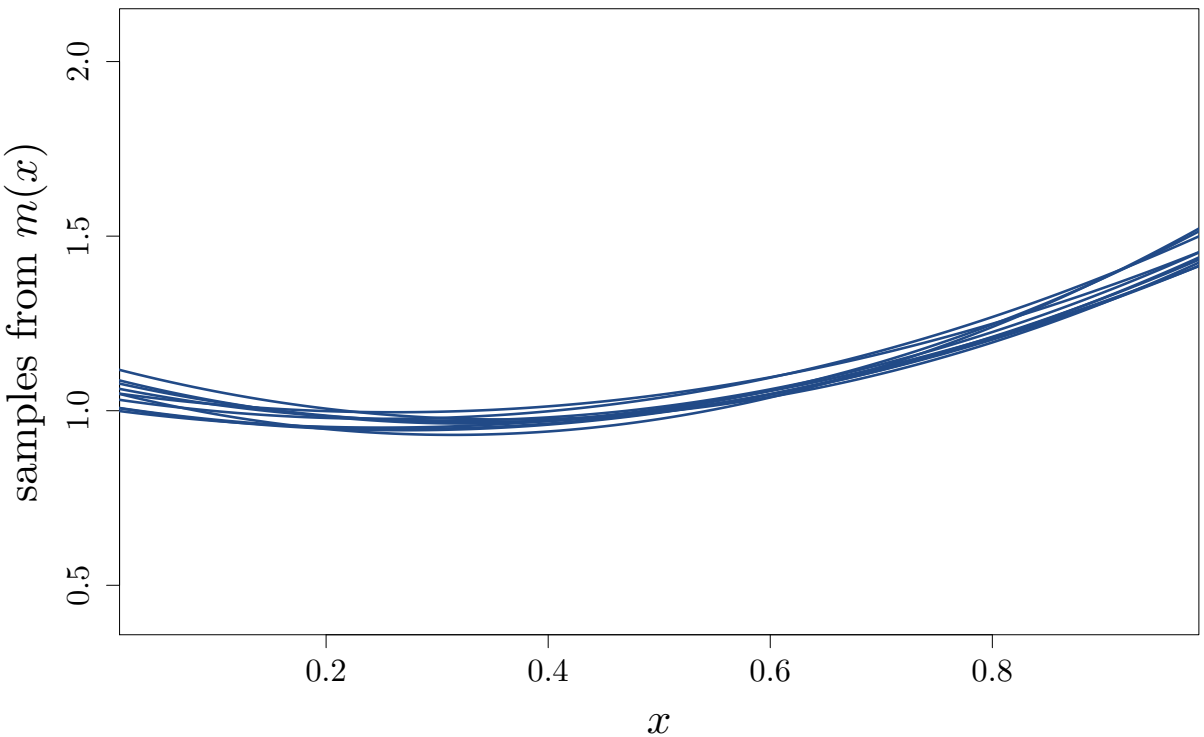
- Its expectation is $\beta \Rightarrow$ The estimator is unbiased
- Its covariance matrix is

$$(B(X)^t B(X))^{-1} B(X)^t \operatorname{cov}[\varepsilon, \varepsilon^t] B(X) (B(X)^t B(X))^{-1}$$

- If ε is multivariate normal, then $\hat{\beta}$ is also multivariate normal.

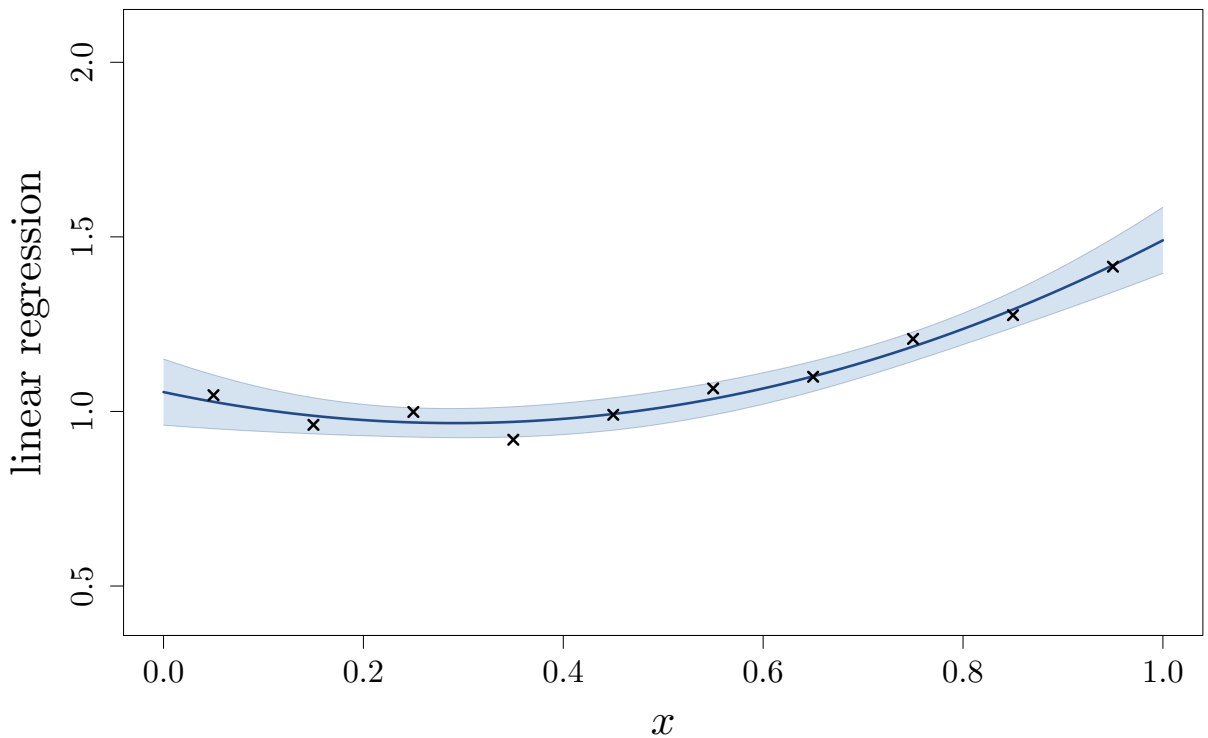
Sampling in the distribution of $\hat{\beta}$ gives us a large variety of models which represent the uncertainty about our estimation:

Back to the example



Back to the example

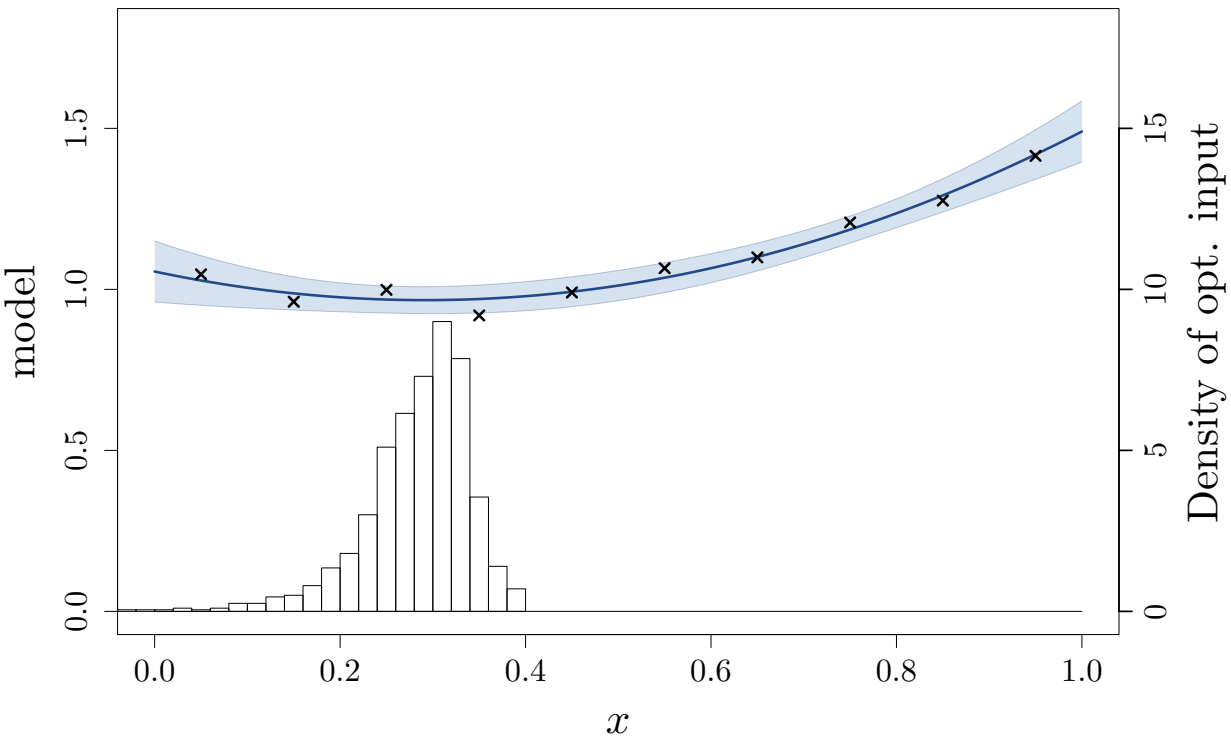
The previous picture can be summarized by showing the mean of m and 95% confidence intervals



This statistical model can be used for **uncertainty quantification**:

Back to the example

If we are interested in the value $x^\star$ minimizing $f(x)$:



we obtain a distribution for $x^\star$.

We could dedicate the entire course to linear regression models...

- model validation
- choice of basis functions
- influence of input locations
- ...

We will just stress a few **pros and cons of these models**:

+ provide a good noise filtering
+ are easy to interpret
− are not flexible (need to choose the basis functions)
− do not interpolate
− may explode when using high order polynomials (over-fitting)

Conclusion

Three things to remember:

- Statistical models are useful when little data is available. they allow to
  - interpolate or approximate functions
  - Compute quantities of interests (such as mean value, optimum, ...)
  - Get an error measure
- GPR is similar to linear regression but the assumption is much weaker (not a finite dimensional space)
- The GPR equations are

$$m(x) = k(x, X)k(X, X)^{-1}F$$
$$c(x, y) = k(x, y) - k(x, X)k(X, X)^{-1}k(X, y)$$

We still have many things to discuss about such models:

- How to choose the observation points?
- How to validate the model?
- How to estimate the model (ie kernel) parameters?

This will be discussed during the next courses.

Reference

Carl Edward Rasmussen and Chris Williams, *Gaussian processes for machine learning*, MIT Press, 2006. (free version online).