Introduction aux modèles prédictifs

Cas continu

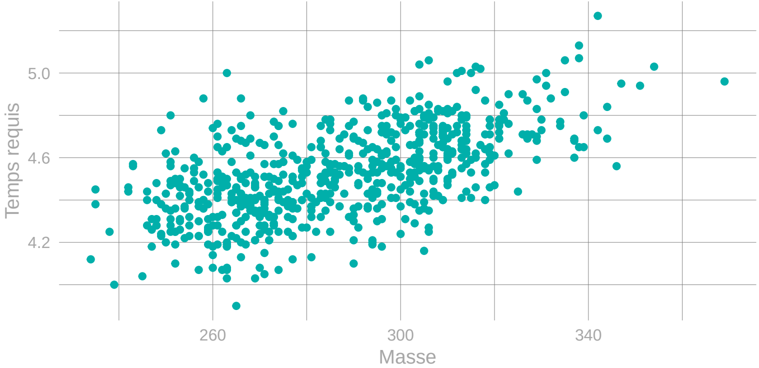
Véronique Tremblay

Cette capsule sert de soutien à la lecture du chapitre 2 de Hastie, Tibshirani, and Friedman (2009).

Qu’est-ce qu’on cherche à faire?

De façon générale...

On veut prédire Y à partir de X.



Plus précisément

On cherche à estimer $f(X)$ la fonction qui relie X et Y.

$$Y = f(X) + \epsilon$$

Fonction de perte
La fonction de perte (L) est une mesure de l'écart par rapport à ce qu'on souhaite mesurer

Par exemple, dans le cas continu, on peut utliser la fonction de perte quadratique.

$$L(Y, f(X)) = (Y - f(X))^2$$

Fonction de risque

La fonction de risque est ce que l'on cherche à minimiser. En général, il s'agit de l'espérance de la fonction de perte.

Par exemple, pour la fonction de perte quadratique, on cherche $\hat{f}(X)$ qui minimise:

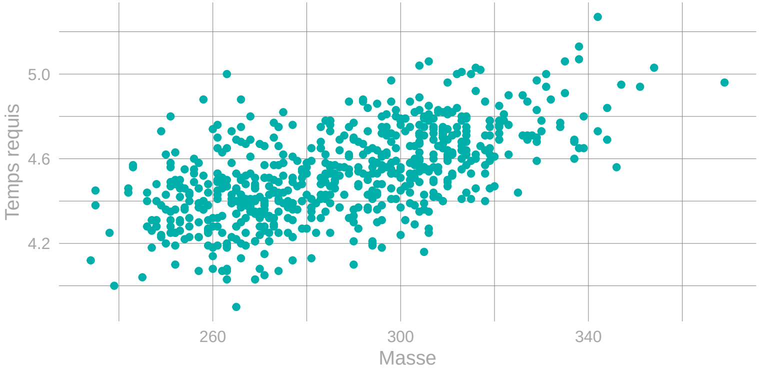
$$\text{EPE}(f) = \mathbb{E}(Y - f(X))^2$$

Comment?

En conditionnant sur X, on constate que l'on minimise EPE(f) avec

$$f(x) = \mathbb{E}(Y|X = x)$$

.



Une autre approche

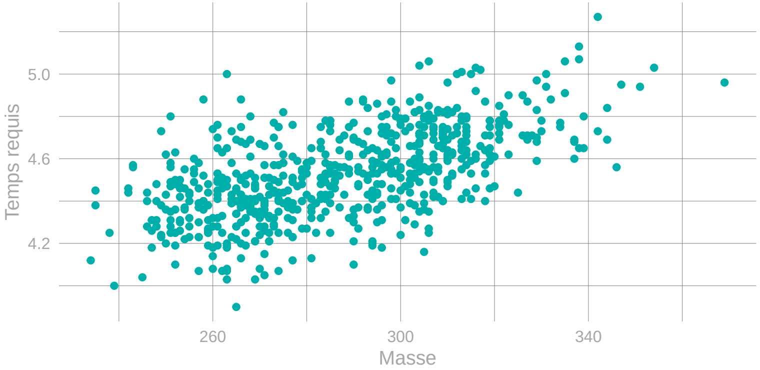
Supposer une certaine forme pour $f(X)$ et minimiser la fonction de perte de façon analytique ou numérique.

La régression linéaire simple

Par exemple, poser $f(X) = \beta_0 + \beta_1 X$ et trouver les valeurs de β_0 et β_1 qui minimisent

$$\mathbb{E}(Y - (\beta_0 + \beta_1 X))^2$$

.



Résumé

- On veut trouver la fonction qui relie X et Y
- Il y a plusieurs façons de trouver $f(X)$
- Dans les prochaines capsules, nous allons voir:
  - Certaines façons d'estimer $f(X)$
  - Le compromis biais-variance
  - Comment évaluer la performance de notre modèle

Références

Lire le chapitre 2 de:

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.