
Méthodes statistiques de prévision

# Prédiction de salaire

**Auteurs :** Ghiles Kemiche

Nom de l'organisme : Institut Mathématique d'Orsay
*Professeur Encadrant :* Marie-Anne Poursat

Licence double diplôme Mathématiques - Informatique

2022 - 2023

# Table des matières

**1 Introduction** **1**

1.1 Données . . . . . 1

1.2 Modèles sélectionnés . . . . . 1

**2 Modèle de régression** **2**

2.1 Régression linéaire simple . . . . . 2

2.2 Régression linéaire avec un terme quadratique . . . . . 3

2.3 Estimateur de Nadaraya-Watson . . . . . 3

2.4 Sélection de la largeur de bande avec la CVLOO pour Nadaraya-Watson 4

2.5 Métrique d'évaluation . . . . . 4

**References** **5**

# 1

# Introduction

Les salaires des travailleurs américains sont influencés par divers facteurs, tels que l'expérience professionnelle, l'éducation, l'industrie et l'âge. Dans cette étude, nous nous concentrons sur l'impact de l'âge sur les salaires des travailleurs américains en utilisant un extrait des données Wage du package ISLR. L'objectif principal de cette analyse est de comprendre l'évolution des salaires en fonction de l'âge et de développer un modèle de regression pour prédire les salaires en fonction de l'âge des travailleurs.

## 1.1 Données

Les données de notre étude comprennent les informations suivantes :

- **Âge :** Il s'agit de la covariable de notre analyse. Cette variable est continue.
- **Salaire :** Cette variable représente la variable à prédire dans notre étude. Elle est égalemment quantitative continue.

## 1.2 Modèles sélectionnés

Pour cette analyse, nous examinerons deux approches de modélisation différentes :

- **Modèle paramètrique :** On prendra le modèle de régression linéaire.
- **Modèle non-paramètrique :** Ensuite, nous explorerons un modèle de régression non paramétrique en utilisant l'estimateur de Nadaraya-Watson.

# 2

# Modèle de régression

Dans notre problème, on considère que $X$ représente l'âge et $Y$ le salaire. Nous proposons le modèle de régression général de la forme :

$$Y_i = m(X_i) + \epsilon_i$$

où $X_i \in \mathbb{R}$ est notre variable explicative (ici, seulement l'âge), et $Y_i$ est la variable réponse (le salaire). Les $X_i$ sont les observations et les $Y_i$ sont les valeurs observées. La fonction $m$ est une fonction de régression inconnue, et $\epsilon_i$ sont des erreurs aléatoires i.i.d. centrées de variance $\sigma^2$. Les $X_i$ et les $\epsilon_i$ sont supposés indépendants.

***Remarque :*** Nous supposerons de plus que $X$ est déterministe.

## 2.1 Régression linéaire simple

Dans le cas de la régression linéaire simple, la fonction de régression est une fonction linéaire de la variable explicative $X_i$. Le modèle est le suivant :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Le paramètre de notre modèle est le vecteur $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$. Pour estimer ce paramètre nous allons utiliser la méthode des moindres carrés ordinaires qui consiste à minimiser $S$ la somme des carrés des résidus, défenie par :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

En résolvant les équations :

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0$$
$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

On obtient finalement :

$$\hat{\beta}1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

Il est facile de verifier que la hessiene en ce point est défenie positive, ce qui valide nos formules (le point est un minimum).

## 2.2 Régression linéaire avec un terme quadratique

Dans ce cas, nous ajoutons un terme quadratique à la fonction de régression, et le modèle devient :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

One pose :

$$X = \begin{bmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

On essaiera de minimiser $\sum_{i=1}^{n}(Y_i - (X_i\beta))^2$ avec une derivée matricielle , on obtient :

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$$

## 2.3 Estimateur de Nadaraya-Watson

L'estimateur de Nadaraya-Watson est un exemple de régression à noyau non paramétrique. La fonction de régression est estimée par moyenne locale :

$$\hat{m}(x) = \sum_{i=1}^{n} W_{hi}(x) Y_i = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

où $K$ est le noyau d'epanachnikov défini par $K(u) = (1 - u^2)\mathbf{1}_{|u| \leq 1}$

, et $h$ est un hyper paramètre qui représente la largeur de la fenêtre (bandwith).

## 2.4 Sélection de la largeur de bande avec la CVLOO pour Nadaraya-Watson

Comme expliqué dans (1), nous allons utiliser le méthode de la validation croisée par Leave-One-Out (CVLOO) pour sélectionner la meilleure largeur de fenêtre (h) dans l'estimateur de Nadaraya-Watson. Cette méthode minimise la formule de la CVLOO donnée par :

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{-i}(X_i))^2 \tag{2.1}$$

où $\hat{m}_{-i}(X_i)$ est l'estimation de la fonction de régression en $X_i$ en laissant le point $i$ hors de l'échantillon, on prend donc :

$$h_{best} = \underset{h>0}{\operatorname{argmin}}\, CV(h)$$

## 2.5 Métrique d'évaluation

Pour évaluer la performance de nos modèles de régression, il existe plusieurs métriques d'évaluation (cités dans (2)), on choisira celle qui convient le mieux :

- **MSE :** définie la moyenne des erreurs quadratique. Une MSE petite signifie que les erreurs entre les valeurs prédites et observées sont petites. Cependant, il est important de garder à l'esprit que la MSE peut être sensible et peut conduire à l'overfiting (sur-apprentissage) car c'est un estimateur biaisé de de l'erreur de prédiction.
- **$\mathbf{R}^2$ :** défini par $R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}$ cette mesure est utilisée pour évaluer (préférable qu'elle soit proche de 1 ) la proportion de la variabilité des données.
- **Validation croisée :** est un estimateur non biaisé de de l'erreur de prédiction et consiste à minimiser (2.1). En comparant les résultats de la CVLOO pour les différents modèles, nous pouvons choisir celui qui a la meilleure performance en termes d'erreur de prédiction sur les données non vues, ça sera donc notre métrique d'évaluation et pour laquelle on se base pour le choix de notre modèle.

# References

[1] P. Cornillon and E. Matzner-Lober. *Régression, théorie et applications.* Springer, Paris, 2011. 4

[2] G. Saporta. *Probabilités, analyse des données et Statistique.* Technip, Paris, 2006. 4