Régression lineaire

Lorenzo Segoni

21 novembre 2025

Table des matières

Chapitre 1

La régression linéaire simple

1.1 Introduction

Le but de la régression linéaire est de modéliser un nuage de points, et donc de trouver la meilleure droite possible pour résumer un nuage de points. On peut avoir deux objectifs différents :

— **Prédiction de l'avenir** (pour avoir une stratégie) : on va le faire en minimisant une perte (*Loss*). On parlera de **Machine Learning**.

— **Trouver la valeur la *plus probable*** (avec des intervalles de confiance) : on va chercher le maximum de vraisemblance. On posera l'hypothèse des moments = **inférence statistique**.

L'hypothèse fondamentale est que le modèle est linéaire.

1.1.1 Le modèle

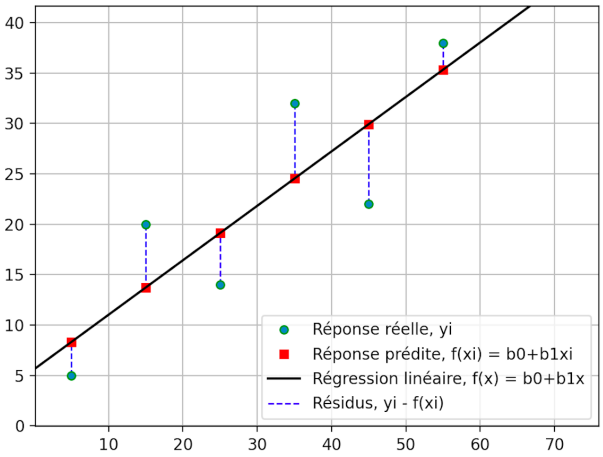

FIGURE 1.1 – Régression Linéaire

Notation 1.1 (Régression linéaire sans aléa)**.** On suppose que la relation entre la variable à prédire y_i et la variable prédictive x_i peut s'écrire sous la forme :

$$y_i \approx \beta_0 + \beta_1 x_i$$

Où β_0 et β_1 sont les paramètres du modèle (ordonnée à l'origine et pente). Ces paramètres sont les paramètres inconnus que l'on cherche.

Limitation de ce modèle : Ce modèle est très limité car il ne prend pas en compte que y_i est affecté par de l'aléa.

Notation 1.2 (Régression linéaire avec aléa)**.** On va introduire cette nouvelle notation :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Où ε_i est la variable aléatoire représentant l'erreur (le *bruit de fond*). Ce que l'on comprend, c'est que ε_i est une variable aléatoire et Y_i dépend directement de ce bruit de fond.

Définition 1.1 (Modèle de régression linéaire simple)**.** Un modèle de régression linéaire simple est défini par une équation de la forme :

$$\forall i \in \{1, ..., n\}, \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

1.1.2 Hypothèses sur l'erreur

On fait les hypothèses suivantes sur les termes d'erreur ε_i (connues sous le nom d'**hypothèses de Gauss-Markov**) : $\forall i \in \{1, ..., n\}$

- $\mathbb{E}[\varepsilon_i] = 0$
- $\mathbb{V}[\varepsilon_i] = \sigma^2$ (même variance finie)
- $\forall j \neq i, \text{Cov}[\varepsilon_i, \varepsilon_j] = 0$ (indépendance)

Les hypothèses de l'espérance et de la variance s'appellent **Homoscédasticité**.

La variance et la covariance peuvent s'écrire de façon condensée :

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{i,j}\sigma^2 \text{ pour tout couple } (i, j)$$

Car $\text{Cov}(\varepsilon_i, \varepsilon_i) = \mathbb{V}[\varepsilon_i]$

Le symbole de Kronecker est défini par :

$$\delta_{i,j} = \begin{cases} 1 \text{ si } j = i \\ 0 \text{ si } i \neq j \end{cases}$$

Définition 1.2 (Hypothèses sur l'erreur)**.** Les hypothèses de **Gauss-Markov** sont :

$$(\mathcal{H}) : \begin{cases} (\mathcal{H}_1) : & \mathbb{E}[\varepsilon_i] = 0 \quad \text{pour tout indice } i \\ (\mathcal{H}_2) : & \text{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{i,j}\sigma^2 \quad \text{pour tout couple } (i,j) \end{cases}$$

Objectif 1.1. Trouver les valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$. On note que la notation $\hat{}$ (chapeau) indique un **estimateur** calculé à partir des données observées.

1.2 Moindres carrée ordinaires

Notation 1.3. Les points (x_i, y_i) étant donnés, le but est de trouver un moyen pour calculer la distance entre les points et la prédiction. Cette fonction permet de trouver la fonction affine d telle que :

$$\sum_{i=1}^{n} d\left((x_i, y_i), (x_i, \beta_0' + \beta_1' x_i)\right)$$

La fonction d doit être choisie préalablement pour bien résoudre le problème.

On a :

— **Moindre Valeur Absolue** $\sum_{i=1}^{n} |\beta_0' + \beta_1' x_i - y_i|$: Ceci est une des fonctions avec les meilleurs scores, mais le problème est que la fonction est très difficile à traiter car il y a un problème de dérivabilité au point $(0, 0)$.

— **Moindres Carrés** [1] $\sum_{i=1}^{n} (\beta_0' + \beta_1' x_i - y_i)^2$: Ceci a l'avantage de rendre les calculs faciles, car c'est une fonction facilement dérivable et continue, sans point *problématique*. On va la noter $L(y_i, \beta_0', \beta_1') = \sum_{i=1}^{n} (\beta_0' + \beta_1' x_i - y_i)^2$

1. Terminologie due à Legendre dans un article de 1805 sur la détermination des orbites des comètes

Définition 1.3 (Estimateurs des Moindres Carrés Ordinaires)**.** On appelle estimateurs des Moindres Carrés Ordinaires (en abrégé MCO) $\hat{\beta}_0$ et $\hat{\beta}_1$ les valeurs minimisant la quantité :

$$L(y_i, \beta_0', \beta_1') = \sum_{i=1}^{n} (\beta_0' + \beta_1' x_i - y_i)^2$$

1.2.1 Calcul des estimateurs de β_0 et β_1

On a vu donc les moindres carrés ordinaires. On a compris donc que l'on veut minimiser la fonction L. On a donc :

Définition 1.4. Les estimateurs de β_0 et β_1 sont :

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0' \in \mathbb{R}, \beta_1' \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta_0' - \beta_1' x_i)^2$$

Une première chose est de montrer :

Proposition 1.1. *Soit*

$$L(\beta_0', \beta_1') = \sum_{i=1}^{n} (\beta_0' + \beta_1' x_i - y_i)^2,$$

considérée comme fonction de $(\beta_0', \beta_1') \in \mathbb{R}^2$ (les x_i et y_i étant fixés). Alors L est convexe. De plus L est strictement convexe si et seulement si les abscisses $x_1, \ldots, x_n$ ne sont pas toutes égales.

Ceci n'est pas important dans le cours, mais nous donne des spécificités importante de la fonction

Démonstration. Posons $\beta' = (\beta_0', \beta_1')^\top \in \mathbb{R}^2$ et pour tout i définissons le vecteur $z_i = (1, x_i)^\top$. On a

$$L(\beta') = \sum_{i=1}^{n} (z_i^\top \beta' - y_i)^2.$$

Calculons la matrice Hessienne de L (matrice des dérivées secondes) : pour tout β',

$$\nabla^2 L(\beta') = 2 \sum_{i=1}^{n} z_i z_i^\top = 2 \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

Cette matrice est de la forme $2X^\top X$ (où X est la matrice de conception à deux

colonnes), donc est symétrique et positive semi-définie. En effet, pour tout vecteur non nul $u = (u_0, u_1)^\top \in \mathbb{R}^2$,

$$u^\top \nabla^2 L(\beta')\, u = 2\sum_{i=1}^{n}(u_0 + u_1 x_i)^2 \geq 0.$$

Ainsi $\nabla^2 L(\beta') \geq 0$, ce qui montre que L est convexe.
Par ailleurs, l'inégalité est stricte (i.e. $u^\top \nabla^2 L(\beta')\, u > 0$ pour tout $u \neq 0$) si et seulement si il n'existe pas de vecteur non nul u tel que $u_0 + u_1 x_i = 0$ pour tout i. Cela équivaut exactement au fait que les x_i ne soient pas tous égaux. Donc sous l'hypothèse « les x_i ne sont pas tous égaux », la matrice $\nabla^2 L(\beta')$ est définie positive et L est strictement convexe. □

Si on développe cette expression, on arrive à cette proposition :

Proposition 1.2 (Estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$)**.** *Les estimateurs des MCO ont pour expressions :*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\, \bar{x},$$

et

$$\hat{\beta}_1 = \frac{s_{x,y}^2}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

avec :

$$\begin{cases} \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \quad (\textit{moyenne empirique de } x) \\ \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i \quad (\textit{moyenne empirique de } y) \\ s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \quad (\textit{(co-)variance empirique}) \\ s_{x,y}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (\textit{(co-)variance empirique}) \end{cases}$$

On a :

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \\ &= \sum_{i=1}^{n} x_i - n \cdot \bar{x} \\ &= \sum_{i=1}^{n} x_i - n \cdot \frac{1}{n} \sum_{i=1}^{n} x_i \\ &= 0\end{aligned}$$

De même pour : $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$

Démonstration. La premiere methode utilise donc ce que on a montré avant, c'est à dire que L est strictement convexe.

Puisque L est structement convexe, elle adment un minimum en un unique point $(\hat{\beta}_1, \hat{\beta}_2)$, lequel est déterminé en annulant les dérivées partielles de L.

On obtient les équations :

$$\begin{cases} \frac{\partial L}{\partial \beta_0'}(\beta_0', \beta_1') = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \\ \frac{\partial L}{\partial \beta_1'}(\beta_0', \beta_1') = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \end{cases}$$

La première équation donne :

$$\begin{aligned}\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad &\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \\ &\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

La seconde équation donne :

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

En remplaçant $\hat{\beta}_0$ par son expression, nous obtenons :

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_i x_i y_i - \sum_i x_i \bar{y}}{\sum_i x_i^2 - \sum_i x_i \bar{x}} \\ &= \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})} \\ &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{s_{x,y}^2}{s_x^2}\end{aligned}$$

La seconde méthode consiste à appliquer la technique de Gauss de réduction des formes quadratiques, c'est-à-dire à décomposer $L(\beta_1, \beta_2)$ en somme de carrés. Après calculs, on obtient :

$$S(\beta_0, \beta_1) = n\left(\beta_0 - (\bar{y} - \beta_1 \bar{x})\right)^2 + \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\beta_1 - \hat{\beta}_1\right)^2 + \left(\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Le dernier terme est indépendant de β_0 et β_1. Le second est nul si et seulement si $\beta_1 = \hat{\beta}_1$, puis le premier est nul si et seulement si $\beta_0 = \hat{\beta}_0$.

□

Important : ces expressions sont indépendantes de l'hypothèse $(\mathcal{H})$. Celles-ci vont en fait servir dans la suite à expliciter les propriétés statistiques de ces estimateurs.

1.2.2 Quelques propriétés des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$.

Théorème 1.1 (Estimateurs sans biais)**.** *$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1 sous l'hypothèse de $\mathbb{E}[\varepsilon_i] = 0$, c'est à dire :*

$$\mathbb{E}[\hat{\beta}_0] = \beta_0 \quad \mathbb{E}[\hat{\beta}_1] = \beta_1$$

Démonstration. Partons par $\hat{\beta}_1$:

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}[(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Or

$$\begin{cases} \mathbb{E}[y_i] = \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 x_i + \mathbb{E}[\varepsilon_i] = \beta_0 + \beta_1 x_i \\ \mathbb{E}[y_i - \bar{y}] = \mathbb{E}[y_i] - \mathbb{E}[\bar{y}] = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \bar{x}) = \beta_1(x_i - \bar{x}) \end{cases}$$

Donc

$$\mathbb{E}[\hat{\beta}_1] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}[(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_1(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1$$

On en déduit que :

$$\begin{aligned}\mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\bar{y} - \hat{\beta}_1\bar{x}] \\ &= \mathbb{E}[\bar{y}] - \mathbb{E}[\hat{\beta}_1]\bar{x} \\ &= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} \\ &= \beta_0\end{aligned}$$

□

On peut aussi montrer que

$$\begin{aligned}\hat{\beta}_1 &= \frac{1}{n \cdot s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n \cdot s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i - [\beta_0 + \beta_1\bar{x} + \bar{\varepsilon}]) \\ &= \beta_1 + \frac{1}{n \cdot s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\end{aligned}$$

Ce qui nous permet de montrer que $\hat{\beta}_1$ est un estimateur sans biais

Théorème 1.2 (Variances et covariance)**.** *Les variances des estimateurs sont (sous les hypothèses $(\mathcal{H})$) :*

$$\mathbb{V}[\hat{\beta}_0] = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right) \quad \mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{n}\frac{1}{s_x^2}$$

tandis que leur covariance vaut :

$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\sigma^2}{n}\frac{\bar{x}}{s_x^2}$$

Démonstration. On a :

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_1] &= \mathbb{V}\left[\beta_1 + \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right] \\
&= \frac{1}{(ns_x^2)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\mathbb{V}[\varepsilon_1] \quad \text{par independence des } \varepsilon_i \\
&= \sigma^2\frac{ns_x^2}{(ns_x^2)^2} \\
&= \frac{\sigma^2}{n}\frac{1}{s_x^2}
\end{aligned}$$

Et on a aussi :

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_0] &= \mathbb{V}[\bar{y} - \hat{\beta}_1\bar{x}] \\
&= \mathbb{V}[\bar{y}] + \mathbb{V}[\hat{\beta}_1\bar{x}] - 2\text{Cov}[\bar{y}, \hat{\beta}_1\bar{x}] \\
&= \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n}\varepsilon_i\right] \\
&= \bar{x}^2\mathbb{V}[\hat{\beta}_1] \\
&= \frac{1}{n}\sigma^2 + \frac{\bar{x}^2\sigma^2}{ns_x^2} \\
&= \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right)
\end{aligned}$$

Attention $\bar{y}$ et $\hat{\beta}_1$ peuvent être corrélé (à cause ε_i)

$$\begin{aligned}
\text{Cov}[\bar{y}, \hat{\beta}_1] &= \text{Cov}\left[\beta_0 + \beta_1\bar{x} + \bar{\varepsilon}, \beta_1 + \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right] \\
&= \text{Cov}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i, \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right] \\
&= \frac{1}{(ns_x^2)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}[\varepsilon_i, (x_j - \bar{x})\varepsilon_j] \\
&= \frac{1}{(ns_x^2)^2}\sum_{i=1}^{n}\text{Cov}[\varepsilon_i, (x_i - \bar{x})\varepsilon_i] \\
&= 0
\end{aligned}$$

Et on a aussi :

$$\begin{aligned}\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] &= \mathrm{Cov}[\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1] \\ &= \mathrm{Cov}[\bar{y}, \hat{\beta}_1] - \bar{x}\mathrm{Cov}[\hat{\beta}_1, \hat{\beta}_1] \\ &= 0 - \bar{x}\frac{\sigma^2}{ns_x^2} \\ &= -\frac{\sigma^2}{n}\frac{\bar{x}}{s_x^2}\end{aligned}$$

□

1.2.3 Théorème Gauss-Markov

Théorème 1.3 (Gauss-Markov)**.** *Sous les hypothèses du modèle, les estimateurs MCO* $(\hat{\beta}_0, \hat{\beta}_1)$ *sont les estimateurs* ***linéaires*** *et* ***biaisés*** *de* β_0, β_1 *ayant la variance minimale.*

Autrement dit, on dit que les estimateurs OLS sont ***BLUE*** *:*

Best Linear Unbiased Estimators.

Démonstration. On peut écrire que :

$$\hat{\beta}_1 = \sum_{i=1}^{n} p_i y_i \quad \text{avec} \quad p_i = \frac{x_i - \bar{x}}{ns_x^2}$$

On va d'abord construire un autre $\tilde{\beta}_1$ comme un autre estimateur sans biais linéaire. Pour que soit linéaire on doit construire

$$\tilde{\beta}_1 = \sum_{i=1}^{n} q_i y_i$$

Montrons que $\sum_{i=1}^{n} q_i = 0$ et $\sum_{i=1}^{n} q_i x_i = 1$.

$$\begin{aligned}\mathbb{E}[\tilde{\beta}_1] &= \sum_{i=1}^{n} \mathbb{E}[q_i y_i] \\ &= \sum_{i=1}^{n} \mathbb{E}[q_i(\beta_0 + \beta_1 x_i + \varepsilon_i)] \\ &= \beta_0 \sum_{i=1}^{n} q_i + \beta_1 \sum_{i=1}^{n} q_i x_i \\ &= \beta_1\end{aligned}$$

Pour que cette estimateur soit sans bias, on doit donc poser $\sum_{i=1}^{n} q_i = 0$ et $\sum_{i=1}^{n} q_i x_i =$

1.

Montrons que $\mathbb{V}[\tilde{\beta}_1] \geq \mathbb{V}[\hat{\beta}_1]$. Donc on a :

$$\begin{aligned}\mathbb{V}[\tilde{\beta}_1] &= \mathbb{V}[\tilde{\beta}_1 - \hat{\beta}_1 + \hat{\beta}_1] \\ &= \mathbb{V}[\tilde{\beta}_1 - \hat{\beta}_1] + \mathbb{V}[\hat{\beta}_1] + 2\mathrm{Cov}[\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1] \\ &= \mathbb{V}[\tilde{\beta}_1 - \hat{\beta}_1] + \mathbb{V}[\hat{\beta}_1]\end{aligned}$$

Car

$$\begin{aligned}\mathrm{Cov}[\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1] &= \mathrm{Cov}[\tilde{\beta}_1, \hat{\beta}_1] - \mathbb{V}[\hat{\beta}_1] \\ &= \sum_{i=1}^{n} p_i q_i \sigma^2 - \frac{\sigma^2}{n s_x^2} \\ &= \frac{\sigma^2}{n s_x^2}\left(\sum_{i=1}^{n} q_i x_i - \sum_{i=1}^{n} q_i \bar{x} - 1\right) \\ &= 0\end{aligned}$$

Puisque $\mathbb{V}[\tilde{\beta}_1 - \hat{\beta}_1] \geq 0$ par la propriété de la variance, on a alors que :

$$\mathbb{V}[\tilde{\beta}_1] \geq \mathbb{V}[\hat{\beta}_1]$$

On peut faire de même pour $\hat{\beta}_0$. □

Remarque 1.1. Dans le cadre du modèle linéaire simple $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, avec $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ indépendants, les estimateurs des moindres carrés ordinaires (MCO) présentent des propriétés de dispersion étroitement liées à la répartition des observations en x_i.

La variance de l'estimateur de la pente s'écrit :

$$\mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{n s_{\mathbf{x}}^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Ainsi, plus les x_i sont dispersés autour de leur moyenne, plus la variance de $\hat{\beta}_1$ est faible : la pente est alors estimée avec davantage de précision. On dit que les observations exercent un *leverage* plus fort sur l'estimation du coefficient directeur.

Concernant l'ordonnée à l'origine, on a :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \mathbb{V}[\hat{\beta}_0] = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_{\mathbf{x}}^2}\right).$$

Lorsque la moyenne des x_i est centrée en zéro, c'est-à-dire $\bar{x} = 0$, l'estimateur devient $\hat{\beta}_0 = \bar{y}$ et sa variance se simplifie en $\text{Var}[\hat{\beta}_0] = \sigma^2/n$. Dans ce cas, l'ordonnée à l'origine correspond à l'estimation directe de la moyenne de la variable Y dans la population.

Enfin, les deux estimateurs sont corrélés négativement :

$$\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] \leq 0.$$

En effet, pour $\bar{x} > 0$, une augmentation de la pente $\hat{\beta}_1$ tend à diminuer l'ordonnée à l'origine $\hat{\beta}_0$. Par ailleurs, lorsque la dispersion des x_i (mesurée par $s_{\mathbf{x}}^2$) est fixe, les variances de $\hat{\beta}_0$ et $\hat{\beta}_1$ décroissent proportionnellement à $1/n$, illustrant le gain de précision avec la taille de l'échantillon.

1.2.4 Prévision

Notation 1.4. À partir de l'échantillon d'apprentissage (x_i, y_i) pour $i = 1, \ldots, n$, on construit les estimateurs MCO $(\hat{\beta}_0, \hat{\beta}_1)$.

On considère une nouvelle valeur x_{n+1} et la variable aléatoire associée y_{n+1} (inconnue). Le prédicteur naturel est :

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}.$$

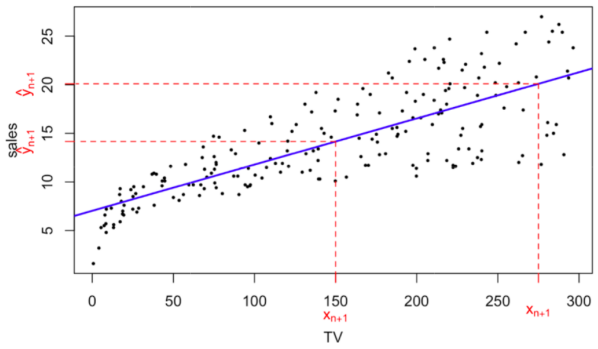

FIGURE 1.2 – Estimation du point n+1

Hypothèse 1.1. On suppose que (x_{n+1}, y_{n+1}) suit le même modèle que les données d'apprentissage, c'est-à-dire :

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1},$$

avec

$$(\mathbb{h}) : \begin{cases} \mathbb{E}[\varepsilon_{n+1}] = 0 \\ \mathbb{V}(\varepsilon_{n+1}) = \sigma^2 \\ \mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0 \quad (i = 1, \ldots, n). \end{cases}$$

Proposition 1.3 (Erreur de prédiction)**.** *On définit*

$$\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1}.$$

Où :

— y_{n+1} *est la vrai valeur inconnue*

— $\hat{y}_{n+1}$ *est le predicteur*

On a :

$$\begin{cases} \mathbb{E}[\hat{\varepsilon}_{n+1}] & = 0 \\ \mathbb{V}[\hat{\varepsilon}_{n+1}] & = \theta^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\, s_x^2}\right) \end{cases}$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ *et* $s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.

Démonstration. **Espérance.**

On a

$$\mathbb{E}[\hat{\varepsilon}_{n+1}] = \mathbb{E}[y_{n+1}] - \mathbb{E}[\hat{y}_{n+1}].$$

Or,

$$\begin{cases} \mathbb{E}[y_{n+1}] = \mathbb{E}[\beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}] = \beta_0 + \beta_1 x_{n+1} + \mathbb{E}[\varepsilon_{n+1}] = \beta_0 + \beta_1 x_{n+1}. \\ \mathbb{E}[\hat{y}_{n+1}] = \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}] = \mathbb{E}[\hat{\beta}_0] + x_{n+1}\,\mathbb{E}[\hat{\beta}_1] = \beta_0 + \beta_1 x_{n+1}. \end{cases}$$

Ainsi :

$$\boxed{\mathbb{E}[\hat{\varepsilon}_{n+1}] = 0.}$$

Variance.

$$\mathbb{V}(\hat{\varepsilon}_{n+1}) = \mathbb{V}(y_{n+1} - \hat{y}_{n+1}) = \mathbb{V}(y_{n+1}) + \mathbb{V}(\hat{y}_{n+1}) - 2\,\mathrm{Cov}(y_{n+1}, \hat{y}_{n+1}).$$

Or y_{n+1} dépend uniquement de ε_{n+1} et $\hat{y}_{n+1}$ dépend uniquement de $(\varepsilon_1, \ldots, \varepsilon_n)$, donc

$$\mathrm{Cov}(y_{n+1}, \hat{y}_{n+1}) = 0.$$

On a aussi :

$$\begin{cases} \mathbb{V}(Y_{n+1}) = \mathbb{V}(\beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}) = \mathbb{V}(\varepsilon_{n+1}) = \sigma^2. \\ \mathbb{V}(\hat{Y}_{n+1}) = \mathbb{V}(\hat{\beta}_0 + x_{n+1}\hat{\beta}_1). = \mathbb{V}(\hat{\beta}_0) + x_{n+1}^2\,\mathbb{V}(\hat{\beta}_1) + 2x_{n+1}\,\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1). \end{cases}$$

D'après les formules connues :

$$\mathbb{V}(\hat{\beta}_0) = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right), \quad \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} \cdot \frac{1}{s_x^2}, \quad \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2}{n} \cdot \frac{\bar{x}}{s_x^2}.$$

Ainsi :

$$\begin{aligned} \mathbb{V}(\hat{Y}_{n+1}) &= \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right) + x_{n+1}^2 \cdot \frac{\sigma^2}{n s_x^2} - 2x_{n+1} \cdot \frac{\sigma^2}{n}\frac{\bar{x}}{s_x^2} \\ &= \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2} + \frac{x_{n+1}^2}{s_x^2} - \frac{2\bar{x}x_{n+1}}{s_x^2}\right) \\ &= \frac{\sigma^2}{n}\left(1 + \frac{(x_{n+1} - \bar{x})^2}{s_x^2}\right) \end{aligned}$$

Enfin :

$$\boxed{\mathbb{V}(\hat{\varepsilon}_{n+1}) = \sigma^2 + \mathbb{V}(\hat{Y}_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\, s_x^2}\right).}$$

□

Interprétation : Le terme

$$(x_{n+1} - \bar{x})^2$$

contrôle l'incertitude de la prédiction :

- Il est **petit** lorsque x_{n+1} est proche de la moyenne $\bar{x}$: le point à prédire est "semblable" aux données déjà observées. $\Rightarrow$ la variance de la prédiction est réduite.
- Il est **grand** lorsque x_{n+1} est loin de $\bar{x}$: le point à prédire est très différent de l'échantillon d'apprentissage. $\Rightarrow$ l'incertitude de la prédiction est plus élevée.

1.2.5 Calcul des résidus et de la variance résiduelle

Définition 1.5. On définit les valeurs ajustées et les résidus :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad \hat{\varepsilon}_i = Y_i - \hat{Y}_i.$$

Proposition 1.4. *Avec cette construction, la somme des résidus est nulle, c'est à dire,*

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

Démonstration. On a :

$$\begin{aligned}
\sum_{i=1}^{n} \hat{\varepsilon}_i &= \sum_{i=1}^{n} y_i - \hat{y}_i \\
&= \sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{car } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\
&= \sum_{i=1}^{n} (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \\
&= 0
\end{aligned}$$

□

Notation 1.5. On appelle **SCR** (somme des carrés des résidus) :

$$\begin{aligned}
\text{SCR} &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \min_{\beta_0', \beta_1'} \sum_{i=1}^{n} (y_i - \beta_0' - \beta_1' x_i)^2.
\end{aligned}$$

Définition 1.6. La variance empirique des x_i est

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Donc l'estimateur de la variance de l'erreur est donné par

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-2}\,\text{SCR}.$$

Théorème 1.4. *La statistique $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.*

Démonstration. Ce qui nous intéresse le plus est $\mathbb{E}[\sum_{i=1}^{n}\hat{\varepsilon}_i^2]$:

On a :

$$\begin{aligned}
\mathbb{V}[\hat{\varepsilon}_i^2] &= \mathbb{V}[y_i - \hat{y}_i] \\
&= \mathbb{V}[\beta_0 + \beta_1 x_i + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] \\
&= \mathbb{V}[\varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] \\
&= \mathbb{V}[\varepsilon_i] + \mathbb{V}[\hat{\beta}_0 + \hat{\beta}_1 x_i] - 2\text{Cov}[\varepsilon_i, \hat{\beta}_0 + \hat{\beta}_1 x_i]
\end{aligned}$$

On calcule :

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_0 + \hat{\beta}_1 x_i] &= \mathbb{V}[\bar{y} + \hat{\beta}_1(x_i - \bar{x}) \\
&= \mathbb{V}[\bar{y}] + \mathbb{V}[\hat{\beta}_1(x_i - \bar{x})] \quad \text{car } \text{Cov}[\bar{y}, \hat{\beta}_1] = 0 \\
&= \frac{\sigma^2}{n} + (x_i - \bar{x})^2 \frac{\sigma^2}{n}\frac{1}{s_x^2}
\end{aligned}$$

et

$$\begin{aligned}
\text{Cov}[\varepsilon_i, \hat{\beta}_0 + \hat{\beta}_1 x_i] &= \text{Cov}[\bar{y}, \varepsilon_i] + \text{Cov}[\hat{\beta}_1(x_i - \bar{x}), \varepsilon_i] \\
&= \frac{\sigma^2}{n} + (x_i - \bar{x})\frac{1}{n s_x^2}(x_i - \bar{x})\sigma^2
\end{aligned}$$

Ceci nous donne :

$$\begin{aligned}
\mathbb{V}[\hat{\varepsilon}_i^2] &= \mathbb{V}[\varepsilon_i] + \mathbb{V}[\hat{\beta}_0 + \hat{\beta}_1 x_i] - 2\text{Cov}[\varepsilon_i, \hat{\beta}_0 + \hat{\beta}_1 x_i] \\
&= \sigma^2 - \frac{\sigma^2}{n} - \frac{\sigma^2(x_i - \bar{x})^2}{n s_x^2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{n} \hat{\varepsilon}_i^2] &= \sum_{i=1}^{n} \mathbb{E}[\hat{\varepsilon}_i^2] \\
&= \sum_{i=1}^{n} \mathbb{V}[\hat{\varepsilon}_i] \\
&= \sum_{i=1}^{n} \left(\sigma^2 - \frac{\sigma^2}{n} - \frac{\sigma^2 (x_i - \bar{x})^2}{n s_x^2} \right) \\
&= n\sigma^2 - \sigma^2 - \sigma^2 \\
&= (n-2)\sigma^2
\end{aligned}$$

Ceci nous donne :

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}[\frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2] \\
&= \frac{1}{n-2} \mathbb{E}[\sum_{i=1}^{n} \hat{\varepsilon}_i^2] \\
&= \boxed{\sigma^2}
\end{aligned}$$

□

Attention : le dénominateur est bien $n-2$ (et non n), car deux paramètres $\hat{\beta}_0$ et $\hat{\beta}_1$ ont été estimés.

1.3 interprétations géométrique

1.3.1 Formulation du modèle

On considère le modèle linéaire simple reliant une variable expliquée y_i à une variable explicative x_i :

$$\begin{cases}
y_1 = \beta_0 \cdot 1 + \beta_1 x_1 + \varepsilon_1, \\
y_2 = \beta_0 \cdot 1 + \beta_1 x_2 + \varepsilon_2, \\
\vdots \\
y_i = \beta_0 \cdot 1 + \beta_1 x_i + \varepsilon_i, \\
\vdots \\
y_n = \beta_0 \cdot 1 + \beta_1 x_n + \varepsilon_n.
\end{cases}$$

Notation 1.6. Pour simplifier cette écriture, on introduit les notations vectorielles suivantes :

— $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$,

— $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$,

— $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

— $\mathbb{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$

Avec ces notations, le modèle s'écrit de façon compacte :

$$y = \beta_0 \cdot \mathbb{1} + \beta_1 \cdot x + \varepsilon.$$

Définition 1.7. On définit le sous-espace vectoriel engendré par les deux vecteurs explicatifs :

$$\mathcal{M}(x) = \text{Vect}(\mathbb{1}, x).$$

Toute combinaison linéaire $\tilde{y} = \beta_0 \mathbb{1} + \beta_1 x$ appartient à cet espace.

Définition 1.8. La projection orthogonale du vecteur y sur $M(x)$ correspond alors au vecteur des valeurs ajustées :

$$\text{Proj}_{\mathcal{M}(x)} y = \arg \min_{\tilde{y} \in \mathcal{M}(x)} \|y - \tilde{y}\|^2.$$

Démonstration. En effet :

$$\begin{aligned}
\text{Proj}_{\mathcal{M}(x)} y &= \arg \min_{\tilde{y} \in \mathcal{M}(x)} \|y - \tilde{y}\|^2. \\
&= \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \|y - (\beta_0 \mathbb{1} + \beta_1 x)\|^2 \\
&= \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \\
&= \hat{y}
\end{aligned}$$

□

On a donc que $\hat{y}$ est la projection de y sur $\mathcal{M}(x)$

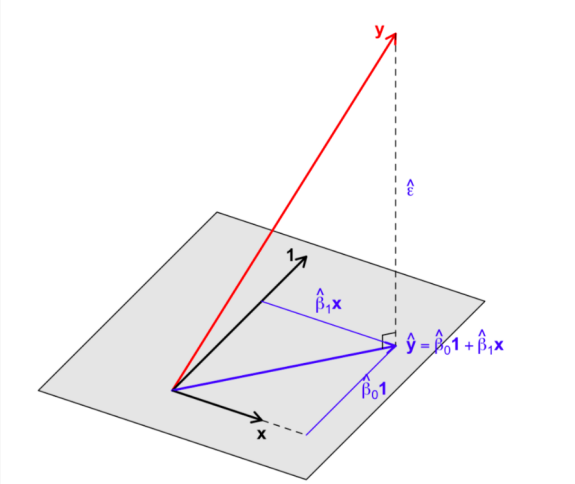

FIGURE 1.3 – Projection

Rappel 1.1. On note :

$$\|z\|^2 = \sum_{i=1}^{n} z_i^2 \quad \text{et} \quad \langle z, w \rangle = \sum_{i=1}^{n} z_i w_i$$

le produit scalaire et la norme euclidienne usuels de $\mathbb{R}^n$.

Notation 1.7. Les estimateurs $(\hat{\beta}_0, \hat{\beta}_1)$ sont définis comme les valeurs des paramètres minimisant la somme des carrés des écarts entre les observations et le modèle :

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0', \beta_1' \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \beta_0' - \beta_1' x_i)^2.$$

Cette expression peut se reformuler à l'aide du vecteur $\mathbb{1}$ et de la norme eucli-

dienne :

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0', \beta_1' \in \mathbb{R}} \|y - \beta_0' \mathbb{1} - \beta_1' x\|^2,$$

1.3.2 Le coefficient de détermination R^2

Avec la notation d'avant on a :

Notation 1.8. — **SCT (Somme Totale des Carrés)** est la mesure de la variabilité totale dans la réponse y avant que la régression ne soit effectuée.

$$\text{SCT} = \|y - \bar{y}\,\mathbb{1}\|^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

— **SCR (Somme des Carrés des Résidus)** est la mesure de la variabilité qui reste après avoir effectué la régression (c'est la variabilité non expliquée).

$$\text{SCR} = \|y - \hat{y}\|^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

— **SCE (Somme des Carrés Expliquée)** est la mesure de la variabilité qui est expliquée par la régression.

$$\text{SCE} = \|\hat{y} - \bar{y}\,\mathbb{1}\|^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Proposition 1.5. *On a :*

$$\text{SCT} = \text{SCE} + \text{SCR}$$

Démonstration. On a par le théorème de Pythagore :

$$\begin{aligned}
\text{SCT} &= \|y - \bar{y}\mathbb{1}\|^2 \\
&= \|\hat{y} - \bar{y}\mathbb{1} + y - \hat{y}\|^2 \\
&= \|\hat{y} - \bar{y}\mathbb{1} + \hat{\varepsilon}\|^2 \\
&= \|\hat{y} - \bar{y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2 \\
&= \text{SCE} + \text{SCR}
\end{aligned}$$

□

Définition 1.9. Le coefficient de détermination R^2 est définie par :

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = \frac{\|\hat{y} - \bar{y}\mathbb{1}\|^2}{\|y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{\text{SCR}}{\text{SCT}}$$

R^2 est la proportion de la variabilité de y qui peut être expliquée par la régression. (ou encore le cosinus carrées de θ)

De façon schématique, on peut différencier les cas suivants :

— Si $R^2 = 1$, le modèle explique tout, l'angle θ vaut zéro et y est dans $\mathcal{M}(X)$, c'est-à-dire que $y_i = \beta_0 + \beta_1 x_i$ pour tout i : les points de l'échantillon sont parfaitement alignés sur la droite des moindres carrés ;

— Si $R^2 = 0$, cela veut dire que $\sum(\hat{y}_i - \bar{y})^2 = 0$, donc $\hat{y}_i = \bar{y}$ pour tout i. Le modèle de régression linéaire est inadapté puisqu'on ne modélise rien de mieux que la moyenne ;

— Si R^2 est proche de zéro, cela veut dire que y est quasiment dans l'orthogonal de $\mathcal{M}(X)$, le modèle de régression linéaire est inadapté, la variable x n'explique pas bien la variable réponse y (du moins pas de façon affine).