Régression linéaire

Frédéric Sur
Mines Nancy

30 mai 2026

Note : les *remarques* sont des compléments d'information.

1 Introduction

La modélisation statistique a pour objectif de formaliser et de quantifier les relations existant entre différentes variables mesurées sur une population d'individus. L'intérêt est, d'une part, la modélisation en elle-même, qui permet de mieux comprendre un phénomène de par les relations établies, et, d'autre part, la capacité de faire des prédictions (des **inférences**) grâce au modèle. Parmi les méthodes statistiques, la **régression linéaire** occupe une place prépondérante pour des raisons historiques et fondamentales. Elle est sans doute le modèle le plus simple d'inférence d'une variable quantitative, et, est, à ce titre, le premier modèle vu dans tout cours d'« apprentissage automatique » (la science de l'« intelligence artificielle »).

L'idée est d'expliquer les variations d'une variable quantitative d'intérêt, appelée **variable expliquée**, **dépendante** ou **cible** (notée Y car il s'agira d'une variable aléatoire), à l'aide d'une ou plusieurs autres variables quantitatives ou qualitatives, appelées **variables explicatives**, **indépendantes** ou **régresseurs**. Le qualificatif « linéaire » signifie que le modèle est construit comme une combinaison linéaire des *paramètres* à estimer. Les relations entre la variable expliquée et les variables explicatives ne sont elles-mêmes pas nécessairement linéaires. Cette hypothèse permet des développements mathématiques élégants et d'une puissance prédictive très intéressante. Dans le cadre de ce cours, on se limitera au cas d'une seule variable explicative, appelé **régression linéaire simple**. Lorsque plusieurs variables explicatives sont considérées, on parle de **régression linéaire multiple**.

Nous nous intéresserons donc à des relations du type : $y = \beta_0 + \beta_1 x$, ou $y = \beta_0 + \beta_1 f(x)$ où f est une fonction de x (se ramène au premier cas en posant $x' = f(x)$), mais pas à des relations comme par exemple $y = \beta_0 e^{\beta_1 x}$ qui n'est pas linéaire en les coefficients β_0 et β_1. Bien entendu, dans ce dernier exemple il est possible de se ramener au cas linéaire par une transformation logarithmique de y.

La figure 1 donne un exemple d'observations (x_i, y_i), et un modèle linéaire qui permettrait d'expliquer y en fonction de x. Évidemment, les observations ne suivent pas rigoureusement le modèle linéaire. Le bon cadre d'étude est celui de la statistique : on va modéliser les écarts au modèle linéaire idéal par des fluctuations aléatoires. L'intérêt d'un tel modèle est double : tout d'abord en terme de modélisation du phénomène si on arrive à le valider sous certaines hypothèses statistiques, et aussi en terme d'inférence car le modèle permettrait de prédire de nouvelles valeurs y correspondant à des x pour lesquelles on n'a pas d'observations.

2 La régression linéaire simple

Le cas de la **régression linéaire simple** met en jeu une unique variable explicative x pour expliquer la variable Y. Cette deuxième variable est écrite en majuscule car nous allons la considérer comme aléatoire.

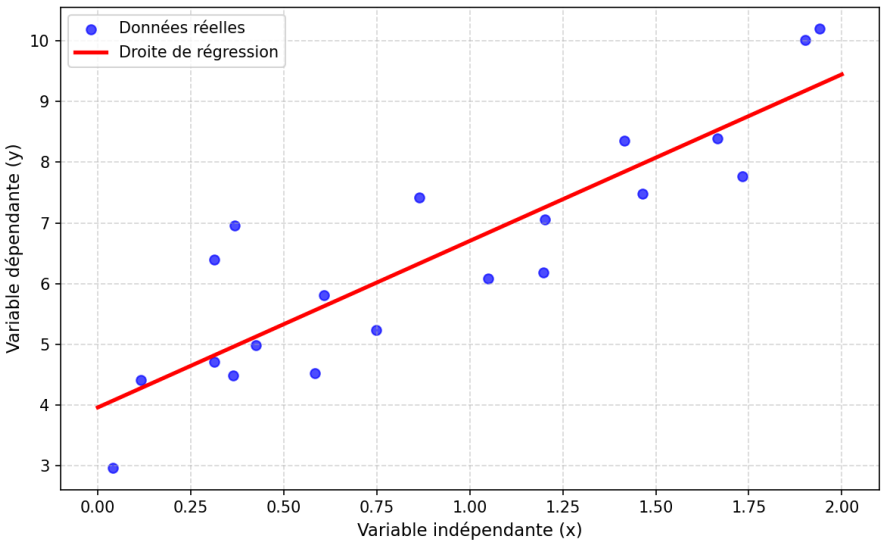

FIGURE 1 – Un exemple de régression linéaire.

2.1 Définition mathématique du modèle

On suppose disposer d'un n-échantillon, noté $(x_i, Y_i)_{1 \leqslant i \leqslant n}$: autrement dit, on supposera (ce sera une conséquence des hypothèses H1 à H4 ci-dessous) les variables aléatoires indépendantes entre elles et identiquement distribuées.

Le modèle de régression linéaire simple postule que pour chaque individu (ou observation) i, la relation entre Y_i et x_i s'écrit :

$$\forall\, i \in \{1, \ldots, n\}, \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

où :

- x_i est la valeur prise par la variable explicative pour l'observation i. Elle est ici supposée déterministe (fixée par l'expérimentateur ou observée sans erreur) ;
- β_0 est la constante du modèle (ordonnée à l'origine, *intercept*) ;
- β_1 est le paramètre de pente, traduisant l'effet marginal de x sur Y ;
- ε_i désigne la **perturbation aléatoire** ou le terme d'erreur. Cette variable aléatoire capte tout ce qui s'écarte de la relation linéaire exacte : erreurs de mesure, omission d'autres variables, résidu intrinsèquement aléatoire du comportement humain ou physique…
- Y_i est la valeur prise par la variable expliquée pour l'observation i : Y_i est donc considéré comme une variable aléatoire, fonction de ε_i.

Ici, les paramètres du modèle β_0 et β_1 sont fixés mais inconnus. Tout l'enjeu est d'en construire des estimateurs statistiques.

Comme d'habitude, en pratique on observe une réalisation $(x_i, y_i)_{1 \leqslant i \leqslant n}$ du n-échantillon, comme dans la figure 1, à partir de laquelle on va déterminer des estimations des paramètres du modèle. Si β_0' et β_1' sont ces estimations à partir des observations, la **droite de régression** a pour équation : $y = \beta_0' + \beta_1' x$, c'est celle qui est tracée sur la figure 1.

2.2 Modélisation probabiliste

Pour mener à bien l'estimation des paramètres du modèle et l'inférence (prédiction), on formule un ensemble d'hypothèses sur le comportement des perturbations aléatoires ε_i :

H1 (moyenne nulle des perturbations) L'espérance des erreurs est nulle : $E(\varepsilon_i) = 0$ pour tout i. Cela signifie que le modèle linéaire est exact en moyenne : $\forall\, 1 \leqslant i \leqslant n,\ E(Y_i) = \beta_0 + \beta_1 x_i$.

H2 (homoscedasticité). La variance du terme d'erreur est constante pour toutes les observations :

$$\forall\, i \in \{1,\dots,n\}, \quad \mathrm{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2 \tag{2}$$

H3 (absence d'autocorrélation). Les erreurs associées à deux observations distinctes sont non corrélées. La corrélation entre les termes d'erreurs étant $\mathrm{Cov}(\varepsilon_i, \varepsilon_k)/\sigma^2$, Cov désignant la covariance, et l'espérance des ε_i étant nulle, on a :

$$\forall\, 1 \leqslant i \neq j \leqslant n, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \tag{3}$$

Les hypothèses H1, H2, H3 sont les **hypothèses de Gauss-Markov**.

2.3 Estimation par la méthode des moindres carrés ordinaires (MCO)

Le principe des moindres carrés ordinaires (MCO) consiste à rechercher les valeurs des paramètres β_0 et β_1 qui minimisent la somme des carrés des écarts entre les valeurs observées y_i et les valeurs prédites par la **droite de régression** d'équation $\hat{y}_i = \beta_0 + \beta_1 x_i$. Considérer les carrés des écarts (plutôt que leur valeur absolue, ou toute autre fonction des écarts) simplifie les développements mathématiques suivants.

On cherche donc à minimiser en fonction de β_0 et β_1 la fonction suivante :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{4}$$

Proposition 2.1 *Les estimateurs des MCO, notés $\hat{\beta}_0$ et $\hat{\beta}_1$, sont donnés par les formules suivantes :*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{5}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{6}$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{Y} = \frac{1}{n}\sum_{Y=1}^{n} Y_i$ désignent les moyennes (moyenne empirique pour $\bar{Y}$).

Notons que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont, comme les Y_i, des variables aléatoires.

La prédiction de la i-ème observation selon ce modèle est donnée par $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: il s'agit également d'une variable aléatoire.

Démonstration

La fonction $S(\beta_0, \beta_1)$ est strictement convexe. Ses dérivées partielles par rapport aux deux paramètres s'écrivent :

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \tag{7}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) \tag{8}$$

L'annulation des dérivées partielles donne le système de ce qu'on appelle les **équations normales** qui définissent les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$:

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \tag{9}$$

$$\sum_{i=1}^{n} x_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{10}$$

De l'équation (9), en divisant par n, on obtient : $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, ce qui démontre la relation pour l'ordonnée à l'origine : $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$.

En injectant cette expression dans l'équation (10), il vient :

$$\begin{aligned}
\sum_{i=1}^{n} x_i Y_i - (\bar{Y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} + \hat{\beta}_1 n\bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} &= \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)
\end{aligned}$$

En utilisant les identités remarquables de la covariance et de la variance empirique, à savoir $\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$ et $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$, on obtient l'expression finale de $\hat{\beta}_1$. □

Comme $Y_i - \bar{Y} = \beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$, on a également :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11}$$

en développant le numérateur, car $\sum_i (x_i - \bar{x}) = 0$.

Ainsi,

$$\hat{\beta}_0 \;=\; \bar{Y} - \beta_1 \bar{x} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x} \;=\; \beta_0 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x} + \bar{\varepsilon} \tag{12}$$

Ces deux relations seront utiles dans la section suivante.

Remarque. En pratique, on observe une réalisation du n-échantillon $(x_i, y_i)_{1 \leqslant i \leqslant n}$, qui permet de calculer une réalisation de $\bar{Y}$, $\hat{\beta}_0$, $\hat{\beta}_1$, et la valeur moyenne $\bar{x}$ (qui est déterministe). La relation $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ signifie que la droite de régression passe toujours par le point moyen $(\bar{x}, \bar{y})$ du nuage des observations (x_i, y_i).

2.4 Propriétés statistiques des estimateurs simples

Sous les hypothèses H1 à H3, on démontre les propriétés essentielles suivantes :

1. $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des **estimateurs sans biais** de β_0 et β_1 : $E(\hat{\beta}_1) = \beta_1$ et $E(\hat{\beta}_0) = \beta_0$.
2. La variance des estimateurs s'écrit :

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{13}$$

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \tag{14}$$

Démonstration

Ce sont des conséquences directes des équations 11 et 12, de la linéarité de l'espérance, de l'absence d'autocorrélation entre les ε_i, et de $E(\varepsilon_i) = E(\bar{\varepsilon}) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ et $\mathrm{Var}(\bar{\varepsilon}) = \sigma^2/n$. □

Ces formules indiquent que, à amplitude σ^2 des perturbations aléatoires fixée, la précision des estimateurs augmente avec la taille de l'échantillon n et avec la dispersion des valeurs de la variable explicative x.

Sauf dans des cas pathologiques où $\sum_i (x_i - \bar{x})^2$ converge vers une valeur non-nulle, **ces estimateurs sont convergents** (la variance tend vers 0 quand $n \to +\infty$).

Par ailleurs, la covariance de $\hat{\beta}_0$ et $\hat{\beta}_1$ est (après des calculs basés sur la bilinéarité de la covariance et la non-corrélation des aléas) :

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \mathrm{Cov}\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})\bar{x}\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} + \bar{\varepsilon}, \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{15}$$

Ces deux estimateurs ne sont donc pas indépendants.

Remarque. Les estimateurs donnés par les équations 5 et 6 sont des **estimateurs linéaires**, dans le sens où ils sont formés comme une combinaison linéaire des données observées Y_i. Le **théorème de Gauss-Markov** (admis) énonce que les estimateurs des MCO sont ceux de variance minimale parmi les estimateurs linéaires sans biais. On dit que cet estimateur est le BLUE (*Best Linear Unbiased Estimator*).

2.5 Estimation sans biais de la variance des résidus

Pour exploiter de manière pratique les formules des variances des estimateurs des MCO, $\mathrm{Var}(\hat{\beta}_1)$ et $\mathrm{Var}(\hat{\beta}_0)$, données par les équations 13 et 14, il est indispensable d'estimer le paramètre de variance inconnu σ^2. On définit pour chaque observation i le résidu empirique $\hat{e}_i$ par :

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{16}$$

Attention, le résidu empirique $\hat{e}_i$ est une variable aléatoire, comme Y_i, $\hat{\beta}_0$ et $\hat{\beta}_1$, mais ce n'est pas l'erreur ε_i dans le modèle (théorique) de l'équation 1, qui reste inconnue.

La **somme des carrés des résidus** (SCR), égale à $\sum_i \hat{e}_1{}^2$, correspond exactement au minimum de la fonction de coût $S(\hat{\beta}_0, \hat{\beta}_1)$. Bien que la variance théorique de chaque perturbation soit σ^2, l'estimateur naturel $\frac{1}{n}\sum_i \hat{e}_i^2$ est biaisé, ce que nous allons démontrer à présent en déterminant l'espérance de $\sum_i \hat{e}_i^2$.

D'une part : $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$; et d'autre part : $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, soit $\bar{Y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$. Donc :

$$\begin{cases} Y_i - \bar{Y} = \beta_1 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \\ \hat{Y}_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x}) \end{cases} \tag{17}$$

Puisque $\hat{e}_i = Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})$, on obtient par substitution :

$$\hat{e}_i = (\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \tag{18}$$

Élevons cette expression au carré et sommons sur i :

$$\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n \left[(\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \right]^2 \tag{19}$$

En développant l'identité remarquable, on obtient trois termes :

$$\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 - 2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x}) + (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n (x_i - \bar{x})^2 \tag{20}$$

On sait d'après l'équation 11 que :

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{21}$$

Ce qui implique que $\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) = (\hat{\beta}_1 - \beta_1) \sum_{i=1}^n (x_i - \bar{x})^2$.
En remplaçant cette somme dans le terme central de l'équation 20, l'expression se simplifie en :

$$\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 - (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n (x_i - \bar{x})^2 \tag{22}$$

Par linéarité de l'espérance :

$$E\left(\sum_{i=1}^{n}\hat{e}_i^2\right) = E\left(\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2\right) - \sum_{i=1}^{n}(x_i-\bar{x})^2 E\left((\hat{\beta}_1-\beta_1)^2\right) \tag{23}$$

Le premier bloc vaut de manière classique :

$$E\left(\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2\right) = (n-1)\sigma^2 \tag{24}$$

Pour le second bloc, puisque $E\left((\hat{\beta}_1-\beta_1)^2\right) = \text{Var}(\hat{\beta}_1)$:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 E\left((\hat{\beta}_1-\beta_1)^2\right) = \sum_{i=1}^{n}(x_i-\bar{x})^2\text{Var}(\hat{\beta}_1) = \sum_{i=1}^{n}(x_i-\bar{x})^2\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sigma^2 \tag{25}$$

En combinant les résultats :

$$E\left(\sum_{i=1}^{n}\hat{e}_i^2\right) = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2 \tag{26}$$

Proposition 2.2 *L'estimateur sans biais de la variance des perturbations σ^2, noté s^2 (ou $\hat{\sigma}^2$), est défini par :*

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \frac{SCR}{n-2} \tag{27}$$

Il s'agit bien d'une variable aléatoire, comme Y_i et $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

En substituant σ^2 inconnu par son estimation s^2 dans les expressions théoriques (équations 13 et 14), on obtient les estimateurs des écarts-types des coefficients, notés $\hat{\sigma}_{\hat{\beta}_1}$ et $\hat{\sigma}_{\hat{\beta}_0}$.

Remarque. Les résidus $\hat{e}_i$ (dits **résidus ordinaires**) sont de moyenne nulle (car $E(Y_i-\hat{Y}_i) = E(Y_i) - \beta_0 - \beta_1 x_i = 0$), mais de variance exprimée (après calculs) par $\text{Var}(\hat{e}_i) = E(\hat{e}_i^2) = \sigma^2(1-h_{ii})$ où $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_j (x_j-\bar{x})^2}$. Les résidus ordinaires $\hat{e}_i$ ne sont donc pas distribués de manière homoscédastique, contrairement aux termes d'erreur ε_i.

Les **résidus de Pearson** $\hat{r}_i$ sont définis par :

$$\hat{r}_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}} \tag{28}$$

Les résidus de Pearson sont des versions standardisées des résidus ordinaires, globalement distribués entre -2 et 2, de manière à pouvoir les comparer entre eux.

2.6 Inférence statistique sous hypothèse de normalité

Afin de mener à bien l'inférence (construction de tests et d'intervalles de confiance), on adjoint aux hypothèses de Gauss-Markov (H1 à H3) une hypothèse sur la loi des perturbations :

H4 (normalité). Les perturbations ε_i suivent des lois normales indépendantes et identiquement distribuées : $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$.

Sous H1–H4, les variables aléatoires Y_i sont normales par linéarité (voir remarque à la fin de la section 2.4). Les estimateurs des MCO $\hat{\beta}_0$ et $\hat{\beta}_1$, formés par combinaison linéaire des Y_i, suivent également des lois normales caractérisées par les moyennes et variances établies en section 2.4 :

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \tag{29}$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right) \tag{30}$$

Comme la variance σ^2 est inconnue et substituée par son estimateur s^2, la « Studentisation »[1] de ces variables aléatoires conduit aux statistiques fondamentales suivantes :

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim T(n-2) \quad \text{et} \quad \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim T(n-2) \tag{31}$$

où $T(n-2)$ désigne la loi de Student à $n-2$ degrés de liberté.

Le justification est que la SCR suit la loi du χ^2 à $n-2$ degrés de liberté. C'est une conséquence admise du théorème de Cochran déjà évoqué dans ce cours.

Remarque. Sous l'hypothèse H4, les observations $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ suivent une loi normale de moyenne $\beta_0 + \beta_1 x_i$ et de variance σ^2. La log-vraisemblance des observations est donc :

$$-\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n \left(\frac{Y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2 \tag{32}$$

En annulant les dérivées de la log-vraisemblance par rapport à β_0 et β_1, on trouve après calcul que les **estimateurs du maximum de vraisemblance** de ces deux paramètres sont les estimateurs des MCO. De la même manière, on trouve que l'estimateur du maximum de vraisemblance de σ^2 est $\sum_i e_i^2/n$, dont on a vu qu'il était biaisé.

2.6.1 Intervalles de confiance des estimateurs et test de significativité des paramètres

L'exploitation des lois de Student définies à la section précédente permet d'encadrer les paramètres inconnus du modèle avec un niveau de confiance fixé à $1-\alpha$ (généralement 95%).

Proposition 2.3 *L'intervalle de confiance au niveau $1-\alpha$ pour le paramètre de pente β_1 est donné par :*

$$IC_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 - t_{1-\alpha/2}^{n-2} \cdot \hat{\sigma}_{\hat{\beta}_1} \; ; \; \hat{\beta}_1 + t_{1-\alpha/2}^{n-2} \cdot \hat{\sigma}_{\hat{\beta}_1}\right] \tag{33}$$

et l'intervalle de confiance pour la constante β_0 s'écrit :

$$IC_{1-\alpha}(\beta_0) = \left[\hat{\beta}_0 - t_{1-\alpha/2}^{n-2} \cdot \hat{\sigma}_{\hat{\beta}_0} \; ; \; \hat{\beta}_0 + t_{1-\alpha/2}^{n-2} \cdot \hat{\sigma}_{\hat{\beta}_0}\right] \tag{34}$$

où $t_{1-\alpha/2}^{n-2}$ représente le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté.

Comme d'habitude, les intervalles de confiance sont estimés sur une réalisation des (x_i, Y_i).

Il est également possible de tester l'hypothèse nulle $\mathcal{H}_0 : \beta_0 = 0$. Autrement dit, on se demande si la valeur $\hat{\beta}_0$ est possiblement due aux aléas et reste compatible avec une « vraie valeur » β_0 qui serait nulle. D'après l'équation 31, cela revient à former la statistique de test $T = \hat{\beta}_0/\hat{\sigma}_{\hat{\beta}_0}$ et à faire un test bilatéral selon la loi de Student à $n-2$ degrés de liberté : si $|T|$ est supérieure à la valeur critique $t_{1-\alpha/2}^{n-2}$, alors on rejette $\mathcal{H}_0$, et $|T| < t_{1-\alpha/2}^{n-2}$, on ne rejette pas $\mathcal{H}_0$. Cette dernière condition se traduit par $|\hat{\beta}_0| < t_{1-\alpha/2}^{n-2} \cdot \hat{\sigma}_{\hat{\beta}_0}$. Autrement dit, cela revient exactement à regarder si la valeur 0 est dans l'intervalle de confiance $IC_{1-\alpha}(\beta_0)$: si ce n'est pas le cas, on rejette $\mathcal{H}_0$. On dit alors que la valeur de l'ordonnée à l'origine du modèle est significative.

Le raisonnement est le même pour $\hat{\beta}_1$.

1. Cela désigne l'opération consistant à centrer et réduire une variable aléatoire avec une estimation de la moyenne et une estimation de l'écart-type, ce qui conduit, sous hypothèse de normalité, à une loi de Student. Plusieurs exemples de ce type ont été vus dans le cours.

2.6.2 Intervalle de confiance d'une prédiction

Soit x_0 une nouvelle valeur de la variable explicative. La prédiction ponctuelle fournie par le modèle estimé est $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Deux cas de figure se présentent selon la nature de l'inférence désirée :

1. **Intervalle de confiance de la valeur moyenne $E(\hat{Y}_0)$.** Il s'agit d'encadrer la vraie position de la droite de régression en x_0 (car $E(\hat{Y}_0) = \beta_0 + \beta_1 x_0$, où β_0 et β_1 sont les « vraies » valeurs, inconnues). La variance de cet estimateur $\hat{Y}_0$ vaut $\mathrm{Var}(\hat{\beta}_0) + x_0^2 \mathrm{Var}(\hat{\beta}_1) + x_0 \mathrm{Covar}(\hat{\beta}_0, \hat{\beta}_1)$ (rappelons que les deux estimateurs des MCO des paramètres ne sont pas indépendants). On trouve :

$$\mathrm{Var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{35}$$

   En estimant σ^2 par s^2, l'intervalle de prédiction à un niveau de confiance $1-\alpha$ pour un point de la droite de régression est défini par :

$$IP_{1-\alpha}(\bar{Y}_0) = \left[\hat{Y}_0 \pm t_{1-\alpha/2}^{n-2} \cdot s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right] \tag{36}$$

   où $t_{1-\alpha/2}^{n-2}$ représente toujours le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté.

2. **Intervalle de prédiction d'une observation individuelle Y_0 :** Il s'agit d'anticiper la valeur future prise par un nouvel individu, ce qui requiert d'intégrer la perturbation aléatoire $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$, car $Y_0 = \hat{Y}_0 + \varepsilon_0$. La variance de l'erreur de prévision globale devient :

$$\mathrm{Var}(Y_0 - \hat{Y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{37}$$

   En estimant σ^2 par s^2, l'intervalle de prédiction à un niveau de confiance $1-\alpha$ pour un individu est défini par :

$$IP_{1-\alpha}(Y_0) = \left[\hat{Y}_0 \pm t_{1-\alpha/2}^{n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right] \tag{38}$$

On voit que l'amplitude de ces intervalles est minimale au centre de gravité des données ($\bar{x}$) et s'accroît de façon hyperbolique à mesure que l'on s'éloigne du point moyen, mettant en évidence la perte de précision du modèle en extrapolation.

Remarque. Lorsque x_0 est « proche » de $\bar{x}$ (on néglige alors $(x_0 - \bar{x})^2$ devant $\sum_i (x_i - \bar{x})^2$) et n est grand, l'intervalle de confiance de la prévision devient négligeable et les observations individuelles sont réparties dans une bande de part et d'autre de la droite de régression d'épaisseur $2 t_{1-\alpha/2} s$.

3 Exemple

En pratique, on observe une réalisation $(x_i, y_i)_{1 \leqslant i \leqslant n}$ du n-échantillon (x_i, Y_i), à partir duquel on calcule des estimations numériques des paramètres de la droite de régression, de la variance du bruit, des intervalles de confiance des prévisions… vues comme des réalisations des variables aléatoires introduites précédemment.

La figure 2 montre une droite de régression calculée sur des données, et les deux types d'intervalles de confiance.

L'intervalle de confiance de la moyenne (zone bleue). Il entoure la droite de régression rouge. Il représente l'incertitude liée à l'estimation de la vraie relation moyenne entre x et y. Si l'on répétait l'expérience à l'infini en collectant de nouveaux échantillons de données, la « vraie » droite théorique se trouverait dans cette zone bleue avec une probabilité de 0,95. La zone bleue est légèrement plus resserrée au centre qu'aux extrémités : c'est parce que le modèle est toujours plus précis autour de la moyenne des données observées. Les frontières de cette zone sont des arcs d'hyperbole.

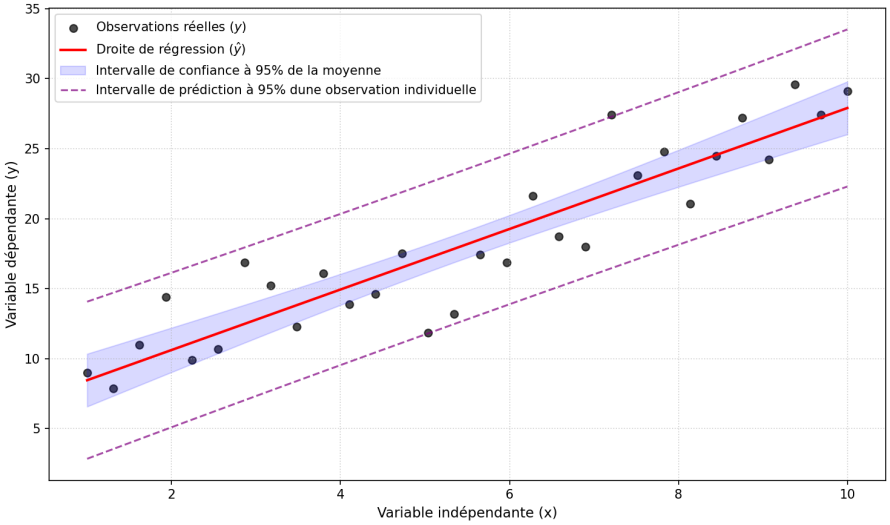

FIGURE 2 – Régression linéaire : intervalles de confiance et intervalles de prédiction.

L'intervalle de prédiction des observations (courbes pointillées). Il est beaucoup plus large et englobe la quasi-totalité des points. Il représente l'incertitude liée à la prédiction d'une valeur individuelle future. Si on choisit une nouvelle valeur de x, il y a 95% de chances que l'observation réelle y se situe entre ces deux lignes violettes. Il est plus large que l'intervalle précédent car il doit cumuler deux types d'incertitudes : l'imprécision de la droite elle-même (la zone bleue) et la variabilité naturelle / le bruit aléatoire de chaque individu (les fluctuations des points noirs autour de la droite).

4 Qualité de l'ajustement et diagnostics avancés

Une fois les calculs effectués, le statisticien doit valider la qualité globale de l'ajustement (les tests de Student précédents sont une manière de s'assurer que le modèle est statistiquement significatif), et s'assurer que les hypothèses théoriques imposées au modèle (hypothèses H1 à H4) se vérifient empiriquement.

4.1 Coefficient de détermination R^2

Le **coefficient de détermination**, noté R^2, est défini comme :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{39}$$

On démontre aussi (c'est un calcul semblable à celui qu'on a fait pour l'analyse de la variance) que :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{40}$$

R^2 mesure donc la proportion de la variance totale de la variable cible expliquée par le régresseur du modèle.

Le coefficient de détermination R^2 est un indicateur compris entre 0 et 1. Plus il est proche de 1, meilleur est l'ajustement global du modèle aux données. Le cas où $R^2 = 1$ correspond à $\forall\, i,\ Y_i - \hat{Y}_i$: le modèle linéaire prédit parfaitement les observations.

Remarque. Dans le cas d'autres modèles de régression, la définition (équation 39) n'implique pas la propriété de l'équation 40. On peut alors avoir des valeurs de R^2 négatives (pour un modèle très mauvais), mais on a toujours $R^2 \leqslant 1$.

4.2 Analyse des résidus et tests de validation

La validation d'un modèle de régression linéaire passe par l'étude des propriétés des résidus observés $e_i = Y_i - \hat{Y}_i$. Les diagnostics suivants seront vus dans des cours spécialisés.

— **Diagnostic d'homoscedasticité**. On trace graphiquement les résidus en fonction des valeurs prédites $\hat{Y}$. Si le nuage de points présente une forme d'entonnoir, l'hypothèse de variance constante est violée. Différents tests statistiques sont disponibles.

— **Diagnostic de normalité**. L'analyse visuelle s'appuie sur un graphique Quantile-Quantile QQ-plot. Les points doivent s'aligner sur une droite. Les tests statistiques de Shapiro-Wilk ou de Jarque-Bera permettent de tester la normalité des résidus.

— **Diagnostic d'indépendance**. En présence de données chronologiques (l'indice i désigne alors le temps), les erreurs peuvent être corrélées dans le temps. On utilise le test de Durbin-Watson pour détecter une éventuelle autocorrélation d'ordre 1.

5 Exercices

5.1 Un exercice d'application directe avec des calculs simples

Un professeur souhaite analyser la relation entre le nombre d'heures de révision de ses étudiants et la note qu'ils ont obtenue (sur 10) lors d'un test. Il a sélectionné un échantillon représentatif de 5 étudiants.

Voici les données récoltées :

Étudiant	**Heures de révision** (x_i)	**Note obtenue** (y_i)
A	1	3
B	2	4
C	3	6
D	4	8
E	5	9

Travail à faire

1. **Droite de régression :** Déterminez l'équation de la droite de régression linéaire de y en x, sous la forme $y = ax + b$.
2. **Intervalle de confiance de la pente :** Sachant que pour un niveau de confiance de 95% et avec 3 degrés de liberté ($n-2$), la valeur critique de la loi de Student est de $t = 3{,}182$, calculez l'intervalle de confiance à 95% pour la pente a. La relation entre heures de révision et note est-elle significative ?
3. **Prédiction :** Si un étudiant révise pendant 3,5 heures, quelle note peut-il espérer obtenir selon ce modèle ?
4. **Intervalle de prédiction :** Calculez l'intervalle de prédiction à 95% pour la note exacte de cet étudiant ayant révisé 3,5 heures.

5.2 Étude sur ordinateur de ce jeu de données réduit

Le carnet Jupyter sur Arche donne le code permettant de faire une régression linéaire et de valider le modèle par les statistiques du cours et la visualisation de différents graphiques. Il est basé sur la bibliothèque Python statsmodels.

Interprétez les sorties du carnet.

5.3 Une étude à l'aide de la bibliothèque statsmodels

Voir sur Arche.