MTH 8302 - Modèles de Régression et d'Analyse de Variance

Leçon 1 : Régression Linéaire

Polytechnique Montréal - Hiver 2025

Chiheb Trabelsi

February 5, 2025

Table des Matières

Introduction à la Régression Linéaire

Introduction à la Régression Linéaire : Introduction

Intro : Qu'est-ce que la Régression Linéaire ?

- La **régression linéaire** est une technique statistique fondamentale pour établir une relation linéaire entre une variable dépendante Y et une ou plusieurs variables indépendantes X_i.
- **Formule du modèle**:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \quad \text{où :}$$

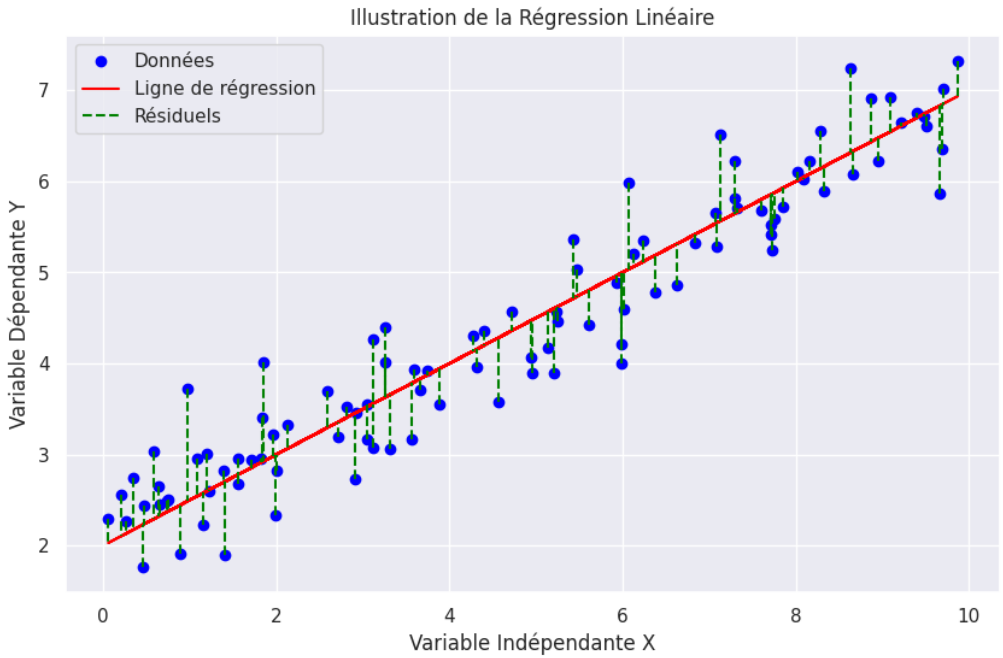
  - ϵ l'erreur aléatoire du modèle.
  - $X_1, X_2, ..., X_p$ les variables explicatives, appelées variables indépendantes.
  - Y la variable dépendante, appelée aussi réponse.
  - β_0 est l'intercept du modèle, qui représente la valeur attendue de Y lorsque toutes les variables indépendantes X_i sont égales à zéro.
  - $\beta_1, \beta_2, ..., \beta_p$ sont les coefficients de pente associés à chaque variable indépendante $X_1, X_2, ..., X_p$. Chaque coefficient β_i mesure le changement attendu dans Y pour une unité de changement dans X_i, en tenant tous les autres facteurs constants.
  - Ces coefficients permettent de quantifier l'effet de chaque variable indépendante sur la variable dépendante.

Intro : Modèle de Regression Linéaire Simple (à 1 Variable)

Expression du Modèle

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 est l'intercept de la régression, la valeur de Y lorsque $X = 0$.
- β_1 est le coefficient de pente, indiquant combien Y change pour chaque unité de changement dans X.
- ϵ représente le terme d'erreur, ajoutant de la variabilité aléatoire.

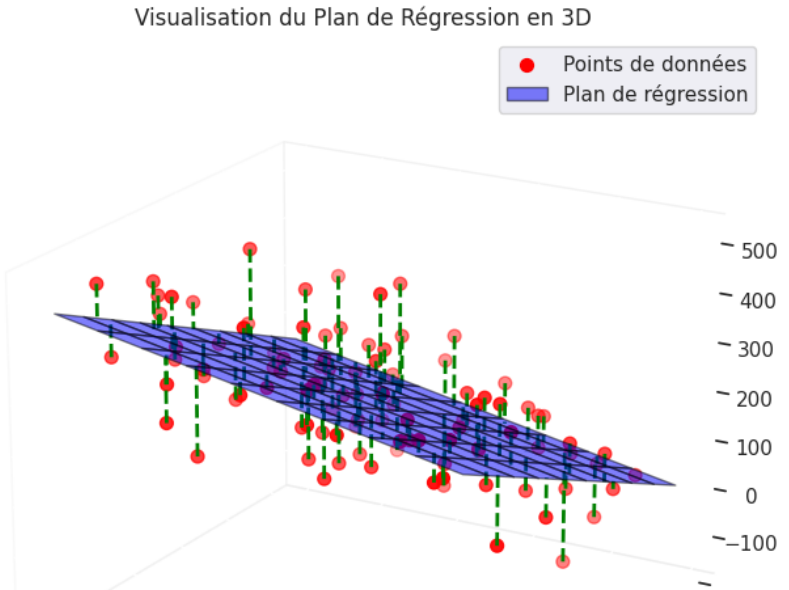


Intro : Modèle de Regression Linéaire Multiple (à 2 Vars)

Expression du Modèle

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- β_0 est l'intercept de la régression, la valeur de Y quand X_1 et X_2 sont nuls.
- β_1 et β_2 sont les coefficients des pentes pour X_1 et X_2, montrant l'impact de chaque unité de changement sur Y.
- ϵ est le terme d'erreur, incorporant la variabilité aléatoire.



Intro : Quelques Applications de la Régression Linéaire

- **Économie**:
  - **Exemple** : Prédiction du PIB basée sur des facteurs tels que les dépenses de consommation, les investissements des entreprises, et les dépenses publiques.
  - **Intuition** : Comprendre comment différentes composantes économiques contribuent au PIB peut aider à formuler des politiques économiques plus efficaces.
- **Médecine**:
  - **Exemple** : Estimation de l'effet d'un nouveau médicament sur la réduction du taux de cholestérol par rapport à un placebo.
  - **Intuition** : Identifier l'efficacité d'un traitement permet de prendre des décisions éclairées sur son utilisation clinique.
- **Finance**:
  - **Exemple** : Modélisation de l'impact des taux d'intérêt et des indices boursiers sur les prix des obligations.
  - **Intuition** : Les investisseurs peuvent utiliser ces informations pour optimiser leurs stratégies de portefeuille, en minimisant les risques et maximisant les rendements.

Intro : Importance de la Régression Linéaire

- Permet une compréhension profonde des relations entre variables, essentielle pour la prise de décision basée sur des données.
- Facilite la prévision et la planification en fournissant des estimations quantitatives.
- Sert de point de départ pour des modèles statistiques plus complexes et des analyses multivariées.
- La régression linéaire a beaucoup de domaines d'applications et on établit certaines hypothèses concernant le modèle pour pouvoir l'appliquer.
- Ces hypothèses sont essentielles pour plusieurs raisons importantes qui concernent:
  - La validité des résultats.
  - L'efficacité de l'interprétation.
  - La précision des prédictions.

Notation en Régression Linéaire

Notation en Régression Linéaire

- **Variable Aléatoire vs. Valeur Observée**
  - Y: Désigne la variable aléatoire. Notation formelle pour le concept d'une variable dépendante.
  - Y_i: Valeur de la variable dépendante comme variable aléatoire pour la i-ème observation.
  - y: Représente la valeur observée spécifique que prend la variable aléatoire Y.
  - y_i: Valeur observée spécifique de Y pour la i-ème observation.
- **Matrice des Données et Coefficients du Modèle**
  - $\mathbf{X}$: Matrice contenant les variables indépendantes.
  - $\mathbf{X}_i$: Vecteur contenant les variables indépendantes pour la i-ème observation.
  - X_{ij}: Variable aléatoire représenant de la j-ème variable indépendante pour la i-ème observation.
  - x_{ij}: Valeur de la Variable aléatoire représenant de la j-ème variable indépendante pour la i-ème observation.
  - $\boldsymbol{\beta}$: Vecteur des coefficients du modèle, incluant l'intercept β_0 et les pentes $\beta_1, \ldots, \beta_p$.

Notations en Régression Linéaire

- $\hat{\boldsymbol{\beta}}$: Estimateur du vecteur des coefficients du modèle, incluant l'intercept β_0 et les pentes $\beta_1, \ldots, \beta_p$.
- $\mathbf{Y}$: Vecteur aléatoire qui représentant les variables aléatoires dépeandantes $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_n]^{\mathrm{T}}$.
- ϵ: Erreurs comme variables aléatoires indiquant l'écart entre la prédiction parfaite (du modèle parfait) et la valeur réelle de Y.
- e ou e_i: Résidus observés, calculés comme la différence entre y_i et la prédiction $\hat{y}_i$: $(e_i = y_i - \hat{y}_i)$.
- **Calculs Principaux en Régression**
  - $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{Y}$
  - Prédiction : $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$
  - Vecteur des erreurs : $\boldsymbol{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$
  - Vecteur des résidus: $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, vecteur des résidus.

Notation en Régression Linéaire

- **Matrice de conception ($\mathbf{X}$)**
  - $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$.
  - n nombre d'observations.
  - p nombre de variable indépendantes.
  - Pourquoi $(p+1)$? $\mathbf{X}$ inclut un vecteur de 1 à la 1ère colonne pour la multiplication avec l'intercept β_0 lors du calcul de la prédiction.
- **Vecteur des coefficients ($\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_p]^{\mathrm{T}} \in \mathbb{R}^{p+1}$)**

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$$

- Cette opération projette les données observées sur le plan (ou la ligne) de régression estimé (par exemple par la méthode les moindres carrés). $\hat{\mathbf{Y}} \in \mathbb{R}^n$

Notations en Régression Linéaire : Exemple Pratique

Si nous avons un modèle avec 2 variables indépendantes et cinq observations, la matrice $\mathbf{X}$ et le vecteur $\mathbf{Y}$ peuvent ressembler à ceci:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & X_{31} & X_{32} \\ 1 & X_{41} & X_{42} \\ 1 & X_{51} & X_{52} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix}$$

Ici, le vecteur $\mathbf{1}$ à la 1ère colonne de $\mathbf{X}$ est multiplié par l'intercept β_0 du modèle lors de la prédiction de $\mathbf{Y}$:

$$Y_i = \beta_0 \times 1 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i,$$

Hypothèses du Modèle Linéaire de la Regression

Hypothèses du Modèle de Régression Linéaire

- **1. Linéarité**
  - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
- **2. Indépendance des erreurs des observations**
  - $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad$ pour tout $i \neq j$
- **3. Homoscédasticité**
  - $\text{Var}(\epsilon_i) = \sigma^2 \quad$ pour tout i
- **4. Normalité des erreurs des observations**
  - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **5. Absence de multicollinéarité**
  - Les variables explicatives ne doivent pas être linéairement dépendantes. (la matrice $\mathbf{X}$ est plein rang.)
- **6. Additivité**
  - Les effets des différentes variables explicatives sur la variable dépendante sont additifs.
- **7. Fixité des X_i**
  - Les X_i sont traitées comme fixes (déterministes).

Hypothèses du Modèle de Régression Linéaire

- **1. Linéarité :** $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
  - L'expression mathématique montre que la relation entre les variables dépendantes et indépendantes est modélisée comme une combinaison linéaire. (**Note importante :** Un modèle de régression linéaire est linéaire en $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_p]^{\mathrm{T}}$ et peut utilser des transformation non linéaires linéaires des variables indépendantes X_i.
- **2. Indépendance des erreurs :** $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0 \quad$ pour tout $i \neq j$
  - Cette condition est cruciale pour éviter l'autocorrélation (correlation entre les observations dans $\mathbf{X}$), qui peut fausser les résultats des tests statistiques utilisés pour évaluer le modèle.
- **3. Homoscédasticité :** $\mathrm{Var}(\epsilon_i) = \sigma^2 \quad$ pour tout i
  - La constance de la variance des erreurs est nécessaire pour que les estimations des erreurs standard des coefficients soient valides, ce qui affecte les intervalles de confiance et les tests d'hypothèses.
- **4. Normalité des erreurs :** $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - Cette hypothèse est particulièrement importante pour la validité des tests d'hypothèses qui supposent la normalité, comme le test t de Student et le test F de Fisher.

Hypothèses du Modèle de Régression Linéaire

- **5. Absence de multicollinéarité :** Les variables explicatives ne doivent pas être linéairement dépendantes.
- **6. Additivité :** Les effets des différentes variables explicatives sur la variable dépendante sont additifs.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Dans cette équation, chaque terme $\beta_i X_i$ représente l'effet additionnel de la variable X_i sur la variable dépendante Y, assumant que les effets des autres variables restent constants.
- L'additivité implique que l'effet d'une augmentation d'une unité dans n'importe quelle variable explicative X_i sur Y est constant, indépendamment des niveaux des autres variables.
- **7. Hypothèse de fixité :** les X_i sont considérées commes fixes, càd, déterministes et connues à l'avance.

Hypothèses du Modèle de Régression Linéaire

- Les variables indépendantes X_i dans la régression linéaire sont souvent traitées comme si elles étaient déterministes ou fixes. On parle alors **d'hypothèse de fixité**
- Les variables indpendantes X_i sont alors considérées comme des quantités non aléatoires, fixées par le design de l'étude et non sujettes à des variations aléatoires.
- Dans les applications classiques (économétrie, analyse de données expérimentales), les X_i sont prises comme données connues à l'avance et non affectées par l'erreur de mesure.
- **Implications**:
  - Simplifie l'analyse et la formulation des estimateurs des moindres carrés.
  - Les tests statistiques standard reposent sur cette hypothèse pour la signification des coefficients.

Hypothèses du Modèle de Régression Linéaire

- **Quand est-ce qu'on considère X_i aléatoires?**
  - En statistique bayésienne ou dans certains modèles de régression, traitent les X_i comme aléatoires.
  - Ceci est souvent le cas lorsque les données proviennent de processus sujets à variation ou incertitude.
- **En Résumé pour l'hypothèse de fixité des X_i:**
  - **L'approche traditionnelle en régression linéaire traite souvent les variables indépendantes comme déterministes**.
  - Il est crucial de comprendre le contexte et les données spécifiques pour déterminer si cette hypothèse est appropriée.
  - Dans les cas où les X_i sont sujets à des variations aléatoires, des modèles statistiques plus complexes sont nécessaires.

Importance des Hypothèses du Modèle Linéaire de la Regression

Importance des Hypothèses de la Régression Linéaire

- **Validité des Estimations :** Les hypothèses de base de la régression linéaire garantissent que les estimations des paramètres (les coefficients β_i) sont les meilleurs estimateurs linéaires non biaisés (Best Linear Unbiased Estimator (BLUE)). Cela signifie qu'en moyenne, les estimations obtenues à partir de l'échantillon représentent correctement la population réelle.
- L'indépendance et la normalité des erreurs, en particulier, sont nécessaires pour utiliser des tests statistiques classiques tels que les tests t et F, qui supposent que les résidus sont distribués normalement.
- **Interprétation Correcte des Coefficients :** Si les variables explicatives sont fortement corrélées (multicollinéarité), cela peut rendre difficile l'interprétation des coefficients individuels. Une relation linéaire correcte et des erreurs indépendantes garantissent que les changements observés dans la variable dépendante Y peuvent être correctement attribués aux variables indépendantes X_i.

Importance des Hypothèses de la Régression Linéaire

- **Efficacité des Prédictions :** Une hypothèse de **homoscédasticité (càd, variance constante des erreurs)** est importante. Autrement, certaines prédictions pourraient être systématiquement plus incertaines que d'autres, ce qui réduit l'utilité pratique du modèle.
- Une telle hypothèse permet de construire un modèle qui est précis sur les données historiques mais qui est également fiable pour la prévision sur de nouvelles données.
- **Évaluation Générale du Modèle :** En vérifiant ces hypothèses, cela aide à déterminer si des ajustements du modèle ou des techniques de modélisation alternatives pourraient être nécessaires. Par exemple, si les résidus ne sont pas normalement distribués, cela peut suggérer la nécessité d'utiliser des transformations des variables.
- **Confiance dans la Prise de Décision Basée sur le Modèle :** Les décideurs qui utilisent des modèles de régression pour orienter les politiques économiques, les stratégies commerciales ou les décisions médicales doivent être confiants que les modèles sont précis et fiables. La vérification des hypothèses est un pas crucial pour établir cette confiance.

Table des Matières

Modèle de Régression Linéaire simple

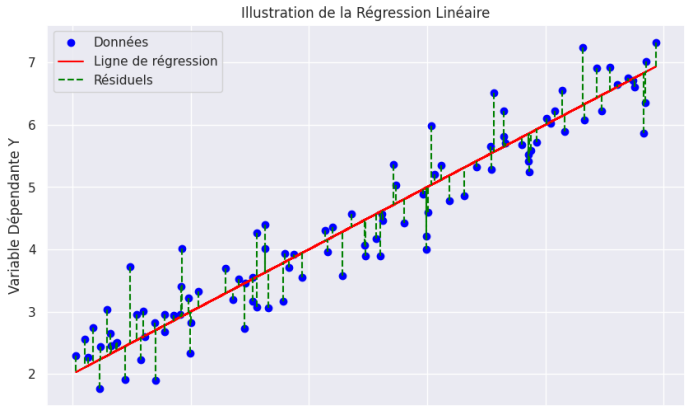
Modèle de Régression Linéaire simple : Introduction

Modèle de Regression Linéaire Simple

Expression du Modèle

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

- La régression linéaire simple est un modèle statistique qui cherche à expliquer la relation entre deux variables :
  - Une variable dépendante Y, par exemple, l'espérance de vie.
  - Une variable indépendante X, par exemple, le PIB d'un pays.
- β_0 est l'intercept de la régression, la valeur de Y_i lorsque $X_i = 0$.
- β_1 est le coefficient de pente, indiquant combien Y_i change pour chaque unité de changement dans X_i.
- ϵ_i représente le terme d'erreur, ajoutant de la variabilité aléatoire.



Modèle de Régression Linéaire simple : $\mathbb{E}[Y_i]$ & Var(Y_i)

Rappel : Hypothèses du Modèle Linéaire Simple

- **1. Linéarité**
  - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
- **2. Indépendance des erreurs des observations**
  - $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad$ pour tout $i \neq j$
- **3. Homoscédasticité**
  - $\text{Var}(\epsilon_i) = \sigma^2 \quad$ pour tout i
- **4. Normalité des erreurs des observations**
  - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **5. Absence de multicollinéarité**
  - Les variables explicatives ne doivent pas être linéairement dépendantes. (la matrice $\mathbf{X}$ est plein rang.)
- **6. Additivité**
  - Les effets des différentes variables explicatives sur la variable dépendante sont additifs.
- **7. Fixité des X_i**
  - Les X_i sont traitées comme fixes (déterministes).

Modèle Linéaire Simple : $\mathbb{E}[Y_i]$ & Var(Y_i)

- **Calcul de l'espérance :**

$$\begin{aligned}\mathbb{E}(Y_i) &= \mathbb{E}(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \mathbb{E}(\beta_0) + \mathbb{E}(\beta_1 X_i) + \mathbb{E}(\epsilon_i) \\ &= \beta_0 + \beta_1 \mathbb{E}(X_i) + 0 \quad (\epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ et } \beta_0,\ \beta_1 \text{ constantes}) \\ &= \beta_0 + \beta_1 X_i, \quad (\text{car } X_i \text{ sont déterministe})\end{aligned}$$

- **Calcul de la variance :**

$$\begin{aligned}\text{Var}(Y_i) &= \text{Var}(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \text{Var}(\beta_0) + \text{Var}(\beta_1 X_i) + \text{Var}(\epsilon_i) \\ &= 0 + 0 + \sigma^2 \quad (\text{car } \beta_0,\ \beta_1, \text{ et } X_i \text{ sont déterministes})\end{aligned}$$

- **Note:**
  - β_0 et β_1 sont des paramètres, pas des variables aléatoires.
  - X_i est considéré comme non-aléatoire dans ce contexte (fixé par conception).
  - ϵ_i est l'unique source de variabilité dans Y_i.

Modèle Linéaire Simple : $\mathbb{E}[Y_i]$ & Var(Y_i)

- **Interprétation du Modèle :**
  - L'espérance de Y_i, $\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i$, montre la dépendance linéaire de Y_i par rapport à X_i.
  - Permet d'interpréter β_1 comme le changement moyen dans Y pour une augmentation unitaire de X.
  - Essentiel pour la prédiction de Y basée sur des valeurs spécifiques de X.
- **Validation des Hypothèses Statistiques :**
  - La variance constante Var$(Y_i) = \sigma^2$ valide l'hypothèse d'homoscédasticité.
  - Crucial pour la validité des tests statistiques sur les coefficients de régression.
  - Assure que l'estimateur des moindres carrés est le (Best Linear Unbiased Estimator (BLUE)).

Modèle de Régression Linéaire simple : Estimation par La Méthode des Moindres Carrés

Estimation par La Méthode des Moindres Carrés

- **Étape 1: Fonction Objectif**
  - L'objectif de la méthode des moindres carrés est de minimiser la somme des carrés des résidus. Le résidu pour chaque observation est la différence entre la valeur observée y_i et la valeur prédite $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Nous formulons cela comme le problème de minimisation suivant :

$$\arg\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} e_i^2$$

$$\equiv \quad \arg\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\equiv \quad \arg\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Cette expression cherche les valeurs de β_0 et β_1 qui minimisent la somme des carrés des écarts entre les valeurs observées et les valeurs prédites.

Estimation par La Méthode des Moindres Carrés

- **Étape 2: Calcul des Dérivées Partielles**
  - On pose $C(\hat{\beta}_0, \hat{\beta}_1) = C = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$.
  - Pour minimiser C, on calcule les dérivées partielles par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$ et on les égalise à zéro pour trouver les valeurs qui minimisent C.
  - Dérivée par rapport à β_0: $\frac{\partial C}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

$$\Rightarrow 0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \Rightarrow n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

  - Dérivée par rapport à $\hat{\beta}_1$:

$$\frac{\partial C}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow 0 = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

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

Estimation par La Méthode des Moindres Carrés

- **Étape 4: Résolution du Système d'Équations**
  - Nous avons les équations linéaires suivantes :

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

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

  - Les équations obtenues à partir des dérivées partielles sont :

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (1)$$

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

Estimation par La Méthode des Moindres Carrés

- **Étape 4: Résolution du Système d'Équations**
  - En multipliant l'équation (1) par $\sum_{i=1}^{n} x_i$ et l'équation (2) par n, nous obtenons deux nouvelles équations qui nous permettent d'isoler $\hat{\beta}_1$:

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

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

  - Soustrayant ces deux équations, nous obtenons :

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

  - En substituant $\hat{\beta}_1$ dans l'équation (1), nous pouvons résoudre pour $\hat{\beta}_0$:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n}$$

  - Ce sont les formules pour les estimateurs des moindres carrés ordinaires (OLS) de $\hat{\beta}_0$ et $\hat{\beta}_1$.

Estimation par La Méthode des Moindres Carrés

- Soustrayant ces deux équations, nous obtenons :

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

- En substituant $\hat{\beta}_1$ dans l'équation (1), nous pouvons résoudre pour $\hat{\beta}_0$:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x}$$

- Ce sont les formules pour les estimateurs des moindres carrés ordinaires (Ordinary Least Squares (OLS)) de β_0 et β_1.

Estimation par La Méthode des Moindres Carrés

- En statistique, S_{xx} et S_{yx} sont des termes que l'on retrouve souvent dans les calculs de régression linéaire.
- S_{xx}: C'est la somme des carrés des écarts des valeurs x_i par rapport à leur moyenne $\overline{x}$. Mathématiquement, cela se formule comme suit:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

où x_i représente chaque valeur de x, $\overline{x}$ est la moyenne des x, et n est le nombre total d'observations.

- S_{yx}: C'est la somme des produits des écarts de y_i par rapport à leur moyenne $\overline{y}$ et des écarts de x_i par rapport à $\overline{x}$. Elle est définie par:

$$S_{yx} = \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})$$

où y_i est chaque valeur correspondante de y, et les autres symboles ont des significations similaires à celles mentionnées plus tôt.

Estimation par La Méthode des Moindres Carrés

- Nous souhaitons montrer que:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{S_{yx}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- Expression de S_{yx}:

$$S_{yx} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

en utilisant que $\sum_{i=1}^n x_i = n\bar{x}$ et $\sum_{i=1}^n y_i = n\bar{y}$.

- Expression de S_{xx}:

$$S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

- Comparaison de $\hat{\beta}_1$ et $\frac{S_{yx}}{S_{xx}}$:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{n}{n}\frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{S_{yx}}{S_{xx}}$$

CQFD

Estimation par La Méthode des Moindres Carrés

- Les estimateurs obtenus par la méthode des moindres carrés sont :

$$\begin{cases} \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad , \\ \hat{\beta}_1 = \dfrac{S_{yx}}{S_{xx}} = \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \end{cases}$$

Implémentation en Python : Régression Linéaire

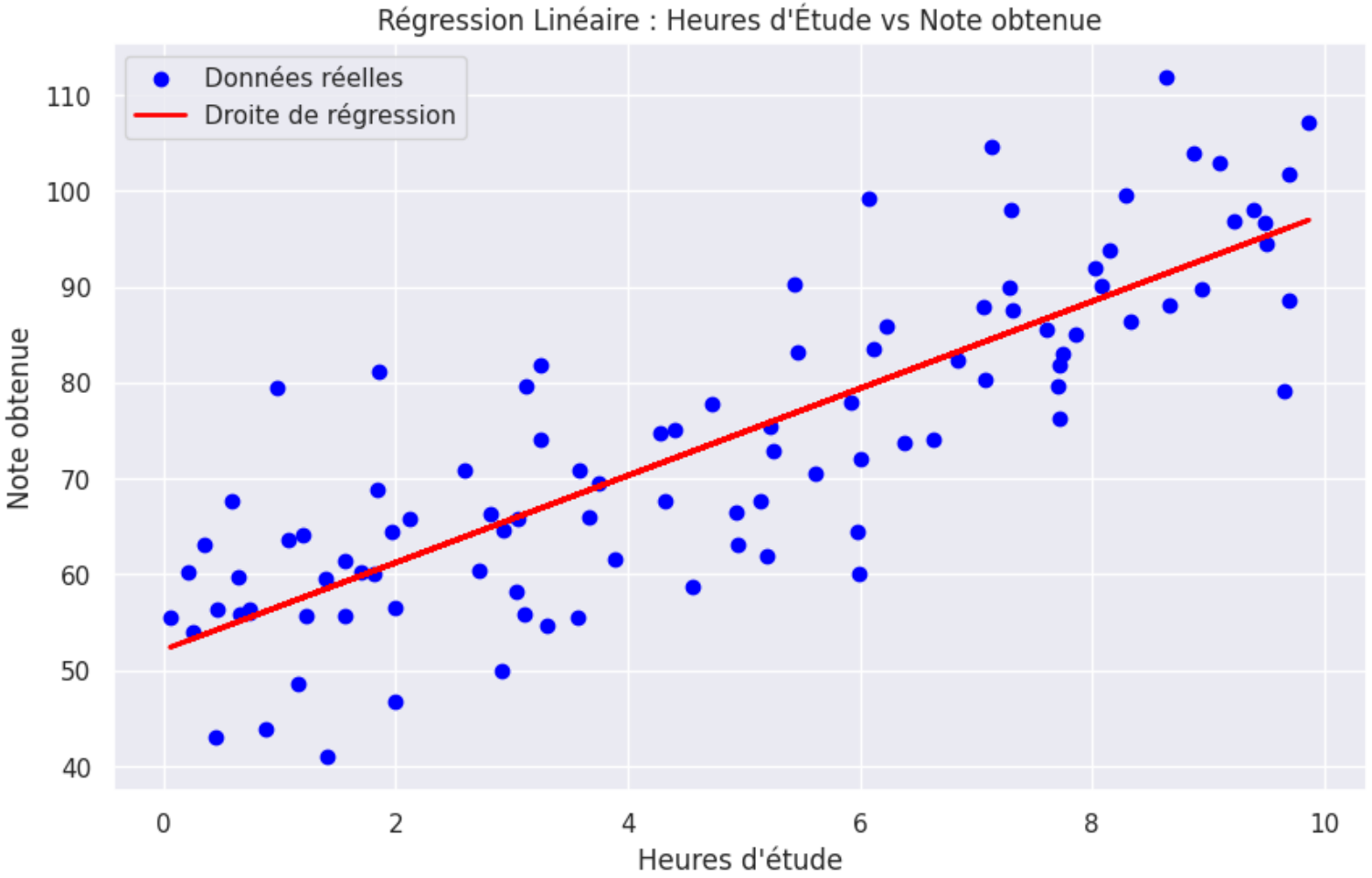
```python
# Importation des bibliothèques nécessaires
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from google.colab import drive
import os
# Étape 1: Monter Google Drive
drive.mount('/content/drive')
# Étape 2: Définir le chemin du fichier CSV dans Google Drive
dossier = "Colab Notebooks"  # Modifier si nécessaire
nom_fichier = "StudentGrades.csv"
chemin_fichier = f"/content/drive/My Drive/{dossier}/{nom_fichier}"
# Étape 3: Charger les données
print("Chargement des données depuis Google Drive...")
data = pd.read_csv(chemin_fichier)
print("Données chargées avec succès.")
print(data.head())  # Afficher les premières lignes du jeu de données
# Étape 4: Extraction des variables
X = data[['Hours Studied']].values
y = data['Grades'].values
```

Implémentation en Python : Régression Linéaire

```
# Étape 5: Création et entraînement du modèle de régression linéaire
print("\nEntraînement du modèle de régression linéaire...")
modele = LinearRegression()
modele.fit(X, y)
print("Modèle entraîné avec succès.") # Étape 6: Faire des prédictions
y_pred = modele.predict(X)
plt.figure(figsize=(10, 6)) # Étape 7: Visualisation
plt.scatter(X, y, color='blue', label='Données réelles')
plt.plot(X, y_pred, color='red', linewidth=2, label='Droite de régression')
plt.title("Régression Linéaire : Heures d'Étude vs Note obtenue")
plt.xlabel("Heures d'étude")
plt.ylabel("Note obtenue")
plt.legend()
plt.grid(True)
plt.show()
intercept = modele.intercept_ # Étape 8: Affichage des coefficients du modèle
pente = modele.coef_[0]
r2 = r2_score(y, y_pred)
print("\nRésultats du Modèle de Régression Linéaire :")
print(f"Intercept (beta0) : {intercept:.2f}")
print(f"Pente (beta1) : {pente:.2f} (Impact de chaque heure d'étude sur la note)")
print(f"Score R² : {r2:.4f} (Indicateur de la qualité de l'ajustement du modèle)")
```

```
Résultats du Modèle de Régression Linéaire :
Intercept (beta0) : 52.15
Pente (beta1) : 4.54 (Impact de chaque heure d'étude sur la note)
Score R² : 0.6913 (Indicateur de la qualité de l'ajustement du modèle)
```

Implémentation en Python : Régression Linéaire



Biais et Variance des Estimateurs des Moindres Carrés

Estimation par La Méthode des Moindres Carrés : Biais

- Nous cherchons à vérifier si les estimateurs des coefficients de régression linéaire obtenus par la méthode des moindres carrés (OLS) sont biaisés ou non.
- Autrement dit, nous voulons vérifier si:

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

- Nous allons étudier cela pour chacun des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_0$ séparément.

Estimation par La Méthode des Moindres Carrés : Biais

- L'estimateur $\hat{\beta}_1$ est donné par :

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

- En remplaçant y_i par son expression dans le modèle linéaire :

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

- Nous obtenons :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i + \epsilon_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Estimation par La Méthode des Moindres Carrés : Biais

- Nous savons que :

$$\overline{y} = \beta_0 + \beta_1 \overline{x} + \overline{\epsilon}, \quad \text{où} \quad \overline{\epsilon} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i.$$

- Donc,

$$y_i - \overline{y} = \beta_1 (x_i - \overline{x}) + (\epsilon_i - \overline{\epsilon}).$$

- En substituant cette expression dans $\hat{\beta}_1$, nous avons :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})[\beta_1 (x_i - \overline{x}) + (\epsilon_i - \overline{\epsilon})]}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Estimation par La Méthode des Moindres Carrés : Biais

- Développons cette expression :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x})(\epsilon_i - \overline{\epsilon})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

- En prenant l'espérance :

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 + \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \overline{x})(\epsilon_i - \overline{\epsilon})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right].$$

- Sous l'hypothèse que $\mathbb{E}[\epsilon_i] = 0$ et que les ϵ_i sont indépendants des X_i :

$$\mathbb{E}\left[\sum_{i=1}^n (x_i - \overline{x})(\epsilon_i - \overline{\epsilon})\right] = 0.$$

- Donc :

$$\mathbb{E}[\hat{\beta}_1] = \beta_1.$$

Estimation par La Méthode des Moindres Carrés : Biais

- L'estimateur $\hat{\beta}_0$ est donné par :

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

- En prenant l'espérance :

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\overline{y}] - \mathbb{E}[\hat{\beta}_1]\overline{x}.$$

Estimation par La Méthode des Moindres Carrés : Biais

- Nous avons :

$$\mathbb{E}[\overline{y}] = \beta_0 + \beta_1 \overline{x}.$$

- Donc :

$$\mathbb{E}[\hat{\beta}_0] = (\beta_0 + \beta_1 \overline{x}) - \mathbb{E}[\hat{\beta}_1]\overline{x}.$$

- Puisque $\mathbb{E}[\hat{\beta}_1] = \beta_1$, on a :

$$\mathbb{E}[\hat{\beta}_0] = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0.$$

Estimation par La Méthode des Moindres Carrés : Biais

- Nous avons démontré que :

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

- Cela prouve que les estimateurs des moindres carrés ordinaires (OLS) sont **non biaisés**.

Estimation par La Méthode des Moindres Carrés : Variance

- Nous avons établi que l'estimateur des moindres carrés pour β_1 est donné par :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- En remplaçant y_i par son expression dans le modèle de régression linéaire :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

- On obtient :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})[\beta_0 + \beta_1 x_i + \epsilon_i - \overline{y}]}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Estimation par La Méthode des Moindres Carrés : Variance

- On sait que :

$$\overline{y} = \beta_0 + \beta_1 \overline{x} + \overline{\epsilon}, \quad \text{où} \quad \overline{\epsilon} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i.$$

- Ainsi, on peut écrire :

$$y_i - \overline{y} = \beta_1 (x_i - \overline{x}) + (\epsilon_i - \overline{\epsilon}).$$

- En substituant cette expression dans $\hat{\beta}_1$, nous avons :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})[\beta_1 (x_i - \overline{x}) + (\epsilon_i - \overline{\epsilon})]}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Estimation par La Méthode des Moindres Carrés : Variance

- Décomposons les termes :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x})(\epsilon_i - \overline{\epsilon})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

- En prenant la variance des deux côtés :

$$\text{Var}(\hat{\beta}_1) = \text{Var}\left(\frac{\sum_{i=1}^n (x_i - \overline{x})(\epsilon_i - \overline{\epsilon})}{\sum_{i=1}^n (x_i - \overline{x})^2} \right).$$

Estimation par La Méthode des Moindres Carrés : Variance

- Nous utilisons la propriété de la variance :

$$\text{Var}(aX) = a^2\text{Var}(X).$$

- Sous l'hypothèse d'homoscédasticité :

$$\text{Var}(\epsilon_i) = \sigma^2.$$

- Et sous l'hypothèse d'indépendance des erreurs, la variance d'une somme de termes indépendants est la somme de leurs variances :

$$\text{Var}\left(\sum_{i=1}^{n} a_i\epsilon_i\right) = \sum_{i=1}^{n} a_i^2\text{Var}(\epsilon_i).$$

Estimation par La Méthode des Moindres Carrés : Variance

- Dans notre cas, les coefficients sont $(x_i - \overline{x})$, donc :

$$\text{Var}\left(\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i\right) = \sum_{i=1}^{n}(x_i - \overline{x})^2\sigma^2.$$

- En divisant par $\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2$:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2}.$$

- En simplifiant :

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Estimation par La Méthode des Moindres Carrés : Variance

- La variance de $\hat{\beta}_1$ est inversement proportionnelle à $\sum_{i=1}^{n}(x_i - \overline{x})^2$.
- Plus les valeurs de x_i sont dispersées, plus la variance est faible, et donc l'estimation de $\hat{\beta}_1$ est plus précise.
- À l'inverse, si les x_i sont très regroupés autour de leur moyenne, la variance de $\hat{\beta}_1$ est plus grande, ce qui rend l'estimation moins fiable.

Estimation par La Méthode des Moindres Carrés : Variance

- Nous avons :

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

- En utilisant la variance :

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\overline{y} - \hat{\beta}_1 \overline{x}).$$

- Comme $\overline{y}$ et $\hat{\beta}_1$ sont corrélés, nous appliquons la propriété de la variance pour une somme :

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\overline{y}) + \overline{x}^2 \text{Var}(\hat{\beta}_1) - 2\overline{x}\text{Cov}(\overline{y}, \hat{\beta}_1).$$

Estimation par La Méthode des Moindres Carrés : Variance

- En utilisant les propriétés des variances et en supposant que X_i est déterministe :

$$\text{Var}(\overline{y}) = \frac{\sigma^2}{n}.$$

- La covariance entre $\overline{y}$ et $\hat{\beta}_1$ est donnée par :

$$\text{Cov}(\overline{y}, \hat{\beta}_1) = \frac{\sigma^2 \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- Ainsi, nous obtenons :

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right).$$

Estimation par La Méthode des Moindres Carrés : Variance

- Nous avons démontré les expressions de variance des estimateurs OLS :

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right).$$

- Ces formules montrent que la précision des estimateurs augmente lorsque la variance des x_i augmente et lorsque le nombre d'observations n augmente.
- Ces résultats sont essentiels pour évaluer la qualité des estimations et pour construire des intervalles de confiance dans l'analyse de régression linéaire.

Equivalence de la Méthode des Moindres Carrés et du Maximum de Vraisemblance

Méthode du Maximum de Vraisemblance

- Nous cherchons à vérifier si l'estimation des coefficients d'un modèle de régression linéaire par la méthode des moindres carrés (OLS) est équivalente à l'estimation par la méthode du maximum de vraisemblance (MLE).
- On se repose sur l'hypothèse que les erreurs suivent une loi normale pour vérifier cela.
- Soit le modèle de régression linéaire simple :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, ..., n$$

- Hypothèse sur les erreurs :

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \text{indépendantes et identiquement distribuées (i.i.d)}$$

Méthode du Maximum de Vraisemblance

- La méthode des moindres carrés consiste à minimiser la somme des carrés des résidus :

$$C(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Les estimateurs OLS sont obtenus en annulant les dérivées partielles :

$$\frac{\partial C}{\partial \hat{\beta}_0} = 0, \quad \frac{\partial C}{\partial \hat{\beta}_1} = 0.$$

- Cela donne les estimateurs :

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

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

Méthode du Maximum de Vraisemblance

- La vraisemblance du modèle sous l'hypothèse de normalité des erreurs est donnée par :

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)$$

- En prenant le logarithme, on obtient la log-vraisemblance :

$$\log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2.$$

- Maximiser cette fonction revient à minimiser :

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

Méthode du Maximum de Vraisemblance

- Puisque la maximisation de la vraisemblance revient à minimiser la somme des carrés des résidus,
- Les estimateurs obtenus par MLE sont les mêmes que ceux obtenus par OLS.
- Ainsi, sous l'hypothèse de normalité des erreurs, OLS et MLE produisent des estimateurs identiques.