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

Leçon 1 : Régression Linéaire

Polytechnique Montréal - Hiver 2025

Chiheb Trabelsi

February 12, 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**:

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

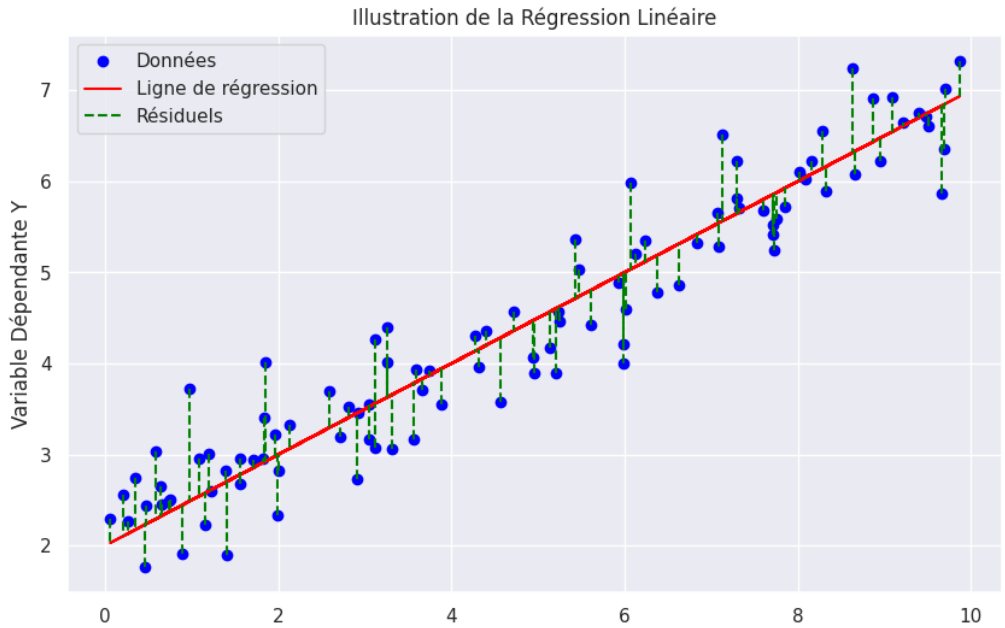
  - ϵ l'erreur aléatoire du modèle.
  - $X_1, X_2, ..., X_p$ les variables explicatives, appelées variables indépendantes.
  - Y la variable dépendante, appelée aussi réponse.
  - β_0 est l'intercept du modèle, qui représente la valeur attendue de Y lorsque toutes les variables indépendantes X_i sont égales 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

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.

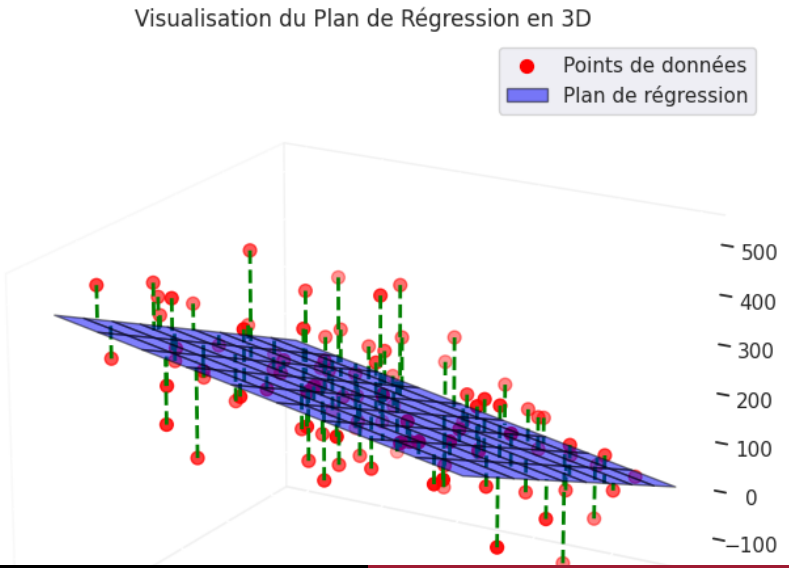


Intro : Modèle de Regression Linéaire Multiple (à 2 Vars)

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

Table des Matières

Modèle de Régression Linéaire simple

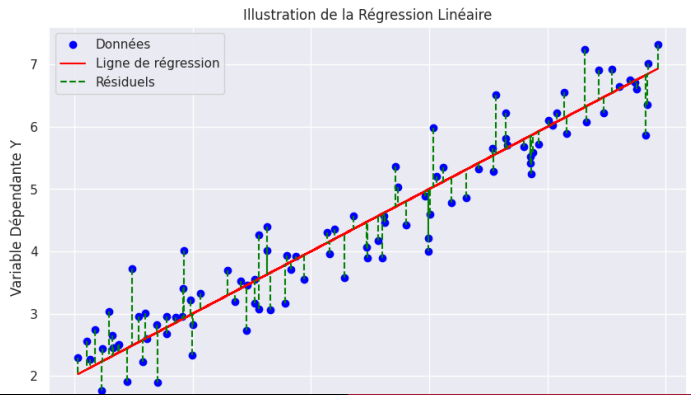
Modèle de Régression Linéaire simple : Introduction

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 $\frac{S_{yx}}{S_{xx}}$:

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

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

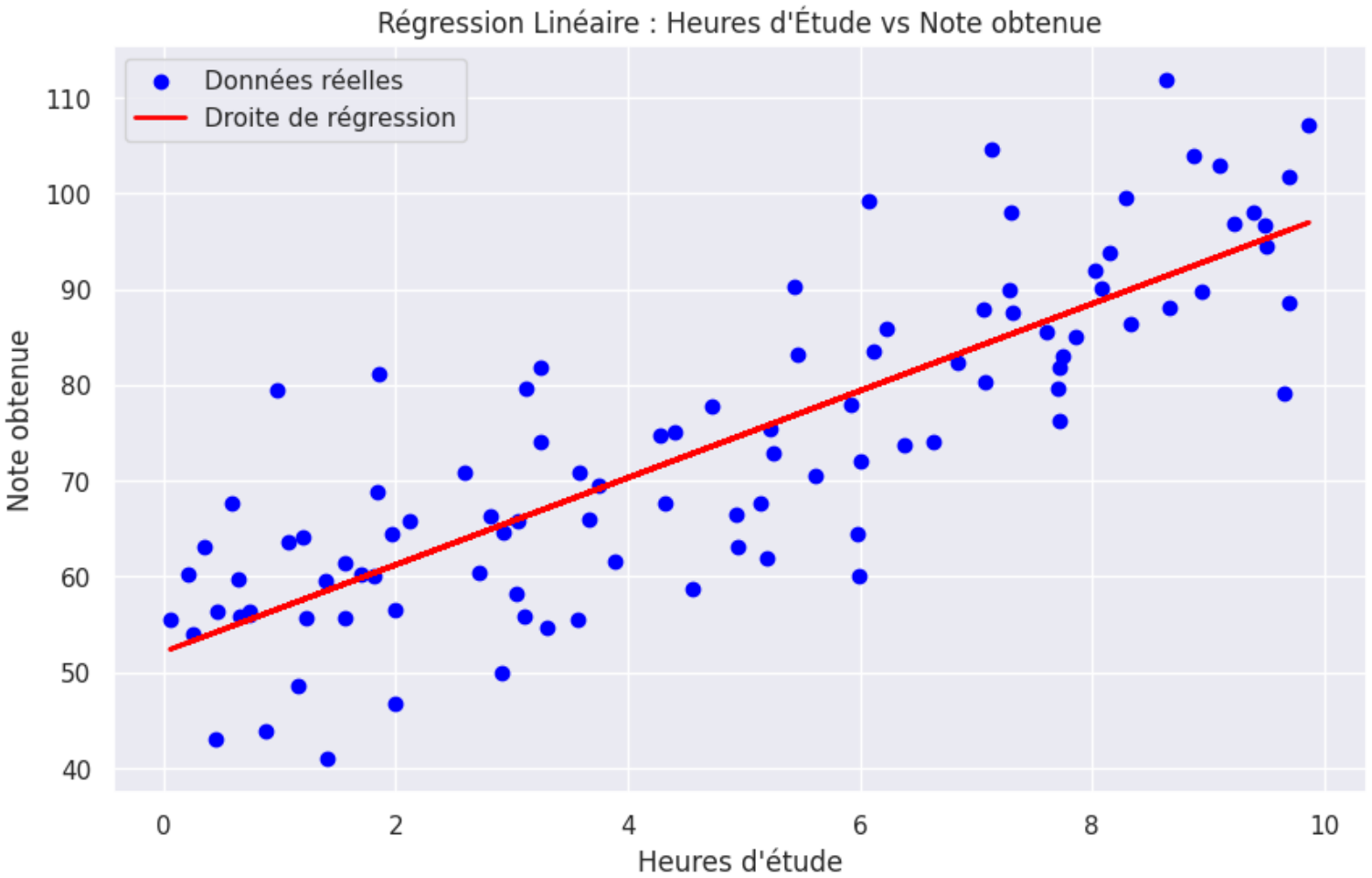
```python
# Importation des bibliothèques nécessaires
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from google.colab import drive
import os
# Étape 1: Monter Google Drive
drive.mount('/content/drive')
# 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

```python
# 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)
```

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 S_{xx} 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.

Résumé des Points Importants et Normalité des Y_i

Résumé : 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).

Résumé : Résultats Importants

- Les estimateurs obtenus par la méthode des moindres carrés (ou aussi la méthode du maximum de vraisemblance) sont :

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

- Moyenne et Variance de Y_i:

$$\begin{cases} \mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i, \\ \text{Var}(Y_i) = \sigma^2. \end{cases}$$

- Moyennes et Variances des Estimateurs :

$$\begin{cases} \mathbb{E}[\hat{\beta}_0] = \beta_0 \quad \text{et} \quad \text{Var}(\hat{\beta}_0) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) = \sigma^2 (\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}), \\ \mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \text{et} \quad \text{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \dfrac{\sigma^2}{S_{xx}}. \end{cases}$$

Distribution de Y_i

- Nous avons le modèle de régression linéaire simple donné par :
  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$
- X_i sont considérés comme **fixes (déterministes)**.
- Les erreurs ϵ_i suivent une loi normale : $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, tels que :
  - **L'hypothèse d'homoscédasticité** est vérifiée :
    $\text{Var}(\epsilon_i) = \sigma^2$ pour tout i.
  - **Les erreurs ϵ_i sont non corrélées** entre elles :
    $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ pour tout $i \neq j$
- Moyenne et Variance de Y_i:

$$\begin{cases} \mathbb{E}(Y_i \mid X_i) = \mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i, \\ \text{Var}(Y_i \mid X_i) = \text{Var}(Y_i) = \sigma^2. \end{cases} \quad \Rightarrow \quad Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i,\, \sigma^2)$$

- Puisque la somme d'une variable normale avec une constante reste normale, nous obtenons : $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i,\, \sigma^2)$

Distribution de Y_i

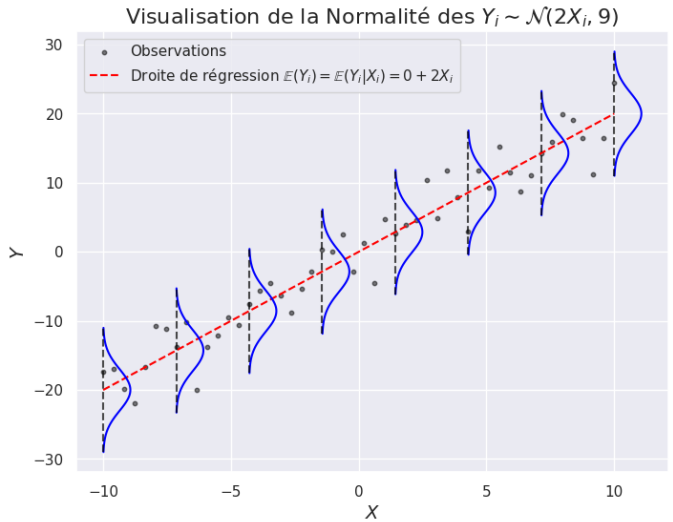
- Y_i suit une loi normale dont la moyenne **dépend linéairement de** X_i: $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$.
- La variance σ^2 de Y_i est égale à celle des erreurs. Elle est constante pour tout i ce qui est une conséquence de l'homoscédasticité des erreurs ϵ_i.
- Cette normalité est essentielle car elle permet d'appliquer :
  - Tests d'hypothèses (test de Student, test de Fisher, etc).
  - Intervalles de confiance sur les coefficients β_0 et β_1.
  - Intervalles de prédiction pour de nouvelles observations.

Distribution des Y_i

- Pour chaque valeur de X_i, la distribution de Y_i est une loi normale :

$$Y_i \sim \mathcal{N}(2X_i, 9).$$

- Ces distributions sont représentées par les courbes en bleu superposées verticalement à différents X_i.
- Chaque courbe illustre la dispersion de Y_i autour de la moyenne conditionnelle $\mathbb{E}(Y_i) = \mathbb{E}(Y_i|X_i)$.
- La variance fixe $\sigma^2 = 9$ montre que la dispersion autour de la droite de régression est constante.



Distribution des Y_i

- Pour l'exemple où $Y_i \sim \mathcal{N}(2X_i, 1)$, chaque distribution est centrée sur $\mathbb{E}(Y_i|X_i) = 2X_i$ et montre la dispersion de Y_i autour de cette valeur.
- Pour $X_i = 1$, la zone en orange représente $P(Y \geq 2)$.
- Pour $X_i = 2$, la zone en rouge représente $P(Y \leq 4)$.
- On peut ainsi mieux comprendre et interpréter la signification des intervalles de confiance.

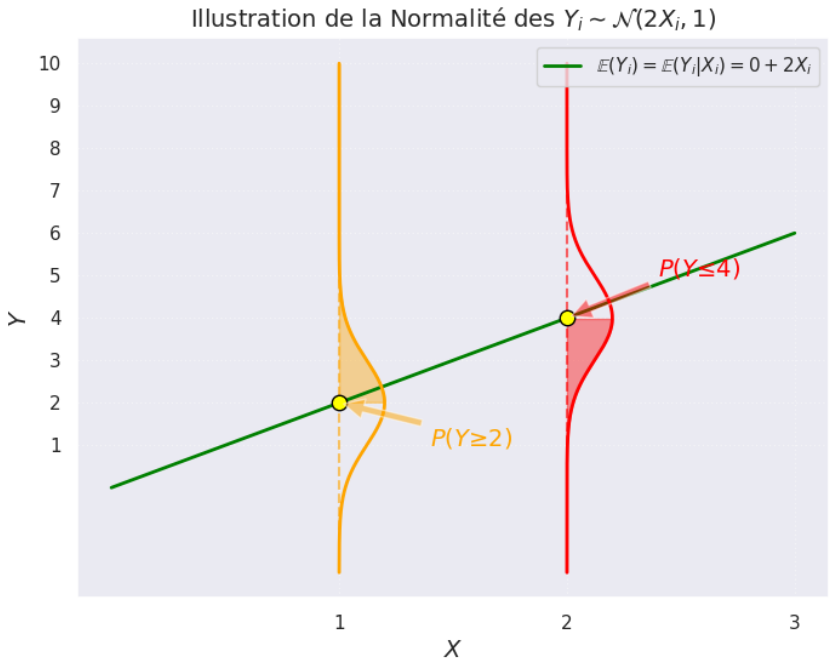


Table des Matières

Décomposition de la Variabilité Totale

Décomposition de la Variabilité Totale : Introduction

- Dans l'analyse de la régression linéaire, on cherche à expliquer la dispersion des valeurs observées de y_i par rapport à leur moyenne $\overline{y}$.
- Cette dispersion est exprimée par la **variabilité totale** notée SC_{totale} qui est la somme des carrés des différences entre les valeurs observées y_i et leur moyenne $\overline{y}$:

$$SC_{totale} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- L'idée centrale repose sur le fait que toute valeur observée y_i peut être décomposée en deux parties :

$$y_i = \hat{y}_i + e_i$$

où :

  - $\hat{y}_i$ est la valeur prédite par le modèle de régression : la partie expliquée par le modèle.
  - $e_i = y_i - \hat{y}_i$ est le résidu qui est la partie non expliquée par le modèle.

Décomposition de la Variabilité Totale : Introduction

- L'objectif est donc de quantifier :
  - La partie de la dispersion expliquée par le modèle de régression.
  - La partie due à d'autres facteurs non inclus dans le modèle, ce qui est reprsentée par les résidus.
- On peut alors exprimer l'écart entre y_i et sa moyenne en le décomposant en 2 parties :

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + e_i$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - \overline{y}) = \sum_{i=1}^{n}(\hat{y}_i - \overline{y}) + \sum_{i=1}^{n}(e_i)$$

où :

- $\hat{y}_i$ est la valeur prédite par le modèle de régression : la partie expliquée par le modèle.
- $e_i = y_i - \hat{y}_i$ est le résidu qui est la partie non expliquée par le modèle.

Décomposition de la Variabilité Totale : Introduction

- On peut démontrer que (**voir Annexe pour la démonstration**) :

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SC_{totale}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{SC_{reg}}+\underbrace{\sum_{i=1}^{n}(e_i)^2}_{SC_{res}} \quad , \quad \text{où :}$$

- $SC_{totale}=\sum_{i=1}^{n}(y_i-\overline{y})^2$ est la variabilité (ou dispersion) totale des y_i autour de leur moyenne $\overline{y}$.
- $SC_{reg}=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$ la variabilité (ou dispersion) des valeurs observées y_i et qui est expliquée par le modèle de régression.
- $SC_{res}=\sum_{i=1}^{n}(e_i)^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ la variabilité ou dispersion des résidus e_i et qui est **notre fonction à minimiser par la méthode des moindre carrés**.

Décomposition de la Variabilité Totale : R^2

- La proportion de la variabilité totale qui expliquée par la régression est exprimée par le **coefficient de détermination** R^2 :

$$R^2 = \frac{SC_{reg}}{SC_{totale}} = 1 - \frac{SC_{res}}{SC_{totale}}$$

- Dans le cas où l'on minimise SC_{res}, ceci se traduit par:
  - La minimisation de la proportion de SC_{res} qui contribue à SC_{totale}.
  - La maximisation de la propostion de SC_{reg} qui contribue à SC_{totale}.
- Le coefficient de détermination R^2 permet de quantifier la qualité de l'ajustement du modèle *(model fitting)* via le coefficient de détermination R^2, qui mesure la proportion de la variabilité totale expliquée par la régression :
- Elle permet aussi d'analyser dans quelle mesure la régression linéaire permet de comprendre et prédire les variations de Y, et quelle part de l'information est perdue ou non expliquée.

Décomposition de la Variabilité Totale : R^2

- $0 \leq R^2 = \frac{SC_{reg}}{SC_{totale}} \leq 1.$
- Si $R^2 \approx 1$, cela signifie que presque toute la variabilité est expliquée par la régression. Le modèle est parfaitement ajusté aux données (*100% model fitting*).
- Si $R^2 \approx 0$, cela signifie que le modèle n'explique presque rien ($\approx 0\%$ *model fitting*).
- Si SC_{res} est trop grand, il faut peut-être ajouter d'autres variables explicatives.
- On peut effectuer des tests d'hypothèse en utilisant l'ANOVA (Analyse de la Variance) pour voir si la régression explique significativement les y_i.
- Si la variabilité expliquée (SC_{reg}) est faible par rapport à la variabilité totale (SC_{totale}), alors X n'est peut-être pas une bonne variable explicative pour Y.

Décomposition de la Variabilité Totale : Limitations de R^2

- **Ne mesure pas la causalité**
  - On cherche à savoir si X n'est peut-être pas une bonne variable explicative pour Y. Un R^2 élevé n'implique pas que X cause Y, mais seulement qu'ils sont corrélés.
  - La relation peut être influencée par des variables omises ou des relations fallacieuses.
- **Sensible au nombre de variables explicatives**
  - Ajouter des variables dans le modèle augmente systématiquement R^2, même si elles n'ont pas un véritable impact sur Y.
  - **Solution** : utiliser le R^2 ajusté :

$$R^2_{\text{ajusté}} = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}$$

    où p est le nombre de variables explicatives et n la taille de l'échantillon.
- **Ne prend pas en compte la qualité de l'ajustement**
  - Un R^2 élevé peut masquer un ajustement médiocre si les résidus présentent des structures non captées par le modèle (ex : tendance non linéaire). Une visualisation des résidus est alors indispensable.

Décomposition de la Variabilité Totale : Limitations de R^2

- **Non adapté aux modèles non linéaires**
  - Dans des modèles non linéaires (régression logistique, arbres de décision, etc.), R^2 perd son interprétation et n'est souvent pas utilisé.
- **Ne reflète pas nécessairement un bon pouvoir prédictif**
  - Un modèle peut avoir un R^2 élevé mais être inefficace pour prédire de nouvelles observations si le modèle est **surajusté à l'échantillon d'entraînement(model overfitting)**.
- **Influence des points extrêmes et des valeurs aberrantes**
  - Un petit nombre de points extrêmes peut fausser R^2, le rendant artificiellement élevé ou bas.
- **Impact des grandes valeurs de la pente β_1**
  - Un modèle avec une pente très élevée peut donner un R^2 artificiellement grand sans pour autant garantir un bon ajustement du modèle.
  - Une grande amplitude des valeurs de X peut accentuer ce phénomène et exagérer la proportion de variance expliquée.
  - Il est donc important de normaliser les variables ou d'examiner les coefficients pour éviter des illusions sur la performance du modèle.

Décomposition de la Variabilité Totale : Exemple Pratique

- Imaginons qu'on veux comprendre pourquoi les notes des étudiants varient dans une classe.
- La variabilité totale, $SC_{totale} = \sum_{i=1}^{n}(y_i - \overline{y})^2$, représente l'ensemble des variations des notes autour de la moyenne.
- La variabilité expliquée, $SC_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$, est la part des variations des notes qui s'explique par le nombre d'heures d'étude (variable X).
- La variabilité inexpliquée (par le modèle), $SC_{res} = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, correspond aux différences dues à d'autres facteurs :
  - stress,
  - sommeil,
  - méthode d'apprentissage, etc.
- En décomposant la variabilité totale, on peut voir dans quelle mesure notre modèle basé sur le nombre d'heures d'étude aide à expliquer les notes.

Table des Matières

Analyse de la Variance (ANOVA) dans le Contexte de la Régression Linéaire Simple

ANOVA en Régression Linéaire Simple

ANOVA en Régression Linéaire Simple

- **L'analyse de variance (ANOVA)** en régression linéaire vise à comprendre dans quelle mesure un modèle de régression est capable d'expliquer la variabilité d'une variable dépendante Y à partir d'une variable indépendante X.
- L'ANOVA permet donc d'évaluer si l'inclusion de la variable X améliore significativement les prédictions de Y ou si les variations de Y sont majoritairement dues au hasard.
- L'idée principale est d'exploiter la décomposition de la variabilité totale pour **évaluer si la variabilité expliquée par le modèle de régression SC_{reg} est significativement plus grande que la variabilité résiduelle SC_{res}.**
- Rappel de la L'égalité fondamentale donnée par la décomposition de la variabilité totale :

$$SC_{\text{tot}} = SC_{\text{reg}} + SC_{\text{res}}.$$

ANOVA (Rappel) : χ^2 pour la Variance Empirique

- **Relation entre χ^2 et les variables normales :**
  - La somme des carrés de k variables normales standards indépendantes suit une distribution χ^2 avec k degrés de liberté : $X \sim \chi^2_k, \quad \text{où } X = \sum_{i=1}^{k} Z_i^2, \quad Z_i \sim N(0,1).$
- **Lien avec les écarts quadratiques et la variance :**
  - Pour n variables alétoires $X_1, X_2, \ldots, X_n$ suivant $\mathcal{N}(\mu, \sigma^2)$, les écarts $X_i - \mu$ sont également normalement distribués, où $(X_i - \mu) \sim \mathcal{N}(0, \sigma^2)$.
  - En standardisant les écarts : $Z_i = \frac{X_i - \mu}{\sigma}, \quad Z_i \sim \mathcal{N}(0,1).$
  - La somme des carrés de ces écarts : $\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n.$
- **Variance de l'échantillon :**
  - La variance empirique de l'échantillon s^2 est reliée aux écarts quadratiques : $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, où $\bar{X}$ est la moyenne de l'échantillon.
  - $\Rightarrow$ Sous l'hypothèse de la normalité des X_i, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$

ANOVA en Régression Linéaire Simple

- Dans un modèle de régression linéaire simple :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{où} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Les erreurs ϵ_i sont supposées i.i.d.
- La variabilité inexpliquée est la somme des carrés des résidus :

$$SC_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  Où $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ est la valeur ajustée.
- Puisque les ϵ_i sont indépendants et normaux, alors leur somme des carrés suit une loi du χ^2 avec $n-2$ degrés de liberté :

$$\frac{SCres}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{e_i - 0}{\sigma} \right)^2 \sim \chi^2_{n-2}.$$

- le degré de liberte est $n-2$ et non n car β_0 et β_1 sont estimés

ANOVA en Régression Linéaire Simple

- Sous l'hypothèse de normalité des erreurs :

$$\frac{SC_{res}}{\sigma^2} \sim \chi^2_{n-2}$$

- Sous l'hypothèse que $\beta_1 = 0$:

$$\frac{SC_{reg}}{\sigma^2} \sim \chi^2_1$$

- Ces résultats sont fondamentaux pour les tests d'hypothèses en régression linéaire, en particulier pour le test F, qui repose sur le rapport de ces deux statistiques du χ^2.

ANOVA (Rappel) : Distribution F de Fisher

- **Définition :**
  - Soit U et V deux variables aléatoires indépendantes où l'on a:

$$\begin{cases} U \sim \chi^2_{d_1} & \text{avec } d_1 \text{degrés de liberté,} \\ V \sim \chi^2_{d_2} & \text{avec } d_2 \text{degrés de liberté.} \end{cases}$$

  - La variable aléatoire continue F définie par

$$F = \frac{U/d_1}{V/d_2}$$

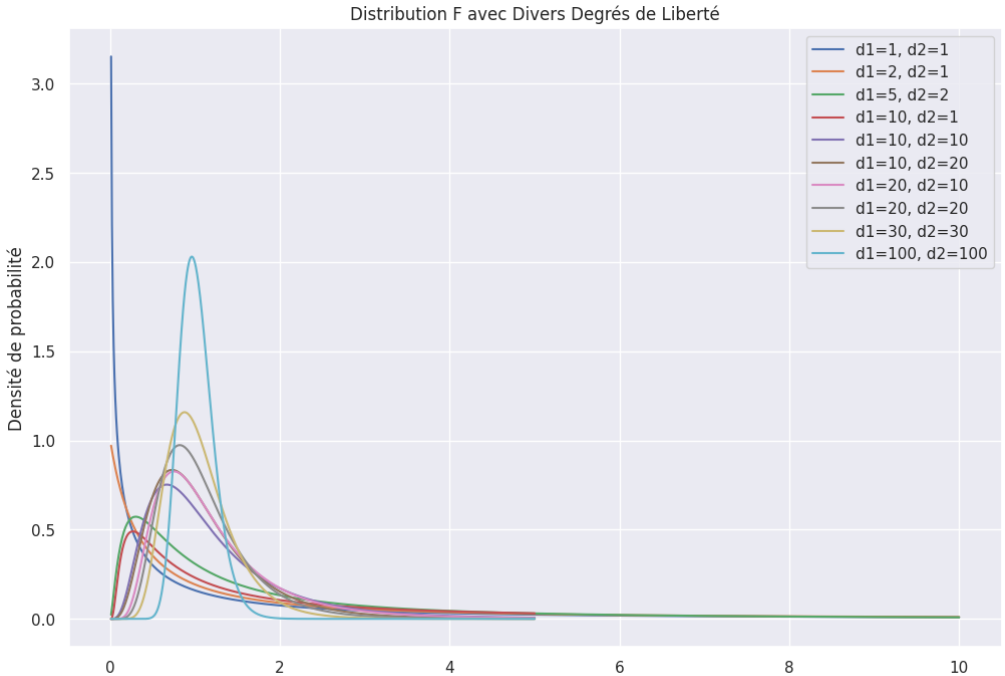
    suit une loi de Fisher avec d_1 et d_2 degrés de liberté : $F \sim F_{d_1,d_2}$.
- **Propriétés**
  - Utilisée pour modéliser le ratio de deux variances échantillonnales avec des degrés de liberté, d_1 et d_2, liés aux 2 échantillons comparés.
  - Utilisée principalement pour comparer deux variances et dans l'analyse de la variance (ANOVA).
  - Espérance : $\mathbb{E}[F] = \frac{d_2}{d_2-2}$ pour $d_2 > 2$.
  - Variance : $\text{Var}[F] = \frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$, pour $d_2 > 4$.

ANOVA (Rappel) : Distribution F de Fisher

- La densité de probabilité de $F \sim F_{d_1,d_2}$ est donnée par :

$$f_F(f; d_1, d_2) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} f^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}f\right)^{-\frac{d_1+d_2}{2}},$$

  - f est la valeur que prend la variable aléatoire F.
  - Asymétrique, avec un support de $[0, \infty)$.



ANOVA (Rappel) : Test F de Fisher

- Le test F est utilisé pour comparer les variances de deux populations indépendantes.
- **Hypothèses :**
  - H_0: Les variances des deux populations sont égales ($\sigma_1^2 = \sigma_2^2$).
  - H_a: Les variances des deux populations sont différentes ($\sigma_1^2 \neq \sigma_2^2$).
- **Statistique de test :**

$$F = \frac{s_1^2}{s_2^2}$$

où s_1^2 et s_2^2 sont les variances échantillonnales des deux groupes.

- **Degrés de liberté :** $d_1 = n_1 - 1, \quad d_2 = n_2 - 1$ où n_1 et n_2 les tailles respectives des 2 échantillons.

ANOVA (Rappel) : Test F de Fisher

- **Étape 1. Formulation des hypothèses :**
  - $H_0 : \sigma_1^2 = \sigma_2^2$
  - $H_a : \sigma_1^2 \neq \sigma_2^2$
- **Étape 2. Calcul de la statistique de test :**

$$F = \frac{s_1^2}{s_2^2}$$

- **Étape 3. Degrés de liberté :**

$$d_1 = n_1 - 1, \quad d_2 = n_2 - 1$$

- **Étape 4. Décision :**
  - Utiliser la distribution F avec (d_1, d_2) degrés de liberté pour trouver la ou les valeurs critiques à un niveau de signification donné α, ou calculer la p-valeur.
  - Rejeter H_0 si la valeur observée de F est plus extrême que la ou les valeurs critiques.

ANOVA : Test F de Fisher

- L'ANOVA utilise le test F pour évaluer si la variabilité expliquée par la régression est significativement plus grande que la variabilité résiduelle.
- **Interprétation** :
  - Si la variabilité expliquée est beaucoup plus grande que la variabilité résiduelle, cela signifie que **la variable X apporte potentiellement une information utile pour prédire Y**.
  - Si la variabilité expliquée est faible, cela signifie que **le modèle ne fait pas mieux qu'une simple moyenne**, donc X n'est peut-être pas une bonne variable explicative.
- L'hypothèse nulle testée est : $H_0 : \beta_1 = 0$.
- Nous avons : $\begin{cases} \frac{SC_{res}}{\sigma^2} \sim \chi^2_{n-2} \\ \frac{SC_{reg}}{\sigma^2} \sim \chi^2_1 \end{cases}$ sous l'hypothèse $\beta_1 = 0$.

ANOVA : Test F de Fisher

- L'hypothèse nulle testée est : $H_0 : \beta_1 = 0$.
- On a $\frac{SC_{res}}{\sigma^2} \sim \chi^2_{n-2}$ et $\frac{SC_{reg}}{\sigma^2} \sim \chi^2_1$ sous l'hypothèse $\beta_1 = 0$
- La variable aléatoire continue F définie par

$$F = \frac{U/d_1}{V/d_2}$$

suit une loi de Fisher avec d_1 et d_2 degrés de liberté : $F \sim F_{d_1,d_2}$ où

$$\begin{cases} U \sim \chi^2_{d_1} & \text{avec } d_1 \text{degrés de liberté,} \\ V \sim \chi^2_{d_2} & \text{avec } d_2 \text{degrés de liberté.} \end{cases}$$

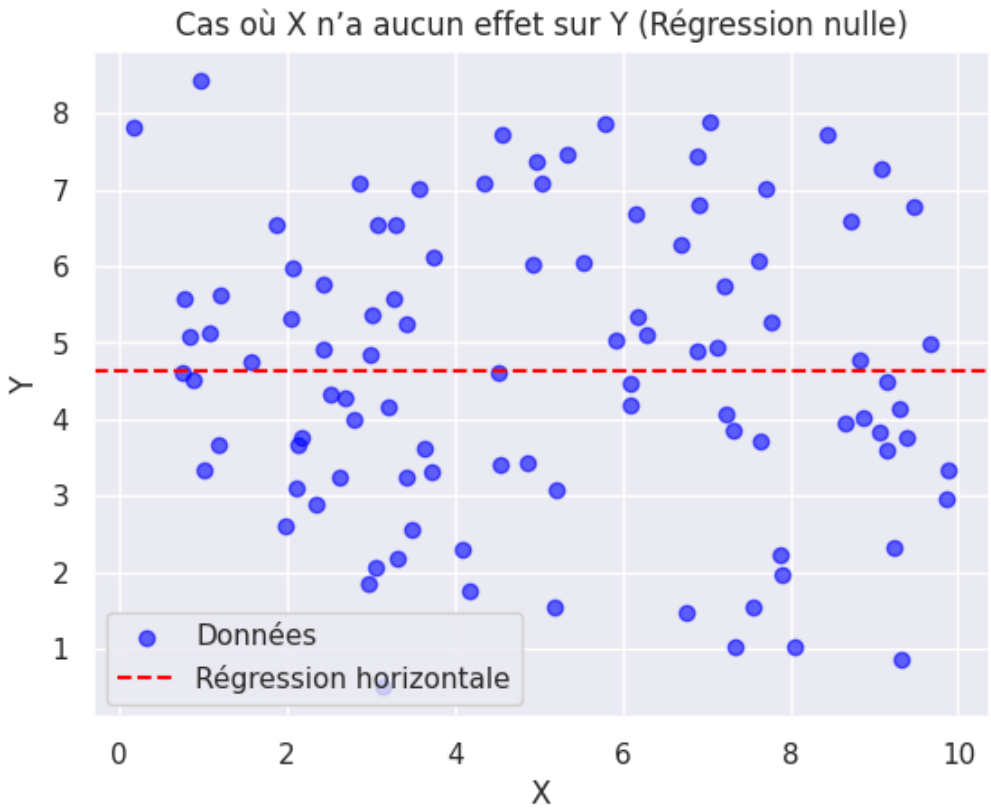
- La F_{stat} qui est calculée dans le contexte de l'ANOVA en régression linéaire simple est :

$$F_{stat} = \frac{(SC_{reg}/\sigma^2)/1}{(SC_{res}/\sigma^2)/(n-2))} = \frac{SC_{reg}/1}{SC_{res}/(n-2)} = \frac{MC_{reg}}{MC_{res}}.$$

- Où MC_{reg} et MC_{res} sont respectivement SC_{reg} et SC_{res} normalisées par leurs degrés de liberté respectives 1 et $n-2$.

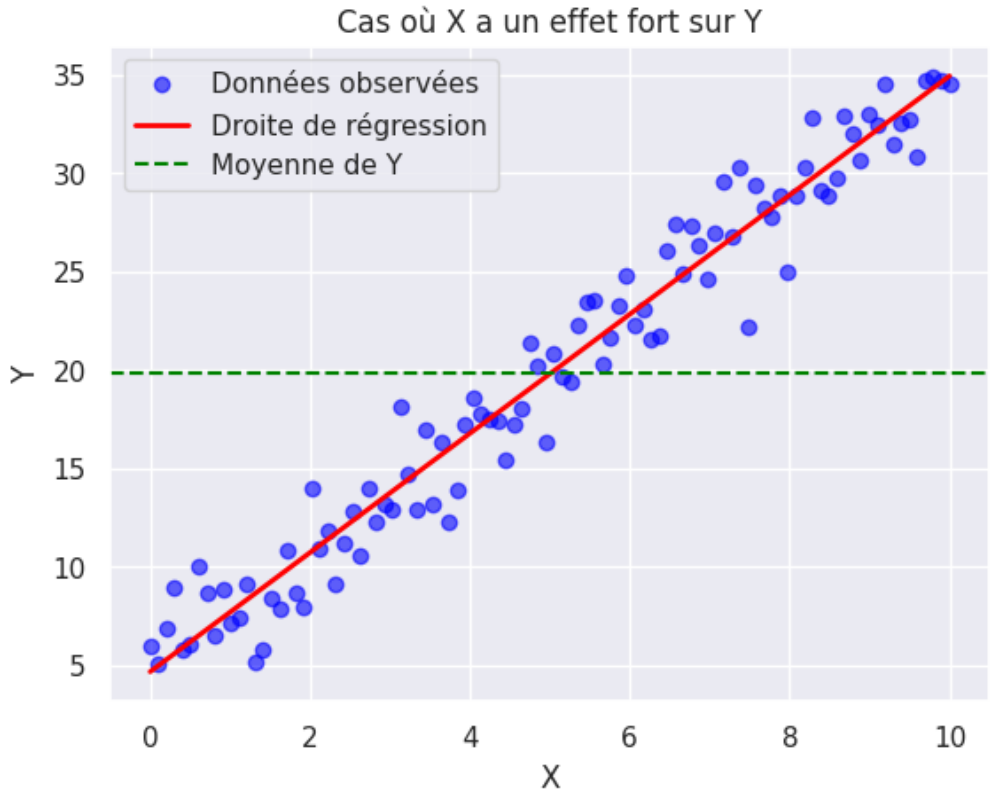
ANOVA : Test F de Fisher

- **Cas où X n'a aucun effet sur Y (régression nulle)**
  - La droite de régression est horizontale ($\beta_1 = 0$).
  - $SC_{reg} \approx 0$ et $SC_{res} \approx SC_{totale}$.
  - $F \approx 1$, donc on ne rejette pas H_0.



ANOVA : Test F de Fisher

- **Cas où X a un effet fort sur Y (bonne régression)**
  - La droite de régression suit bien la tendance des données.
  - SC_{reg} est grand et SC_{res} est petit.
  - F est élevé, donc on rejette H_0.



ANOVA : Test F de Fisher

Tableau de l'ANOVA pour la régression linéaire

Source	Somme Carrés	dl	Moyennes Carrés	F_{stat}
Régression	SC_{reg}	1	$MC_{reg} = \frac{SC_{reg}}{1}$	$F = \frac{MC_{reg}}{MC_{res}}$
Résiduel	SC_{res}	$n-2$	$MC_{res} = \frac{SC_{res}}{n-2}$	
Total	SC_{totale}	$n-1$		

- L'ANOVA en régression linéaire sert à **mesurer combien de la variabilité totale peut être attribuée** à la variable explicative X.
- Elle permet de tester si le modèle de régression est significatif par rapport à un modèle trivial (constante seulement).
- Le test F **compare la variabilité expliquée et la variabilité résiduelle** pour évaluer la pertinence du modèle.

ANOVA Exemple Numérique : Heure d'Études vs Notes Obtenues

ANOVA Exemple Numérique : Heure d'Études

- **Chargement des données :**
  - On charge les heures d'étude (X) et les notes obtenues (Y) à partir du fichier `StudentGrades.csv`.
- **Régression Linéaire :**
  - Entraînement du modèle $Y = \beta_0 + \beta_1 X$.
  - Prédiction des valeurs ajustées $\hat{Y}$.
- **Calcul des Sommes des Carrés :**
  - Variabilité totale : $SC_{totale} = \sum_{i=1}^{n} (y_i - \overline{y})^2$
  - Variabilité expliquée par la régression : $SC_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$
  - Variabilité résiduelle : $SC_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- **Construction du Tableau ANOVA :**
  - Degrés de liberté : $dl_{reg} = 1, \quad dl_{res} = n - 2$
  - Moyenne des carrés : $MC = \frac{SC}{dl}$
  - Statistique F : $F_{stat} = \frac{MC_{reg}}{MC_{RES}}$
- **Interprétation du Test F :**
  - Si $p < 0.05$, alors la régression est significative.
  - Sinon, X n'explique pas significativement Y.
- **Visualisation :** Droite de régression et les points réels.

ANOVA Heure d'Études vs Notes : Code Python

```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
import scipy.stats as stats
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"  # Assurez-vous que le nom du fichier est correct
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
```

ANOVA Heure d'Études vs Notes : Code Python

```
# Étape 4: Extraction des variables
X = data[['Hours Studied']].values  # Variable indépendante (heures étudiées)
y = data['Grades'].values  # Variable dépendante (note obtenue)

# Étape 5: Création et entraînement du modèle de régression linéaire
modele = LinearRegression()
modele.fit(X, y)
y_pred = modele.predict(X)
```

ANOVA Heure d'Études vs Notes : Code Python

```python
# Étape 6: Calcul des sommes des carrés pour l'ANOVA
SS_tot = np.sum((y - np.mean(y))**2)  # Somme des carrés totale
SS_reg = np.sum((y_pred - np.mean(y))**2)  # Somme des carrés expliquée (régression)
SS_res = np.sum((y - y_pred)**2)  # Somme des carrés résiduelle

# Degrés de liberté
df_reg = 1  # Une seule variable explicative
df_res = len(y) - 2  # n - 2 car on estime beta_0 et beta_1

# Moyennes des carrés
MS_reg = SS_reg / df_reg
MS_res = SS_res / df_res

# Calcul de la statistique F et p-valeur
F_stat = MS_reg / MS_res
p_value = 1 - stats.f.cdf(F_stat, df_reg, df_res)  # p-valeur associée
```

ANOVA Heure d'Études vs Notes : Code Python

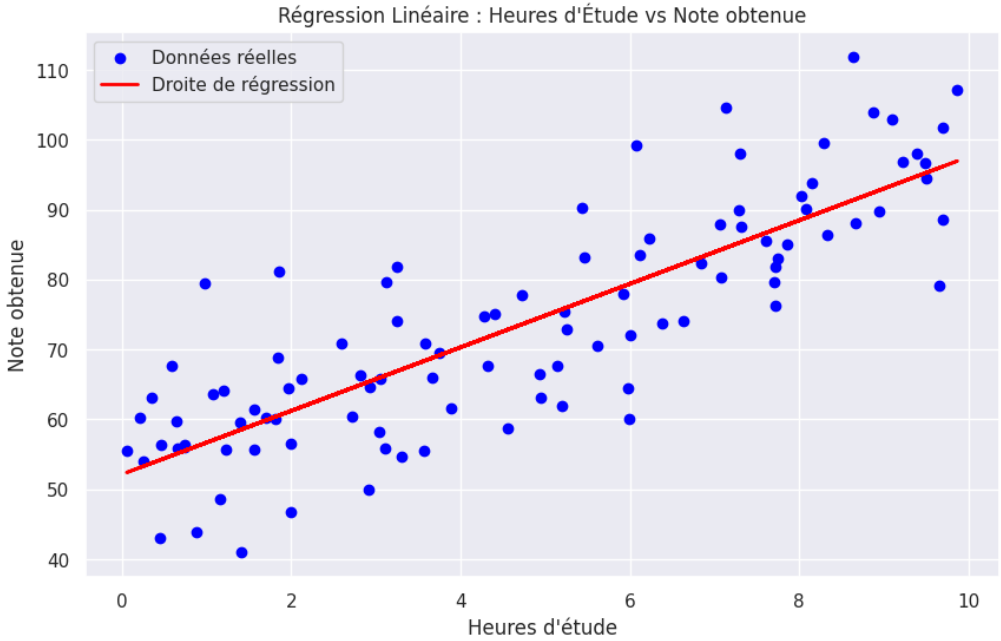
```python
# Étape 7: Affichage du tableau ANOVA
anova_table = pd.DataFrame({
    "Source": ["Régression", "Résiduelle", "Totale"],
    "Somme des Carrés": [SS_reg, SS_res, SS_tot],
    "Degrés de Liberté": [df_reg, df_res, df_reg + df_res],
    "Moyenne des Carrés": [MS_reg, MS_res, None],
    "F-stat": [F_stat, None, None],
    "p-value": [p_value, None, None]
})

print("\n Tableau ANOVA :")
print(anova_table)

# Étape 8: Interprétation du test F
print("\n Interprétation du Test F :")
if p_value < 0.05:
    print(f"Le test F indique que X pourrait avoir un effet significatif sur "
          f"Y (p = {p_value:.5f}).")
else:
    print(f"Le test F n'est pas significatif (p = {p_value:.5f}). X pourrait "
          f"ne pas être un bon prédicteur de Y.")
```

ANOVA Heure d'Études vs Notes : Visualisation

```
Drive already mounted at /content/drive; to attempt to forcibly remount,
call drive.mount("/content/drive", force_remount=True).
Chargement des données depuis Google Drive...
Données chargées avec succès.
   Hours Studied     Grades
0       3.745401  69.597477
1       9.507143  94.545642
2       7.319939  87.517305
3       5.986585  60.057235
4       1.560186  55.604213
Tableau ANOVA :
       Source  Somme des Carrés  Degrés de Liberté  Moyenne des Carrés
0  Régression      18060.647192                  1        18060.647192
1  Résiduelle       8065.845640                 98           82.304547
2      Totale      26126.492831                 99                 NaN
       F-stat       p-value
0  219.436808  1.110223e-16
1         NaN           NaN
2         NaN           NaN
Interprétation du Test F :
Le test F indique que X pourrait avoir un effet significatif sur Y (p = 0.00000).
```

ANOVA Heure d'Études vs Notes : Visualisation

```
# Étape 9: Visualisation de la régression
plt.figure(figsize=(10, 6))
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()
```



Résultats du Modèle de Régression Linéaire

Intercept (beta0) : 52.15
Pente (beta1) : 4.54
Score R^2 : 0.6913

ANOVA Exemple Numérique : PIB par Habitant vs Espérance de Vie

ANOVA Exemple Numérique : PIB vs Espérance de vie

- **Chargement des données :**
  - On charge le PIB par habitant (X) et l'espérance de vie (Y) à partir du fichier `Esperance_vie_pib.csv`.
- **Régression Linéaire :**
  - Entraînement du modèle $Y = \beta_0 + \beta_1 X$.
  - Prédiction des valeurs ajustées $\hat{Y}$.
- **Calcul des Sommes des Carrés :**
  - Variabilité totale : $SC_{\text{totale}} = \sum_{i=1}^{n} (y_i - \overline{y})^2$
  - Variabilité expliquée par la régression : $SC_{\text{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$
  - Variabilité résiduelle : $SC_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- **Construction du Tableau ANOVA :**
  - Degrés de liberté : $dl_{\text{reg}} = 1, \quad dl_{\text{res}} = n - 2$
  - Moyenne des carrés : $MC = \frac{SC}{dl}$
  - Statistique F : $F_{\text{stat}} = \frac{MC_{\text{reg}}}{MC_{\text{res}}}$
- **Interprétation du Test F :**
  - Si $p < 0.05$, alors la régression est significative.
  - Sinon, X n'explique pas significativement Y.
- **Visualisation :** Droite de régression et les points réels.

ANOVA PIB vs Espérance de Vie : Code Python

```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
import scipy.stats as stats
from google.colab import drive
import os

# Étape 1: Monter Google Drive
drive.mount('/content/drive')

# Étape 2: Définir le chemin du fichier CSV
dossier = "Colab Notebooks"
nom_fichier = "Esperance_vie_pib.csv"
chemin_fichier = f"/content/drive/My Drive/{dossier}/{nom_fichier}"

# Étape 3: Chargement des 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
```

ANOVA PIB vs Espérance de Vie : Code Python

```
# Étape 4: Extraction des variables
X = data[['GDP per capita (current US$)']].values  # Variable indépendante (PIB par habitan
y = data['Life Expect 2024'].values  # Variable dépendante (espérance de vie)

# Étape 5: Création et entraînement du modèle de régression linéaire
modele = LinearRegression()
modele.fit(X, y)
y_pred = modele.predict(X)
```

ANOVA PIB vs Espérance de Vie : Code Python

```python
# Étape 6: Calcul des sommes des carrés pour l'ANOVA
SS_tot = np.sum((y - np.mean(y))**2)  # Somme des carrés totale
SS_reg = np.sum((y_pred - np.mean(y))**2)  # Somme des carrés expliquée (régression)
SS_res = np.sum((y - y_pred)**2)  # Somme des carrés résiduelle

# Degrés de liberté
df_reg = 1  # Une seule variable explicative
df_res = len(y) - 2  # n - 2 car on estime beta0 et beta1

# Moyennes des carrés
MS_reg = SS_reg / df_reg
MS_res = SS_res / df_res

# Calcul de la statistique F et p-valeur
F_stat = MS_reg / MS_res
p_value = 1 - stats.f.cdf(F_stat, df_reg, df_res)  # p-valeur associée
```

ANOVA PIB vs Espérance de Vie : Code Python

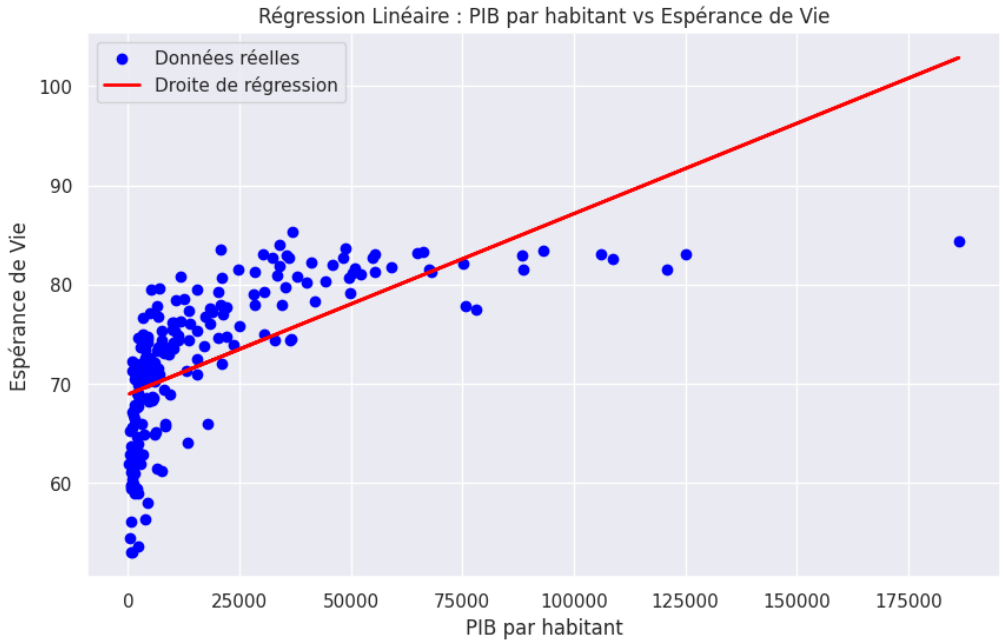
```python
# Étape 7: Affichage du tableau ANOVA
anova_table = pd.DataFrame({
    "Source": ["Régression", "Résiduelle", "Totale"],
    "Somme des Carrés": [SS_reg, SS_res, SS_tot],
    "Degrés de Liberté": [df_reg, df_res, df_reg + df_res],
    "Moyenne des Carrés": [MS_reg, MS_res, None],
    "F-stat": [F_stat, None, None],
    "p-value": [p_value, None, None]
})

print("\n Tableau ANOVA :")
print(anova_table)

# Étape 8: Interprétation du test F
if p_value < 0.05:
    print(f"Le test F indique que le PIB a un effet significatif sur l'espérance de vie "
          f"(p = {p_value:.5f}).")
else:
    print(f"Le test F n'est pas significatif (p = {p_value:.5f}). PIB pourrait ne pas "
          f"être un bon prédicteur.")
```

ANOVA PIB vs Espérance de Vie : Visualisation

```
Drive already mounted at /content/drive; to attempt to forcibly remount,
call drive.mount("/content/drive", force_remount=True).
Chargement des données depuis Google Drive...
Données chargées avec succès.
           Country Name  Life Expect 2024  GDP per capita (current US$)
0                 Aruba            74.992                  30559.533530
1           Afghanistan            62.879                    357.261153
2                Angola            61.929                   2929.694455
3               Albania            76.833                   6846.426143
4  United Arab Emirates            79.196                  49899.065300

Tableau ANOVA :
       Source  Somme des Carrés  Degrés de Liberté  Moyenne des Carrés  \
0  Régression       4896.398451                  1         4896.398451
1  Résiduelle       6993.508713                202           34.621330
2      Totale      11889.907163                203                 NaN

       F-stat       p-value
0  141.427219  1.110223e-16
1         NaN           NaN
2         NaN           NaN

Interprétation du Test F :
Le test F indique que le PIB pourrait avoir un effet significatif sur l'espérance de vie
(p = 0).
```

ANOVA PIB vs Espérance de Vie : Visualisation

```
# Étape 9: Visualisation de la régression
plt.figure(figsize=(10, 6))
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 : PIB par habitant vs Espérance de Vie")
plt.xlabel("PIB par habitant")
plt.ylabel("Espérance de Vie")
plt.legend()
plt.grid(True)
plt.show()
```



Résultats du Modèle

Intercept (beta0) : 68.9431
Pente (beta1) : 0.0002
Score R^2 : 0.4118

ANOVA PIB vs Espérance de Vie : Analyse

- **Les résultats obtenus montrent une relation positive** entre le PIB par habitant et l'espérance de vie, comme l'indiquent les valeurs suivantes :
  - **Intercept** (β_0) : 68.9431.
  - **Pente** (β_1) : 0.0002, ce qui signifie qu'une augmentation du PIB de 10 000$ entraîne une augmentation attendue de 2 ans d'espérance de vie.
  - **Score** R^2 : 0.4118, ce qui indique que 41.18% de la variabilité de l'espérance de vie est expliquée par le PIB.

ANOVA PIB vs Espérance de Vie : Analyse

- **Signification de la Régression**
  - La statistique F obtenue est très élevée : $F_{stat} = 141.43$ avec une p-valeur extrêmement faible ($p < 10^{-16}$).
  - Cela signifie que l'hypothèse nulle $H_0 : \beta_1 = 0$ est fortement rejetée, indiquant que le PIB a un effet statistiquement significatif sur l'espérance de vie.
- **Qualité du Modèle**
  - Un $R^2 = 0.4118$ indique que 41.18% de la variation de l'espérance de vie est expliquée par le PIB.
  - Bien que ce soit un effet significatif, il reste une grande partie de la variabilité qui n'est pas expliquée, suggérant que d'autres facteurs influencent l'espérance de vie.

ANOVA PIB vs Espérance de Vie : Analyse

- **Non-linéarité potentielle :**
  - On observe une courbe en éventail au début des valeurs faibles de PIB.
  - Cela suggère que l'effet du PIB sur l'espérance de vie n'est pas strictement linéaire.
  - Une transformation logarithmique ($\log(\text{PIB})$) pourrait améliorer l'ajustement du modèle.
  - Il serait aussi pertinent d'examiner d'autres variables explicatives (ex. accès aux soins de santé, éducation, inégalités sociales) pour améliorer la prédiction.
- **Hétéroscédasticité (variance non constante des résidus) :**
  - La variance des erreurs semble plus grande pour les valeurs élevées de PIB, ce qui viole une hypothèse clé de la régression linéaire.
  - Une correction typique consiste à utiliser une transformation des variables.
  - La forte hétéroscédasticité et la non-linéarité potentielle indiquent que l'utilisation brute du PIB pourrait ne pas être optimale.

Table des Matières

Test t pour la significativité des coefficients de Régression

Test t de Student : Rappel

- Le test t de Student permet de vérifier si la moyenne d'une population suit une valeur théorique donnée (μ_0).
- **Hypothèses :**
  - H_0: La moyenne de la population est égale à la moyenne théorique ($\mu = \mu_0$).
  - H_a: La moyenne de la population est différente de la moyenne théorique ($\mu \neq \mu_0$).
- **Statistique de test :**

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}},$$

où $\overline{X}$ est la moyenne de l'échantillon, s est l'écart-type de l'échantillon, et n est la taille de l'échantillon.

- **Degrés de liberté :**

$$k = n - 1,$$

où n est la taille de l'échantillon.

Test t de Student : Rappel des Étapes

- **Étape 1. Formulation des hypothèses :**
  - $H_0 : \mu = \mu_0$
  - $H_a : \mu \neq \mu_0$
- **Étape 2. Calcul de la statistique de test :**

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

- **Étape 3. Degrés de liberté :**

$$k = n - 1$$

- **Étape 4. Décision :**
  - Utiliser la distribution t_k avec k degrés de liberté pour trouver la valeur critique à un niveau de signification donné α, ou calculer la p-valeur.
  - Rejeter H_0 si la valeur observée de t est plus extrême que la valeur critique, selon les seuils définis par la distribution t.

Test t pour la significativité des coefficients

- Dans une régression linéaire simple, nous avons un modèle de la forme :

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

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

- Les estimateurs des coefficients $\hat{\beta}_0$ et $\hat{\beta}_1$ sont obtenus par la méthode des moindres carrés et suivent des distributions normales :

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right),$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right).$$

Test t pour la significativité des coefficients

- Nous souhaitons tester si un coefficient de régression est significativement différent de zéro.
- **Hypothèses pour β_1 :**
  - H_0 : $\beta_1 = 0$ (le prédicteur X n'a pas d'effet sur Y).
  - H_a : $\beta_1 \neq 0$ (le prédicteur X a un effet significatif sur Y).
- **Statistique de test :**

$$t = \frac{\hat{\beta}_1 - 0}{s/\sqrt{S_{xx}}}$$

où s est l'estimation de l'écart-type des résidus.

- Sous H_0, la statistique suit une loi de Student avec $n-2$ degrés de liberté :

$$t \sim t_{n-2}$$

Test t pour la significativité des coefficients

- De même, on peut tester si l'ordonnée à l'origine β_0 est significativement différente d'une valeur donnée β_0^*.
- **Hypothèses :**
  - $H_0 : \beta_0 = \beta_0^*$.
  - $H_a : \beta_0 \neq \beta_0^*$.
- **Statistique de test :**

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}}$$

où s est l'estimation de l'écart-type des résidus.

- Sous H_0, la statistique suit une loi de Student avec $n-2$ degrés de liberté :

$$t \sim t_{n-2}$$

Test t pour la significativité des coefficients

- **Avant l'ajustement du modèle** : Il y a n données indépendantes.
- **Après l'ajustement du modèle** :
  - Deux paramètres (β_0 et β_1) sont estimés à partir des données.
  - Cela retire donc 2 degrés de liberté.
- **Degrés de liberté pour l'erreur (résidus)** : $n - 2$.
- **Comparaison avec un test t classique** :
  - Si nous étions en train d'estimer uniquement une moyenne (comme dans un test t classique sur une moyenne),
  - Nous n'aurions qu'un seul paramètre estimé (μ),
  - Donc les degrés de liberté seraient $n - 1$.

Merci!

E-mail: chiheb.trabelsi@polymtl.ca

Table des Matières

Annexe

Preuve Décomposition de la Variabilité Totale

Décomposition de la Variabilité Totale : Preuve

- **Égalité fondamentale de la décomposition de la variabilité totale en régression linéaire simple :**

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{SC_{totale}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{SC_{reg}} + \underbrace{\sum_{i=1}^{n}(e_i)^2}_{SC_{res}} \quad .$$

- Nous voulons démontrer cette égalité. Nous partons du fait que :

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + \underbrace{(y_i - \hat{y}_i)}_{e_i}.$$

- En élevant au carré et en sommant sur i :

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i).$$

- L'objectif est de démontrer que le dernier terme s'annule, ce qui permet d'obtenir l'égalité fondamentale de la décomposition.

Décomposition de la Variabilité Totale : Preuve

- Nous avons la décomposition fondamentale :

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + \underbrace{(y_i - \hat{y}_i)}_{e_i}.$$

- En multipliant chaque côté par $(\hat{y}_i - \overline{y})$ et en sommant sur i :

$$\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i).$$

- Nous cherchons à prouver que :

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 0.$$

Décomposition de la Variabilité Totale : Preuve

- La droite de régression est donnée par :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

et la moyenne des prédictions est :

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

- Ainsi, on peut exprimer :

$$\hat{y}_i - \overline{y} = \hat{\beta}_1 (x_i - \overline{x}).$$

- De plus, la différence entre la valeur réelle et la prédiction peut être réécrite comme :

$$y_i - \hat{y}_i = (y_i - \overline{y}) - (\hat{y}_i - \overline{y}).$$

Décomposition de la Variabilité Totale : Preuve

- En substituant dans l'équation précédente :

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

- En développant :

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

- L'objectif est maintenant de montrer que ce terme s'annule.

Décomposition de la Variabilité Totale : Preuve

- Nous utilisons la définition du coefficient de régression dans une régression linéaire simple :

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

- Maintenant, nous remplaçons ce $\hat{\beta}_1$ dans le premier terme de notre équation :

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

- En substituant $\hat{\beta}_1$:

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

- Nous reconnaissons ici que le numérateur apparaît deux fois, donc nous obtenons : $\frac{\left(\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$.

Décomposition de la Variabilité Totale : Preuve

- Examinons le deuxième terme de l'équation :

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

- En remplaçant $\hat{\beta}_1$:

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

- Ce qui nous donne :

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

- Nous avons donc les deux termes suivants :

$$\frac{\left(\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} - \frac{\left(\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = 0.$$

Décomposition de la Variabilité Totale : Preuve

- Nous avons montré que :

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 0.$$

- Ce qui signifie que le dernier terme de la décomposition de la variabilité totale s'annule, et nous obtenons :

$$SC_{totale} = SC_{reg} + SC_{res}.$$