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

## Leçon 2 : Analyse des Résidus et Régression Linéaire Multiple

Polytechnique Montréal - Hiver 2025

Chiheb Trabelsi

26 Février 2025

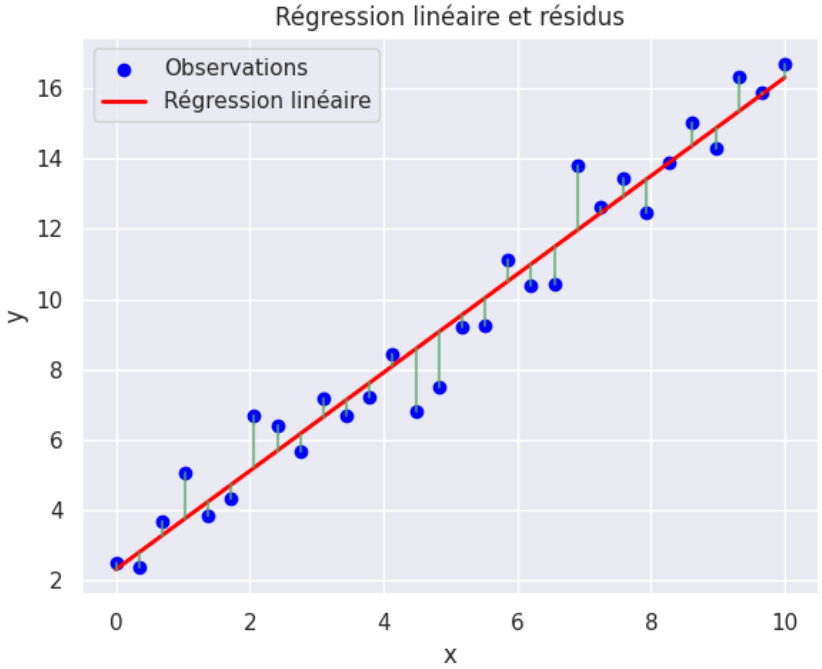
# Table des Matières

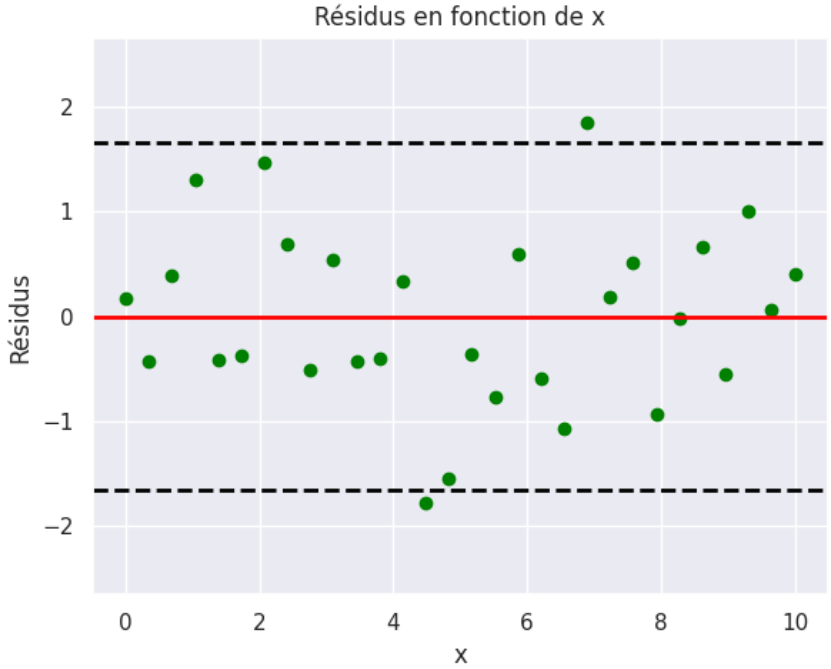
1. Analyse des Résidus
2. Régression Linéaire Multiple
3. Décomposition de la Variabilité Totale
4. ANOVA en Régression Linéaire Multiple
5. Exercice : Boston Housing Prices Dataset
6. Intervalle de Confiance et Intervalle de Prédiction
7. Annexe

# Analyse des Résidus pour Vérifier les Hypothèses du Modèle de Régression Linéaire

# Analyse des Résidus pour Vérifier les Hypothèses

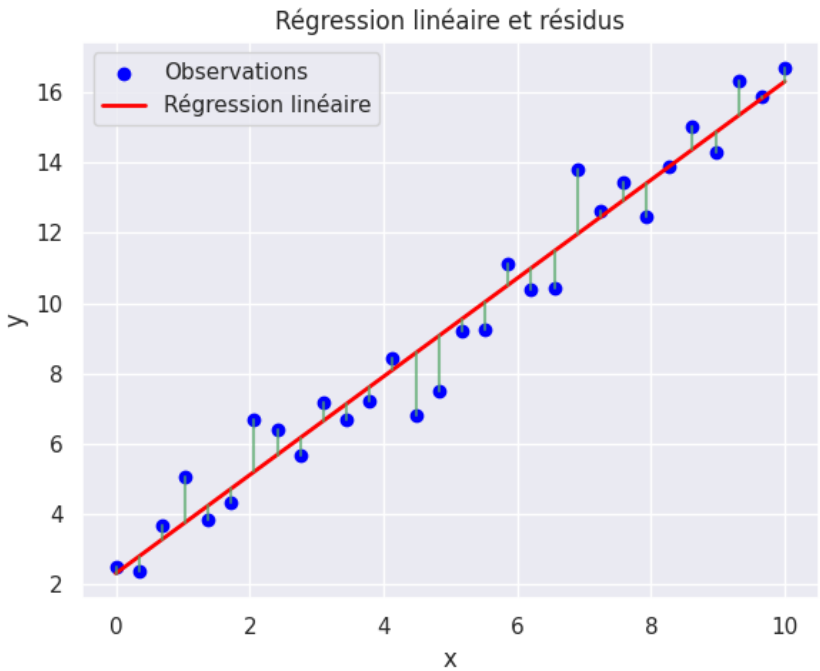
- **L'espérance des erreurs est nulle :** $E(\epsilon_i) = 0$.
- **Homoscédasticité :** La variance des erreurs est constante, $\text{Var}(\epsilon_i) = \sigma^2 \quad \forall\, i \in \{1, \cdots, n\}$.
- **Indépendance des erreurs :** Les erreurs $\epsilon_i$ ne sont pas corrélées entre elles $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad$ pour tout $i \neq j$.

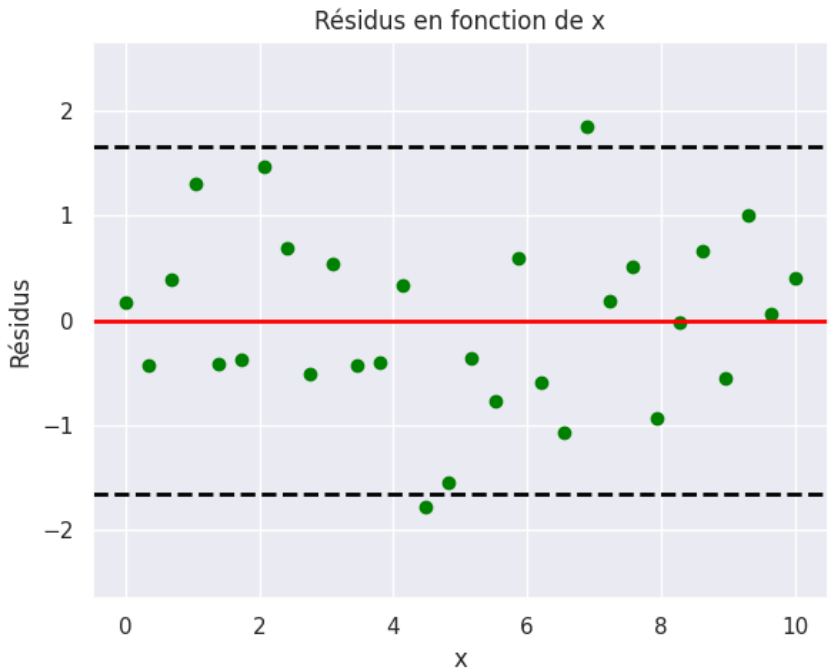


# Analyse des Résidus pour Vérifier les Hypothèses

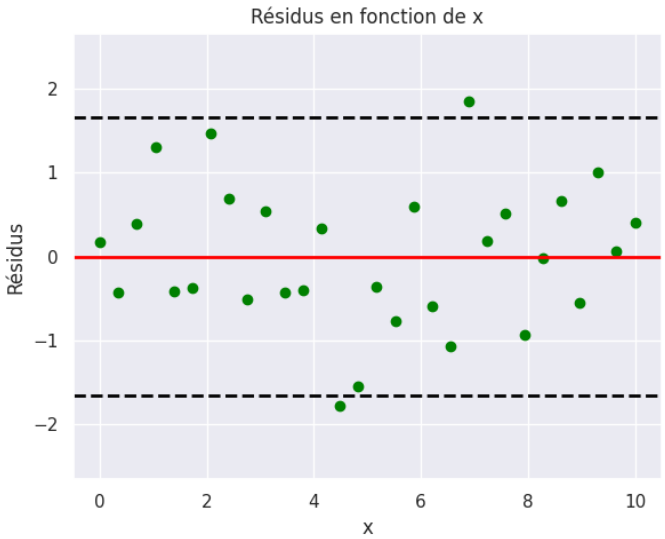
- **Graphique de gauche :** Représente la régression linéaire avec les résidus en vert.
- **Graphique de droite :** Affiche les résidus en fonction de $x$ avec des seuils supérieur et inférieur (bandes noires) pour examiner leur dispersion.



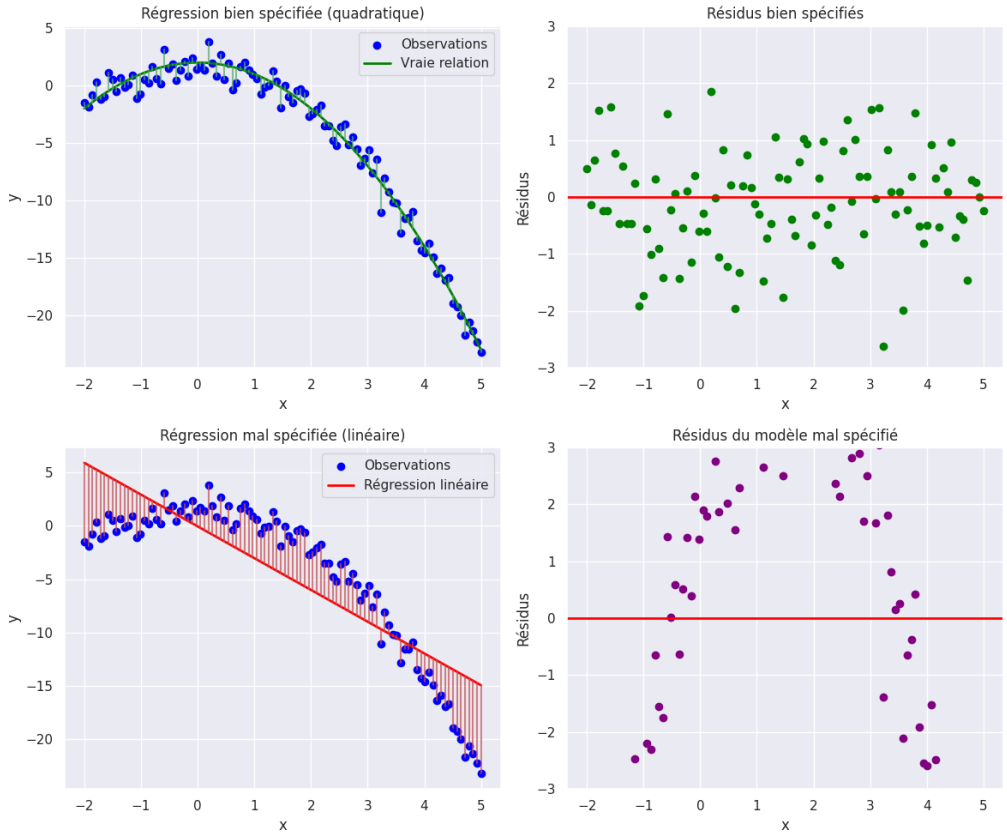
# Analyse des Résidus pour Vérifier les Hypothèses

- Les résidus doivent être **centrés autour de zéro** sans structure apparente.
- La bande noire illustre la dispersion des résidus. Une distribution homogène indique que l'hypothèse d'homoscédasticité est respectée.
- Si la dispersion des résidus augmente ou diminue selon $x$, cela peut indiquer une hétéroscédasticité.
- Une tendance dans les résidus peut indiquer un problème de spécification du modèle.

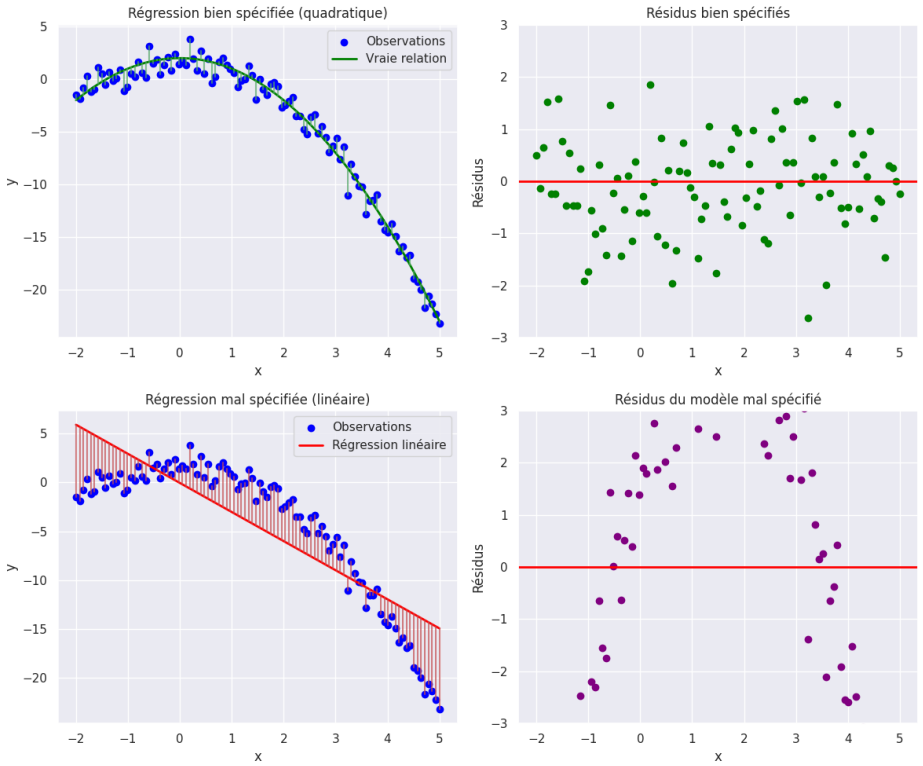
# Analyse des Résidus pour Vérifier les Hypothèses

- **Graphiques en haut :**
  - À gauche : Régression linéaire avec des résidus bien distribués.
  - À droite : Résidus bien centrés autour de $0$, avec dispersion homogène.



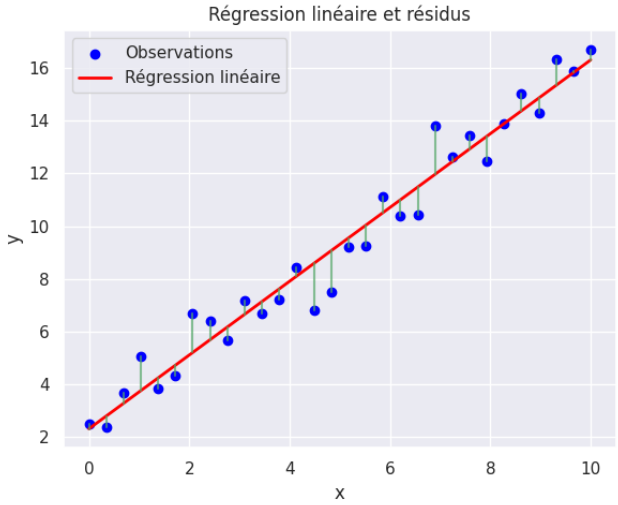
# Analyse des Résidus pour Vérifier les Hypothèses

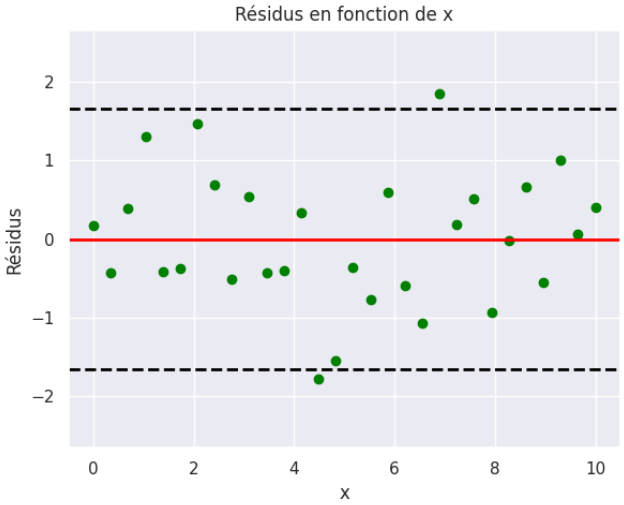
- **Graphiques en bas :**
  - À gauche : Mauvaise spécification du modèle, absence d'un terme quadratique.
  - À droite : Résidus en courbe, indiquant la non prise en compte du terme quadratique.



# Analyse des Résidus pour Vérifier les Hypothèses

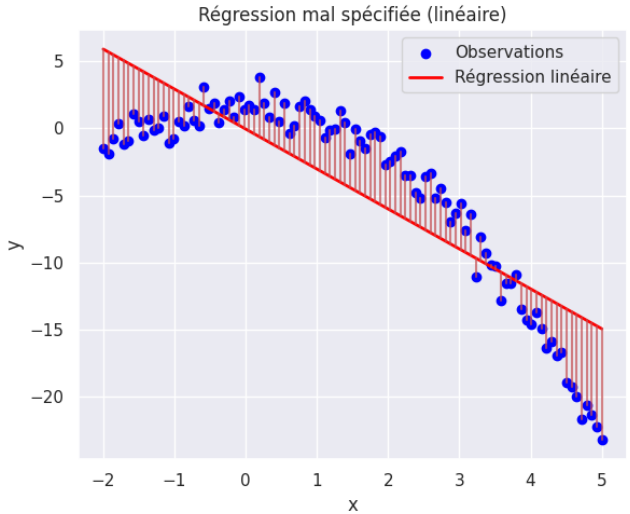
- Un modèle bien spécifié présente des résidus sans structure et bien répartis autour de $0$.
- Un modèle mal spécifié peut mener à :
  - Une tendance dans les résidus, indiquant qu'une variable importante est omise.
  - Une dispersion non homogène des résidus, signalant une non-linéarité ignorée.
- L'ajout de termes (par exemple $x^2$) dans la régression permettrait de corriger la non-linéarité observée.

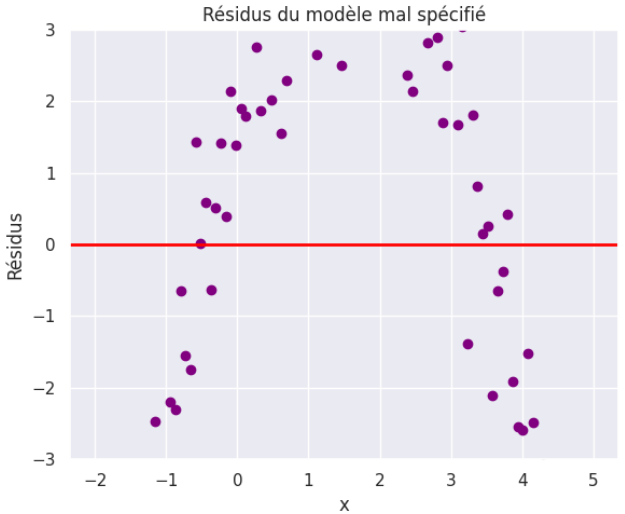


# Analyse des Résidus pour Vérifier les Hypothèses

- Un modèle bien spécifié présente des résidus sans structure et bien répartis autour de $0$.
- Un modèle mal spécifié peut mener à :
  - Une tendance dans les résidus, indiquant qu'une variable importante est omise.
  - Une dispersion non homogène des résidus, signalant une non-linéarité ignorée.
- L'ajout de termes (par exemple $x^2$) dans la régression permettrait de corriger la non-linéarité observée.



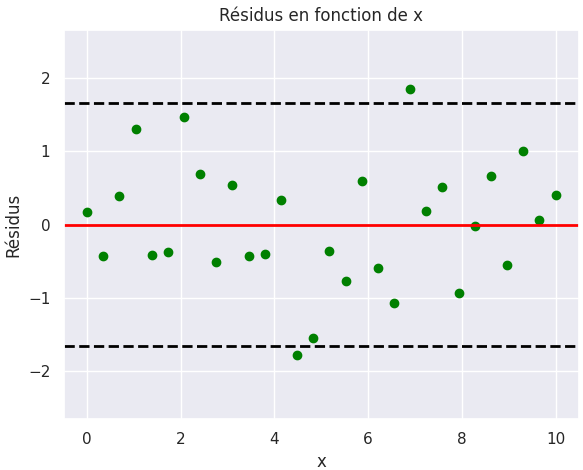
# Analyse des Résidus avec Histogrammes et QQ-Plots

# Analyse des Résidus avec Histogrammes et QQ-Plots

- L'analyse des résidus est essentielle en régression linéaire.
- La normalité des résidus est une hypothèse clé pour plusieurs tests statistiques.
- Nous explorons différentes formes distributions de résidus.
- Le modèle de régression est défini par :

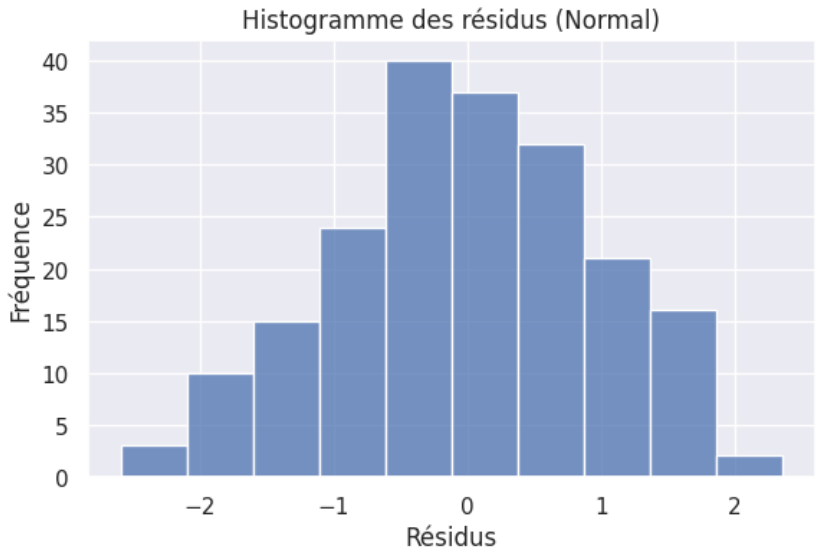
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

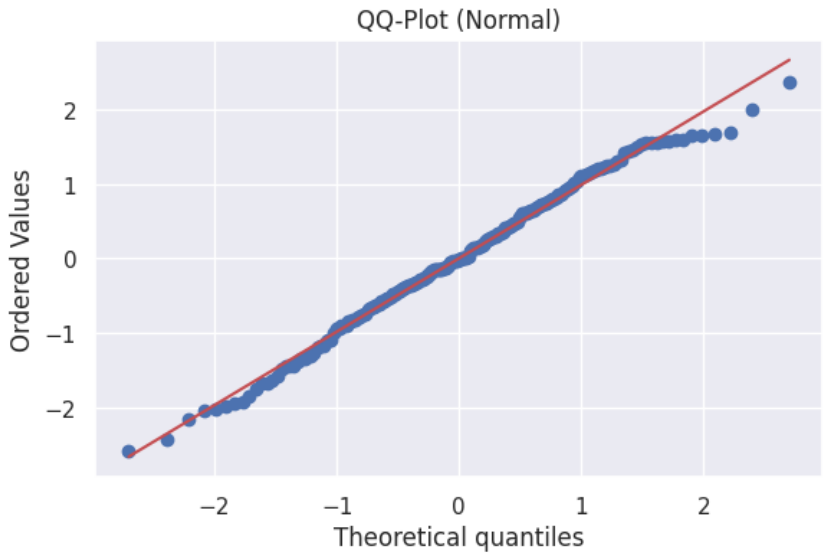
- Les erreurs sont supposées indépendantes et suivent : $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Les résidus sont calculés comme suit : $e_i = y_i - \hat{y}_i$

# Analyse des Résidus avec Histogrammes et QQ-Plots

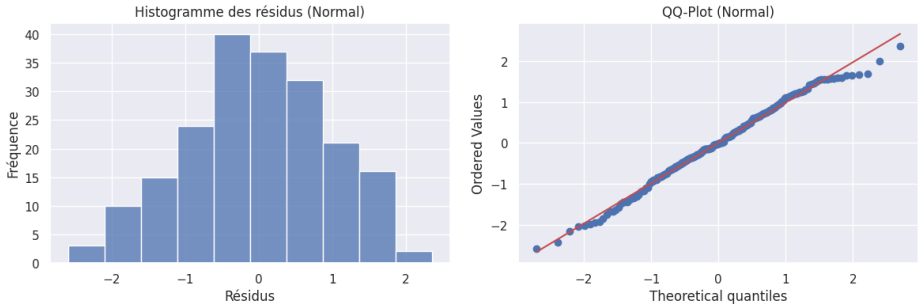
- Le QQ-Plot compare la distribution empirique des résidus à une normale.
- Ordonnée : résidus triés, abscisse : quantiles théoriques.
- Position des quantiles : $\Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right)$



# Analyse des Résidus avec Histogrammes et QQ-Plots

- Un **QQ-Plot (Quantile-Quantile Plot)** est un graphique permettant de comparer la distribution d'un ensemble de données empiriques à une distribution théorique, en particulier la distribution normale.
- Axe des abscisses (x) : quantiles théoriques d'une distribution normale standard $\mathcal{N}(0, 1)$.
- Axe des ordonnées (y) : valeurs des résidus triés par ordre croissant.
- **Objectif** : Si les points suivent une droite diagonale, cela suggère que les résidus suivent une distribution normale.
  - Une courbure en S indique une asymétrie (positive ou négative).
  - Une dispersion excessive aux extrémités indique une distribution à queue lourde.



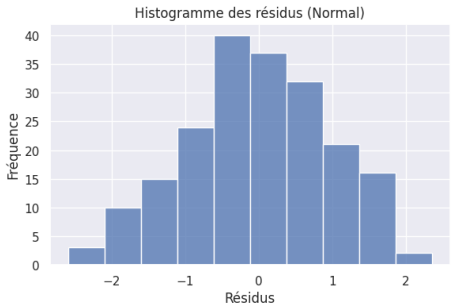
# Analyse des Résidus avec Histogrammes et QQ-Plots

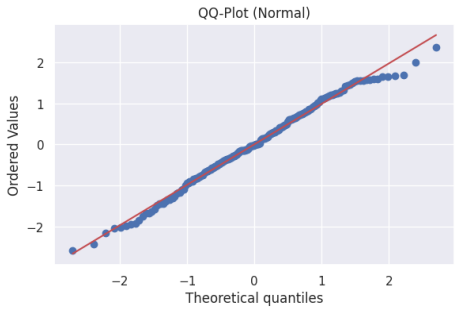
- La fonction $\Phi^{-1}$ est l'inverse de la fonction de répartition de la loi normale standard, aussi appelée **fonction quantile de la distribution normale**.
- **Définition de** $\Phi(x)$ :

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

  Cette fonction donne la probabilité qu'une variable aléatoire normale standard soit inférieure ou égale à $x$.
- **Inverse** $\Phi^{-1}(p)$ : Cette fonction renvoie la valeur $z$ telle que :

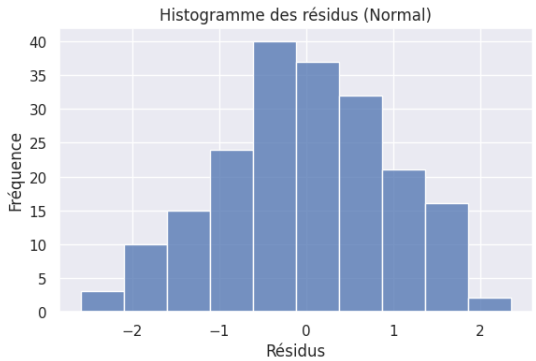
$$\Phi^{-1}(p) = z \quad \text{tel que} \quad P(X \leq z) = p$$

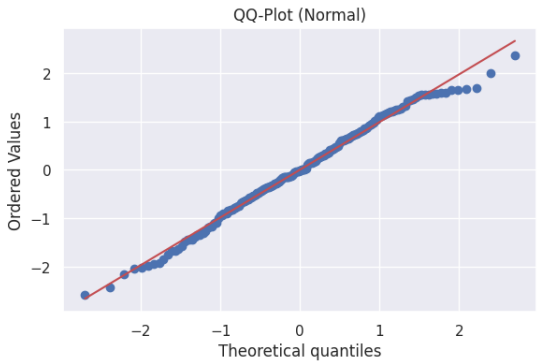


# Analyse des Résidus avec Histogrammes et QQ-Plots

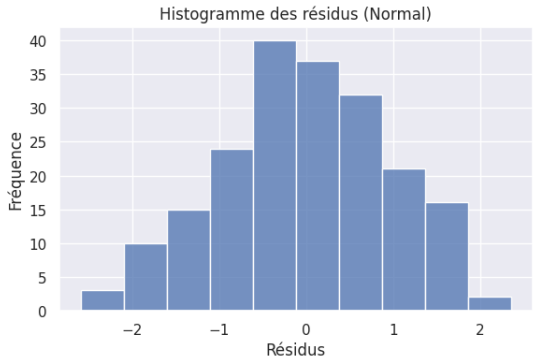
- La formule utilisée pour positionner les points sur l'axe des abscisses du QQ-Plot est : $\Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right)$
- **Explication des termes** :
  - $k$ : Indice de l'observation après tri des résidus (du plus petit au plus grand).
  - $n$ : Nombre total d'observations.
  - $0.375$ et $0.25$ : Ajustements empiriques souvent utilisés pour éviter des valeurs extrêmes et améliorer la robustesse de l'estimation.
- Cette formule donne les quantiles théoriques de la loi normale standard correspondant aux positions des observations dans l'échantillon.

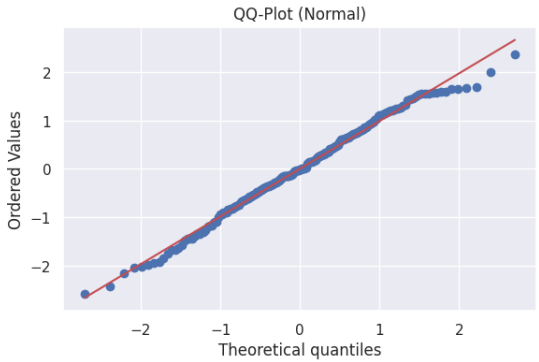


# Analyse des Résidus avec Histogrammes et QQ-Plots

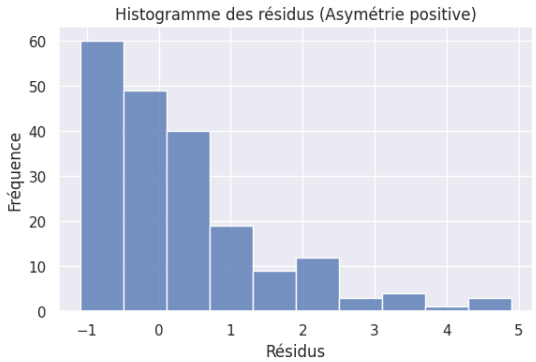
- Analyse des résidus via histogrammes et QQ-Plots.
- Différents types de distributions :
  - **Normal : résidus bien répartis autour de 0 et les points du QQ-Plot alignés.**
  - Asymétrie positive : biais vers la droite et courbure au-dessus de la diagonale.
  - Asymétrie négative : biais vers la gauche et courbure au-dessous de la diagonale.
  - Queue lourde : extrêmes plus fréquents que dans une normale et dispersion importante aux extrémités.

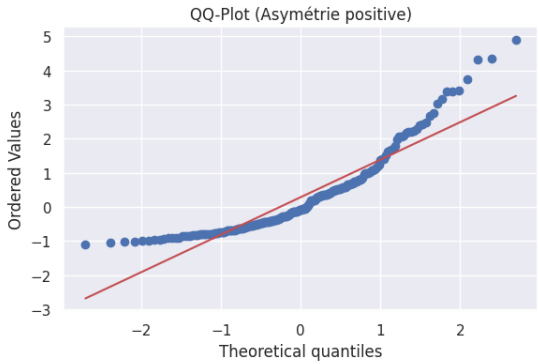


# Analyse des Résidus avec Histogrammes et QQ-Plots

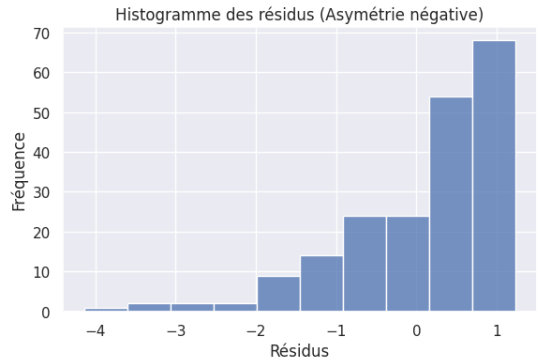
- Analyse des résidus via histogrammes et QQ-Plots.
- Différents types de distributions :
  - Normal : résidus bien répartis autour de 0 et le points du QQ-Plot alignés.
  - **Asymétrie positive : biais vers la droite et courbure au-dessus de la diagonale.**
  - Asymétrie négative : biais vers la gauche et courbure au-dessous de la diagonale.
  - Queue lourde : extrêmes plus fréquents que dans une normale et dispersion importante aux extrémités.

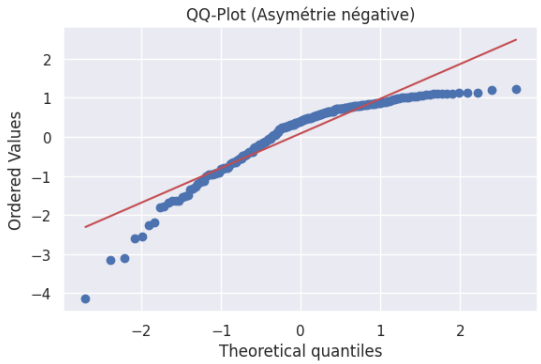


# Analyse des Résidus avec Histogrammes et QQ-Plots

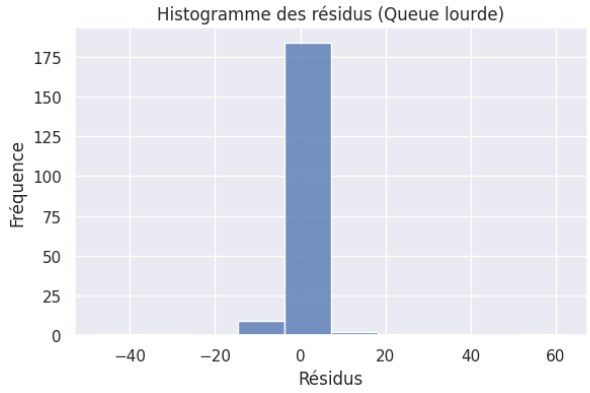
- Analyse des résidus via histogrammes et QQ-Plots.
- Différents types de distributions :
  - Normal : résidus bien répartis autour de 0 et le points du QQ-Plot alignés.
  - Asymétrie positive : biais vers la droite et courbure au-dessus de la diagonale.
  - **Asymétrie négative : biais vers la gauche et courbure au-dessous de la diagonale.**
  - Queue lourde : extrêmes plus fréquents que dans une normale et dispersion importante aux extrémités.

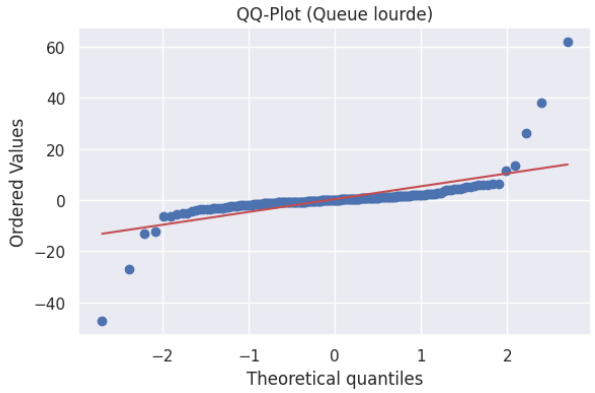


# Analyse des Résidus avec Histogrammes et QQ-Plots

- Analyse des résidus via histogrammes et QQ-Plots.
- Différents types de distributions :
  - Normal : résidus bien répartis autour de 0 et le points du QQ-Plot alignés.
  - Asymétrie positive : biais vers la droite et courbure au-dessus de la diagonale.
  - Asymétrie négative : biais vers la gauche et courbure au-dessous de la diagonale.
  - **Queue lourde : extrêmes plus fréquents que dans une normale et dispersion importante aux extrémités.**



# Analyse des Résidus avec Histogrammes et QQ-Plots

- Un **QQ-Plot** compare la distribution empirique d'un échantillon à une distribution théorique.
- $\Phi^{-1}$ est la fonction quantile de la loi normale standard, utilisée pour placer les quantiles théoriques en abscisse.
- Si les résidus sont normaux, les points du QQ-Plot sont alignés.
- **Effet des distributions** :
  - **Asymétrie positive** : courbure des points au-dessus de la diagonale.
  - **Asymétrie négative** : courbure des points en dessous.
  - **Queue lourde** : dispersion importante aux extrémités.
- Vérifier la normalité permet d'ajuster le modèle en conséquence.

# Exercice : Analyse des Résidus pour les Problèmes des Notes Obtenues et l'Espérance de Vie

# Analyse des Résidus : Code Python

```python
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from google.colab import drive
# Étape 0.a: Monter Google Drive
drive.mount('/content/drive')
# Étape 0.b: Définir le chemin du fichier CSV dans Google Drive
dossier = "Colab Notebooks"  # Modifier si nécessaire
nom_fichier1 = "StudentGrades.csv"  # Assurez-vous que le nom du fichier est correct
chemin_fichier1 = f"/content/drive/My Drive/{dossier}/{nom_fichier1}"
nom_fichier2 = "Esperance_vie_pib.csv"  # Assurez-vous que le nom du fichier est correct
chemin_fichier2 = f"/content/drive/My Drive/{dossier}/{nom_fichier2}"
# Étape 0.c: Charger les données
print("Chargement des données 1 depuis Google Drive...")
data_etudes = pd.read_csv(chemin_fichier1)
print("Données chargées avec succès.")
print(data_etudes.head())  # Afficher les premières lignes du jeu de données
print("Chargement des données 2 depuis Google Drive...")
data_life_expect = pd.read_csv(chemin_fichier2)
print("Données chargées avec succès.")
print(data_life_expect.head())  # Afficher les premières lignes du jeu de données
```

# Analyse des Résidus : Code Python

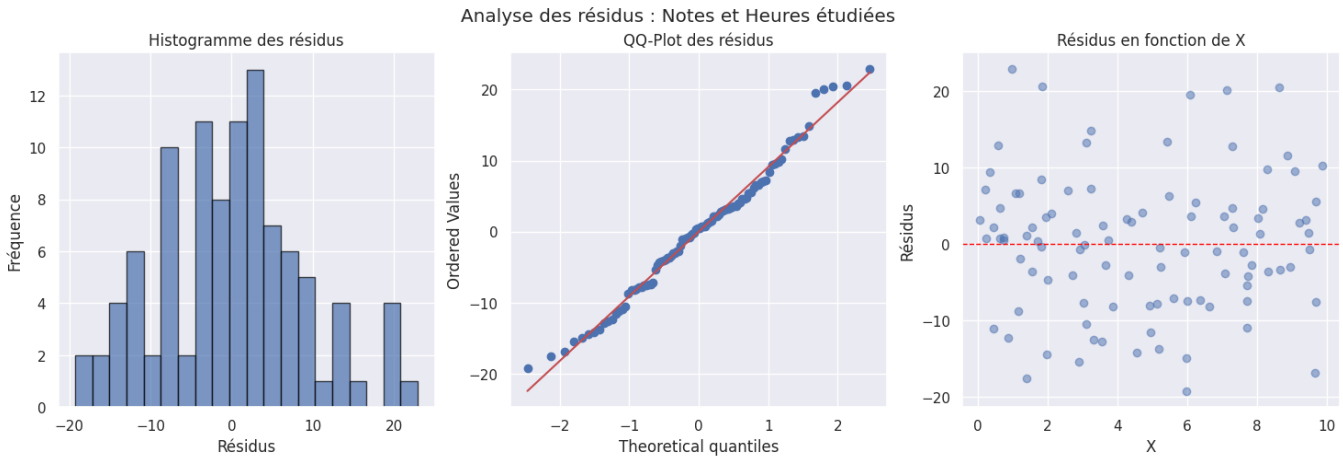
```python
# Étape 0.d: Extraction des variables
X1 = data_etudes[['Hours Studied']].values  # Variable indépendante (heures étudiées)
y1 = data_etudes['Grades'].values  # Variable dépendante (note obtenue)
X2 = data_life_expect[['GDP per capita (current US$)']].values  # Variable indépendante (PI
y2 = data_life_expect['Life Expect 2024'].values  # Variable dépendante (espérance de vie)
# Fonction pour tracer les graphiques de résidus
def plot_residuals(X, y, title):
    modele = LinearRegression()
    modele.fit(X, y)
    y_pred = modele.predict(X)
    residuals = y - y_pred
    fig, axes = plt.subplots(1, 3, figsize=(18, 5))
    # Histogramme des résidus
    axes[0].hist(residuals, bins=20, edgecolor='black', alpha=0.7)
    axes[0].set_title("Histogramme des résidus")
    axes[0].set_xlabel("Résidus")
    axes[0].set_ylabel("Fréquence")
    stats.probplot(residuals, dist="norm", plot=axes[1]) # QQ-Plot des résidus
    axes[1].set_title("QQ-Plot des résidus")
    # Graphique des résidus
    axes[2].scatter(X, residuals, alpha=0.5)
    axes[2].axhline(0, color='red', linestyle='dashed', linewidth=1)
    axes[2].set_title("Résidus en fonction de X")
    axes[2].set_xlabel("X")
    axes[2].set_ylabel("Résidus")
```

# Analyse des Résidus : Code Python

```python
def plot_residuals(X, y, title):
    modele = LinearRegression()
    modele.fit(X, y)
    y_pred = modele.predict(X)
    residuals = y - y_pred
    fig, axes = plt.subplots(1, 3, figsize=(18, 5))
    # Histogramme des résidus
    axes[0].hist(residuals, bins=20, edgecolor='black', alpha=0.7)
    axes[0].set_title("Histogramme des résidus")
    axes[0].set_xlabel("Résidus")
    axes[0].set_ylabel("Fréquence")
    stats.probplot(residuals, dist="norm", plot=axes[1]) # QQ-Plot des résidus
    axes[1].set_title("QQ-Plot des résidus")
    # Graphique des résidus
    axes[2].scatter(X, residuals, alpha=0.5)
    axes[2].axhline(0, color='red', linestyle='dashed', linewidth=1)
    axes[2].set_title("Résidus en fonction de X")
    axes[2].set_xlabel("X")
    axes[2].set_ylabel("Résidus")
    plt.suptitle(title)
    plt.show()
plot_residuals(X1, y1, "Analyse des résidus : Notes et Heures étudiées")
plot_residuals(X2, y2, "Analyse des résidus : Espérance de Vie et PIB")
```

# Analyse des Résidus : Notes et Heures étudiées

- **Objectif :** Vérifier la normalité des résidus pour la régression des notes en fonction des heures étudiées.
- **Hypothèses** :
  - Les résidus doivent être centrés autour de 0.
  - Ils doivent être homoscédastiques (variance constante).
  - Ils ne doivent pas suivre de structure particulière.
- **Graphiques à Fournir:**
  - Histogramme des résidus.
  - QQ-Plot des résidus.
  - Graphique des résidus en fonction des heures étudiées.



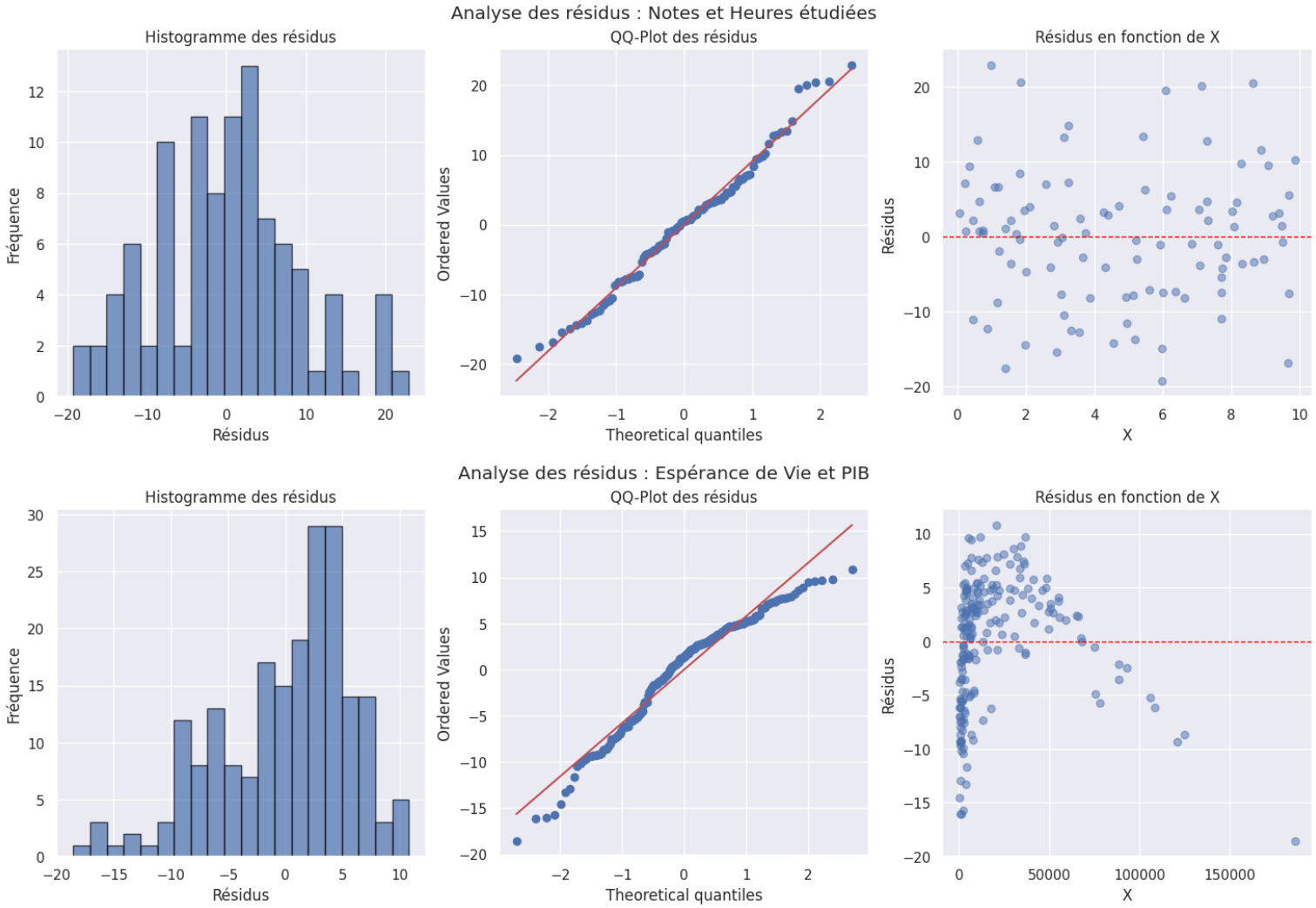
# Analyse des Résidus : Explication des Graphiques

- **Histogramme des résidus :** Permet de visualiser si les erreurs suivent une loi normale.
- **QQ-Plot :** Vérifie si les quantiles des résidus suivent ceux d'une loi normale.
- **Graphique des résidus :** Permet de détecter une éventuelle structure ou hétéroscédasticité.
- **Interprétation :**
  - Si les résidus sont bien répartis autour de 0 et suivent une distribution normale, l'hypothèse de normalité est valide.
  - Si les résidus montrent une structure ou une variance variable, la régression peut être mal spécifiée.

# Analyse des Résidus : Explication des Graphiques

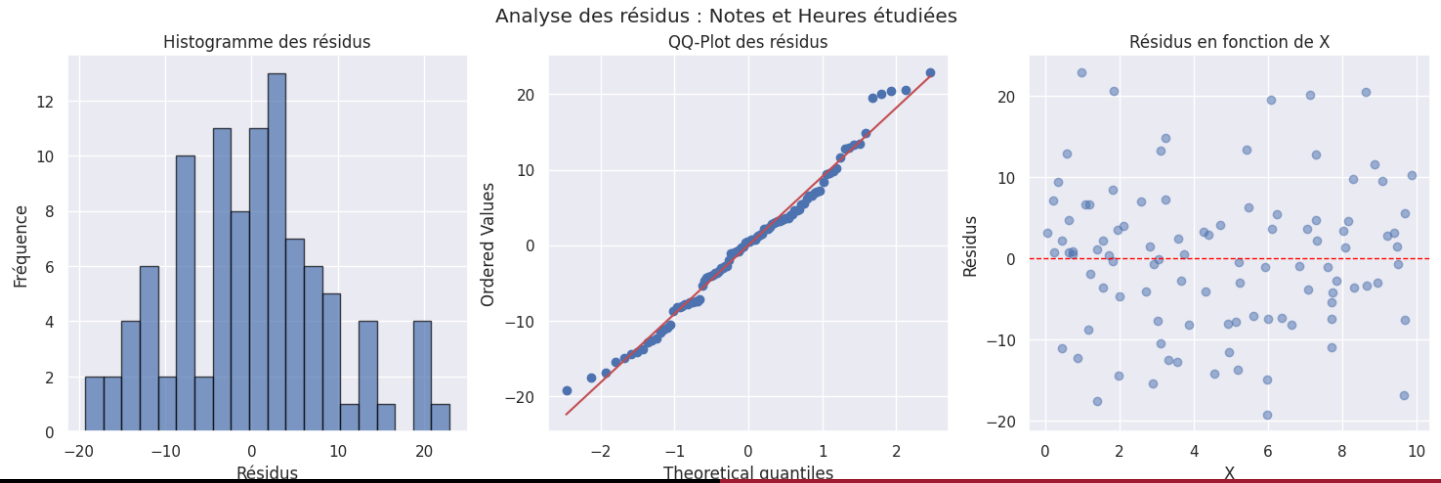
- **Si les hypothèses sont respectées :**
  - La régression linéaire est adaptée.
  - Les inférences statistiques basées sur les tests t et F sont valides.
- **Si les hypothèses sont violées :**
  - Transformation des variables (ex : log transformation pour PIB).
  - Régression non linéaire si la relation n'est pas strictement linéaire.
  - Utilisation de modèles plus robustes.

# Analyse des Résidus : Résultats Obtenus



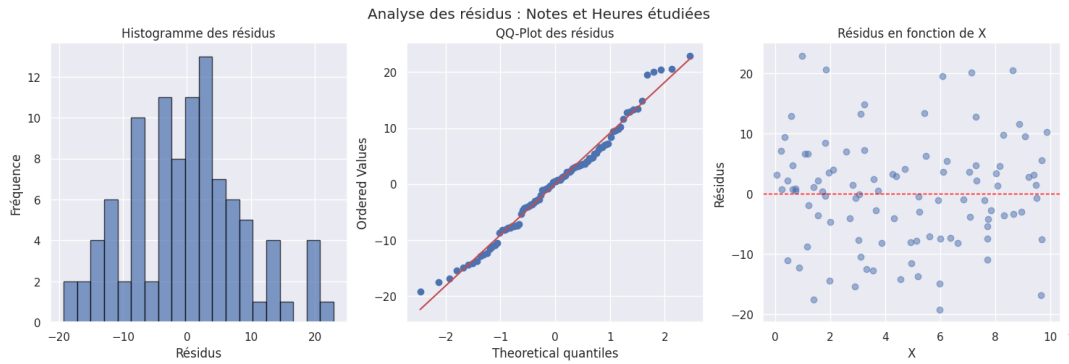
# Analyse des Résidus : Notes et Heures Étudiées

- **Histogramme des résidus :**
  - L'histogramme montre une distribution des résidus qui est globalement centrée autour de zéro.
  - La distribution semble légèrement asymétrique mais reste proche d'une distribution normale.
- **QQ-Plot des résidus :**
  - Les points suivent assez bien la droite de référence, ce qui suggère que la normalité des résidus est approximativement respectée.
  - Quelques écarts aux extrémités pourraient indiquer la présence de légères queues épaisses (léger écart à la normalité).



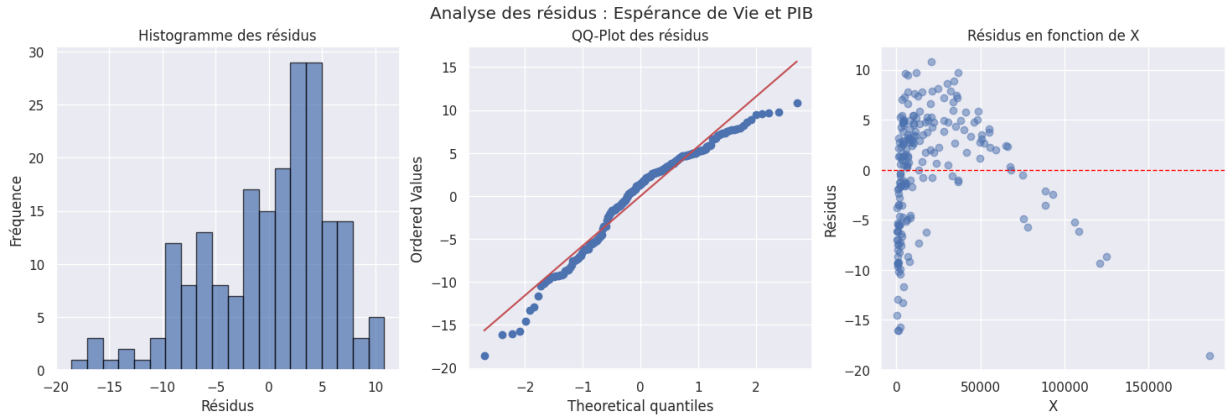
# Analyse des Résidus : Notes et Heures Étudiées

- **Graphique des résidus en fonction des heures étudiées :**
  - Les résidus sont dispersés de manière relativement homogène autour de zéro, sans motif évident.
  - Il n'y a pas de tendance marquée, ce qui indique que l'hypothèse de linéarité semble raisonnable.
  - L'hypothèse d'homoscédasticité (variance constante des résidus) semble être respectée.
- **Conclusion :**
  - La régression linéaire est appropriée pour modéliser la relation entre le nombre d'heures étudiées et les notes obtenues.
  - Légère asymétrie possible dans la distribution des résidus, mais pas de preuve évidente de non-normalité ni de non-linéarité significative.



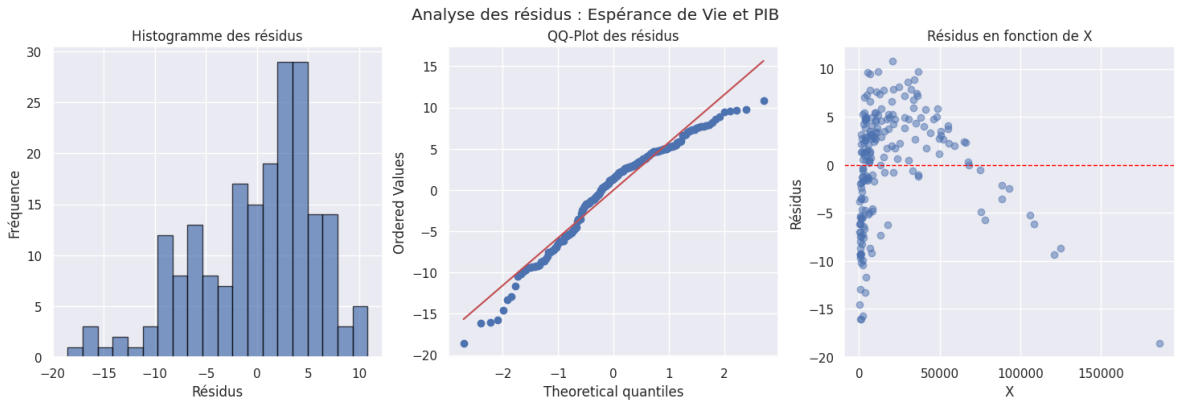
# Analyse des Résidus : Espérance de Vie et PIB

- **Histogramme des résidus :**
  - La distribution des résidus est clairement asymétrique, avec une forte concentration de valeurs proches de zéro et une queue plus allongée du côté négatif.
  - Cela indique que la régression linéaire ne capture pas bien la relation entre l'espérance de vie et le PIB.
- **QQ-Plot des résidus :**
  - Les points s'écartent notablement de la diagonale, en particulier aux extrémités, ce qui indique que les résidus ne suivent pas une distribution normale.
  - Cela suggère que le modèle linéaire n'est pas bien adapté.



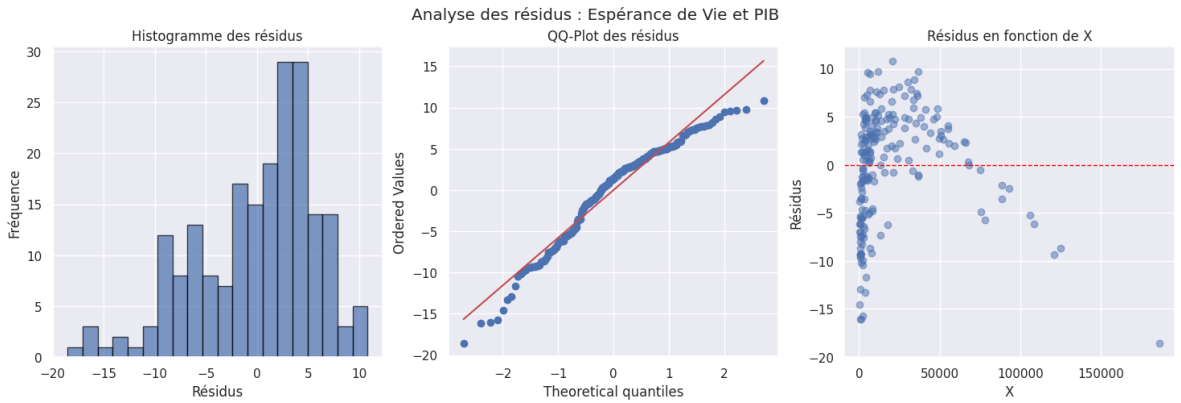
# Analyse des Résidus : Espérance de Vie et PIB

- **Graphique des résidus en fonction du PIB :**
  - Forte hétéroscédasticité : on observe une concentration des résidus autour de zéro pour les petits PIB et une dispersion plus grande pour les PIB élevés.
  - Cela suggère une relation non linéaire entre le PIB et l'espérance de vie.
  - Le modèle linéaire ne semble pas capturer correctement la dynamique entre ces deux variables.



# Analyse des Résidus : Espérance de Vie et PIB

- **Conclusion :**
  - Le modèle linéaire est inadapté pour prédire l'espérance de vie en fonction du PIB.
  - Il serait pertinent d'explorer une transformation logarithmique du PIB, d'utiliser une régression polynomial ou une transformation log-log.
  - L'hétéroscédasticité est forte, ce qui affecte la validité des tests statistiques classiques.



# Transformations en Régression Linéaire

# Exercice : Transformation Logarithmique des Données

# Transformation Logarithmique des Données : Code Python

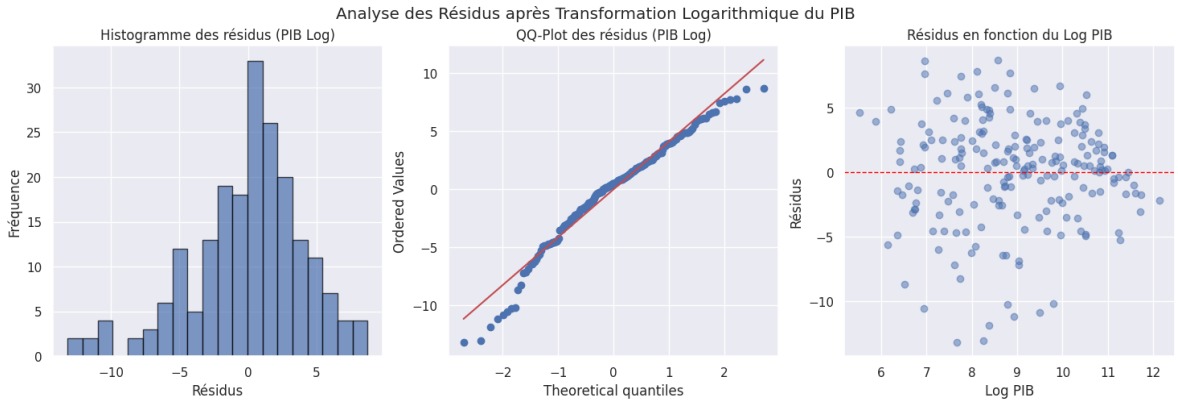
```python
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
# Charger les données
dossier = "Colab Notebooks"
nom_fichier = "Esperance_vie_pib.csv"
chemin_fichier = f"/content/drive/My Drive/{dossier}/{nom_fichier}"
print("Chargement des données...")
data = pd.read_csv(chemin_fichier)
print("Données chargées avec succès.")
print(data.head())
# Transformation logarithmique du PIB
data["Log_GDP_per_capita"] = np.log(data["GDP per capita (current US$)"])
# Définition des variables pour la régression
X_log = data[['Log_GDP_per_capita']].values  # PIB transformé en log
y = data['Life Expect 2024'].values  # Espérance de vie
# Création et entraînement du modèle de régression linéaire
modele_log = LinearRegression()
modele_log.fit(X_log, y)
y_pred_log = modele_log.predict(X_log)
# Calcul des résidus
residuals_log = y - y_pred_log
```

# Transformation Logarithmique des Données : Code Python

```python
# Visualisation des résidus
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
# Histogramme des résidus
axes[0].hist(residuals_log, bins=20, edgecolor='black', alpha=0.7)
axes[0].set_title("Histogramme des résidus (PIB Log)")
axes[0].set_xlabel("Résidus")
axes[0].set_ylabel("Fréquence")
# QQ-Plot des résidus
stats.probplot(residuals_log, dist="norm", plot=axes[1])
axes[1].set_title("QQ-Plot des résidus (PIB Log)")
# Graphique des résidus
axes[2].scatter(X_log, residuals_log, alpha=0.5)
axes[2].axhline(0, color='red', linestyle='dashed', linewidth=1)
axes[2].set_title("Résidus en fonction du Log PIB")
axes[2].set_xlabel("Log PIB")
axes[2].set_ylabel("Résidus")
plt.suptitle("Analyse des Résidus après Transformation Logarithmique du PIB")
plt.show()
# Affichage du R² du modèle transformé
r2_log = r2_score(y, y_pred_log)
print(f"R² après transformation logarithmique du PIB: {r2_log:.4f}")
```

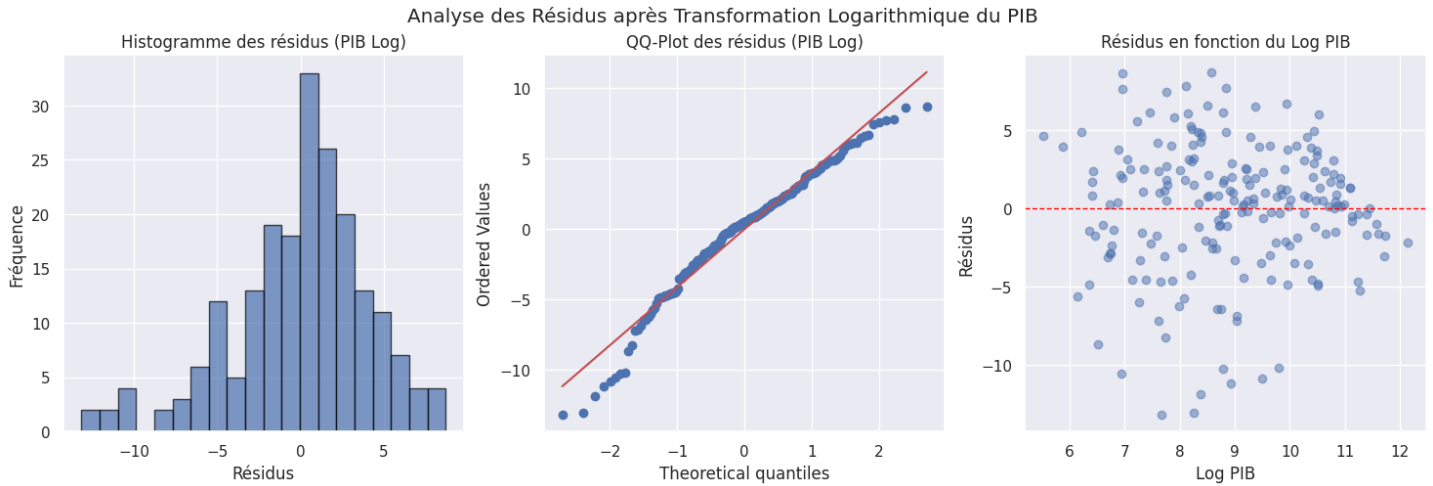
# Analyse des Résidus : Transformation Logarithmique

- **Histogramme des Résidus**
  - **Observations :**
    - La distribution des résidus est plus centrée autour de zéro, comparée à l'histogramme avant transformation.
    - Elle semble moins asymétrique, ce qui indique une amélioration en termes de normalité des résidus.
    - Il reste une légère queue négative, mais la distribution est plus proche d'une normale.
  - **Interprétation :**
    - La transformation logarithmique a permis de mieux ajuster la relation entre le PIB et l'espérance de vie.
    - Il subsiste des écarts, mais ceux-ci sont moins marqués que dans le modèle linéaire sans transformation.



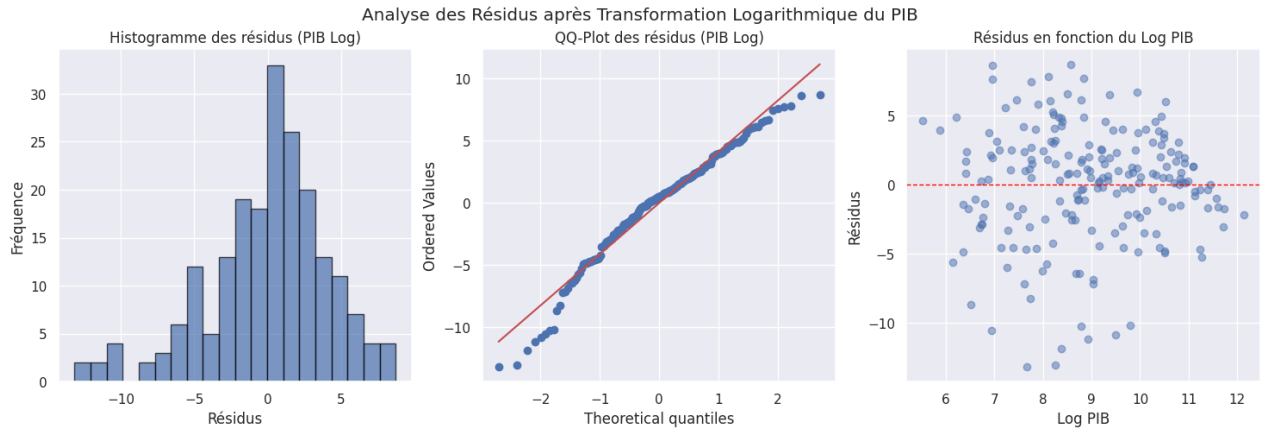
# Analyse des Résidus : Transformation Logarithmique

- **QQ-Plot des Résidus**
  - **Observations :**
    - Les points suivent beaucoup mieux la droite diagonale, ce qui indique une meilleure normalité des résidus.
    - Avant transformation, les extrémités montraient des écarts importants (queues épaisses), alors qu'ici l'alignement est nettement amélioré.
  - **Interprétation :**
    - L'hypothèse de normalité des résidus est plus raisonnable après transformation.
    - Cela signifie que les tests statistiques associés (tests de Student, F) sont plus fiables.



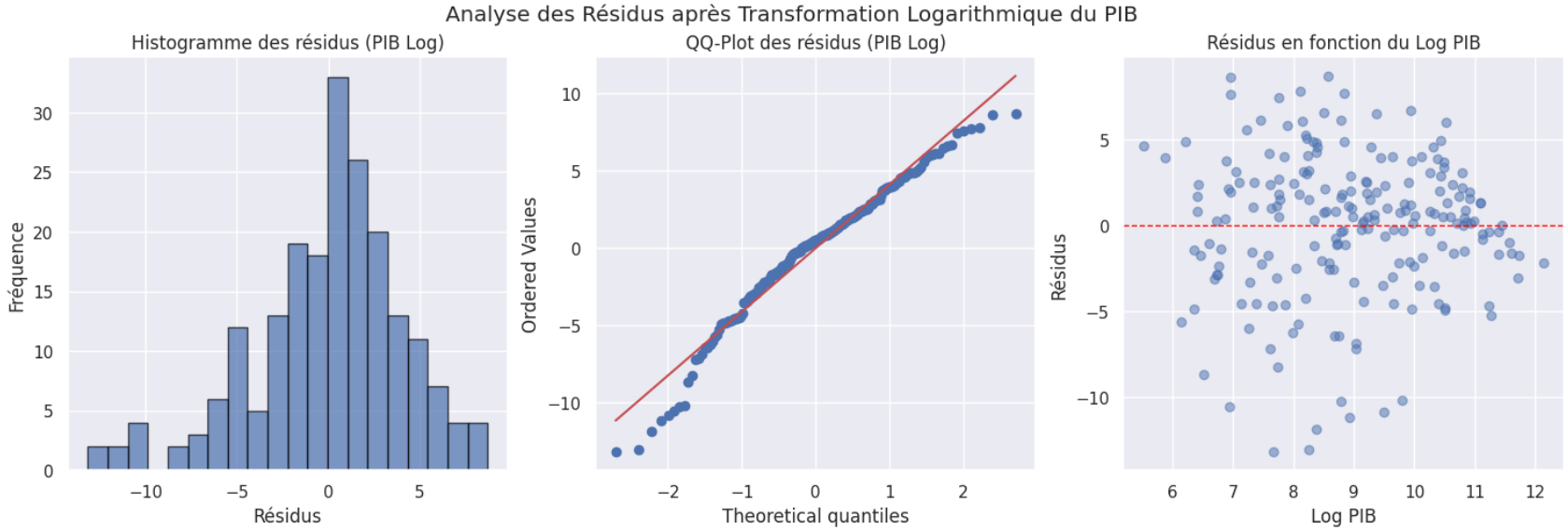
# Analyse des Résidus : Transformation Logarithmique

- **Résidus en Fonction du Log(PIB)**
  - **Observations :**
    - Contrairement au modèle linéaire de départ, la dispersion des résidus est plus homogène.
    - On observe moins de structure évidente, suggérant que l'hypothèse d'homoscédasticité (variance constante) est mieux respectée.
    - Toutefois, une légère variabilité reste visible pour les PIB élevés, mais elle est réduite par rapport au modèle précédent.
  - **Interprétation :**
    - L'effet de hétéroscédasticité a été significativement atténué.
    - Cela indique que la relation PIB - Espérance de vie suit bien une courbe logarithmique plutôt qu'une relation linéaire simple.



# Analyse des Résidus : Transformation Logarithmique

- **Améliorations après transformation logarithmique :**
  - **Normalité des résidus :** QQ-plot montre une meilleure adéquation.
  - **Hétéroscédasticité réduite :** la dispersion des résidus est plus homogène.
  - **Meilleur ajustement du modèle :** les écarts aux extrémités ont diminué.



# Applications Générales des Transformations en Régression Linéaire

# Applications Générales des Transformations

- Dans un modèle de régression linéaire, certaines hypothèses doivent être respectées pour garantir la validité des résultats :
  - **Normalité des résidus**.
  - **Homoscédasticité** (variance constante des résidus).
  - **Linéarité** de la relation entre $X$ et $Y$.
  - **Indépendance** des observations.
- Lorsque ces hypothèses ne sont pas respectées, des transformations mathématiques sont souvent appliquées aux variables dépendantes $(Y)$ et indépendantes $(X)$ pour améliorer l'ajustement du modèle.

# Applications Générales des Transformations : $\log(X)$

**Utilisation :**

- Réduit l'effet des valeurs extrêmes (grandeurs variant sur plusieurs ordres de magnitude).
- Rend une relation exponentielle linéaire.
- Réduit l'hétéroscédasticité.

**Formule :**

$$X^* = \log(X), \quad Y^* = \log(Y)$$

**Exemples d'application :**

- Relation PIB $\rightarrow$ Espérance de Vie
- Relation Revenu $\rightarrow$ Consommation
- Réduction de l'effet des grandes valeurs dans des distributions à queue lourde.

# Applications Générales des Transformations : $\sqrt{(X)}$

**Utilisation :**

- Réduit l'impact des valeurs extrêmes sans trop modifier la structure des données.
- Utile lorsque la variance augmente avec la moyenne (hétéroscédasticité).
- Souvent utilisée pour les variables comptant des fréquences.

**Formule :**

$$X^* = \sqrt{X}, \quad Y^* = \sqrt{Y}$$

**Exemples d'application :**

- Modélisation du nombre de ventes ou d'appels en marketing.
- Variables de comptage comme le nombre d'accidents ou la population.

# Applications Générales des Transformations : $\frac{1}{X}$

**Utilisation :**

- Convient aux relations hyperboliques (décroissance rapide).
- Linéarise des relations où l'effet marginal diminue fortement.

**Formule :**

$$X^* = \frac{1}{X}$$

**Exemples d'application :**

- Temps de réponse $\rightarrow$ performance du système.
- Coût marginal $\rightarrow$ Quantité produite (rendements décroissants).

# Applications Générales des Transformations : (*Box-Cox*)

**Utilisation :**

- Recherche automatiquement la meilleure puissance pour transformer les données.
- Corrige l'hétéroscédasticité et la non-normalité.

**Formule (Box-Cox) :**

$$X^* = \begin{cases} \frac{X^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(X) & \text{si } \lambda = 0 \end{cases}$$

**Exemples d'application :**

- Si les données sont positives et asymétriques.
- Si une transformation simple (log, sqrt) ne fonctionne pas.

# Applications Générales des Transformations : Différences

**Utilisation :**

- Élimine les tendances et rend la série stationnaire en analyse de séries temporelles.
- Peut être utilisée sur $Y$ et/ou $X$ pour capturer des variations plus fines.

**Formule :**

$$Y_t^* = Y_t - Y_{t-1}$$

**Exemples d'application :**

- Prédictions économiques (ex: croissance du PIB, inflation).
- Analyse des séries temporelles (marchés financiers, climatologie).

# Applications Générales des Transformations : log-log

**Utilisation :**

- Rend une relation puissance linéaire.
- Permet d'interpréter les coefficients comme des élasticités.

**Formule :**

$$Y^* = \log(Y), \quad X^* = \log(X)$$

**Exemples d'application :**

- Économie et finance : Relation entre prix et demande.
- Écologie : Relation taille des animaux $\rightarrow$ consommation d'énergie.

# Applications Générales des Transformations : Sigmoïde

**Utilisation :**

- Appliquée quand les valeurs de $Y$ sont bornées (ex: taux de conversion, notation sur 100).
- Appropriée pour des variables qui croissent lentement puis rapidement, avant de saturer.

**Formule :**

$$Y^* = \frac{1}{1 + e^{-Y}}$$

**Exemples d'application :**

- Modèles de croissance (ex: adoption d'une technologie).
- Modélisation des probabilités en régression logistique.

# Tableau Récapitulatif des Transformations

| **Transformation** | **Formule** | **Cas d'utilisation** |
|---|---|---|
| Logarithmique | $X^* = \log(X)$ | Exponentielle, hétéroscédasticité |
| Racine Carrée | $X^* = \sqrt{X}$ | Var croissante avec la moy |
| | $X^* = \frac{1}{X}$ | Rendements décroissants |
| Box-Cox | $\frac{X^\lambda - 1}{\lambda}$ | Optimisation de normalité |
| Différentielle | $Y_t^* = Y_t - Y_{t-1}$ | Séries temporelles stationnaires |
| Log-Log | $X^* = \log(X), Y^* = \log(Y)$ | Élasticité, relations de puissance |
| Sigmoïde | $Y^* = \frac{1}{1+e^{-Y}}$ | Variables bornées (taux, proportion |

# Table des Matières

1 Analyse des Résidus

2 Régression Linéaire Multiple

3 Décomposition de la Variabilité Totale

4 ANOVA en Régression Linéaire Multiple

5 Exercice : Boston Housing Prices Dataset

6 Intervalle de Confiance et Intervalle de Prédiction

7 Annexe

# Régression Linéaire Multiple

# 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 $\beta_0$ 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 $\beta_0$ et les pentes $\beta_1, \ldots, \beta_p$.
- $\mathbf{Y}$: Vecteur aléatoire qui représentant les variables aléatoires dépeandantes $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_n]^{\mathrm{T}}$.
- $\epsilon$: 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 $\beta_0$ 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 $\beta_0$ du modèle lors de la prédiction de $\mathbf{Y}$:

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

# Régression Linéaire Multiple

## Expression du Modèle

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

- $\epsilon$ l'erreur aléatoire du modèle.
- $X_1, X_2, ..., X_p$ les variables explicatives, appelées variables indépendantes.
- $Y$ la variable dépendante, appelée aussi réponse.
- $\beta_0$ 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 régression partiels** associés à chaque variable indépendante $X_1, X_2, ..., X_p$. Chaque coefficient $\beta_i$ 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.

# Formulation Matricielle de la Régression Linéaire Multiple

- **Forme vectorielle :**

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- **Définition des matrices :**

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

# 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. Déterminisme des $X_i$**
  - Les $X_i$ sont traitées comme déterministes (non aléatoires).

# Régression Linéaire Multiple : Modèles Linéaires

- **Modèle Linéaire (en $\beta$): Ajout de termes quadratiques :**

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

- **Modèle Linéaire (en $\beta$): Ajout d'interactions entre variables :**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{2i} + \beta_3 (X_i \cdot X_{2i}) + \epsilon_i$$

- **Modèle Linéaire (en $\beta$): Transformation logarithmique du modèle :**

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_{1i}) + \beta_2 \frac{1}{X_{2i}} + \epsilon_i$$

- **Modèle Non-Linéaire (en $\beta$): Non-linéatité avec une fonction sigmoïde :**

$$Y_i = \frac{\beta_0}{1 + e^{-(\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i})}} + \epsilon_i$$

# Modèle de Régression Linéaire Multiple : Estimation par La Méthode des Moindres Carrés Ordinaires (MCO)

# Rappel : Gradients et Hessiennes

- **Rappel :**
  - Le gradient $\nabla f(x)$ d'une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est un vecteur contenant les dérivées partielles :

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \left( \frac{\partial f}{\partial \mathbf{x}} \right).$$

  - **Dérivée d'une forme quadratique :** Si $\mathbf{x}$ est un vecteur et $\mathbf{A}$ est une matrice symétrique, la dérivée de la forme quadratique $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ par rapport à $\mathbf{x}$ est donnée par :

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$$

  - **Dérivée d'un produit de type vecteur-matrice-vecteur :** Si $\mathbf{b}$ et $\mathbf{x}$ sont des vecteurs et $\mathbf{A}$ est une matrice, alors la dérivée de $\mathbf{b}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ par rapport à $\mathbf{x}$ est :

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{b}) = \mathbf{A}\mathbf{b}$$

# Moindres Carrés Ordinaires - 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 $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i$.
- Le problème d'optimisation est donc défini comme suit :

$$\begin{aligned} & \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} e_i^2 \\ \equiv \quad & \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ \equiv \quad & \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2 \end{aligned}$$

- Cette expression cherche les valeurs de $\boldsymbol{\beta}$ qui minimisent la somme des carrés des écarts entre les valeurs observées et les valeurs prédites

# MCO - Étape 2: Calcul des Dérivées Partielles

- On définit la fonction objectif :

$$\begin{aligned} C(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= ||\boldsymbol{y} - \hat{\boldsymbol{y}}||_2^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}} (\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

- Pour minimiser $C$, on calcule la dérivée partielle par rapport à $\hat{\beta}$ et on l'annule (en sachant que la matrice de Gram $\mathbf{G} = \mathbf{X}^{\mathrm{T}}\mathbf{X} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}$ est toujours symétrique) : $\frac{\partial C}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}$. (**Consulter l'annexe pour les propriétés de $\mathbf{G} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$**)

# MCO - Étape 3: Résolution du Système d'Équations

- En posant cela à zéro, on obtient :

$$-\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

- Nous avons obtenu l'équation normale suivante :

$$\mathbf{X}^{\top}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\top}\boldsymbol{y}$$

- Si $\mathbf{X}^{\top}\mathbf{X}$ est inversible, on peut isoler $\hat{\boldsymbol{\beta}}$ :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{y} = \mathbf{G}^{-1}\mathbf{X}^{\top}\boldsymbol{y}$$

- Cet estimateur est l'estimateur des moindres carrés ordinaires MCO (Ordinary Least Squares - OLS) pour la régression multiple.
- Géométriquement, cela signifie que $\hat{\boldsymbol{\beta}}$ représente la projection orthogonale du vecteur $\boldsymbol{y}$ sur le sous-espace engendré par les colonnes de $\mathbf{X}$.

# MCO - Importance de l'Absence de Multicolinéarité

- L'estimateur des moindres carrés ordinaires est donc donné par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y} = \mathbf{G}^{-1} \mathbf{X}^\top \boldsymbol{y}$$

- Si **les variables explicatives sont linéairement dépendantes, alors $\mathbf{X}^\top \mathbf{X}$ est singulière (non-inversible)**, empêchant le calcul des coefficients (**Veuillez consulter l'annexe pour la démonstration**).
- Il faut donc s'assurer que la matrice $\mathbf{X}^\top \mathbf{X}$ est inversible pour assurer l'existence de l'estimateur des moindres carrés.
- $\Rightarrow$ **d'où l'importance de l'hypothèse de l'absence de multicolinéarité** (la matrice $\mathbf{X}$ est plein rang).
- L'absence de multicolinéarité est alors une hypothèse essentielle en régression linéaire multiple.
- Cela suppose que les variables explicatives ne doivent pas être linéairement dépendantes.
- La matrice $\mathbf{X}$ doit être de plein rang ($\text{rang}(\mathbf{X}) = p$, où $p$ est le nombre de variables explicatives).

# MCO - Importance de l'Absence de Multicolinéarité

- Lorsque deux (ou plusieurs) variables explicatives sont fortement corrélées, les colonnes de $\mathbf{X}$ sont presque linéairement dépendantes.
- Cela rend $\mathbf{X}^\top \mathbf{X}$ proche de la singularité, ce qui signifie que son inverse $(\mathbf{X}^\top \mathbf{X})^{-1}$ a des valeurs élevées.
- La conséquence est une grande variance des coefficients $\hat{\boldsymbol{\beta}}$, rendant leur estimation instable **(Veuillez consulter l'annexe pour une explication intuitive)**.
- Petites variations dans les données peuvent entraîner de grands changements dans les coefficients estimés.
- Difficulté d'interprétation : un coefficient peut devenir très grand ou même changer de signe par rapport aux attentes théoriques.
- L'augmentation de la variance entraîne des intervalles de confiance plus larges, ce qui rend plus difficile la détection de l'effet réel d'une variable et le calcul des tests de significativité (statistique $t$).

# Définition du Nombre de Conditionnement

- Pour analyser le conditionnement et la sensibilité numérique de la matrice de Gram $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$, nous devons raisonner en termes de **nombre de conditionnement** (*condition number*).
- Le nombre de conditionnement de la matrice $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$ permet de mesurer la sensibilité des solutions du système d'équations $\hat{\boldsymbol{\beta}} = \mathbf{G}^{-1}\mathbf{X}^\top \boldsymbol{y}$ aux perturbations des données.

## 1 Définition du Nombre de Conditionnement

- Le nombre de conditionnement d'une matrice $\mathbf{G}$ est défini par :

$$\kappa(\mathbf{G}) = \left| \frac{\lambda_{\max}(\mathbf{G})}{\lambda_{\min}(\mathbf{G})} \right|$$

où :

- $\lambda_{\max}(\mathbf{G})$ est la plus grande valeur propre de $\mathbf{G}$,
- $\lambda_{\min}(\mathbf{G})$ est la plus petite valeur propre de $\mathbf{G}$.

# Interprétation du Nombre de Conditionnement

## 2 Interprétation du Nombre de Conditionnement

- Le nombre de conditionnement $\kappa(\mathbf{G})$ quantifie la sensibilité de la solution du système linéaire aux erreurs dans les données :
  - **Si $\kappa(\mathbf{G}) \approx 1$ :**
    - La matrice $\mathbf{G}$ est bien conditionnée.
    - L'inversion de $\mathbf{G}$ est numériquement stable.
    - L'estimation des coefficients de régression est fiable.
  - **Si $\kappa(\mathbf{G}) \gg 1$ (grande valeur) :**
    - La matrice $\mathbf{G}$ est mal conditionnée.
    - L'inversion de $\mathbf{G}$ est numériquement instable.
    - Une petite variation dans $\mathbf{X}$ peut produire une grande variation dans $\hat{\boldsymbol{\beta}}$.
    - La **multicolinéarité** est un problème majeur.
  - **Si $\kappa(\mathbf{G}) \to \infty$ :**
    - $\lambda_{\min}(\mathbf{G})$ proche de 0
    - La matrice $\mathbf{G}$ est **singulière** (non inversible).
    - Cela signifie que les colonnes de $\mathbf{X}$ sont **linéairement dépendantes**.
    - L'estimateur des moindres carrés ne peut pas être calculé sans régularisation ($L_1$ LASSO ou $L2$ Ridge).

# MCO - Importance de l'Absence de Multicolinéarité

- L'absence de multicolinéarité garantit des estimations **stables**, **interprétables** et **statistiquement fiables**.
- La présence de multicolinéarité peut fausser les résultats et compromettre l'analyse.
- Il est essentiel de diagnostiquer et corriger ce problème lorsque nécessaire.
- **Solutions Potenitelles :**
  - Supprimer une des variables fortement corrélées si elle n'apporte pas d'information supplémentaire **(en calculant par exemple le Facteur d'Inflation de la Variance (FIV). Voir annexe)**.
  - Combiner les variables corrélées en une seule nouvelle variable (par ex. via une analyse en composantes principales, ACP).
  - Utiliser la régression Ridge (L2) qui pénalise les grands coefficients et stabilise l'estimation.
  - Augmenter la taille de l'échantillon $n$, ce qui peut permettre de mieux identifier les effets distincts des variables.

# Biais et Variance de l'Estimateur des Moindres Carrés $\hat{\beta}$

# Biais de l'Estimateur des Moindres Carrés $\hat{\beta}$

# Biais de l'Estimateur des Moindres Carrés $\hat{\boldsymbol{\beta}}$

- On considère le modèle de régression linéaire multiple sous forme matricielle :

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Où :
  - $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ est le vecteur des observations,
  - $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ est la matrice des variables explicatives,
  - $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ est le vecteur des coefficients inconnus,
  - $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$ est le vecteur des erreurs aléatoires.
- On suppose que les erreurs $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n)$ suivent une loi normale centrée avec variance $\sigma^2\mathbf{I}_n$ :

$$\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \text{et} \quad \boldsymbol{\Sigma} = \text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n,$$

où $\boldsymbol{\Sigma}$ est la matrice de covariance de $\boldsymbol{\epsilon}$.

# Biais de l'Estimateur des Moindres Carrés $\hat{\boldsymbol{\beta}}$

- L'estimateur par la méthode des moindres carrés ordinaires de $\boldsymbol{\beta}$ est donné par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

- En remplaçant $\boldsymbol{y}$ par son expression dans le modèle de régression :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}).$$

- En développant :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X})\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}.$$

- En utilisant la propriété $(\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{X}) = \mathbf{I}_{p+1}$ (sous l'hypothèse que $\mathbf{X}$ est de plein rang) :

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}.$$

# Biais de l'Estimateur des Moindres Carrés $\hat{\boldsymbol{\beta}}$

- L'espérance de $\hat{\boldsymbol{\beta}}$ est donnée par :

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left[\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}\right].$$

- En utilisant la linéarité de l'espérance :

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\boldsymbol{\beta}) + \mathbb{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}\right].$$

- Comme $\boldsymbol{\beta}$ est un vecteur de constantes, son espérance est lui-même :

$$\mathbb{E}(\boldsymbol{\beta}) = \boldsymbol{\beta}.$$

- De plus, en utilisant l'hypothèse $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, on obtient :

$$\mathbb{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}\right] = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbb{E}(\boldsymbol{\epsilon}) = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{0} = \mathbf{0}.$$

- Ainsi, on obtient :

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \Rightarrow \text{biais}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = 0.$$

- Cela signifie que l'estimateur $\hat{\boldsymbol{\beta}}$ est sans biais.

# Variance de l'Estimateur des Moindres Carrés $\hat{\beta}$

# Variance de l'Estimateur des Moindres Carrés $\hat{\boldsymbol{\beta}}$

- L'estimateur par la méthode des moindres carrés des coefficients est donné par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

- En remplaçant $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

- En développant :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X})\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}$$

- Comme $(\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) = \mathbf{I}_{p+1}$, on obtient :

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}$$

# Variance de l'Estimateur des Moindres Carrés $\hat{\boldsymbol{\beta}}$

- La variance de $\hat{\boldsymbol{\beta}}$ est donnée par :

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}\left(\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}\right)$$

- Comme $\boldsymbol{\beta}$ est une constante, sa variance est nulle, donc :

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}\left((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}\right)$$

- Utilisant la propriété $\text{Var}(\mathbf{Az}) = \mathbf{A}\text{Var}(\mathbf{z})\mathbf{A}^\top$ et sachant que $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$, on obtient :

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \sigma^2 \mathbf{I}_n \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}$$

- Ce qui simplifie à :

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

- $\text{Var}(\hat{\boldsymbol{\beta}})$ est proportionnelle à $(\mathbf{X}^\top \mathbf{X})^{-1}$. Une forte colinéarité entre les variables explicatives entraîne une augmentation de cette variance, rendant les estimations instables.

# Conclusion

- La régression linéaire multiple est une généralisation de la régression simple avec plusieurs prédicteurs.
- En régression multiple, $S_{xx}$ est analogue à $\mathbf{X}^\top \mathbf{X}$ et $S_{xy}$ est analogue à $\mathbf{X}^\top \boldsymbol{y}$.
- La variance des estimateurs est affectée par la **multicolinéarité** dans le cas multiple, contrairement au cas simple.

| Concept | Régression Linéaire Simple | Régression Linéaire Multiple |
|---|---|---|
| **Estimateur MCO** | $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ | $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$ |
| **Biais** | $\mathbb{E}[\hat{\beta}_1] = \beta_1$ | $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ |
| **Variance** | $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ | $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ |

# Expression de l'Intercept en Régression Multiple

- L'expression générale en régression linéaire multiple :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- L'estimateur des coefficients est donné par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad \text{où} \quad \hat{\boldsymbol{\beta}} = \left[\hat{\beta}_0, \quad \hat{\beta}_1, \quad \cdots, \quad \hat{\beta}_p\right]^\top$$

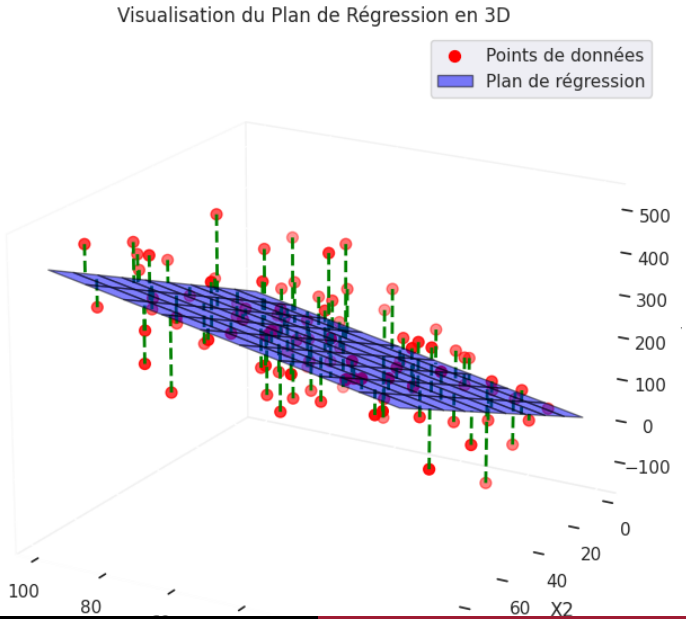
- L'estimateur de $\beta_0$ qui est $\hat{\beta}_0$ et qui est contenu dans $\hat{\boldsymbol{\beta}}$ s'écrit alors :

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j = \bar{y} - \hat{\boldsymbol{\beta}}_{1:p}^\top \bar{\mathbf{x}}, \quad \text{où :}$$

  - $\bar{y}$ est la moyenne de la variable dépendante $y$,
  - $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)^\top$ est le vecteur des moyennes des variables explicatives,
  - $\hat{\boldsymbol{\beta}}_{1:p}$ est le vecteur des coefficients estimés associés aux variables explicatives (excluant $\hat{\beta}_0$).

# Interprétation Géométrique

- En régression simple, la droite de régression passe par le point moyen $(\bar{x}, \bar{y})$.
- En régression multiple, **le plan de régression passe par le point moyen du nuage de points en dimension** $p+1$.
- L'équation $\hat{\beta}_0 = \bar{y} - \hat{\boldsymbol{\beta}}_{1:p}^\top \bar{\mathbf{x}}$ exprime le fait que l'intercept $\beta_0$ est ajusté pour que la régression soit centrée autour des moyennes des données.



# Exercice : Équivalence de la Méthode d'Estimation des Moindres Carrés Ordinaires avec la Méthode d'Estimation par Maximum de Vraisemblance

# Exercice : Équivalence avec le Maximum de Vraisemblance

- **Question :** Montrer que l'estimateur des Moindres Carrés Ordinaires (MCO) est identique à celui obtenu par la méthode du Maximum de Vraisemblance (MV) en régression linéaire multiple.
- **Modèle de Régression Multiple**
  On considère le modèle de régression linéaire multiple :

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

avec :

- $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ : vecteur des observations,
- $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ : matrice des prédicteurs,
- $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)\times 1}$ : vecteur des coefficients inconnus,
- $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ : erreurs supposées normales et indépendantes.

# Solution : Équivalence de la Méthode d'Estimation des Moindres Carrés Ordinaires avec la Méthode d'Estimation par Maximum de Vraisemblance

# Solution : Équivalence avec le Maximum de Vraisemblance

**1 Méthode des Moindres Carrés Ordinaires (MCO)**

Les MCO consistent à minimiser la somme des carrés des résidus :

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

La solution est donnée par :

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

**2 Estimation par Maximum de Vraisemblance (MV)**

Sous l'hypothèse $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, la vraisemblance de l'échantillon est :

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left( -\frac{1}{2\sigma^2} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right).$$

La log-vraisemblance s'écrit :

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

# Solution : Équivalence avec le Maximum de Vraisemblance

- Maximiser $\log L$ revient à minimiser :

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}),$$

qui est exactement le critère MCO.

3 **Conclusion : Équivalence des Estimateurs**

On obtient le même estimateur :

$$\hat{\boldsymbol{\beta}}_{\text{MV}} = \hat{\boldsymbol{\beta}}_{\text{MCO}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

4 **Différences entre MCO et MV**

- Si les erreurs sont normales, les deux méthodes sont équivalentes.
- MV fournit aussi une estimation de la variance $\sigma^2$ :

$$\hat{\sigma}^2 = s^2 = \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2.$$

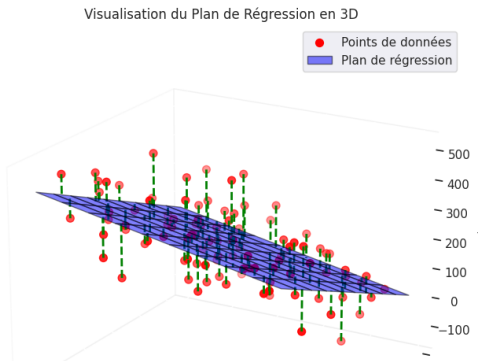
En général, on utilise une version corrigée :

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-p-1} \|\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2.$$

# Solution : Équivalence avec le Maximum de Vraisemblance

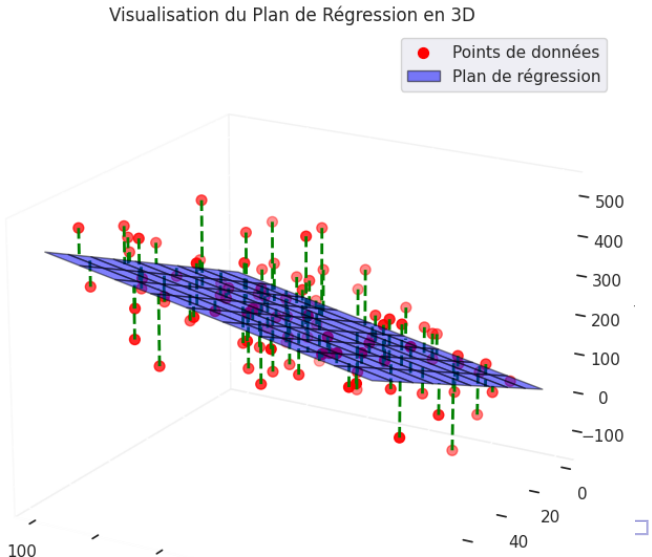
## 5 Interprétation Géométrique

- **Projection orthogonale sur le sous-espace engendré par les prédicteurs:**
- L'hyperplan en bleu représente la **régression linéaire multiple**, c'est-à-dire le sous-espace sur lequel les valeurs prédites $\hat{y}$ sont projetées ($\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$).
- Les **lignes verticales vertes pointillées** représentent les **résidus**, c'est-à-dire la différence entre les points de données réels (rouges) et leurs projections sur l'hyperplan (prédictions).
- La **minimisation des moindres carrés** ajuste cet hyperplan de manière à minimiser la somme des carrés des longueurs de ces segments.



# Solution : Équivalence avec le Maximum de Vraisemblance

- **Minimisation de la distance quadratique aux observations (MV sous hypothèse normale) :**
- Sous l'hypothèse que les erreurs suivent une loi normale $\mathcal{N}(0, \sigma^2)$, la **log-vraisemblance** est maximisée lorsque la somme des carrés des distances entre les **points rouges** (observations) et le **plan bleu** (prédictions) est minimisée.
- Comme dans l'interprétation des MCO, l'objectif du MV est également de **minimiser ces écarts quadratiques**.



# Table des Matières

1 Analyse des Résidus

2 Régression Linéaire Multiple

3 Décomposition de la Variabilité Totale

4 ANOVA en Régression Linéaire Multiple

5 Exercice : Boston Housing Prices Dataset

6 Intervalle de Confiance et Intervalle de Prédiction

7 Annexe

# Décomposition de la Variabilité Totale

# Décomposition de la Variabilité Totale : Régression Multiple

- En régression multiple, on cherche à expliquer la variabilité de $Y$ en fonction de plusieurs variables explicatives $X_1, X_2, \ldots, X_p$.
- La variabilité totale des observations $y_i$ autour de leur moyenne $\overline{y}$ est donnée par :

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

- Tout comme en régression simple, chaque observation $y_i$ peut être décomposée en :

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

où :

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

# Décomposition de la Variabilité Totale : Régression Multiple

- Comme en régression simple, on peut décomposer l'écart entre $y_i$ et $\overline{y}$ en deux termes :

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

- En sommant les carrés, on obtient la relation fondamentale :

$$SC_{totale} = SC_{reg} + SC_{res}, \quad \text{où :}$$

- $SC_{totale} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = ||\boldsymbol{y} - \bar{y}\mathbf{1}||_2^2$ est la variabilité totale. Elle représente le carré de la norme $l_2$ de $\mathbf{y} - \bar{y}\mathbf{1}$.
- $SC_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = ||\mathbf{X}\hat{\boldsymbol{\beta}} - \bar{y}\mathbf{1}||_2^2$est la variabilité expliquée par le modèle. Elle est la norme $l_2$ au carré de la projection de $\mathbf{y}$ sur le sous-espace engendré par $\mathbf{X}$.
- $SC_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = ||\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||_2^2$ est la variabilité résiduelle. Elle est la norme $l_2$ au carré des résidus $\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

# Coefficient de Détermination $R^2$ en Régression Multiple

- Le coefficient de détermination $R^2$ mesure la proportion de la variabilité expliquée :

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

- Si $R^2 \approx 1$, cela signifie que le modèle explique presque toute la variabilité.
- Si $R^2 \approx 0$, cela signifie que la régression multiple n'explique pas mieux $y$ qu'une simple moyenne.
- **Différence avec la régression simple** :
  - En régression multiple, ajouter des prédicteurs augmente $R^2$ même si ces variables sont peu informatives.
  - Pour éviter cela, on utilise souvent 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.

# Table des Matières

1 Analyse des Résidus

2 Régression Linéaire Multiple

3 Décomposition de la Variabilité Totale

4 ANOVA en Régression Linéaire Multiple

5 Exercice : Boston Housing Prices Dataset

6 Intervalle de Confiance et Intervalle de Prédiction

7 Annexe

# ANOVA en Régression Linéaire Multiple

# ANOVA en Régression Linéaire Multiple

- **L'analyse de variance (ANOVA)** en régression linéaire multiple permet d'analyser dans quelle mesure un modèle de régression avec plusieurs prédicteurs est capable d'expliquer la variabilité d'une variable dépendante $Y$.
- L'ANOVA évalue si l'ajout des $p$ variables explicatives $X_1, X_2, \ldots, X_p$ améliore significativement la prédiction de $Y$ ou si les variations de $Y$ sont principalement 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 multiple $SC_{\text{reg}}$ est significativement plus grande que la variabilité résiduelle $SC_{\text{res}}$.**
- Rappel de l'égalité fondamentale donnée par la décomposition de la variabilité totale :

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

# ANOVA (Rappel) : $\chi^2$ pour la Variance Empirique

- **Relation entre $\chi^2$ et les variables normales :**
  - La somme des carrés de $k$ variables normales standards indépendantes suit une distribution $\chi^2$ 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 Multiple

- Dans un modèle de régression linéaire Multiple :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad \text{où} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Les erreurs $\epsilon_i$ 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 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$ est la valeur ajustée.

- Puisque les $\epsilon_i$ sont indépendants et normaux, alors leur somme des carrés suit une loi du $\chi^2$ avec $n - p - 1$ degrés de liberté :

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

- le degré de liberte est $n - (p + 1)$ et non $n$ car $\hat{\boldsymbol{\beta}}$ contient $p + 1$ coefficients estimés.

# 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

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

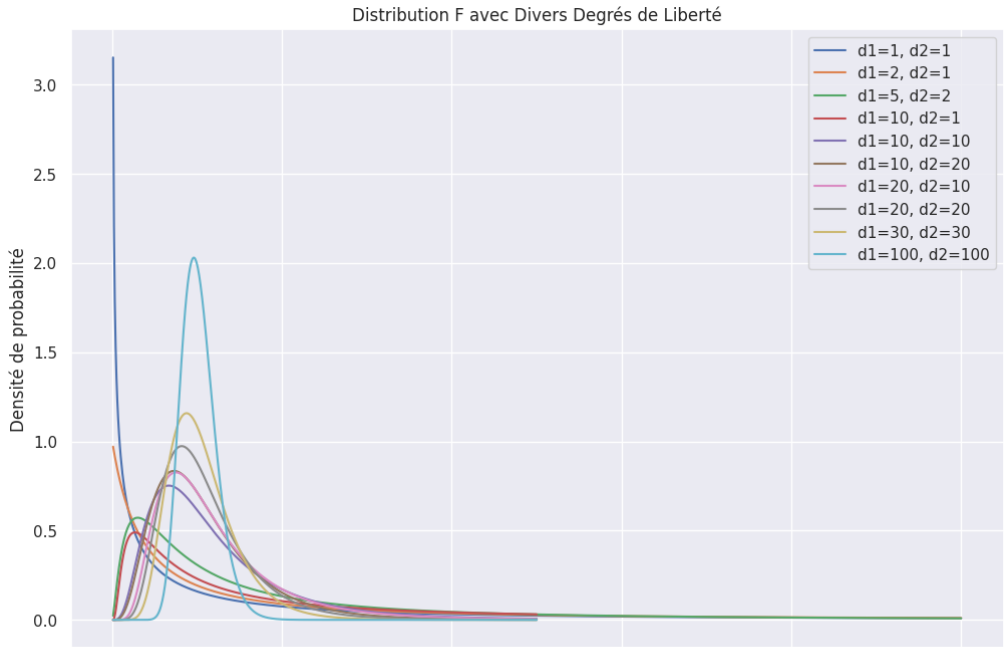
    suit une loi de Fisher avec $d_1$ et $d_2$ degrés de liberté : $F \sim F_{d_1,d_2}$.
- **Propriétés**
  - Utilisée pour modéliser le ratio de deux variances échantillonnales avec des degrés de liberté, $d_1$ et $d_2$, liés aux 2 échantillons comparés.
  - Utilisée principalement pour comparer deux variances 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 $\alpha$, 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 en Régression Linéaire Multiple

- 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 **qu'au moins 1 des variables explicatives parmi** $X_1, X_2, ..., X_p$ apportent 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 les variables $X_1, ..., X_p$ ne sont peut-être pas de bons prédicteurs de $Y$.
- L'hypothèse nulle testée est : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.
- Nous avons : $\begin{cases} \frac{SC_{\text{res}}}{\sigma^2} \sim \chi^2_{n-p-1} \\ \frac{SC_{\text{reg}}}{\sigma^2} \sim \chi^2_p \end{cases}$ sous l'hypothèse $H_0$.

# ANOVA : Test $F$ de Fisher en Régression Linéaire Multiple

- L'hypothèse nulle testée est : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.
- On a :
$$\frac{SC_{\text{res}}}{\sigma^2} \sim \chi^2_{n-p-1}, \qquad \frac{SC_{\text{reg}}}{\sigma^2} \sim \chi^2_p$$
sous l'hypothèse $H_0$.
- 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
$$\begin{cases} U \sim \chi^2_{d_1} & \text{avec } d_1 = p \text{ degrés de liberté,} \\ V \sim \chi^2_{d_2} & \text{avec } d_2 = n-p-1 \text{ degrés de liberté.} \end{cases}$$
- La statistique de test $F_{\text{stat}}$ dans le contexte de l'ANOVA en régression linéaire multiple est donnée par :
$$F_{\text{stat}} = \frac{(SC_{\text{reg}}/\sigma^2)/p}{(SC_{\text{res}}/\sigma^2)/(n-p-1)} = \frac{SC_{\text{reg}}/p}{SC_{\text{res}}/(n-p-1)} = \frac{MC_{\text{reg}}}{MC_{\text{res}}}.$$

# ANOVA en Régression Linéaire Multiple : Test $F$

- **Objectif** : Tester si au moins une des variables explicatives a un effet significatif sur $Y$.
- **Hypothèses** :
  - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (le modèle n'explique pas $Y$).
  - $H_1$ : Au moins un $\beta_j \neq 0$.
- La statistique $F$ est définie comme :

$$F_{\text{stat}} = \frac{(SC_{\text{reg}}/p)}{(SC_{\text{res}}/(n-p-1))} = \frac{MC_{\text{reg}}}{MC_{\text{res}}}$$

où :

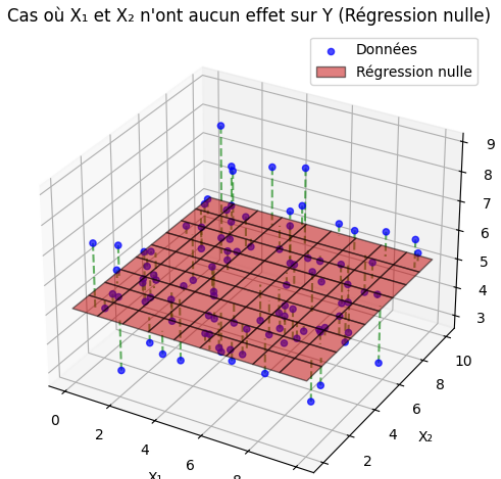
  - $SC_{\text{reg}}$ est la somme des carrés expliquée par la régression.
  - $SC_{\text{res}}$ est la somme des carrés des résidus.
- Sous $H_0$, $F_{\text{stat}}$ suit une loi de Fisher $F_{p,n-p-1}$.

# ANOVA en Régression Linéaire Multiple : Test $F$

- **Étape 1. Formulation des hypothèses :**
  - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (aucune variable explicative n'a d'effet sur $Y$, la régression ne fait pas mieux qu'une constante).
  - $H_1$ : Au moins un $\beta_j \neq 0$ (au moins une variable $X_j$ contribue significativement à la prédiction de $Y$).
- **Étape 2. Calcul de la statistique de test :**
  $F_{\text{stat}} = \frac{(SC_{\text{reg}}/p)}{(SC_{\text{res}}/(n-p-1))}$, où :
  - $SC_{\text{reg}}$ est la somme des carrés expliquée par la régression.
  - $SC_{\text{res}}$ est la somme des carrés des résidus.
- **Étape 3. Degrés de liberté :** $d_1 = p, \quad d_2 = n - p - 1$, où :
  - $d_1 = p$ correspond au nombre de variables explicatives.
  - $d_2 = n - p - 1$ correspond au degré de liberté des résidus (car $p + 1$ paramètres sont estimés, incluant $\beta_0$).
- **Étape 4. Décision :**
  - Sous $H_0$, $F_{\text{stat}}$ suit une loi de Fisher $F_{p,n-p-1}$.
  - Trouver la ou les valeurs critiques à un niveau de signification donné $\alpha$ ou calculer la $p$-valeur. **Rejeter $H_0$ si $F_{\text{stat}}$ grand**: p-valeur $< \alpha$.

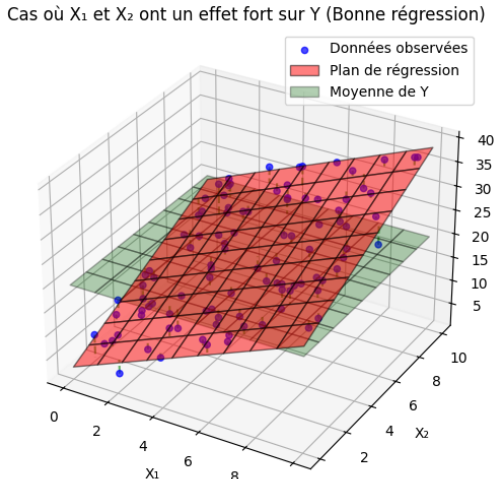
# ANOVA : Test $F$ de Fisher

- **Cas où $X_1, X_2, \ldots, X_p$ n'ont aucun effet sur $Y$ (rég nulle)** :
  - L'hyperplan de régression est **horizontal**, ce qui signifie que toutes les variables explicatives $X_j$ n'ont aucun effet.
  - La variabilité expliquée par la régression est proche de zéro $(SC_{reg} \approx 0)$.
  - La variabilité résiduelle est équivalente à la variabilité totale $(SC_{res} \approx SC_{totale})$.
  - La $F_{stat}$ est proche de 1 : $F \approx 1$ ce qui signifie que le modèle de régression multiple ne fait pas mieux qu'une simple moyenne $(\hat{Y} = \beta_0)$. On ne rejette donc pas $H_0$.



# ANOVA : Test $F$ de Fisher

- **Cas où $X_1, X_2, \ldots, X_p$ ont un effet fort sur $Y$ (bonne rég)** :
  - L'hyperplan de régression suit la tendance des données, ce qui signifie que les variables explicatives $X_j$ ont un impact significatif sur $Y$.
  - La variabilité expliquée par la régression est grande $(SC_{reg} \gg 0)$.
  - La variabilité résiduelle est faible comparée à la variabilité totale $(SC_{res} \ll SC_{totale})$.
  - La $F_{stat}$ est élevée : $F \gg 1$ ce qui signifie que le modèle de régression multiple explique une part importante de la variabilité de $Y$. On **rejette** donc $H_0$, indiquant que les variables explicatives $X_1, X_2, \ldots, X_p$ contribuent significativement au modèle.



# Tableau de l'ANOVA pour la Régression Multiple

## Tableau de l'ANOVA en Régression Multiple

| Source | Somme des Carrés | Degrés de Liberté (dl) | Moyennes Carrées | Statistique $F$ |
|---|---|---|---|---|
| Régression | $SC_{\text{reg}}$ | $p$ | $MC_{\text{reg}} = \frac{SC_{\text{reg}}}{p}$ | $F = \frac{MC_{\text{reg}}}{MC_{\text{res}}}$ |
| Résiduel | $SC_{\text{res}}$ | $n-p-1$ | $MC_{\text{res}} = \frac{SC_{\text{res}}}{n-p-1}$ | |
| Total | $SC_{\text{totale}}$ | $n-1$ | | |

- La $F_{stat} = \frac{MC_{\text{reg}}}{MC_{\text{res}}}$ (noté simplement par $F$) permet de tester si l'ensemble des variables explicatives a un effet significatif sur $Y$.
- Si $F$ est grand, cela signifie que la variabilité expliquée $SC_{\text{reg}}$ est beaucoup plus grande que la variabilité résiduelle $SC_{\text{res}}$, indiquant un bon ajustement du modèle.

# ANOVA : Test $F$ de Fisher (Résumé)

## Tableau de l'ANOVA pour la régression linéaire multiple

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

- L'ANOVA en régression linéaire multiple sert à **mesurer combien de la variabilité totale peut être attribuée** aux $p$ variables explicatives $X_1, X_2, \ldots, X_p$.
- Elle permet de tester si au moins une des variables $X_j$ améliore significativement la prédiction de $Y$ par rapport à un modèle trivial (constante seulement).
- Le test $F$ compare la variabilité expliquée et la variabilité résiduelle.
- L'hypothèse nulle testée est : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.
- Si $F_{stat}$ est suffisamment grand, on rejette $H_0$ et on conclut qu'au moins une des variables $X_j$ a un effet significatif sur $Y$.

# Table des Matières

1 Analyse des Résidus
2 Régression Linéaire Multiple
3 Décomposition de la Variabilité Totale
4 ANOVA en Régression Linéaire Multiple
5 Exercice : Boston Housing Prices Dataset
6 Intervalle de Confiance et Intervalle de Prédiction
7 Annexe

# Boston Housing Prices Dataset

# Exercice : Boston Housing Prices Dataset

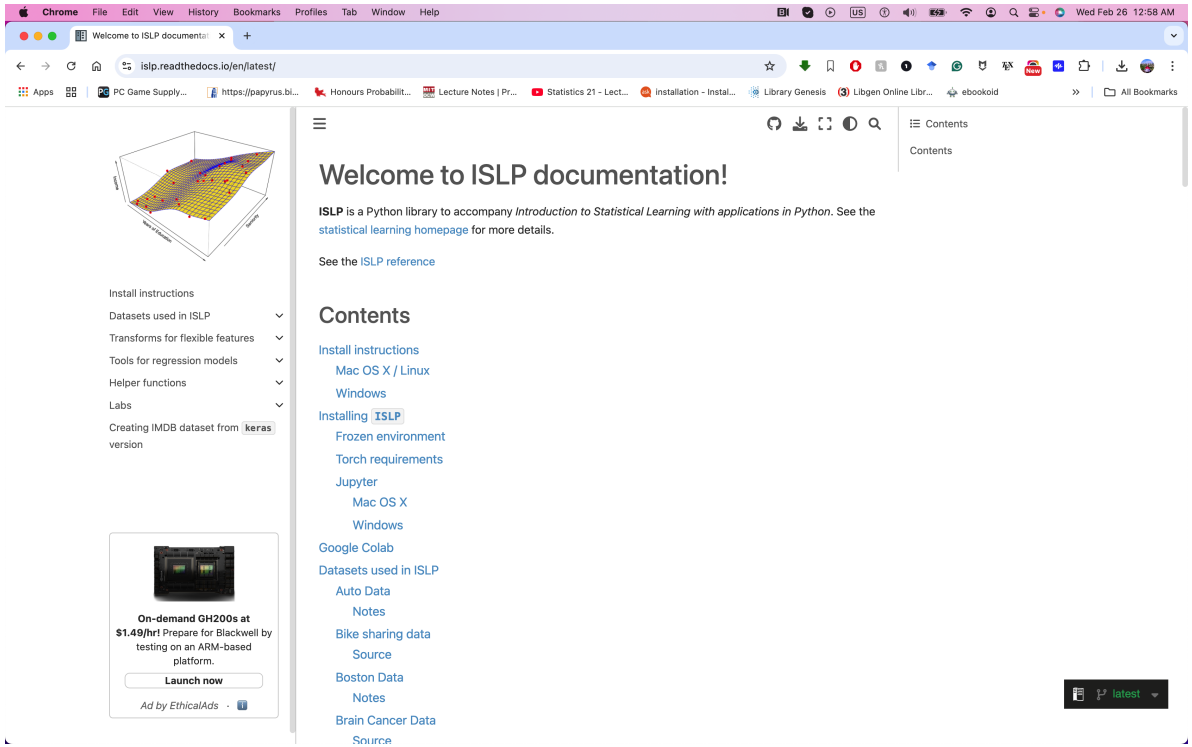
- Vous allez charger et explorer un jeu de données réel.
- Implémenter un modèle de régression linéaire multiple avec 3 manières différentes :
  1. Moindres Carrés Ordinaires (MCO) avec la formule analytique.
  2. Utilisation de `LinearRegression` de `sklearn`.
  3. Utilisation de `OLS` de `statsmodels.api`.
- Comparer les résultats en termes de :
  - Coefficients $\hat{\boldsymbol{\beta}}$.
  - Coefficient $R^2$.
  - Tableau ANOVA.

Introduction à la Librairie ISLP
Partie 1 : Chargement et Exploration des Données
Partie 2 : Régression Multiple avec 3 Méthodes
Partie 3 : Calcul du Tableau ANOVA
Partie 4 : Comparaison des 3 Méthodes

# Introduction à la Librairie ISLP

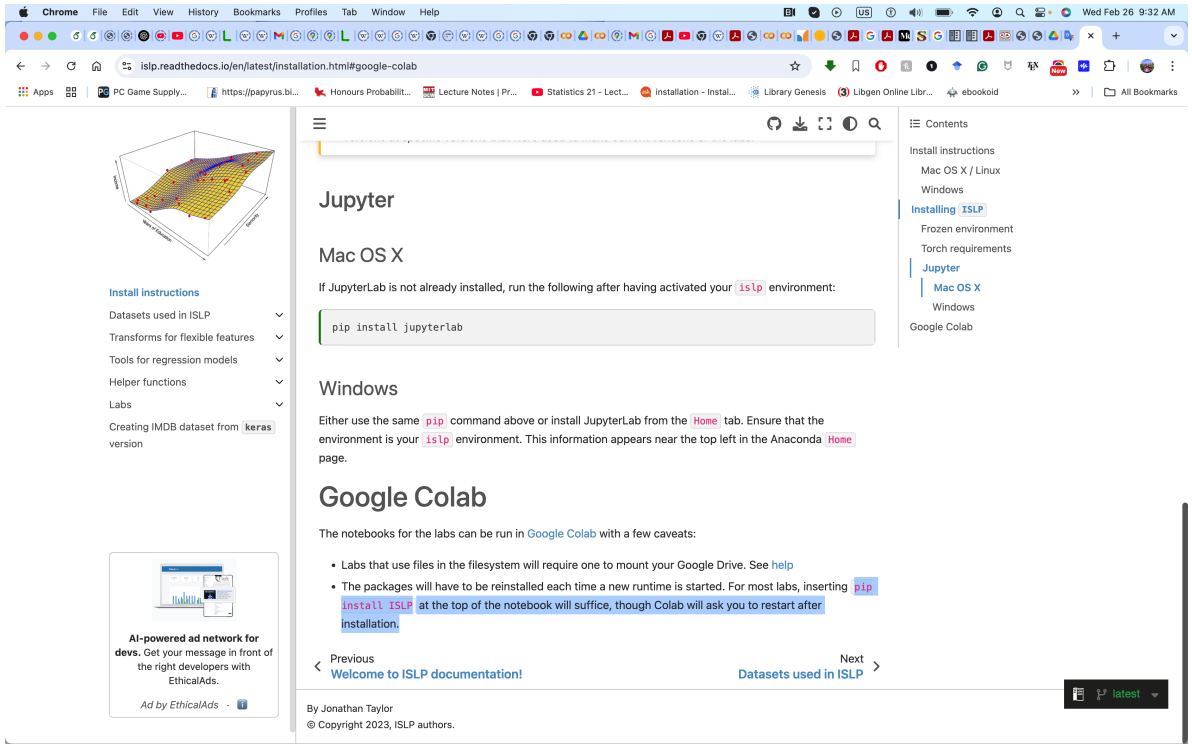
# Exercice : Boston Housing Prices Dataset

- Vous allez utilisez la librairie `ISLP`.
- `ISLP` est une librarie Python qui accompagne le manuel *Introduction to Statistical Learning*.



# Exercice : Boston Housing Prices Dataset

- Vous allez tout d'abord installer la librairie sur Colab avec la commande :
  - `!pip install ISLP`



# Partie 1 : Chargement et Exploration des Données

Introduction à la Librairie ISLP
Partie 1 : Chargement et Exploration des Données
Partie 2 : Régression Multiple avec 3 Méthodes
Partie 3 : Calcul du Tableau ANOVA
Partie 4 : Comparaison des 3 Méthodes

# Partie 1 : Chargement et Exploration des Données

- Vous aller charger les données du jeu **Boston Housing** de la bibliothèque `ISLP`
- Observer les premières lignes des données et identifier les variables explicatives et la variable cible.
- Vérifier s'il y a des valeurs manquantes.

# Partie 1 : Chargement et Exploration des Données

```python
from ISLP import load_data
Boston = load_data('Boston')
print("_____________________________________________________________________________")
print("Affiche les premières lignes du dataset :")
# Afficher les premières lignes du dataset
print(Boston.head())
print("_____________________________________________________________________________")
print("Affiche les noms des colonnes :")
print(Boston.columns)  # Affiche les noms des colonnes
print("_____________________________________________________________________________")
print("Affiche le nombre de lignes et de colonnes :")
print(Boston.shape)  # Affiche le nombre de lignes et de colonnes
print("_____________________________________________________________________________")
print("Vérifie les types des données :")
print(Boston.dtypes)   # Vérifie les types des données
print("_____________________________________________________________________________")
print("Affiche les statistiques descriptives :")
print(Boston.describe())  # Affiche les statistiques descriptives
print("_____________________________________________________________________________")
print("Vérifie s'il y a des valeurs manquantes :")
print(Boston.isnull().sum())  # Vérifie s'il y a des valeurs manquantes
print("_____________________________________________________________________________")
Boston = Boston.dropna()  # Supprime les lignes avec des valeurs manquantes
Boston = Boston.fillna(Boston.mean())  # Remplace les valeurs manquantes par la
↪  moyenne
```

# Partie 1 : Chargement et Exploration des Données

```
--------------------------------------------------------------------------------
Affiche les premières lignes du dataset :
      crim    zn  indus  chas    nox     rm   age     dis  rad  tax  ptratio  \
0  0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296     15.3
1  0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242     17.8
2  0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242     17.8
3  0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222     18.7
4  0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222     18.7

   lstat  medv
0   4.98  24.0
1   9.14  21.6
2   4.03  34.7
3   2.94  33.4
4   5.33  36.2
--------------------------------------------------------------------------------
Affiche les noms des colonnes :
Index(['crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad', 'tax',
       'ptratio', 'lstat', 'medv'],
      dtype='object')
--------------------------------------------------------------------------------
Affiche le nombre de lignes et de colonnes :
(506, 13)
--------------------------------------------------------------------------------
```

# Partie 1 : Chargement et Exploration des Données

```
Vérifie les types des données :
crim       float64
zn         float64
indus      float64
chas         int64
nox        float64
rm         float64
age        float64
dis        float64
rad          int64
tax          int64
ptratio    float64
lstat      float64
medv       float64
dtype: object
--------------------------------------------------------------------------------
```

# Partie 1 : Chargement et Exploration des Données

Affiche les statistiques descriptives :

| | crim | zn | indus | chas | nox | rm | \ |
|---|---|---|---|---|---|---|---|
| count | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | |
| mean | 3.613524 | 11.363636 | 11.136779 | 0.069170 | 0.554695 | 6.284634 | |
| std | 8.601545 | 23.322453 | 6.860353 | 0.253994 | 0.115878 | 0.702617 | |
| min | 0.006320 | 0.000000 | 0.460000 | 0.000000 | 0.385000 | 3.561000 | |
| 25% | 0.082045 | 0.000000 | 5.190000 | 0.000000 | 0.449000 | 5.885500 | |
| 50% | 0.256510 | 0.000000 | 9.690000 | 0.000000 | 0.538000 | 6.208500 | |
| 75% | 3.677083 | 12.500000 | 18.100000 | 0.000000 | 0.624000 | 6.623500 | |
| max | 88.976200 | 100.000000 | 27.740000 | 1.000000 | 0.871000 | 8.780000 | |

| | age | dis | rad | tax | ptratio | lstat | \ |
|---|---|---|---|---|---|---|---|
| count | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | |
| mean | 68.574901 | 3.795043 | 9.549407 | 408.237154 | 18.455534 | 12.653063 | |
| std | 28.148861 | 2.105710 | 8.707259 | 168.537116 | 2.164946 | 7.141062 | |
| min | 2.900000 | 1.129600 | 1.000000 | 187.000000 | 12.600000 | 1.730000 | |
| 25% | 45.025000 | 2.100175 | 4.000000 | 279.000000 | 17.400000 | 6.950000 | |
| 50% | 77.500000 | 3.207450 | 5.000000 | 330.000000 | 19.050000 | 11.360000 | |
| 75% | 94.075000 | 5.188425 | 24.000000 | 666.000000 | 20.200000 | 16.955000 | |
| max | 100.000000 | 12.126500 | 24.000000 | 711.000000 | 22.000000 | 37.970000 | |

# Partie 1 : Chargement et Exploration des Données

```
              medv
count  506.000000
mean    22.532806
std      9.197104
min      5.000000
25%     17.025000
50%     21.200000
75%     25.000000
max     50.000000

--------------------------------------------------------------------------------
Vérifie s'il y a des valeurs manquantes :
crim       0
zn         0
indus      0
chas       0
nox        0
rm         0
age        0
dis        0
rad        0
tax        0
ptratio    0
lstat      0
medv       0
dtype: int64
```

# Partie 1 : Chargement et Exploration des Données

- Vous pouvez enregistrer le dataset `Boston Housing` au format `CSV` et le sauvegarder dans votre **Google Drive**.
- Cela permet de réutiliser les données sans devoir à installer `ISLP` à chaque session.

```
import numpy as np
import pandas as pd
import os
from google.colab import drive
from ISLP import load_data

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

# Étape 2 : Charger le dataset Boston Housing
Boston = load_data('Boston')

# Étape 3 : Définir le chemin de sauvegarde dans Google Drive
folder_name = "Colab_Data"  # Dossier cible dans Google Drive
save_path = f"/content/drive/MyDrive/{folder_name}"  # Modifiez selon votre structure Drive

# Créer le dossier s'il n'existe pas
os.makedirs(save_path, exist_ok=True)
```

# Partie 1 : Chargement et Exploration des Données

```
# Étape 4 : Sauvegarder le dataset en CSV
file_path = os.path.join(save_path, "BostonHousing.csv")
Boston.to_csv(file_path, index=False)

# Étape 5 : Confirmation de la sauvegarde
print(f"Données sauvegardées avec succès : {file_path}")

# Afficher les premières lignes pour vérification
print("Afficher les premières lignes pour vérification :")
print(Boston.head())

# Vérifier si le fichier existe bien dans Google Drive
print("Vérifier si le fichier existe bien dans Google Drive :")
!ls "{save_path}"
```

# Partie 1 : Chargement et Exploration des Données

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount('
Données sauvegardées avec succès : /content/drive/MyDrive/Colab_Data/BostonHousing.csv
Afficher les premières lignes pour vérification :
      crim    zn  indus  chas    nox     rm   age     dis  rad  tax  ptratio  \
0  0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296     15.3
1  0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242     17.8
2  0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242     17.8
3  0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222     18.7
4  0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222     18.7

   lstat  medv
0   4.98  24.0
1   9.14  21.6
2   4.03  34.7
3   2.94  33.4
4   5.33  36.2
Vérifier si le le fichier existe bien dans Google Drive :
BostonHousing.csv
```

# Partie 1 : Chargement et Exploration des Données

- Si vous avez préalablement sauvegardé le dataset `BostonHousing.csv` dans votre **Google Drive**, vous pouvez le recharger sans devoir utiliser `ISLP`.

```
import pandas as pd
import os
from google.colab import drive
# Étape 1 : Monter Google Drive
drive.mount('/content/drive')
# Étape 2 : Définir le chemin du fichier sauvegardé
folder_name = "Colab_Data"  # Dossier où le fichier a été sauvegardé
file_path = f"/content/drive/MyDrive/{folder_name}/BostonHousing.csv"  # Chemin complet
# Étape 3 : Vérifier si le fichier existe
if os.path.exists(file_path):
    # Charger le fichier CSV
    Boston_loaded = pd.read_csv(file_path)
    print("Fichier chargé avec succès !")
    # Afficher les premières lignes du dataset
    print(Boston_loaded.head())
    # Vérifier les informations du dataset
    print("\nInformations sur le dataset :")
    print(Boston_loaded.info())
else:
    print(f"Le fichier {file_path} n'existe pas. Vérifiez le chemin !")
```

# Partie 1 : Chargement et Exploration des Données

```python
# Sélection des variables explicatives et cible
X = Boston_loaded.drop(columns=['medv'])  # Variables explicatives
Y = Boston_loaded['medv'].values  # Variable cible
print("Variables explicatives :")
print(X)
```

Résultats:

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount('
Fichier chargé avec succès !
      crim    zn  indus  chas    nox     rm   age     dis  rad  tax  ptratio  \
0  0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296     15.3
1  0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242     17.8
2  0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242     17.8
3  0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222     18.7
4  0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222     18.7
   lstat  medv
0   4.98  24.0
1   9.14  21.6
2   4.03  34.7
3   2.94  33.4
4   5.33  36.2
Informations sur le dataset :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
```

# Partie 1 : Chargement et Exploration des Données

```
Data columns (total 13 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   crim     506 non-null    float64
 1   zn       506 non-null    float64
 2   indus    506 non-null    float64
 3   chas     506 non-null    int64
 4   nox      506 non-null    float64
 5   rm       506 non-null    float64
 6   age      506 non-null    float64
 7   dis      506 non-null    float64
 8   rad      506 non-null    int64
 9   tax      506 non-null    int64
 10  ptratio  506 non-null    float64
 11  lstat    506 non-null    float64
 12  medv     506 non-null    float64
dtypes: float64(10), int64(3)
memory usage: 51.5 KB
None
```

# Partie 1 : Chargement et Exploration des Données

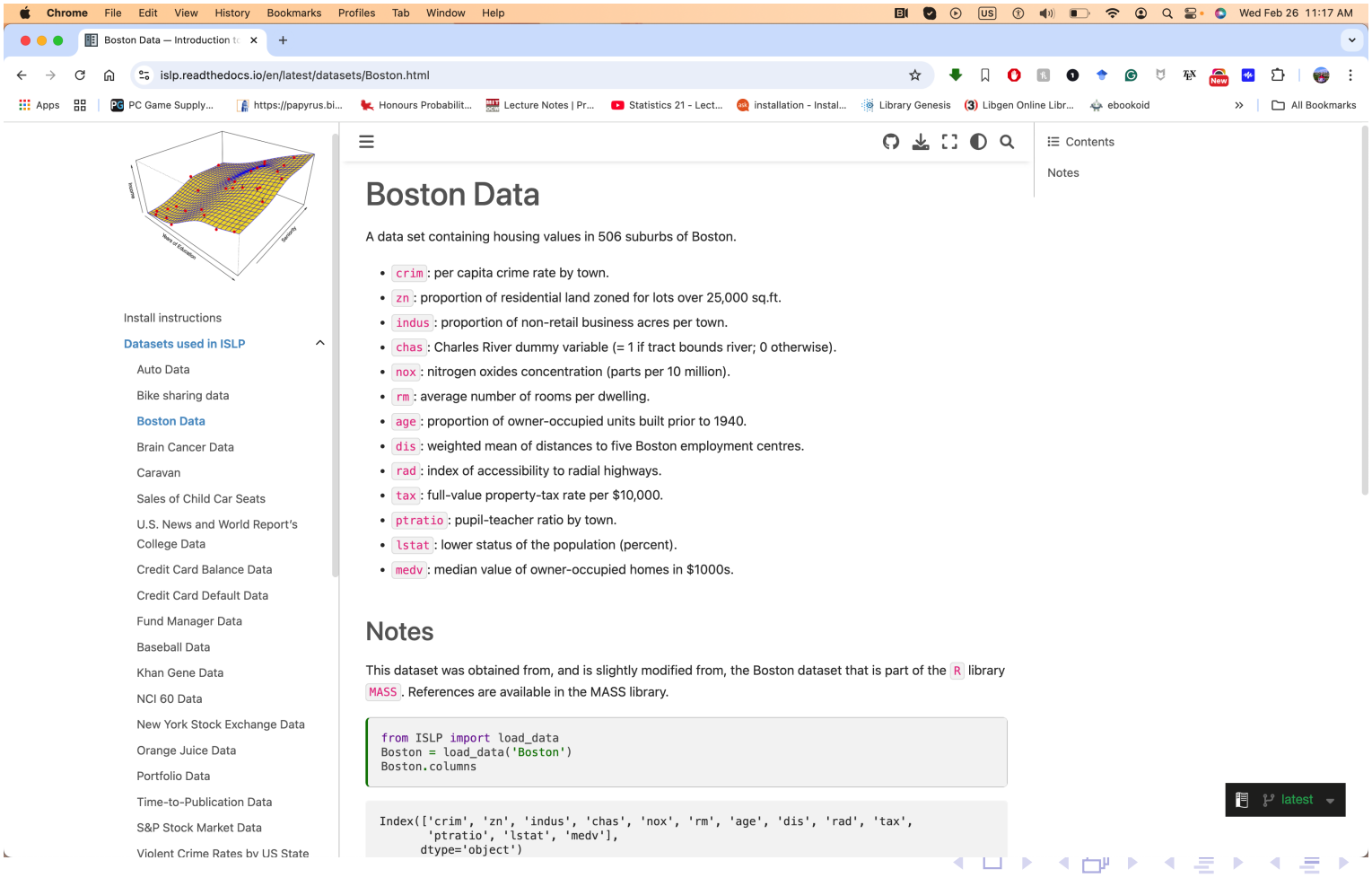
variables explicatives :

| | crim | zn | indus | chas | nox | rm | age | dis | rad | tax | \ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00632 | 18.0 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.0900 | 1 | 296 | |
| 1 | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | |
| 2 | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | |
| 3 | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | |
| 4 | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | |
| .. | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 501 | 0.06263 | 0.0 | 11.93 | 0 | 0.573 | 6.593 | 69.1 | 2.4786 | 1 | 273 | |
| 502 | 0.04527 | 0.0 | 11.93 | 0 | 0.573 | 6.120 | 76.7 | 2.2875 | 1 | 273 | |
| 503 | 0.06076 | 0.0 | 11.93 | 0 | 0.573 | 6.976 | 91.0 | 2.1675 | 1 | 273 | |
| 504 | 0.10959 | 0.0 | 11.93 | 0 | 0.573 | 6.794 | 89.3 | 2.3889 | 1 | 273 | |
| 505 | 0.04741 | 0.0 | 11.93 | 0 | 0.573 | 6.030 | 80.8 | 2.5050 | 1 | 273 | |

| | ptratio | lstat |
|---|---|---|
| 0 | 15.3 | 4.98 |
| 1 | 17.8 | 9.14 |
| 2 | 17.8 | 4.03 |
| 3 | 18.7 | 2.94 |
| 4 | 18.7 | 5.33 |
| .. | ... | ... |
| 501 | 21.0 | 9.67 |
| 502 | 21.0 | 9.08 |
| 503 | 21.0 | 5.64 |
| 504 | 21.0 | 6.48 |
| 505 | 21.0 | 7.88 |

# Partie 1 : Chargement et Exploration des Données

- **Colonnes du Dataset Boston Housing**



# Partie 1 : Chargement et Exploration des Données

- **Colonnes du Dataset Boston Housing**
  - **crim** : Taux de criminalité par zone.
  - **zn** : Proportion de terrains résidentiels pour des lots $>$ 25,000 sq.ft.
  - **indus** : Proportion de terrains non résidentiels par ville.
  - **chas** : Variable binaire (1 = proximité de la rivière Charles).
  - **nox** : Concentration de $NO_2$ (pollution de l'air).
  - **rm** : Nombre moyen de pièces par logement.
  - **age** : Proportion des logements construits avant 1940.
  - **dis** : Distance moyenne aux centres d'emploi de Boston.
  - **rad** : Accessibilité aux autoroutes radiales.
  - **tax** : Taux d'imposition foncière par \$10,000.
  - **ptratio** : Ratio élèves/enseignants par ville.
  - **lstat** : Pourcentage de la population à faible statut socio-économique.
  - **medv** : Valeur médiane des logements (en milliers de dollars).

# Partie 2 : Régression Multiple avec 3 Méthodes d'Implémentation

# Partie 2 : Régression Multiple avec 3 Méthodes

- **Objectif :** Prédire la valeur médiane des logements (**medv**) en fonction des caractéristiques des quartiers et des maisons.
- **Modèle de Régression Multiple :**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \text{où :}$$

  - $Y$ est la valeur médiane des logements (**medv**).
  - $X_1, X_2, \ldots, X_p$ sont les variables explicatives (**crim, zn,** $\cdots$**, lstat**).
  - $\beta_0$ est l'intercept, et $\beta_1, \beta_2, ..., \beta_p$ sont les coefficients de régression.
  - $\epsilon$ représente l'erreur aléatoire.
- **Problématique :**
  - Quelles sont les variables qui influencent le plus le prix des maisons ?
  - Peut-on expliquer la variabilité des prix avec seulement ces facteurs ?
  - Quelle est la qualité de notre modèle? ($R^2$, test $F$, p-valeurs)

# Partie 2 : Régression Multiple avec 3 Méthodes

**Questions :** Implémenter un modèle de régression multiple sur le dataset Boston Housing en utilisant trois approches :

- **1-a.** Utiliser laformule analytique des moindres carrés ordinaires (MCO) pour estimer les coefficients :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{X}})^{-1}\mathbf{X}^{\top}\boldsymbol{y}$$

- **1-b.** Calculer les **prédictions** de $Y$ en utilisant les coefficients estimés.
- **2.** Utiliser `LinearRegression` de `sklearn` pour **entraîner un modèle** et faire des **prédictions**.
- **3.** Utiliser `OLS` de `statsmodels` pour **ajuster un modèle de régression** et obtenir les coefficients.

# Partie 2 : Régression Multiple avec 3 Méthodes

```
import numpy as np
import pandas as pd
from scipy.stats import f
from tabulate import tabulate
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn.linear_model import LinearRegression
from ISLP import load_data
from sklearn.metrics import r2_score
# Sélection des variables explicatives et cible
X = Boston_loaded.drop(columns=['medv'])  # Variables explicatives
Y = Boston_loaded['medv'].values  # Variable cible
# Ajouter une colonne d'intercept manuellement
X_with_intercept = np.c_[np.ones(X.shape[0]), X]
# 1. Estimation des paramètres avec la Formule des Moindres Carrés Ordinaires (MCO)
beta_mco = np.linalg.inv(X_with_intercept.T @ X_with_intercept) @ X_with_intercept.T @ Y
Y_pred_mco = X_with_intercept @ beta_mco
# 2. Estimation des paramètres avec `LinearRegression` de `sklearn`
model_sklearn = LinearRegression()
model_sklearn.fit(X, Y)
Y_pred_sklearn = model_sklearn.predict(X)
# 3. Estimation des paramètres avec `OLS` de `statsmodels.api`
formula = "medv ~ " + " + ".join(X.columns)  # Formule pour statsmodels
model_sm = smf.ols(formula, data=Boston_loaded).fit()
Y_pred_sm = model_sm.fittedvalues
```

# Partie 3 : Calcul du Tableau ANOVA

# Partie 3 : Calcul du Tableau ANOVA

- **Question :** Implémenter une fonction qui calcule et affiche le tableau ANOVA pour un modèle de régression multiple.
  - La fonction doit prendre en entrée :
    - Les valeurs observées de la variable cible $\boldsymbol{y}$.
    - Les valeurs prédites $\hat{\boldsymbol{y}}$.
    - La taille de l'échantillon $n$.
    - Le nombre de variables explicatives $p$.
  - La fonction doit calculer :
    - **Sommes des Carrés :**

$$SC_{totale} = \sum (y_i - \bar{y})^2, \quad SC_{reg} = \sum (\hat{y}_i - \bar{y})^2, \quad SC_{res} = \sum (y_i - \hat{y}_i)^2$$

    - **Degrés de Liberté :** $\text{dl}_{reg} = p, \quad \text{dl}_{res} = n - p - 1$
    - **Moyennes des Carrés :** $MC_{reg} = \frac{SC_{reg}}{p}, \quad MC_{res} = \frac{SC_{res}}{n-p-1}$
    - **Statistique de test F :** $F_{stat} = \frac{MC_{reg}}{MC_{res}}$
    - **p-valeur associée** en utilisant la loi de Fisher.
  - **Affichage du tableau ANOVA** avec une colonne supplémentaire pour la p-valeur.

# Partie 3 : Calcul du Tableau ANOVA

```python
def compute_anova(Y, Y_pred, n, p):
    Y_mean = np.mean(Y)
    # Sommes des Carrés
    SC_tot = np.sum((Y - Y_mean) ** 2)
    SC_reg = np.sum((Y_pred - Y_mean) ** 2)
    SC_res = np.sum((Y - Y_pred) ** 2)
    # Degrés de liberté
    df_reg = p
    df_res = n - p - 1
    # Moyennes des Carrés
    MC_reg = SC_reg / df_reg
    MC_res = SC_res / df_res
    F_stat = MC_reg / MC_res # Statistique de test F
    p_value = 1 - f.cdf(F_stat, df_reg, df_res) # p-valeur associée
    # Création du tableau ANOVA
    anova_table = pd.DataFrame({
        "Source": ["Régression", "Résiduel", "Total"],
        "Somme des Carrés (SC)": [SC_reg, SC_res, SC_tot],
        "Degrés de Liberté (dl)": [df_reg, df_res, n - 1],
        "Moyenne des Carrés (MC)": [MC_reg, MC_res, ""],
        "F_stat": [F_stat, "", ""],
        "p-value": [p_value, "", ""]
    })
    return anova_table
```

# Partie 4 : Comparaison des 3 Méthodes d'Implémentation

# Partie 4 : Comparaison des 3 Méthodes

**Question :** Comparer les résultats obtenus avec les 3 méthodes d'implémentation en termes de :

- Coefficients estimés $\hat{\boldsymbol{\beta}}$.
- Coefficient de détermination $R^2$.
- Tableau ANOVA.

**Tâches demandées aux étudiants :**

- Créer un **DataFrame** pour comparer les coefficients obtenus par :
  - La formule analytique des **moindres carrés ordinaires (MCO)**.
  - La classe `LinearRegression` de `sklearn`.
  - La méthode `OLS` de `statsmodels`.
- Calculer et afficher $R^2$ pour chacune des trois méthodes.
- Générer et afficher le tableau **ANOVA** pour chaque méthode en utilisant la librairie `tabulate` Avec l'aimable autorisation de votre camarade Adam Hsaine :).

# Partie 4 : Comparaison des 3 Méthodes

```python
# obtention des coefficients pour chacune des 3 méthodes
coefficients_sm = model_sm.params
coefficients_sklearn = np.append(model_sklearn.intercept_, model_sklearn.coef_)
coefficients_mco = beta_mco.flatten()

# Créer un DataFrame pour comparer les résultats
comparison_df = pd.DataFrame({
    'Statsmodels': coefficients_sm,
    'Sklearn': coefficients_sklearn,
    'Moindres Carrés': coefficients_mco
}, index=['Intercept'] + list(Boston_loaded.drop(columns=['medv']).columns))
print("\nComparaison des coefficients entre Statsmodels, Sklearn et Moindres Carrés :")
print(comparison_df)

# Vérification de R²
Y_pred_mco = X_with_intercept @ beta_mco
r2_mco = r2_score(Y, Y_pred_mco)
print("\nR² obtenu par chaque méthode :")
print("Statsmodels R² :", model_sm.rsquared)
print("Sklearn R² :", r2_score(Y, model_sklearn.predict(X)))
print("Moindres Carrés R² :", r2_mco)
```

# Partie 4 : Comparaison des 3 Méthodes

```
# Calcul des tableaux ANOVA
anova_mco = compute_anova(Y, Y_pred_mco, X.shape[0], X.shape[1])
anova_sklearn = compute_anova(Y, Y_pred_sklearn, X.shape[0], X.shape[1])
anova_sm = sm.stats.anova_lm(model_sm, typ=1)

# Affichage des résultats sous forme de tableau
# Avec l'aimable autorisation de votre camarade Adam Hsaine :)
print("\nTableau ANOVA avec la formule des Moindres Carrés Ordinaires (MCO)")
print(tabulate(anova_mco, headers="keys", tablefmt="grid", showindex=False, floatfmt=".4f")
print("\nTableau ANOVA avec `LinearRegression` de sklearn")
print(tabulate(anova_sklearn, headers="keys", tablefmt="grid", showindex=False, floatfmt="
print("\nTableau ANOVA avec `OLS` de statsmodels")
print(tabulate(anova_sm, headers="keys", tablefmt="grid", floatfmt=".4f"))
```

# Partie 4 : Comparaison des 3 Méthodes

Comparaison des coefficients entre Statsmodels, Sklearn et Moindres Carrés :

| | Statsmodels | Sklearn | Moindres Carrés |
|---|---|---|---|
| Intercept | 41.617270 | 41.617270 | 41.617270 |
| crim | -0.121389 | -0.121389 | -0.121389 |
| zn | 0.046963 | 0.046963 | 0.046963 |
| indus | 0.013468 | 0.013468 | 0.013468 |
| chas | 2.839993 | 2.839993 | 2.839993 |
| nox | -18.758022 | -18.758022 | -18.758022 |
| rm | 3.658119 | 3.658119 | 3.658119 |
| age | 0.003611 | 0.003611 | 0.003611 |
| dis | -1.490754 | -1.490754 | -1.490754 |
| rad | 0.289405 | 0.289405 | 0.289405 |
| tax | -0.012682 | -0.012682 | -0.012682 |
| ptratio | -0.937533 | -0.937533 | -0.937533 |
| lstat | -0.552019 | -0.552019 | -0.552019 |

$R^2$ obtenu par chaque méthode :
Statsmodels $R^2$ : 0.7343070437613076
Sklearn $R^2$ : 0.7343070437613076
Moindres Carrés $R^2$ : 0.7343070437613076

# Partie 4 : Comparaison des 3 Méthodes

- **1. Comparaison des coefficients estimés**
  - Les trois méthodes (**formule analytique des Moindres Carrés Ordinaires (MCO)**, `LinearRegression` de `sklearn` et `OLS` de `statsmodels`) donnent exactement les mêmes coefficients pour chaque variable explicative.
  - Cela confirme que ces trois approches utilisent la même estimation basée sur la **minimisation de la somme des carrés des erreurs**.
  - Les coefficients interprètent l'effet marginal de chaque variable explicative sur la variable cible **MEDV** (valeur médiane des logements en milliers de dollars), toutes choses égales par ailleurs.

# Partie 4 : Comparaison des 3 Méthodes

- **2. Comparaison du coefficient de détermination $R^2$**
  - Le coefficient $R^2$ est identique pour les trois méthodes et vaut **0.7343**.
  - Cela signifie que **73.43% de la variabilité de la valeur médiane des logements est expliquée par le modèle**.
  - Un $R^2$ de 0.7343 est assez élevé, indiquant que le modèle de régression multiple est bien ajusté sur les donnés observés. Il y a une bonne partie de la variabilité de **MEDV** qui est expliqué par le modèle de régression.

# Partie 4 : Comparaison des 3 Méthodes

- **3. Interprétation des coefficients**
  - $\hat{\beta}_j$ représente l'effet d'une unité d'augmentation de la variable $X_j$ sur la valeur médiane des logements, en supposant que toutes les autres variables sont constantes.
  - **Exemples** :
    - **rm** (nombre moyen de pièces par logement) a un coefficient **positif** (**3.658**), ce qui signifie que **plus une maison a de pièces, plus sa valeur médiane tend à augmenter**.
    - **nox** (pollution $NO_2$) a un coefficient **négatif** (**-18.758**), indiquant qu'**une augmentation de la pollution est associée à une baisse des prix des logements**.
    - **lstat** (pourcentage de population à faible statut socio-économique) a aussi un **effet négatif** (**-0.552**), ce qui signifie que **plus le quartier a une proportion élevée de population défavorisée, plus la valeur médiane des logements baisse**.

# Partie 4 : Comparaison des 3 Méthodes

Tableau ANOVA avec la formule des Moindres Carrés Ordinaires (MCO)

| Source | Somme des Carrés (SC) | Degrés de Liberté (dl) | Moyenne des Carrés (MC) | F_stat | p-value |
|---|---|---|---|---|---|
| Régression | 31366.8766 | 12 | 2613.9063838863754 | 113.543774268359 | 1.1102230246251565e-16 |
| Résiduel | 11349.4188 | 493 | 23.021133485561673 | | |
| Total | 42716.2954 | 505 | | | |

Tableau ANOVA avec `LinearRegression` de sklearn

| Source | Somme des Carrés (SC) | Degrés de Liberté (dl) | Moyenne des Carrés (MC) | F_stat | p-value |
|---|---|---|---|---|---|
| Régression | 31366.8766 | 12 | 2613.9063838864613 | 113.54377426836275 | 1.1102230246251565e-16 |
| Résiduel | 11349.4188 | 493 | 23.02113348556167 | | |
| Total | 42716.2954 | 505 | | | |

Tableau ANOVA avec `OLS` de statsmodels

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| crim | 1.0000 | 6440.7831 | 6440.7831 | 279.7770 | 0.0000 |
| zn | 1.0000 | 3554.3362 | 3554.3362 | 154.3945 | 0.0000 |
| indus | 1.0000 | 2551.2364 | 2551.2364 | 110.8215 | 0.0000 |
| chas | 1.0000 | 1529.8479 | 1529.8479 | 66.4541 | 0.0000 |
| nox | 1.0000 | 76.2476 | 76.2476 | 3.3121 | 0.0694 |
| rm | 1.0000 | 10938.1166 | 10938.1166 | 475.1337 | 0.0000 |
| age | 1.0000 | 90.2679 | 90.2679 | 3.9211 | 0.0482 |
| dis | 1.0000 | 1779.5011 | 1779.5011 | 77.2986 | 0.0000 |
| rad | 1.0000 | 34.1343 | 34.1343 | 1.4827 | 0.2239 |
| tax | 1.0000 | 329.5541 | 329.5541 | 14.3153 | 0.0002 |
| ptratio | 1.0000 | 1309.3093 | 1309.3093 | 56.8742 | 0.0000 |
| lstat | 1.0000 | 2733.5420 | 2733.5420 | 118.7405 | 0.0000 |
| Residual | 493.0000 | 11349.4188 | 23.0211 | nan | nan |

# Partie 4 : Comparaison des 3 Méthodes

- **Observation principale :**
  - Les valeurs des sommes des carrés ($SC_{reg}$ et $SC_{res}$) sont identiques pour **MCO** et **Sklearn**, ce qui confirme que les trois méthodes mènent aux mêmes calculs et résultats.
  - La statistique $F_{stat}$ et la **p-valeur** sont également identiques dans ces deux méthodes, ce qui indique que la variabilité expliquée par la régression est significativement plus grande que la variabilité résiduelle.
  - Le modèle est donc **statistiquement significatif** avec $p \ll 0.05$, ce qui signifie que les variables explicatives ont un impact significatif sur la variable cible.

# Partie 4 : Comparaison des 3 Méthodes

- **Pourquoi les résultats de `statsmodels` sont différents ?**
  - La méthode `anova_lm` de `statsmodels` décompose l'ANOVA par variable, c'est-à-dire qu'elle analyse l'effet individuel de chaque variable.
  - Contrairement aux deux autres méthodes qui ne donnent qu'un seul test global sur l'ensemble du modèle, ici nous avons des tests séparés pour chaque variable.
  - Cela permet de voir quelles variables **contribuent significativement** au modèle.

# Partie 4 : Comparaison des 3 Méthodes

- **Analyse des résultats individuels :**
  - **Variables très significatives** $(p \ll 0.05)$ :
    - `crim`, `zn`, `indus`, `chas`, `rm`, `dis`, `tax`, `ptratio`, `lstat`.
    - Ces variables expliquent une **part importante de la variabilité de la valeur des logements**.
    - Par exemple, `rm` (nombre moyen de pièces) a le **plus grand effet** avec $F = 475.13$, ce qui signifie qu'elle contribue énormément à expliquer la valeur des logements.
  - **Variable modérément significative** $(p \approx 0.05)$ :
    - age $(p = 0.0482)$ : Cela signifie que l'**âge des bâtiments** a un effet, mais il est plus faible.
  - **Variables non significatives** $(p > 0.05)$ :
    - `nox` $(p = 0.0694)$ et `rad` $(p = 0.2239)$.
    - Ces variables n'ont **pas d'effet statistiquement significatif** sur la valeur des logements dans ce modèle.

# Table des Matières

1 Analyse des Résidus

2 Régression Linéaire Multiple

3 Décomposition de la Variabilité Totale

4 ANOVA en Régression Linéaire Multiple

5 Exercice : Boston Housing Prices Dataset

6 Intervalle de Confiance et Intervalle de Prédiction

7 Annexe

# Intervalle de Confiance pour la Réponse Moyenne et Intervalle de Prédiction pour une nouvelle Observation en Régression Linéaire Multiple

# Intervalle de Confiance pour la Réponse Moyenne en Régression Linéaire Multiple

# Intervalle de Confiance pour la Réponse Moyenne

- **Dans le Contexte de la Régression Linéaire Multiple :**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \text{où } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Le but est de construire un intervalle de confiance pour la réponse **moyenne** $\mathbb{E}(Y|X_1 = x_{1,0}, X_2 = x_{2,0}, ..., X_p = x_{p,0}) = \mathbb{E}(Y_0)$.

- **Méthode :** Encadrer la valeur de la réponse moyenne en un point donné $\boldsymbol{x_0} = (1, x_{1,0}, x_{2,0}, \ldots, x_{p,0})^\top$ en tenant compte de l'incertitude sur la régression, càd, Var($\hat{\boldsymbol{\beta}}$).
- **Valeur prédite :** La valeur ajustée est donnée par :

$$\hat{Y}_0 = \boldsymbol{x_0}^\top \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,0} + \hat{\beta}_2 x_{2,0} + \cdots + \hat{\beta}_p x_{p,0}.$$

# Intervalle de Confiance pour la Réponse Moyenne

- **Décomposition de la Variance de $\hat{Y}_0$**
  En utilisant les propriétés de la variance pour les estimateurs des moindres carrés :

$$\text{Var}(\hat{Y}_0) = \text{Var}\left(\boldsymbol{x_0}^\top \hat{\boldsymbol{\beta}}\right).$$

Comme $\hat{\boldsymbol{\beta}}$ est un estimateur linéaire de $\boldsymbol{\beta}$, on a :

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

En appliquant la règle $\text{Var}(\boldsymbol{a}^\top\mathbf{z}) = \boldsymbol{a}^\top\text{Var}(\mathbf{z})\boldsymbol{a}$, on obtient :

$$\text{Var}(\hat{Y}_0) = \boldsymbol{x_0}^\top\text{Var}(\hat{\boldsymbol{\beta}})\,\boldsymbol{x_0}.$$

En remplaçant par l'expression de $\text{Var}(\hat{\beta})$, on obtient :

$$\text{Var}(\hat{Y}_0) = \sigma^2\boldsymbol{x_0}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{x_0}.$$

# Intervalle de Confiance pour la Réponse Moyenne

- **Comparaison avec le cas de la régression linéaire simple**
  En régression simple, nous avions : $\text{Var}(\hat{Y}_0) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$.
  Dans le cas multiple, cette relation devient matricielle, prenant en compte la covariance entre les variables explicatives.
- La variance de la réponse moyenne $\hat{Y}_0$ est donnée par :

$$\text{Var}(\hat{Y}_0) = \sigma^2 \boldsymbol{x_0}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x_0}.$$

- **Distribution de $\hat{Y}_0$ et statistique de test** :
  Puisque les erreurs suivent une loi normale, la valeur prédite suit :

$$\hat{Y}_0 \sim \mathcal{N}\left( \mathbb{E}(Y|\boldsymbol{x_0}), \sigma^2 \boldsymbol{x_0}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x_0}. \right).$$

  En standardisant :

$$\frac{\hat{Y}_0 - \mathbb{E}(Y|\boldsymbol{x_0})}{\sigma \sqrt{\boldsymbol{x_0}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x_0}}} \sim \mathcal{N}(0,1).$$

# Intervalle de Confiance pour la Réponse Moyenne

- En utilisant une estimation de $\sigma^2$ par $s^2 = MC_{res}$, on obtient une distribution de Student :

$$\frac{\hat{Y}_0 - \mathbb{E}(Y|\boldsymbol{x_0})}{s\sqrt{\boldsymbol{x_0}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{x_0}}} \sim t_{n-p-1}.$$

- **Intervalle de confiance à $1-\alpha$ pour la réponse moyenne** $\mathbb{E}(Y|X_0)$ :

$$\hat{Y}_0 \pm t_{\alpha/2,n-p-1} \times s\sqrt{\boldsymbol{x_0}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{x_0}}.$$

- Cet intervalle est un intervalle de confiance pour la **réponse moyenne** $\mathbb{E}(Y|\boldsymbol{x_0})$ et non un intervalle de prédiction, qui inclurait l'incertitude associée à une nouvelle observation individuelle.

# Intervalle de Prédiction pour une Nouvelle Observation en Régression Linéaire Multiple

# Intervalle de Prédiction pour une Nouvelle Observation

- On considère toujours le modèle de régression linéaire multiple :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \text{où } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

- **Distribution de l'erreur de prédiction** $Y_0 - \hat{Y}_0$
  Pour une nouvelle observation $\boldsymbol{x_0} = (1, x_{1,0}, x_{2,0}, ..., x_{p,0})^\top$, l'erreur de prédiction est :

$$Y_0 - \hat{Y}_0 = (\boldsymbol{x_0}^\top \boldsymbol{\beta} + \epsilon_0) - (\boldsymbol{x_0}^\top \hat{\boldsymbol{\beta}}).$$

- En utilisant la variance de $\hat{Y}_0$ :

$$\text{Var}(Y_0 - \hat{Y}_0) = \text{Var}(\epsilon_0) + \text{Var}(\hat{Y}_0) = \sigma^2 + \sigma^2 \boldsymbol{x_0}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x_0}.$$

# Intervalle de Prédiction pour une Nouvelle Observation

- **Standardisation et passage à la loi de Student**
  En remplaçant $\sigma^2$ par son estimateur $s^2 = MC_{res}$, l'erreur normalisée suit une loi de Student à $n - p - 1$ degrés de liberté :

$$\frac{Y_0 - \hat{Y}_0}{s\sqrt{1 + \boldsymbol{x_0}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x_0}}} \sim t_{n-p-1}.$$

- **Construction de l'intervalle de prédiction**

  Pour une nouvelle observation $\boldsymbol{x_0} = (1, x_{1,0}, x_{2,0}, ..., x_{p,0})^\top$, l'intervalle de prédiction à $(1 - \alpha)\%$ est donné par :

$$\hat{Y}_0 \pm t_{\alpha/2, n-p-1} \times s\sqrt{1 + \boldsymbol{x_0}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x_0}}.$$

# Intervalle de Prédiction pour une Nouvelle Observation

- **Conclusion :** L'intervalle de prédiction tient compte de deux sources d'incertitude :
  - L'incertitude sur la réponse moyenne estimée $\text{Var}(\hat{Y}_0)$.
  - L'incertitude intrinsèque des nouvelles observations $\text{Var}(Y_0)$.
- **Relation entre variance de prédiction et variance totale :**

  Comme en régression simple, nous avons :

$$\text{Var}(Y_0 - \hat{Y}_0) = \underbrace{\text{Var}(Y_0)}_{\sigma^2} + \underbrace{\text{Var}(\hat{Y}_0)}_{\sigma^2 \boldsymbol{x_0}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x_0}} .$$

  Cet intervalle est donc plus large que l'intervalle de confiance pour la réponse moyenne, car il inclut l'incertitude liée à la variabilité des nouvelles observations.

# Exercice : Intervalles de Confiance et de Prédiction

# Exercice : Intervalles de Confiance et de Prédiction

```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import statsmodels.api as sm
from ISLP import load_data
# Chargement des données Boston Housing
Boston = load_data('Boston')
# Sélection des variables explicatives et cible
X = Boston.drop(columns=['medv'])  # Variables explicatives
Y = Boston['medv'].values  # Variable cible
# Ajout d'une colonne d'intercept
X_with_intercept = np.c_[np.ones(X.shape[0]), X]
# Ajustement du modèle de régression multiple
model = sm.OLS(Y, X_with_intercept).fit()
# Prédictions sur l'échantillon
Y_pred = model.predict(X_with_intercept)
# Calcul des paramètres statistiques
n, p = X.shape  # Nombre d'observations et nombre de variables explicatives
s_squared = np.sum((Y - Y_pred) ** 2) / (n - p - 1)  # Estimation de la variance des résidu
s = np.sqrt(s_squared)  # Écart-type des résidus
t_value = stats.t.ppf(1 - 0.025, df=n-p-1)  # Valeur critique de Student pour IC à 95%
```

# Intervalles de Confiance et de Prédiction pour Boston Housing

```
XTX_inv = np.linalg.inv(X_with_intercept.T @ X_with_intercept) # Matrice (X'X)^-1
# Choix des points spécifiques pour x0
x0_values = np.array([
    X.iloc[5].values,   # 6e logement
    X.iloc[50].values,  # 51e logement
    X.iloc[100].values, # 101e logement
    X.iloc[200].values  # 201e logement
])
predictions = {}
for i, x_0 in enumerate(x0_values):
    x_0 = np.insert(x_0, 0, 1)  # Ajout de l'intercept
    variance_IC = s_squared * (x_0 @ XTX_inv @ x_0.T)
    variance_IP = s_squared * (1 + x_0 @ XTX_inv @ x_0.T)
    SE_IC = np.sqrt(variance_IC)
    SE_IP = np.sqrt(variance_IP)
    y_0_hat = model.predict([x_0])[0]
    IC_bounds = (y_0_hat - t_value * SE_IC, y_0_hat + t_value * SE_IC)
    IP_bounds = (y_0_hat - t_value * SE_IP, y_0_hat + t_value * SE_IP)
    predictions[f"x0_{i+1}"] = {
        "prediction": y_0_hat,
        "IC_95%": IC_bounds,
        "IP_95%": IP_bounds
    }
```

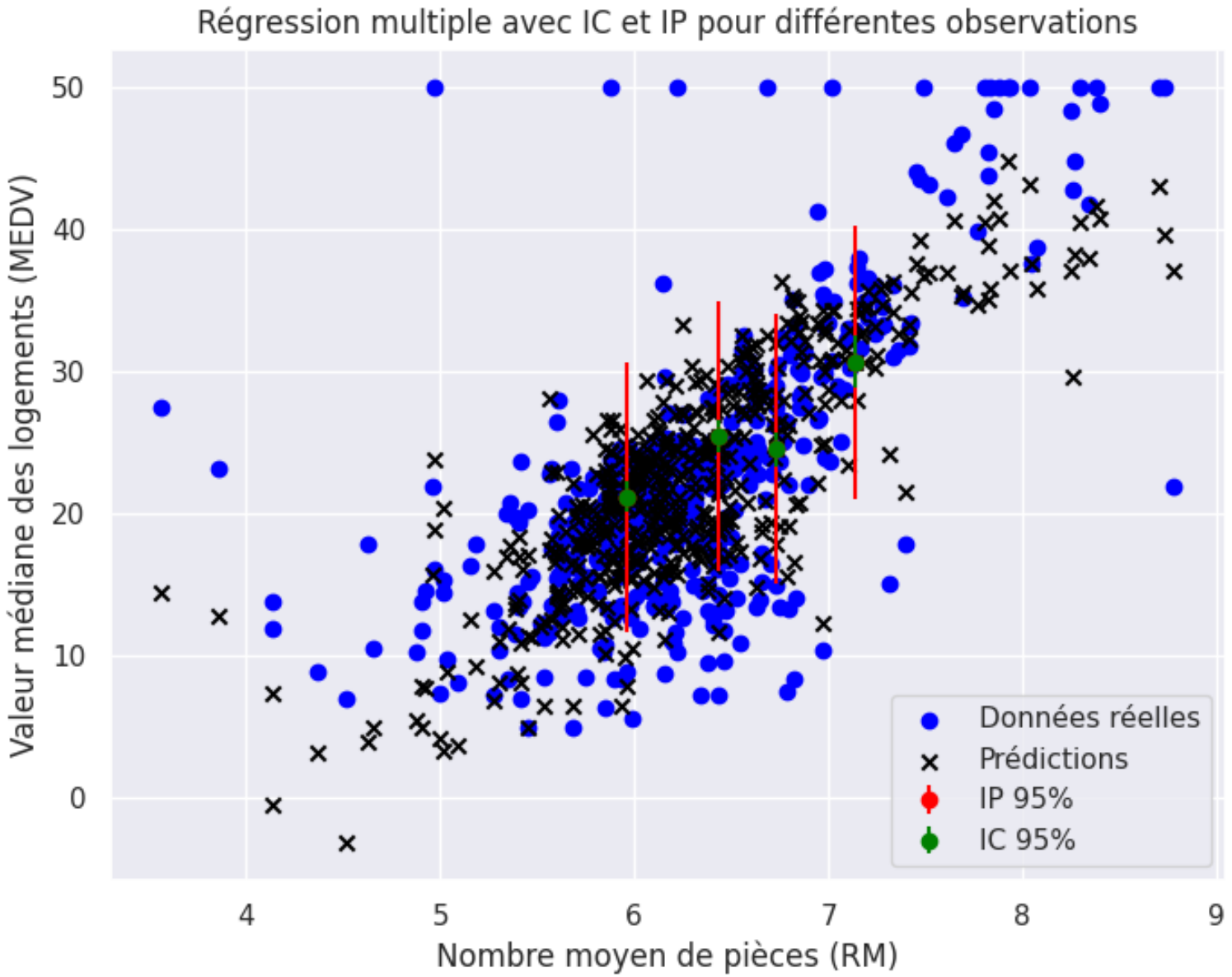
# Résultats des Intervalles de Confiance et de Prédiction

```python
# Affichage des résultats
for x_0, result in predictions.items():
    print(f"Prédiction pour {x_0} : {result['prediction']:.2f}")
    print(f"Intervalle de confiance à 95%: [{result['IC_95%'][0]:.2f}, {result['IC_95%'][1]
    print(f"Intervalle de prédiction à 95%: [{result['IP_95%'][0]:.2f}, {result['IP_95%'][
    print("-" * 50)
```

```
Prédiction pour x0_1 : 25.39
Intervalle de confiance à 95%: [24.25, 26.54]
Intervalle de prédiction à 95%: [15.90, 34.89]
--------------------------------------------------
Prédiction pour x0_2 : 21.15
Intervalle de confiance à 95%: [20.09, 22.22]
Intervalle de prédiction à 95%: [11.67, 30.64]
--------------------------------------------------
Prédiction pour x0_3 : 24.50
Intervalle de confiance à 95%: [23.32, 25.67]
Intervalle de prédiction à 95%: [15.00, 34.00]
--------------------------------------------------
Prédiction pour x0_4 : 30.65
Intervalle de confiance à 95%: [28.80, 32.51]
Intervalle de prédiction à 95%: [21.05, 40.26]
--------------------------------------------------
```

# Visualisation des Intervalles de Confiance et de Prédiction

```
# Tracé de l'intervalle pour la première variable explicative (ex: RM)
plt.figure(figsize=(8, 6))
plt.scatter(X['rm'], Y, label="Données réelles", color='blue')
plt.scatter(X['rm'], Y_pred, label="Prédictions", color='black', marker="x")
# Affichage des intervalles
for x_0 in x0_values:
    y_hat = model.predict([np.insert(x_0, 0, 1)])[0]
    variance_IC = s_squared * (np.insert(x_0, 0, 1) @ XTX_inv @ np.insert(x_0, 0, 1).T)
    variance_IP = s_squared * (1 + np.insert(x_0, 0, 1) @ XTX_inv @ np.insert(x_0, 0, 1).T
    SE_IC = np.sqrt(variance_IC)
    SE_IP = np.sqrt(variance_IP)
    plt.errorbar(x_0[5], y_hat, yerr=t_value * SE_IC, fmt='o', color='green', label="IC 95%
    plt.errorbar(x_0[5], y_hat, yerr=t_value * SE_IP, fmt='o', color='red', label="IP 95%"
plt.xlabel("Nombre moyen de pièces (RM)")
plt.ylabel("Valeur médiane des logements (MEDV)")
plt.legend()
plt.title("Régression multiple avec IC et IP pour différentes observations")
plt.show()
```


Régression multiple avec IC et IP pour différentes observations
Valeur médiane des logements (MEDV)
Nombre moyen de pièces (RM)
0
10
20
30
40
50
4
5
6
7
8
9
Données réelles
Prédictions
IP 95%
IC 95%

# Analyse des Résultats des Intervalles de Confiance et de Prédiction

- **Différence entre IC et IP :**
  - L'**intervalle de confiance (IC)** quantifie l'incertitude sur l'**estimation moyenne** de la variable cible (MEDV) pour un ensemble donné de caractéristiques.
  - L'**intervalle de prédiction (IP)** est toujours plus large car il inclut la **variance des nouvelles observations** en plus de l'incertitude sur l'estimation moyenne.
- **Observations sur les résultats :**
  - Les valeurs prédites sont cohérentes avec la tendance des données.
  - Plus une observation est éloignée de la moyenne des caractéristiques, plus son intervalle est large.
  - Les IP sont plus larges que les IC, confirmant que la variabilité des nouvelles observations est plus grande que l'incertitude sur la moy.

# Analyse des Résultats des IC et IP pour Boston Housing

- **Enseignements pour la régression multiple :**
  - La précision des prédictions dépend à la fois du **modèle** et de la **distribution des données**.
  - Un modèle avec une bonne généralisation aura des **IP relativement étroits**, indiquant une faible variabilité des nouvelles observations.
  - Les intervalles permettent d'évaluer la **fiabilité des prédictions** et doivent toujours être pris en compte pour interpréter les résultats d'un modèle de régression.
- **Interprétation des intervalles :**
  - L'**intervalle de confiance (IC)** encadre l'**estimation moyenne** de la valeur médiane des logements (`MEDV`) pour un ensemble donné de caractéristiques.
  - L'**intervalle de prédiction (IP)** est plus large car il tient compte de la variabilité des nouvelles observations individuelles.

# Analyse des Résultats des IC et IP pour Boston Housing

- **Observations sur les résultats numériques :**
  - Les valeurs prédites sont cohérentes avec les tendances observées dans les données.
  - Les IC sont relativement serrés, indiquant une bonne précision de l'estimation moyenne.
  - Les IP sont nettement plus larges que les IC, ce qui reflète la variabilité naturelle du marché immobilier.
- **Enseignements pour la régression multiple :**
  - Les prédictions sont plus fiables pour des points proches de la moyenne des variables explicatives.
  - Lorsque les caractéristiques du logement (`RM`, `LSTAT`, etc.) s'éloignent des valeurs moyennes observées, l'incertitude augmente, ce qui élargit l'IP.
  - Un bon modèle doit minimiser l'écart entre les observations réelles et les valeurs prédites, tout en gardant des IP raisonnables.

# Merci!

E-mail: chiheb.trabelsi@polymtl.ca

# Table des Matières

1 Analyse des Résidus

2 Régression Linéaire Multiple

3 Décomposition de la Variabilité Totale

4 ANOVA en Régression Linéaire Multiple

5 Exercice : Boston Housing Prices Dataset

6 Intervalle de Confiance et Intervalle de Prédiction

7 Annexe

# Matrice de Gram $\mathbf{X}^\top\mathbf{X}$

# Matrice de Gram $\mathbf{X}^\top\mathbf{X}$

- Une **matrice de Gram** est une matrice symétrique qui contient les produits scalaires entre ses vecteurs et permet d'analyser leur dépendance linéaire.
- Soit $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ un ensemble de $m$ vecteurs dans un espace de dimension $n$.
- La **matrice de Gram** associée à ces vecteurs est définie par :

$$\mathbf{G} = \mathbf{X}^\top\mathbf{X}$$

où $\mathbf{X}$ est la matrice dont les colonnes sont les vecteurs $\mathbf{x}_i$, soit :

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \end{bmatrix}$$

- La matrice de Gram $\mathbf{G}$ a donc pour éléments :

$$G_{ij} = \mathbf{x}_i^\top\mathbf{x}_j$$

# Matrice de Gram $\mathbf{X}^\top\mathbf{X}$

- **Matrice symétrique** :

$$G_{ij} = \mathbf{x}_i^\top \mathbf{x}_j = \mathbf{x}_j^\top \mathbf{x}_i = G_{ji}$$

ce qui implique que $\mathbf{G}$ est toujours symétrique.

- **Définie positive ou semi-définie positive** :
  - Si les vecteurs $\mathbf{x}_i$ sont **linéairement indépendants**, alors $\mathbf{G}$ est **définie positive**, c'est-à-dire :

$$\forall \quad \mathbf{v} \neq 0, \quad \mathbf{v}^\top \mathbf{G}\mathbf{v} = \mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} = (\mathbf{X}\mathbf{v})^\top (\mathbf{X}\mathbf{v}) = ||\mathbf{X}\mathbf{v}||_2^2 > 0.$$

  - Si les vecteurs sont **linéairement dépendants**, alors $\mathbf{G}$ est **semi-définie positive** :

$$\exists \mathbf{v} \neq 0 \text{ tel que } \mathbf{v}^\top \mathbf{G}\mathbf{v} = 0.$$

Ce qui implique que $\mathbf{G}$ est non inversible.

# Preuve : La Multicolinéarité de $\mathbf{X}$ Implique la Singularité de $\mathbf{X}^\top\mathbf{X}$

# Preuve : Multicolinéarité $\Rightarrow$ Singularité de $\mathbf{X}^\top\mathbf{X}$

On considère le modèle de régression linéaire sous forme matricielle :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

où :

- $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ est le vecteur des observations.
- $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ est la matrice des variables explicatives.
- $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ est le vecteur des coefficients.
- $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$ est le vecteur des erreurs.

Si les variables explicatives sont linéairement dépendantes, alors il existe un ensemble de scalaires $\lambda_1, \lambda_2, \ldots, \lambda_p$, non tous nuls, tels que :

$$\sum_{j=1}^{p} \lambda_j \mathbf{x}_j = \mathbf{0}$$

# Preuve : Multicolinéarité $\Rightarrow$ Singularité de $\mathbf{X}^\top\mathbf{X}$

- Si $\mathbf{X}^\top\mathbf{X}$ est inversible, elle doit être de plein rang (rang$(\mathbf{X}^\top\mathbf{X}) = p$).
- Si les colonnes de $\mathbf{X}$ sont linéairement dépendantes, alors il existe un vecteur $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^\top \neq \mathbf{0}$ tel que :

$$\mathbf{X}\boldsymbol{\lambda} = \mathbf{0}.$$

- cela veut dire que $\boldsymbol{\lambda}$ appartient au **noyau** (**espace nul**) de $\mathbf{X}$ qui est noté par $\ker(\mathbf{X})$
- Le **noyau** (aussi appelé **espace nul**) d'une matrice $\mathbf{A}$ est l'ensemble des vecteurs qui sont envoyés sur le vecteur nul par cette matrice.
- Mathématiquement, le noyau de $\mathbf{A}$ est défini comme :

$$\ker(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^p \mid \mathbf{A}\mathbf{v} = \mathbf{0}\}$$

- Autrement dit, le noyau de $\mathbf{A}$ est l'ensemble des solutions de l'équation homogène :

$$\mathbf{A}\mathbf{v} = \mathbf{0}$$

# Preuve : Multicolinéarité $\Rightarrow$ Singularité de $\mathbf{X}^\top\mathbf{X}$

- Dans le contexte de la régression linéaire, la matrice des moindres carrés est donnée par :
$$\mathbf{X}^\top\mathbf{X}$$
- Le **noyau** de $\mathbf{X}^\top\mathbf{X}$ est défini comme l'ensemble des vecteurs $\mathbf{v}$ tels que :
$$\ker(\mathbf{X}^\top\mathbf{X}) = \{\mathbf{v} \in \mathbb{R}^p \mid \mathbf{X}^\top\mathbf{X}\mathbf{v} = \mathbf{0}\}$$
- Cela signifie que tout vecteur appartenant au noyau de $\mathbf{X}^\top\mathbf{X}$ est une combinaison linéaire des colonnes de $\mathbf{X}$ qui s'annule lorsqu'on applique $\mathbf{X}^\top\mathbf{X}$.

# Preuve que $\mathbf{X}^\top\mathbf{X}$ est singulière si $\mathbf{X}$ est linéairement dépendante

- Si les colonnes de $\mathbf{X}$ sont linéairement dépendantes, alors il existe un vecteur $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^\top \neq \mathbf{0}$ tel que $\boldsymbol{\lambda} \in \ker(\mathbf{X})$, c'est à dire :

$$\mathbf{X}\boldsymbol{\lambda} = \mathbf{0}.$$

- En multipliant cette expression à gauche par $\mathbf{X}^\top$, on obtient :

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\lambda} = \mathbf{X}^\top\mathbf{0} = \mathbf{0}$$

- Cela signifie que $\boldsymbol{\lambda}$ appartient au noyau de $\mathbf{X}^\top\mathbf{X}$, donc cette matrice n'est pas inversible.
- Cela signifie que $\ker(\mathbf{X}^\top\mathbf{X})$ contient au moins un vecteur non nul, donc $\mathbf{X}^\top\mathbf{X}$ **n'est pas inversible** CQFD.

# Preuve : Multicolinéarité $\Rightarrow$ Singularité de $\mathbf{X}^\top\mathbf{X}$

- Si les colonnes de $\mathbf{X}$ sont **linéairement indépendantes**, alors $\mathbf{X}^\top\mathbf{X}$ est **de plein rang** et son noyau est réduit au vecteur nul :

$$\ker(\mathbf{X}^\top\mathbf{X}) = \{\mathbf{0}\}$$

.

- Si les colonnes de $\mathbf{X}$ sont **linéairement dépendantes**, alors il existe des vecteurs non nuls $\mathbf{v} \neq \mathbf{0}$ tel que :

$$\mathbf{X}\mathbf{v} = \mathbf{0}.$$

# Explication Intuitive : $\mathbf{X}^\top \mathbf{X}$ Mal Conditionnée Implique une Amplification de la Variance des Coefficients $\hat{\boldsymbol{\beta}}$

# $\mathbf{X}^\top\mathbf{X}$ Mal conditionnée $\Rightarrow$ Amplification de Var$(\hat{\boldsymbol{\beta}})$

- L'estimateur des moindres carrés ordinaires des coefficients est donné par :
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$
- Sa variance est :
$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$
- Où $\sigma^2 = \text{Var}(\epsilon)$ est la variance de l'erreur.
- Si $\mathbf{X}^\top\mathbf{X}$ est mal conditionnée (i.e., proche d'être singulière), son inverse contiendra des valeurs élevées, amplifiant ainsi la variance des coefficients.

# Facteur d'Inflation de la Variance (FIV)

# Facteur d'Inflation de la Variance (FIV)

- Une mesure de la colinéarité est le Facteur d'Inflation de la Variance (FIV) (*Variance Inflation Factor (VIF)*) :

$$VIF_j = \frac{1}{1 - R_j^2}$$

- Où $R_j^2$ est le coefficient de détermination de la régression de la variable $x_j$ sur les autres variables explicatives.
- Si $R_j^2$ est proche de 1, alors $VIF_j$ est très grand, indiquant une forte colinéarité.
- Une règle empirique est que $VIF > 10$ signifie une colinéarité problématique.