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

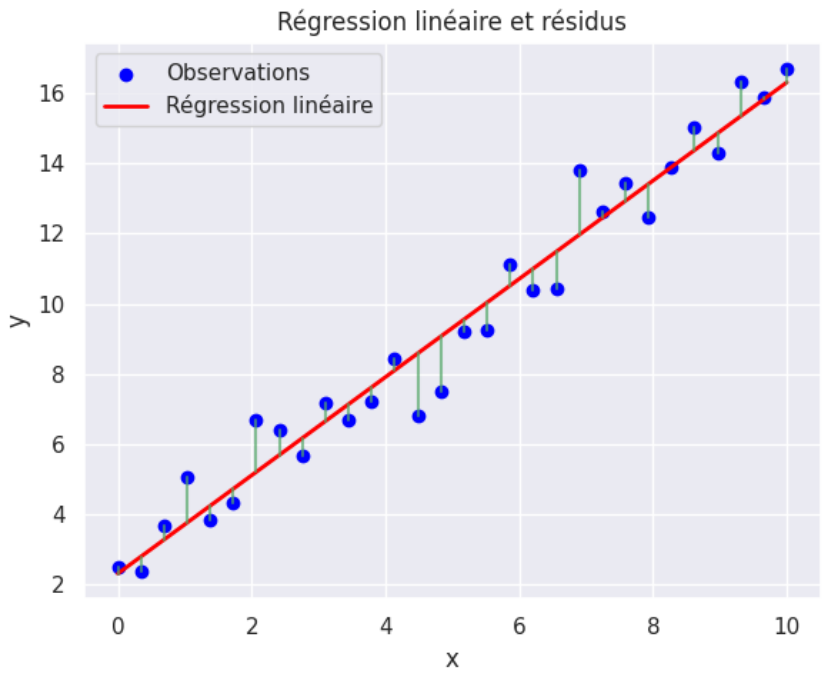
February 19, 2025

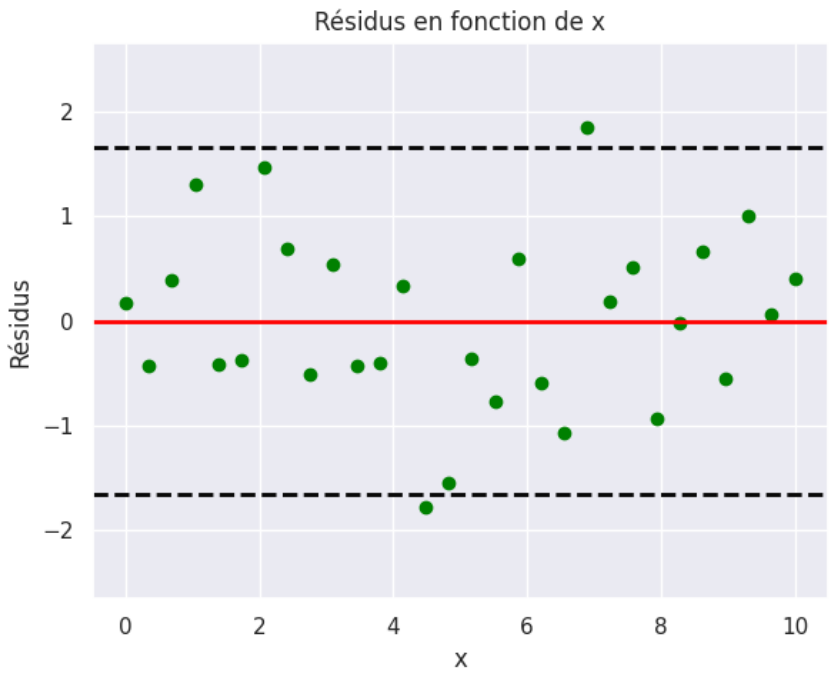
Table des Matières

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

Analyse des Résidus pour Vérifier les Hypothèses

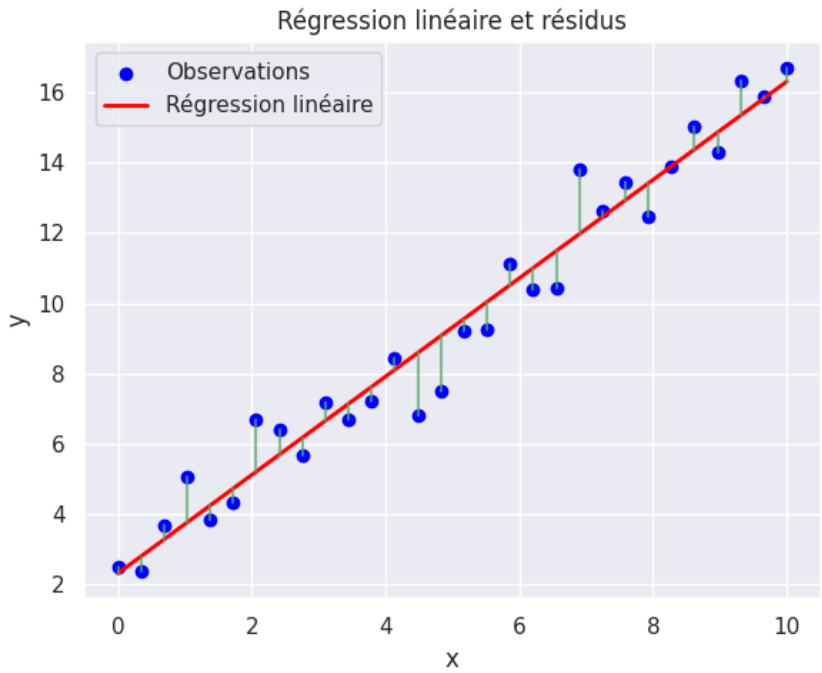
- **L'espérance des erreurs est nulle :** $E(\epsilon_i) = 0$.
- **Homoscédasticité :** La variance des erreurs est constante, $\text{Var}(\epsilon_i) = \sigma^2 \quad \forall\, i \in \{1, \cdots, n\}$.
- **Indépendance des erreurs :** Les erreurs ϵ_i ne sont pas corrélées entre elles $\text{Cov}(\epsilon_i,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

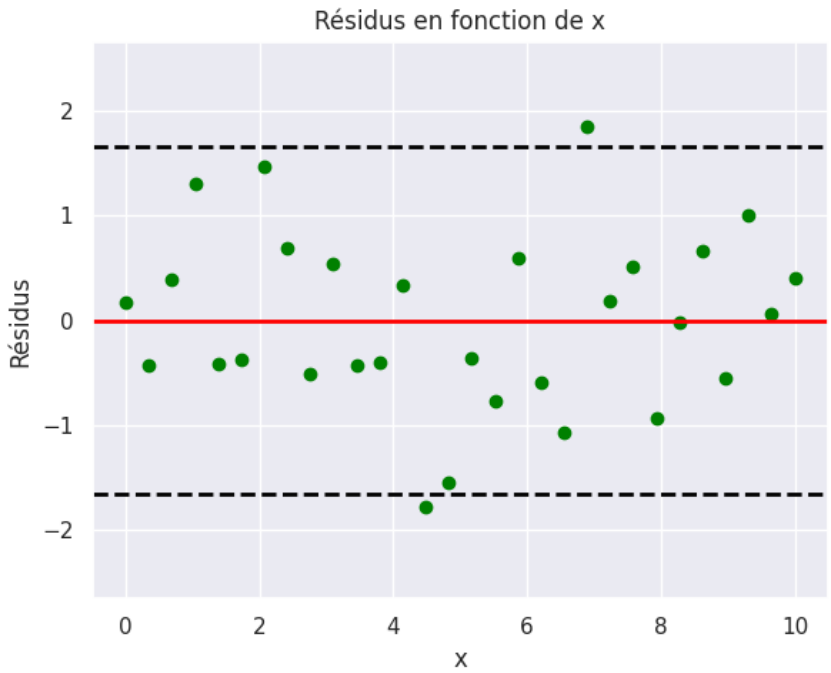




Analyse des Résidus pour Vérifier les Hypothèses

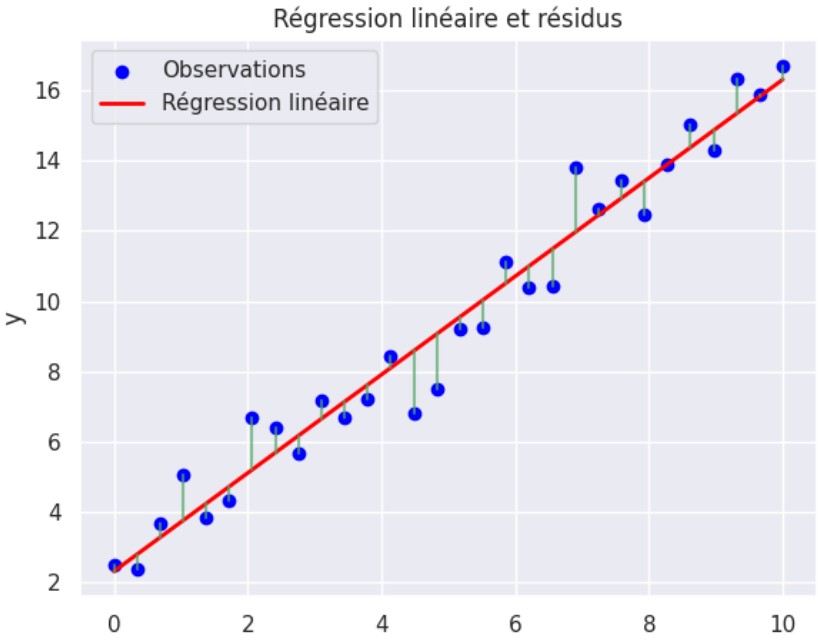
- **Graphique de gauche :** Représente la régression linéaire avec les résidus en vert.
- **Graphique de droite :** Affiche les résidus en fonction de x avec des seuils supérieur et inférieur (bandes noires) pour examiner leur dispersion.

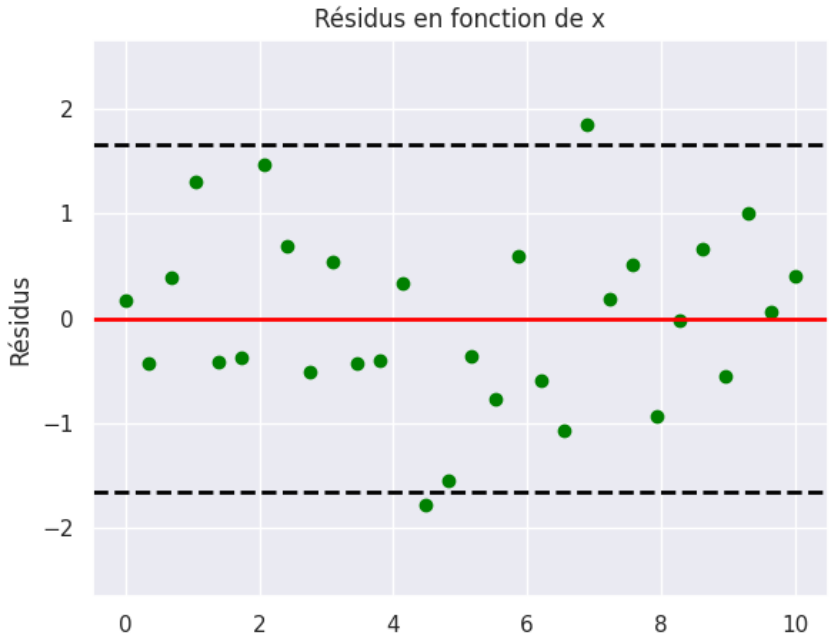




Analyse des Résidus pour Vérifier les Hypothèses

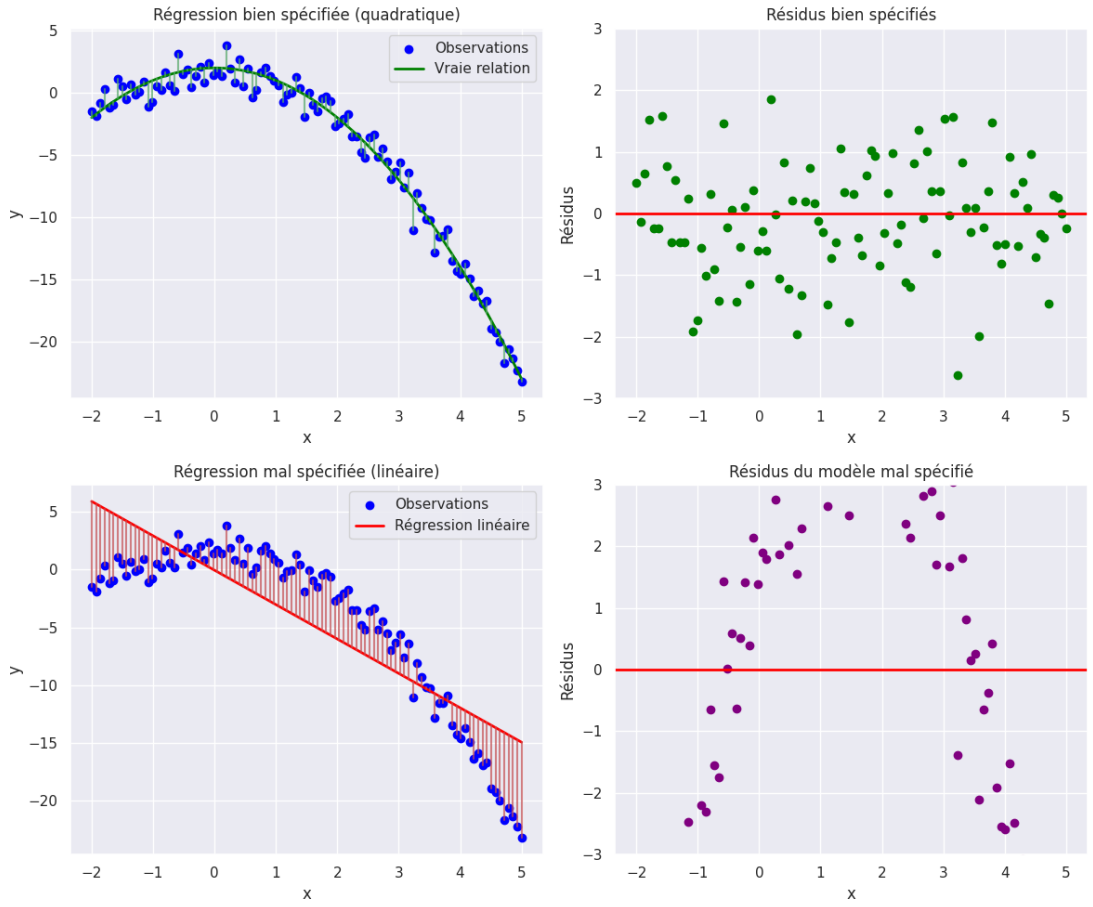
- Les résidus doivent être **centrés autour de zéro** sans structure apparente.
- La bande noire illustre la dispersion des résidus. Une distribution homogène indique que l'hypothèse d'homoscédasticité est respectée.
- Si la dispersion des résidus augmente ou diminue selon x, cela peut indiquer une hétéroscédasticité.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





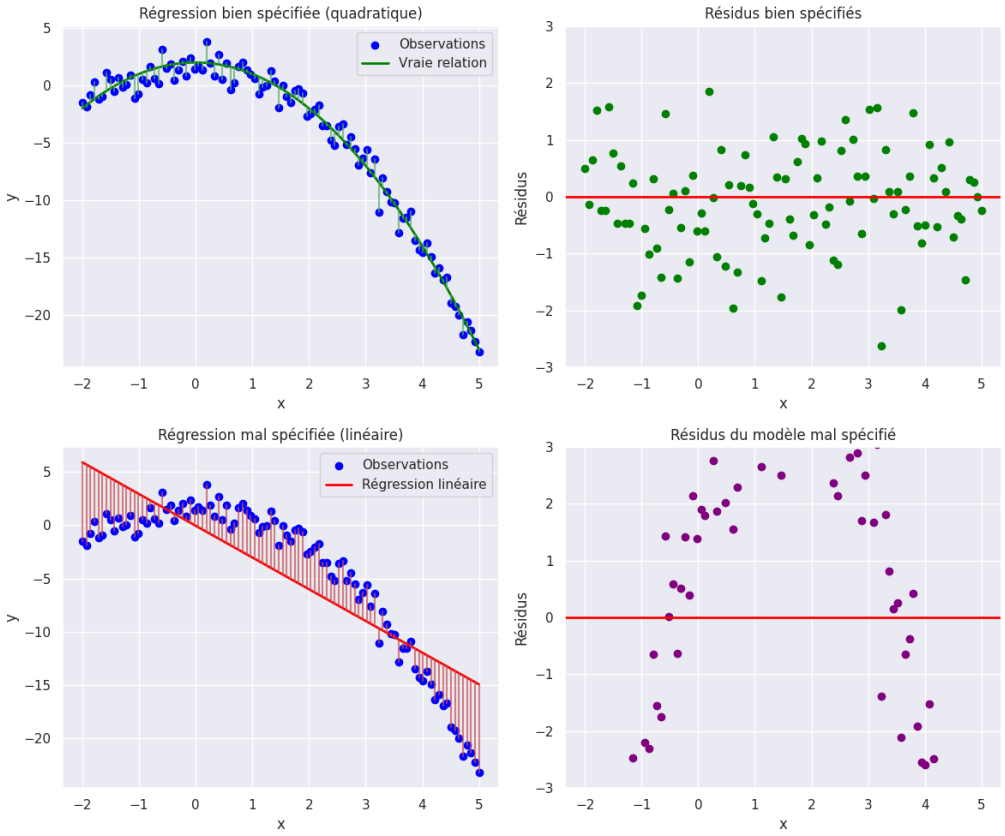
Analyse des Résidus pour Vérifier les Hypothèses

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



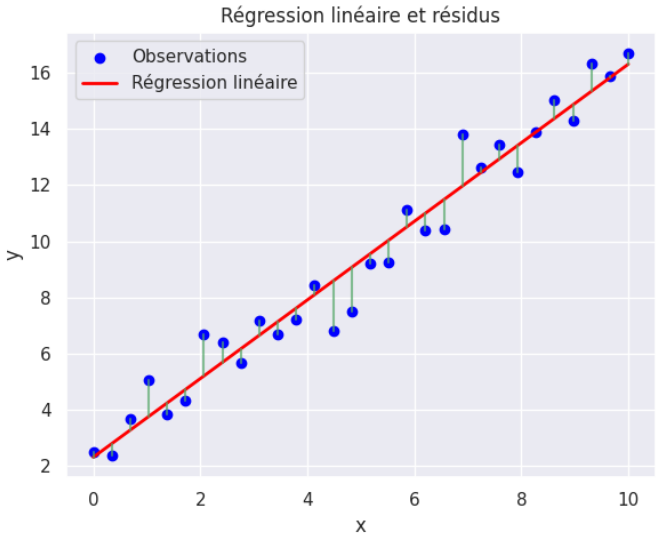
Analyse des Résidus pour Vérifier les Hypothèses

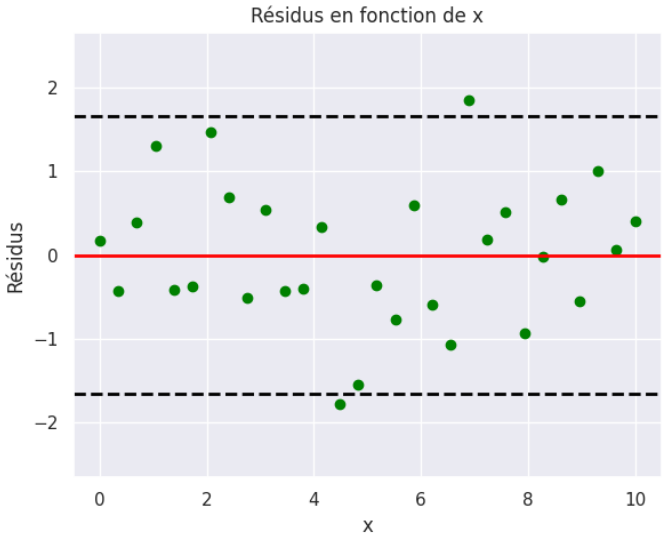
- **Graphiques en bas :**
  - À gauche : Mauvaise spécification du modèle, absence d'un terme quadratique.
  - À droite : Résidus en courbe, indiquant la non prise en compte du terme quadratique.



Analyse des Résidus pour Vérifier les Hypothèses

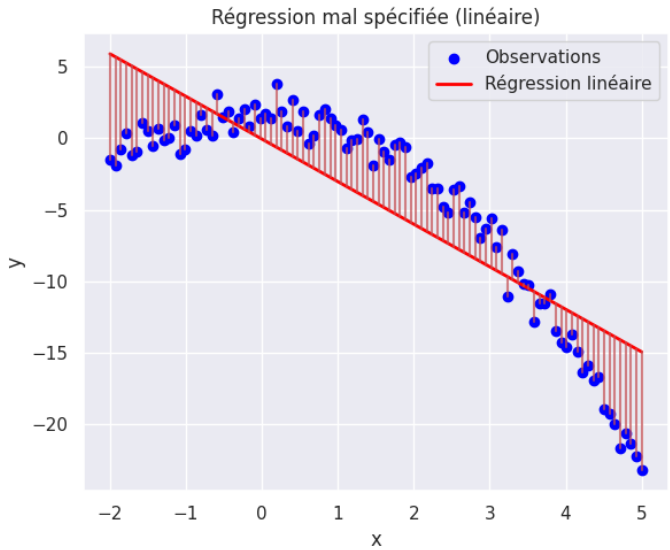
- Un modèle bien spécifié présente des résidus sans structure et bien répartis autour de 0.
- Un modèle mal spécifié peut mener à :
  - Une tendance dans les résidus, indiquant qu'une variable importante est omise.
  - Une dispersion non homogène des résidus, signalant une non-linéarité ignorée.
- L'ajout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

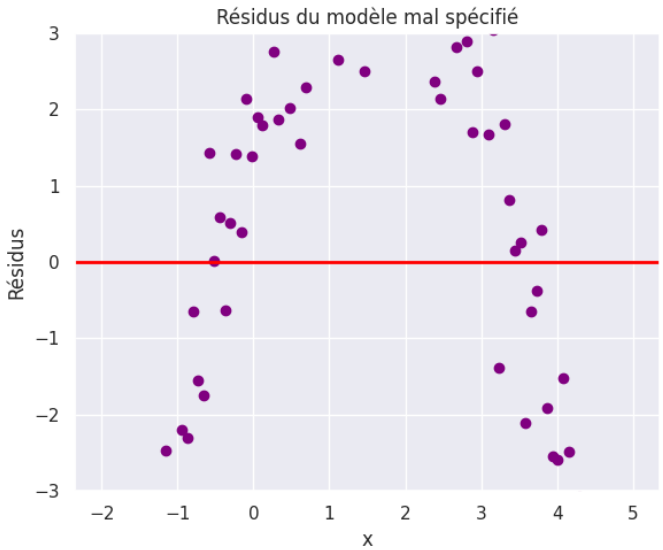




Analyse des Résidus pour Vérifier les Hypothèses

- 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.





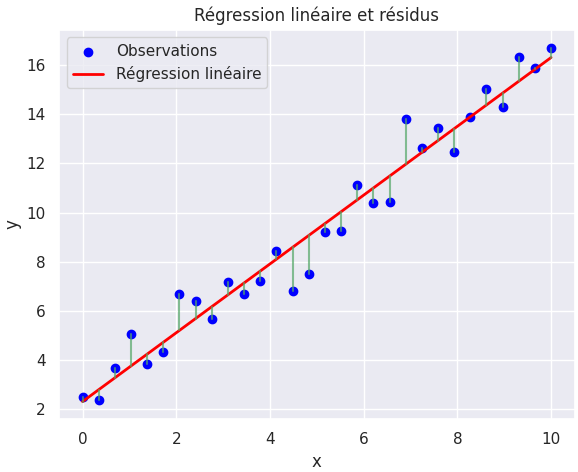
Analyse des Résidus avec Histogrammes et QQ-Plots

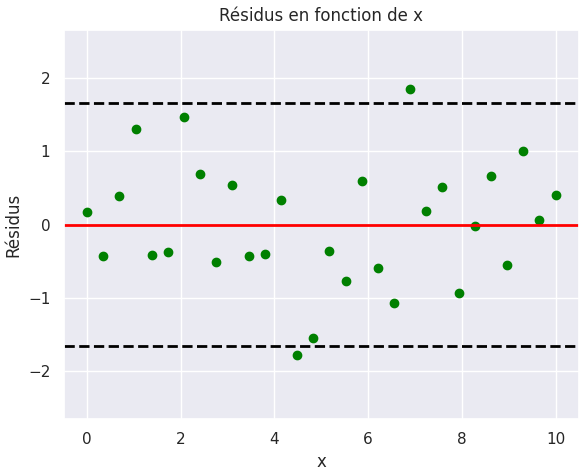
Analyse des Résidus avec Histogrammes et QQ-Plots

- 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 :

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

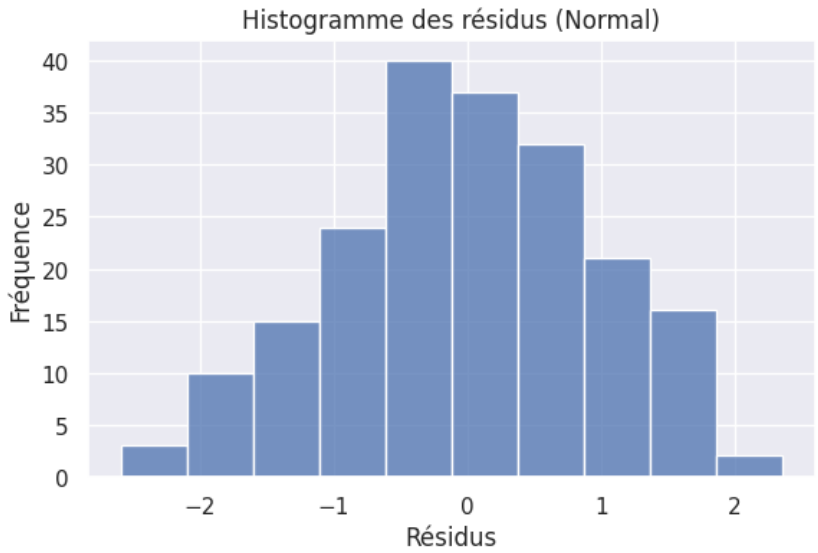
- Les erreurs sont supposées indépendantes et suivent : $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Les résidus sont calculés comme suit : $e_i = y_i - \hat{y}_i$

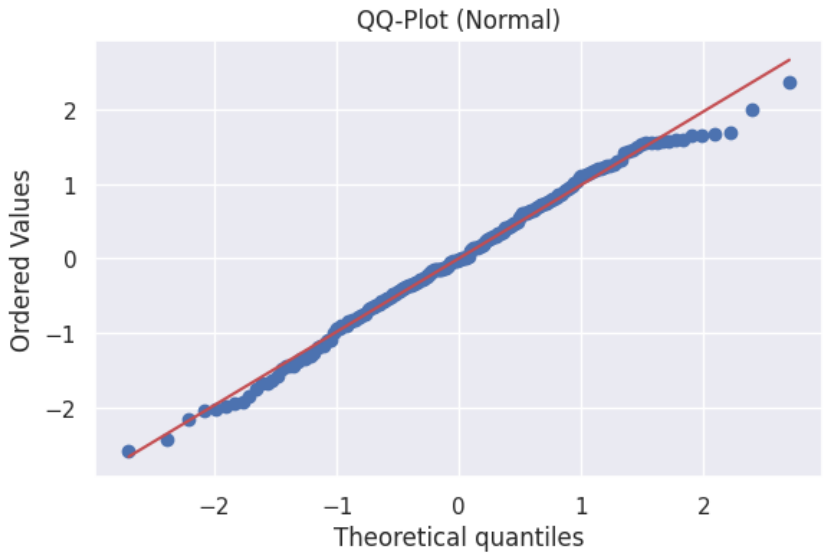




Analyse des Résidus avec Histogrammes et QQ-Plots

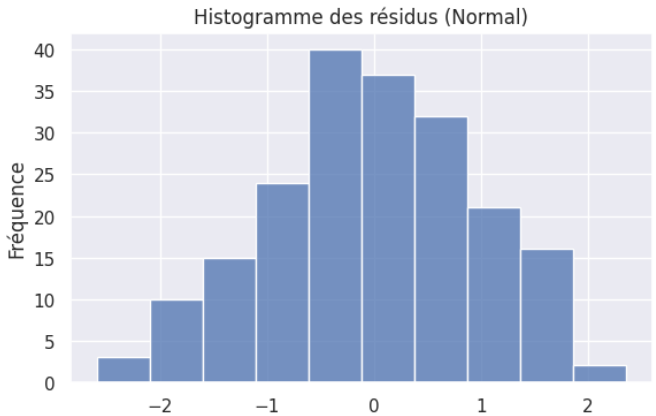
- Le QQ-Plot compare la distribution empirique des résidus à une normale.
- Ordonnée : résidus triés, abscisse : quantiles théoriques.
- Position des quantiles : $\Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right)$

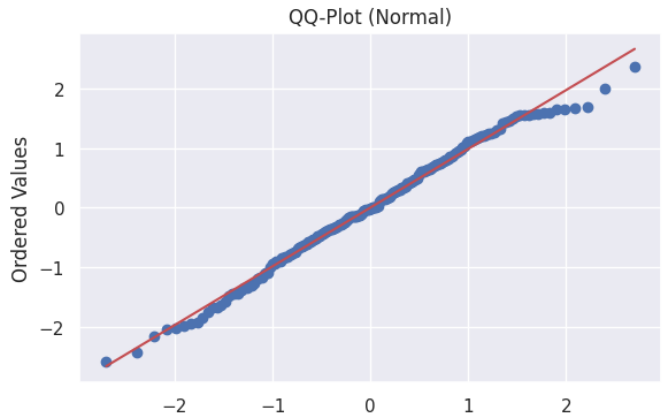




Analyse des Résidus avec Histogrammes et QQ-Plots

- 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

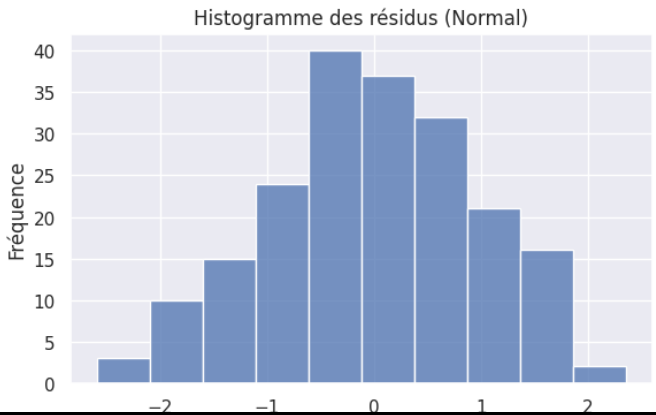
- La fonction Φ^{-1} est l'inverse de la fonction de répartition de la loi normale standard, aussi appelée **fonction quantile de la distribution normale**.
- **Définition de** $\Phi(x)$:

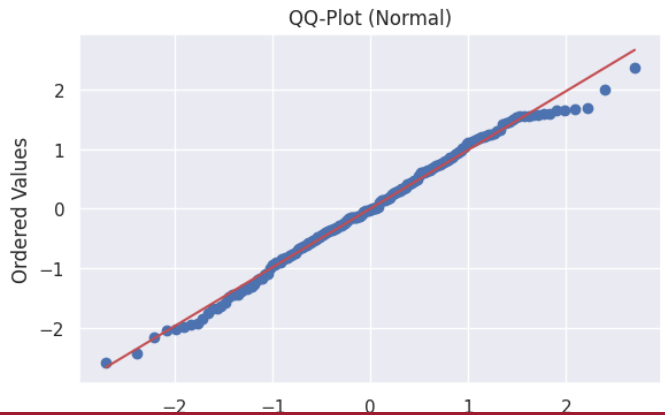
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

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 :

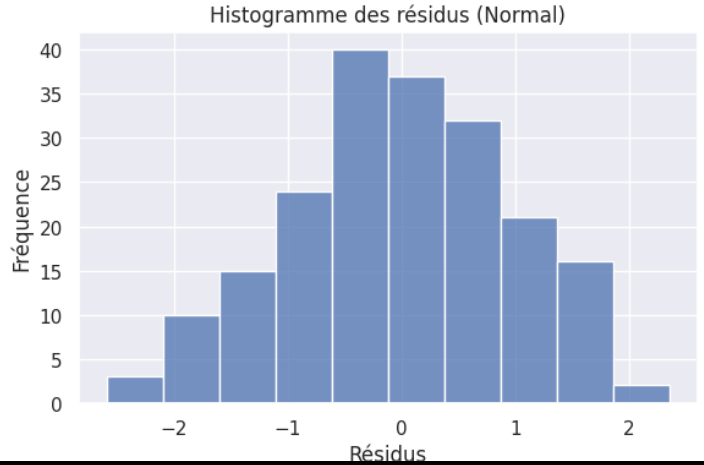
$$\Phi^{-1}(p) = z \quad \text{tel que} \quad P(X \leq z) = p$$

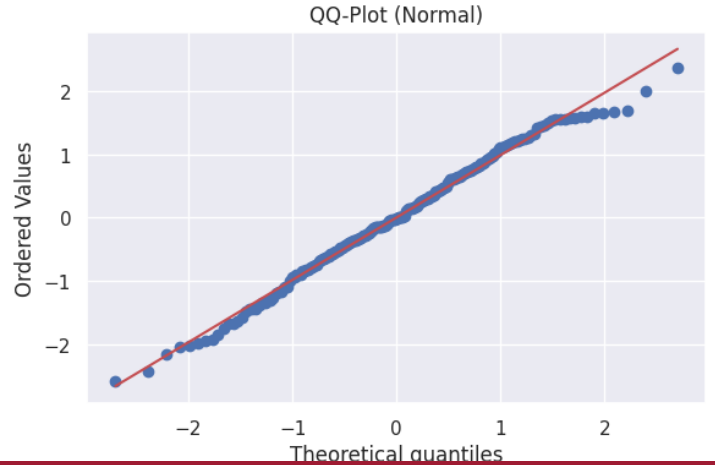




Analyse des Résidus avec Histogrammes et QQ-Plots

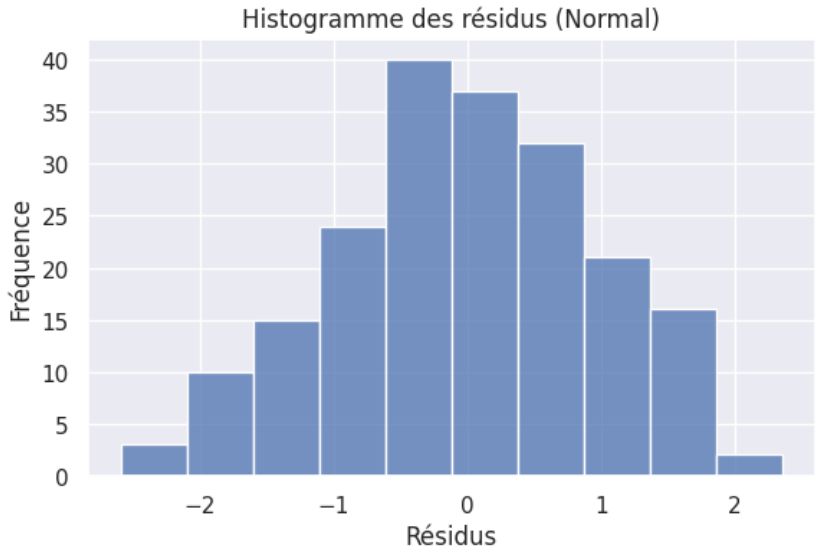
- La formule utilisée pour positionner les points sur l'axe des abscisses du QQ-Plot est : $\Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right)$
- **Explication des termes** :
  - k : Indice de l'observation après tri des résidus (du plus petit au plus grand).
  - n : Nombre total d'observations.
  - 0.375 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.

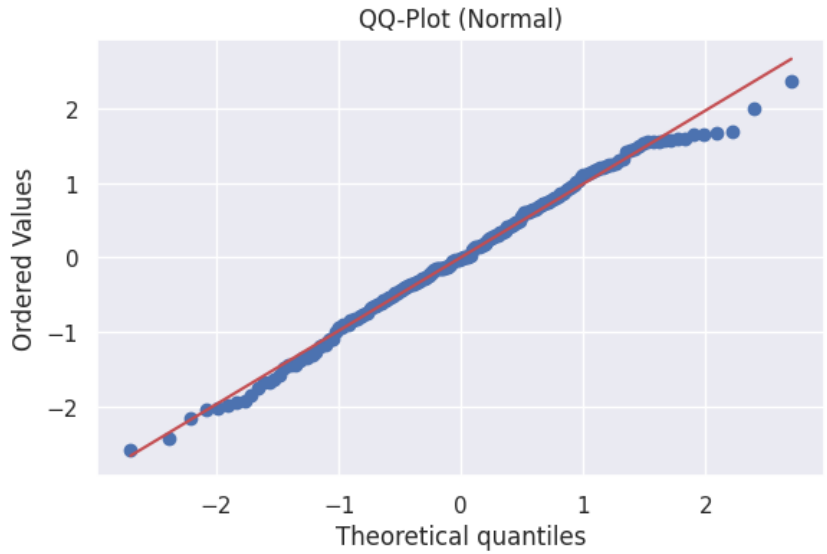




Analyse des Résidus avec Histogrammes et QQ-Plots

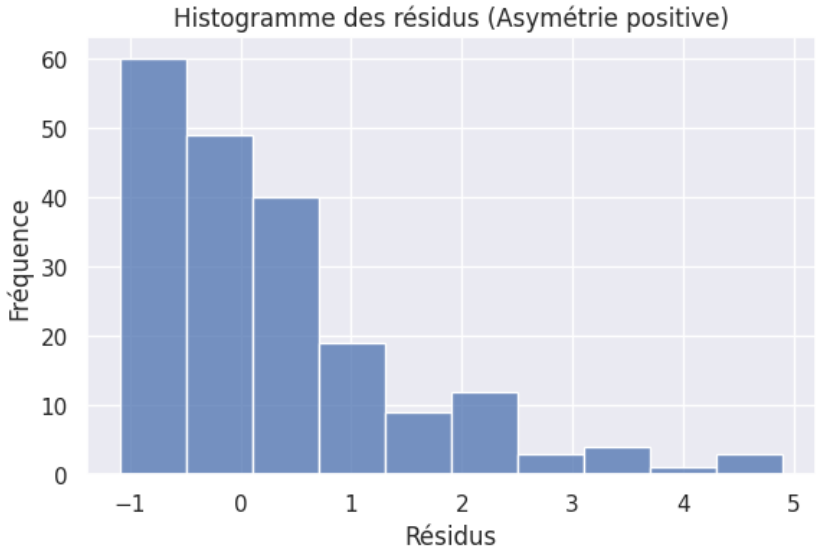
- Analyse des résidus via histogrammes et QQ-Plots.
- Différents types de distributions :
  - **Normal : résidus bien répartis autour de 0 et les points du QQ-Plot alignés.**
  - Asymétrie positive : biais vers la droite et courbure au-dessus de la diagonale.
  - Asymétrie négative : biais vers la gauche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

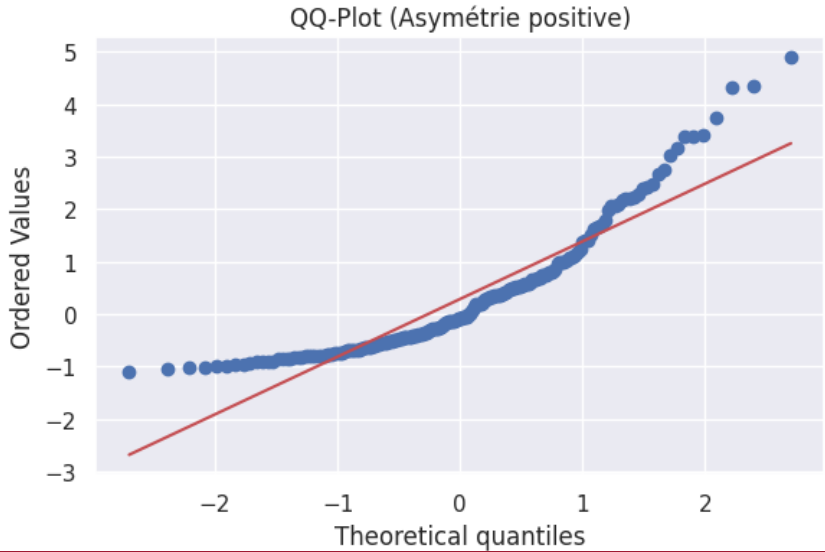




Analyse des Résidus avec Histogrammes et QQ-Plots

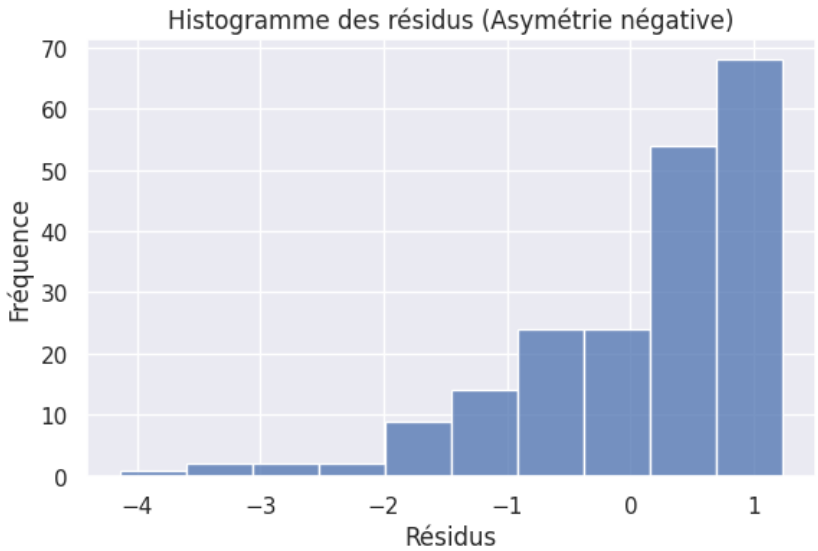
- Analyse des résidus via histogrammes et QQ-Plots.
- Différents types de distributions :
  - Normal : résidus bien répartis autour de 0 et le points du QQ-Plot alignés.
  - **Asymétrie positive : biais vers la droite et courbure au-dessus de la diagonale.**
  - Asymétrie négative : biais vers la gauche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

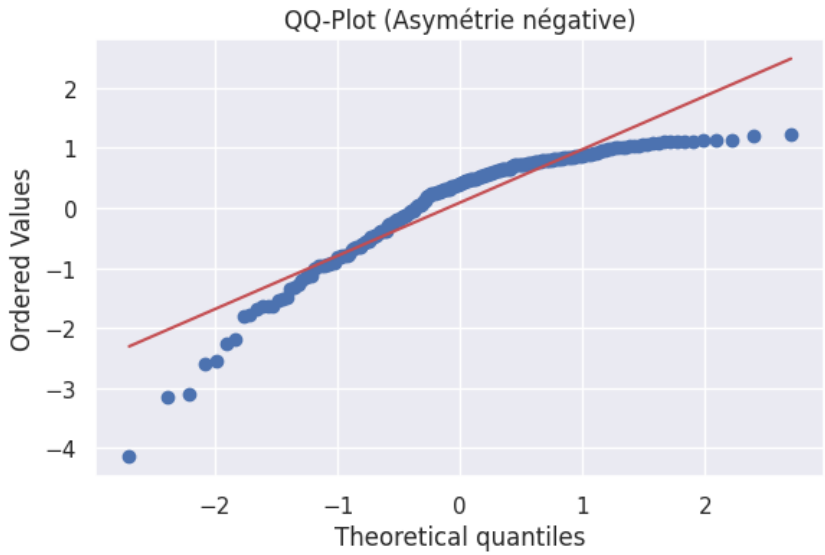




Analyse des Résidus avec Histogrammes et QQ-Plots

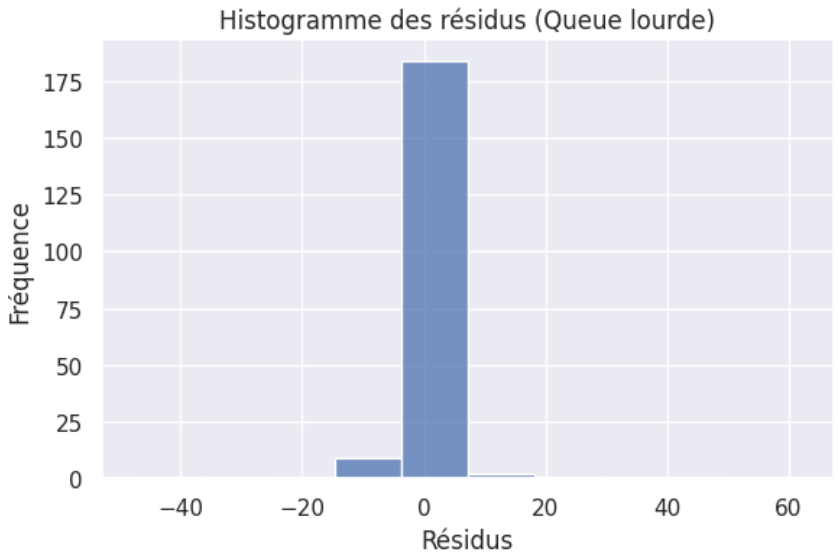
- Analyse des résidus via histogrammes et QQ-Plots.
- Différents types de distributions :
  - Normal : résidus bien répartis autour de 0 et le points du QQ-Plot alignés.
  - Asymétrie positive : biais vers la droite et courbure au-dessus de la diagonale.
  - **Asymétrie négative : biais vers la gauche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

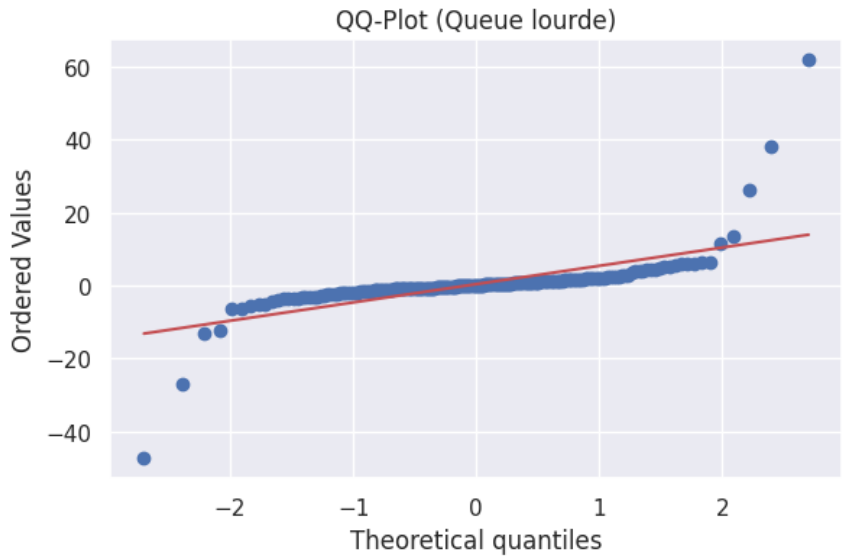




Analyse des Résidus avec Histogrammes et QQ-Plots

- 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.
- Φ^{-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

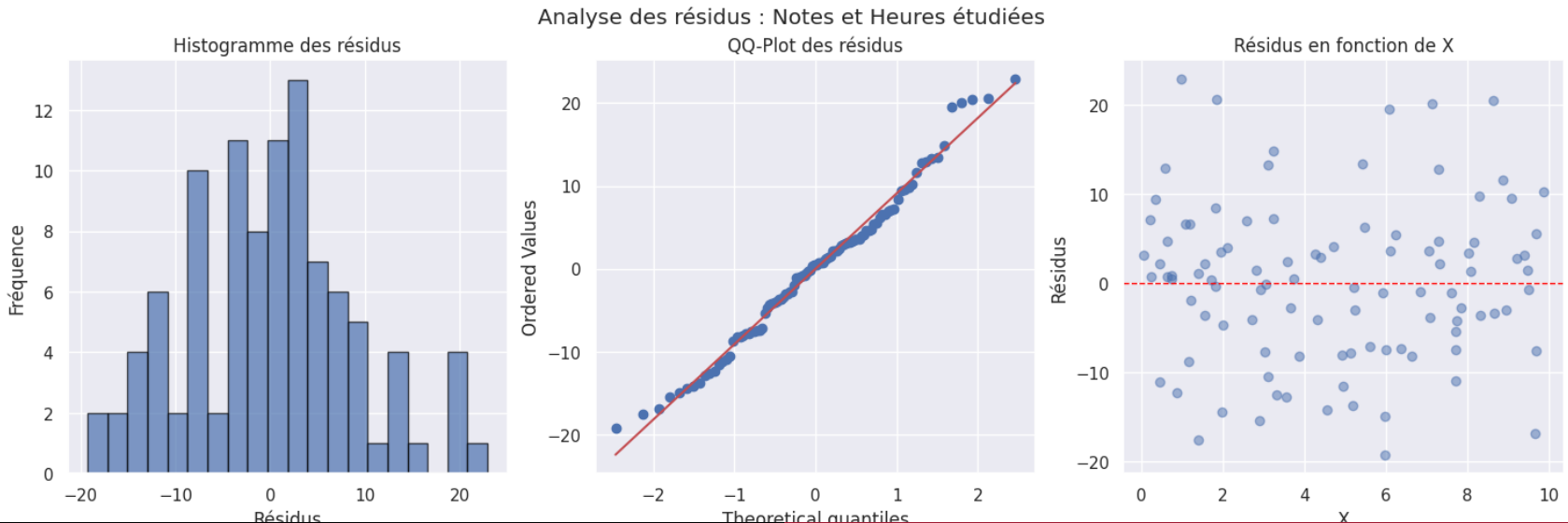
```python
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from google.colab import drive
# Étape 0.a: Monter Google Drive
drive.mount('/content/drive')
# Étape 0.b: 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
# 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)
```

Analyse des Résidus : Code Python

```
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")
    # QQ-Plot des résidus
    stats.probplot(residuals, dist="norm", plot=axes[1])
    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

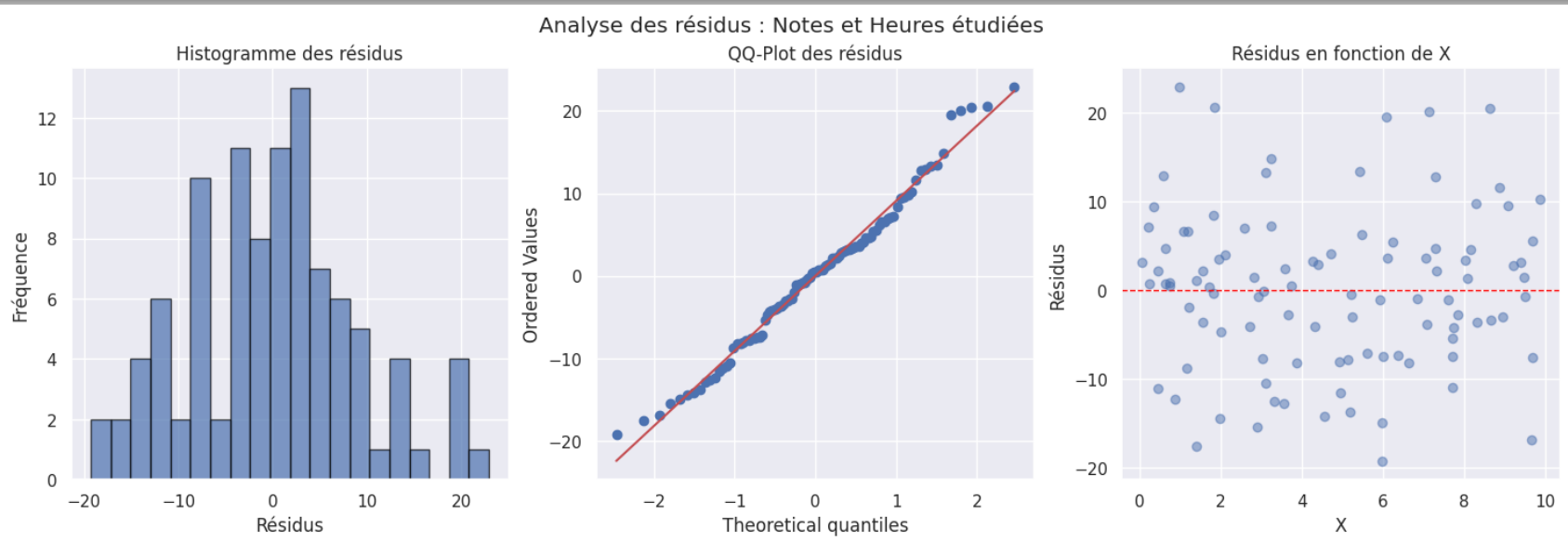
- **Objectif :** Vérifier la normalité des résidus pour la régression des notes en fonction des heures étudiées.
- **Hypothèses** :
  - Les résidus doivent être centrés autour de 0.
  - Ils doivent être homoscédastiques (variance constante).
  - Ils ne doivent pas suivre de structure particulière.
-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

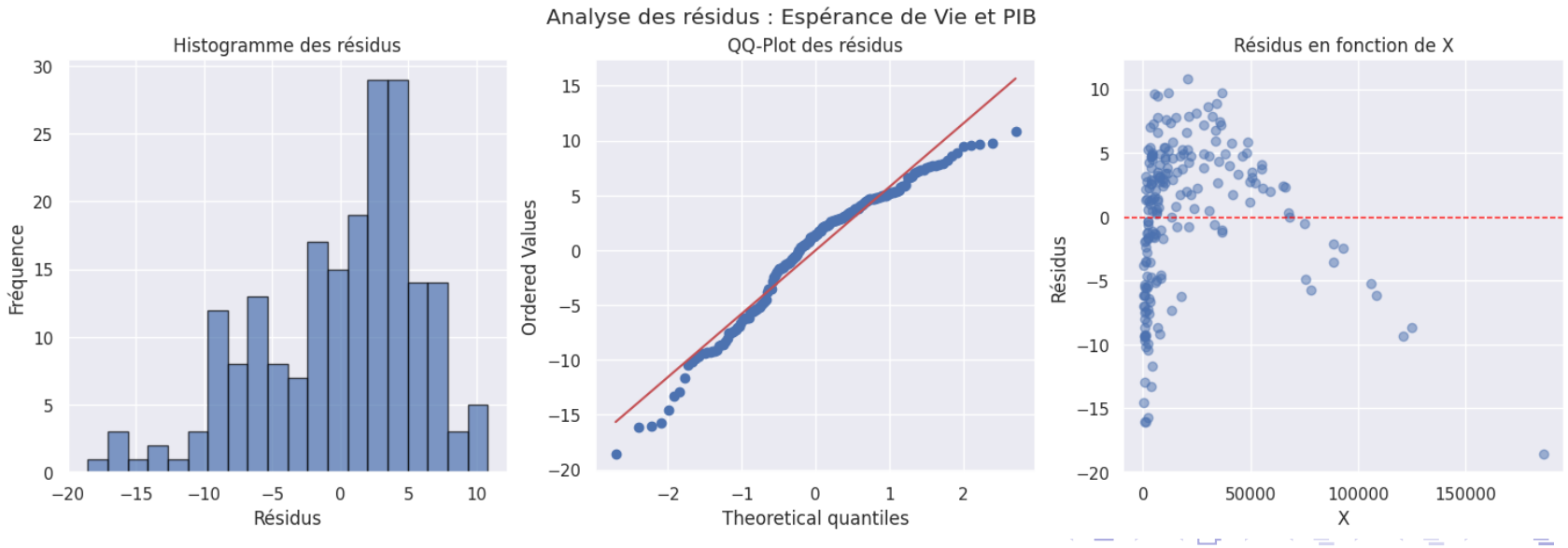


Analyse des Résidus : Explication des Graphiques

- **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.
- **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.

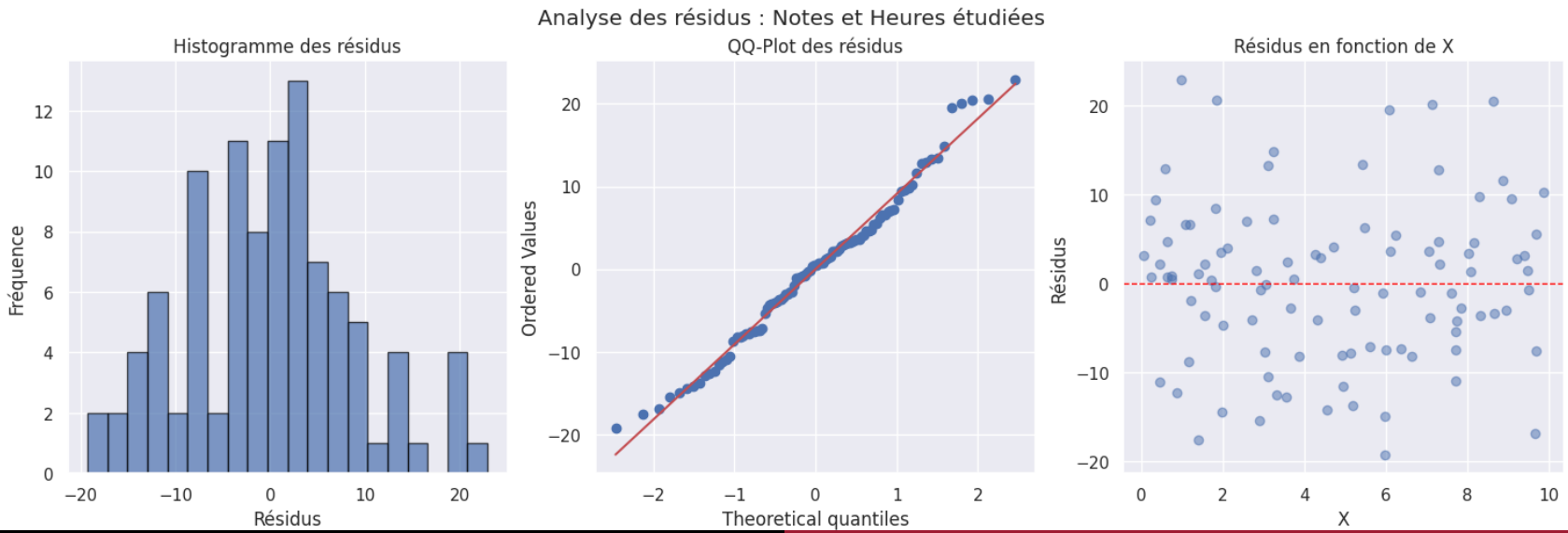
Analyse des Résidus : Résultats Obtenus





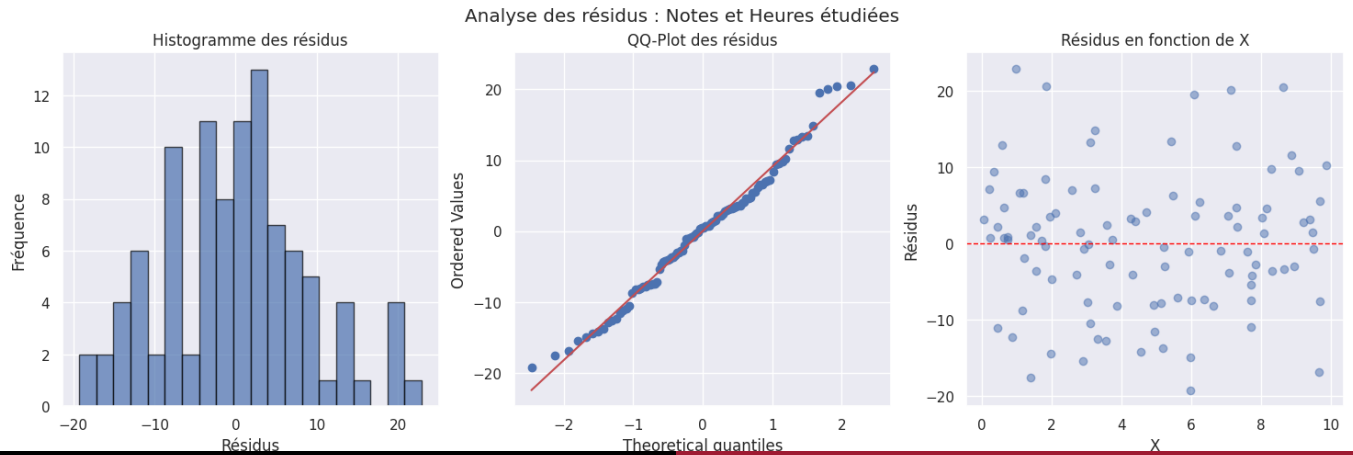
Analyse des Résidus : Notes et Heures Étudiées

- **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).



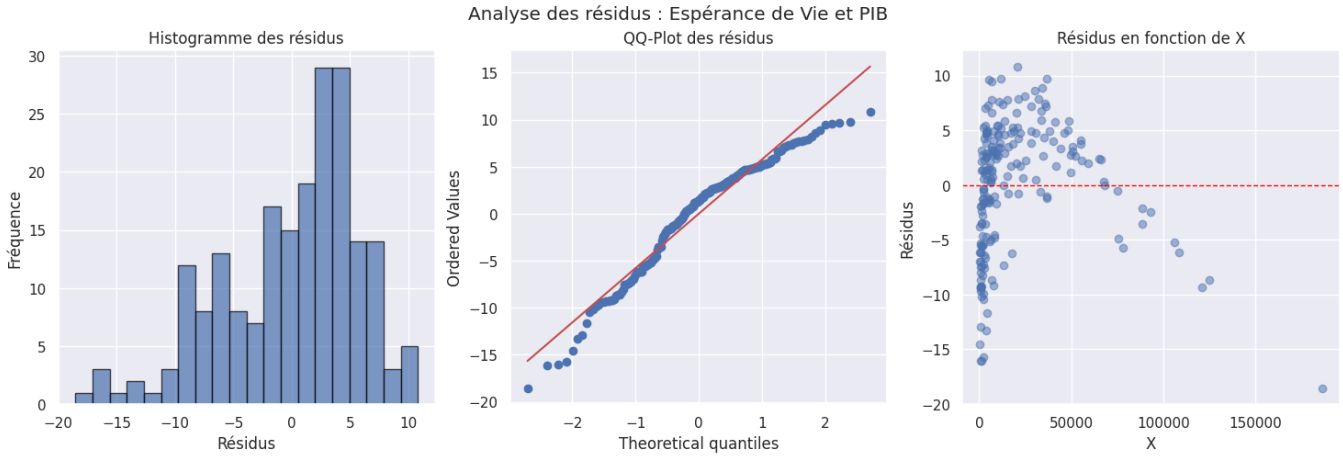
Analyse des Résidus : Notes et Heures Étudiées

- **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.



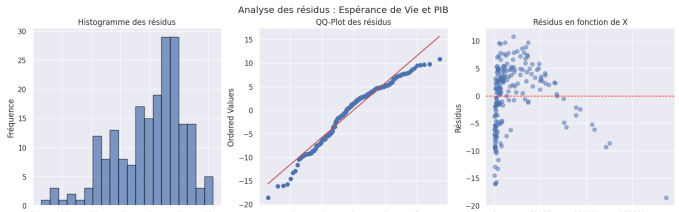
Analyse des Résidus : Espérance de Vie et PIB

- **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.
- **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

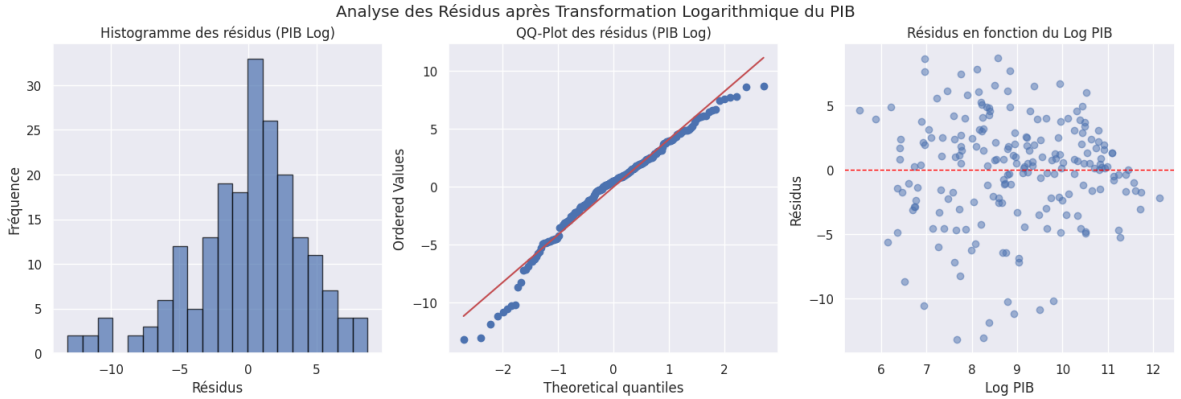
```python
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
# Charger les données
dossier = "Colab Notebooks"
nom_fichier = "Esperance_vie_pib.csv"
chemin_fichier = 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
# Visualisation des résidus
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
```

Transformation Logarithmique des Données : Code Python

```python
# 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}")
```

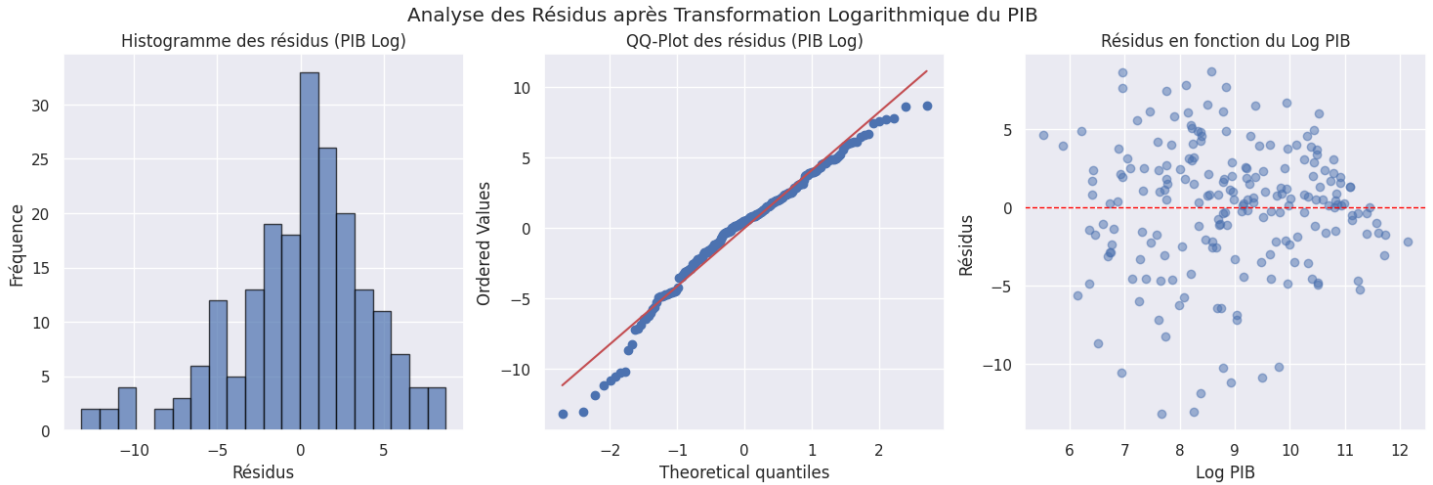
Analyse des Résidus : Transformation Logarithmique

- **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.



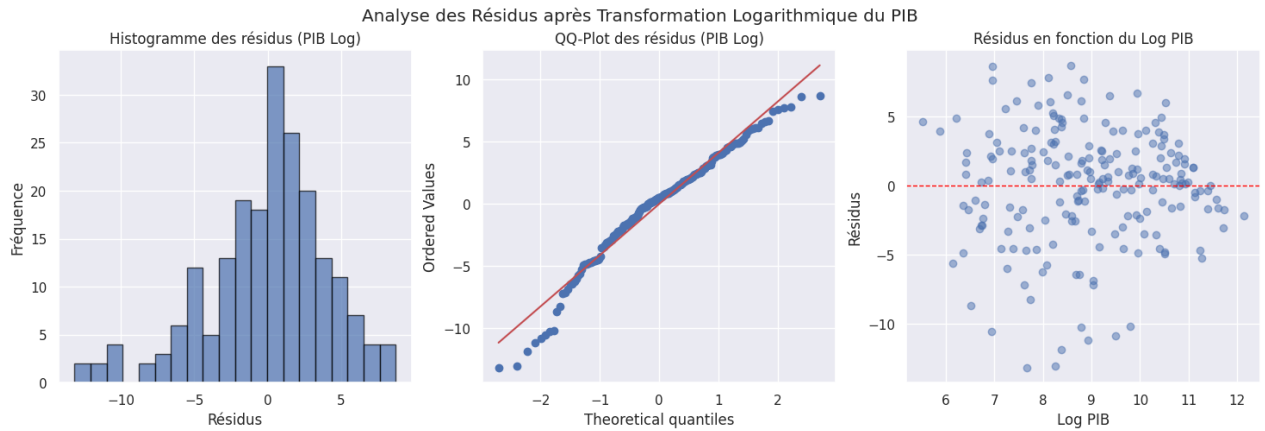
Analyse des Résidus : Transformation Logarithmique

- **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.



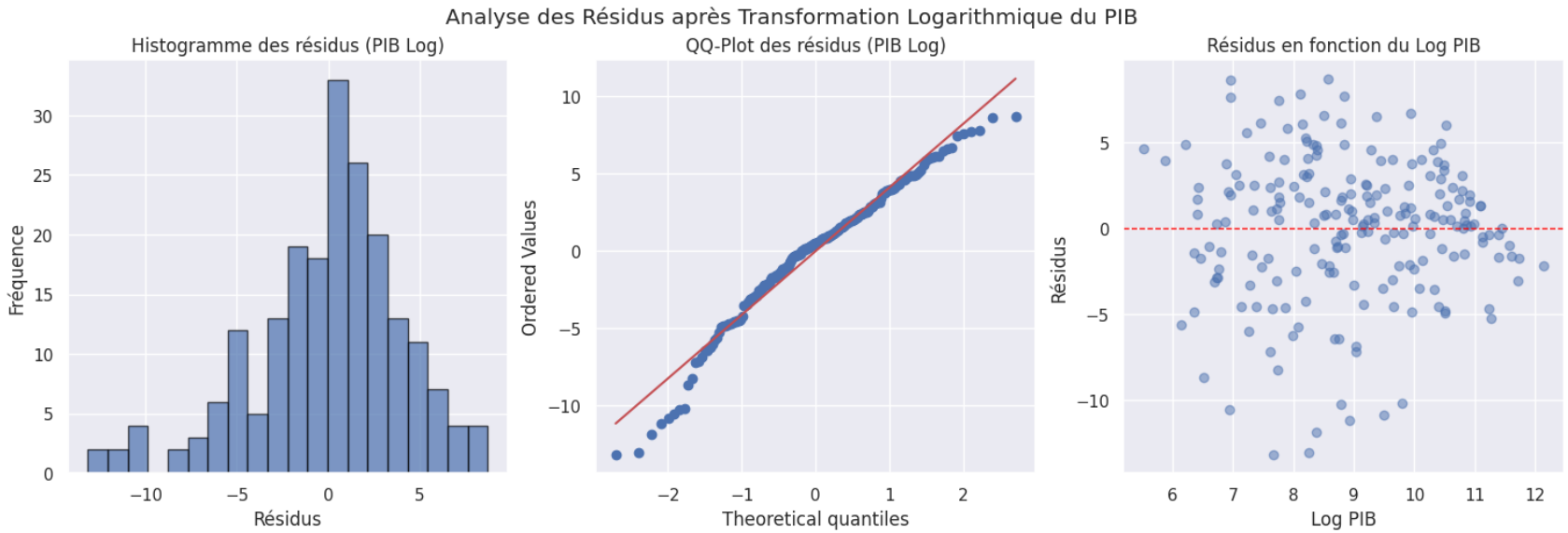
Analyse des Résidus : Transformation Logarithmique

- **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

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 β_0 et les pentes $\beta_1, \ldots, \beta_p$.

Notations en Régression Linéaire

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

Notation en Régression Linéaire

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

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

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

Notations en Régression Linéaire : Exemple Pratique

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

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

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

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

Régression Linéaire Multiple

Expression du Modèle

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

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

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 β): Ajout de termes quadratiques :**

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

- **Modèle Linéaire (en β): 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 β): 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 β): 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

Problème 1 : 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 :

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

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.

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 $\hat{\boldsymbol{\beta}}$ et on l'annule (en sachant que $\mathbf{X}^{\mathrm{T}}\mathbf{X} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}$ toujours symétrique):

$$\frac{\partial C}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}.$$

- En posant cela à zéro, on obtient :

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

Étape 3: Résolution du Système d'Équations

- Nous avons obtenu l'équation normale suivante :

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

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

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

- Cet estimateur est l'estimateur des moindres carrés ordinaires (Ordinary Least Squares - OLS) pour la régression multiple.
- Sous l'hypothèse que les erreurs e_i suivent une distribution normale centrée, $\hat{\boldsymbol{\beta}}$ est un estimateur sans biais et optimal en termes de variance minimale.