# Fiche no. 1 : Modèle linéaire simple

## 1 Intuition graphique

L'intuition graphique est souvent utilisée pour illustrer les relations entre variables dans un modèle de régression. Dans un graphique simple, on peut représenter les données sous forme de nuages de points. Le modèle de régression cherche alors à ajuster une droite (dans le cas d'une régression linéaire) qui minimise l'écart entre les points observés et les valeurs prédites. Cette approche visuelle permet de mieux comprendre la manière dont les variables sont liées et aide à vérifier la pertinence d'un modèle.

## 2 Vocabulaire

Le vocabulaire associé à la régression est crucial pour bien comprendre les concepts sous-jacents. Les termes principaux incluent :

- **Variable dépendante** : Celle que l'on cherche à expliquer ou prédire.
- **Variable(s) explicative(s)** : Ce sont les variables utilisées pour expliquer ou prédire la variable dépendante.
- **Terme d'erreur** : La différence entre la valeur observée et la valeur prédite par le modèle.
- **Estimateur** : Une méthode pour estimer les paramètres inconnus du modèle à partir des données disponibles.

## 3 Modèle de régression simple et propriété du terme d'erreur

### 3.1 Modèle de régression simple

Le modèle de régression simple est une approche statistique qui cherche à modéliser la relation entre une variable dépendante $y$ et une variable explicative $x$. La forme générale du modèle est :

$$y = \beta_0 + \beta_1 x + \varepsilon$$

où $\beta_0$ est l'ordonnée à l'origine, $\beta_1$ est le coefficient de régression, et $\varepsilon$ est le terme d'erreur. Ce modèle permet d'analyser comment la variable $x$ influence $y$ de manière linéaire.

### 3.2 Propriétés du terme d'erreur

Le terme d'erreur $\varepsilon$ représente les facteurs non observés ou non mesurés qui influencent la variable dépendante. Il est supposé répondre à certaines propriétés, comme :

- **Espérance nulle** : $\mathbb{E}(\varepsilon) = 0$, ce qui signifie que, en moyenne, l'erreur ne biaise pas les prédictions du modèle.

- **Variance constante** (homoscédasticité) : Les erreurs ont une variance constante à travers les observations.
- **Indépendance** : Les erreurs ne doivent pas être corrélées entre elles.

# 4 Les Moindres Carrés Ordinaires (MCO)

## 4.1 Principe des MCO

Les Moindres Carrés Ordinaires (MCO) sont une méthode d'estimation des paramètres du modèle de régression. L'objectif des MCO est de minimiser la somme des carrés des erreurs (écarts entre les valeurs observées et les valeurs prédites). Cela permet d'obtenir des estimateurs des coefficients qui rendent l'écart entre les données et le modèle aussi faible que possible.

## 4.2 Estimateurs MCO

Les estimateurs MCO sont obtenus en résolvant un système d'équations qui minimise la somme des carrés des résidus. Matériellement, on résout :

$$\hat{\beta} = (X'X)^{-1}X'y$$

où $\hat{\beta}$ est le vecteur des estimateurs des paramètres, $X$ est la matrice des variables explicatives, et $y$ est le vecteur des variables dépendantes.

## 4.3 Propriétés des estimateurs

Les estimateurs MCO ont plusieurs propriétés importantes :

- **Non-biaisé** : $\mathbb{E}(\hat{\beta}) = \beta$, ce qui signifie que, en moyenne, les estimateurs sont égaux aux vrais paramètres.
- **Efficacité** : Les estimateurs MCO sont les meilleurs estimateurs linéaires non biaisés sous les hypothèses classiques (homoscédasticité et absence d'autocorrélation).
- **Consistance** : Les estimateurs convergent vers les vrais paramètres à mesure que le nombre d'observations augmente.

# 5 Application : la relation entre le prix du pétrole et le DJI

## 5.1 Exposition du problème

Dans cette application, l'objectif est d'étudier la relation entre le prix du pétrole et l'indice boursier DJI (Dow Jones Industrial Average). Le modèle cherche à déterminer si des variations du prix du pétrole influencent le mouvement du marché boursier.

## 5.2 Collecte des données

Les données utilisées pour cette analyse incluent les prix historiques du pétrole et les valeurs quotidiennes du DJI sur une période donnée. Les données peuvent être collectées à partir de bases de données financières ou d'agences spécialisées.

## 5.3 Statistiques descriptives

Avant de modéliser la relation entre les deux variables, il est important de réaliser une analyse descriptive des données. Cela inclut des calculs de moyenne, de médiane, d'écart-type, ainsi que la représentation graphique des séries temporelles pour observer les tendances et les relations apparentes.

## 5.4 Estimation du modèle et interprétation

Une fois les données collectées, un modèle de régression simple peut être estimé pour analyser l'impact du prix du pétrole sur le DJI. L'estimation du modèle fournit les coefficients qui permettent d'évaluer l'effet du prix du pétrole sur l'indice boursier. L'interprétation des résultats inclut l'analyse de la signification statistique des coefficients (p-value) et la vérification de l'adéquation du modèle via les résidus et autres tests diagnostics.

# 6 Hypothèses classiques du modèle de régression linéaire

Les hypothèses classiques sous-jacentes aux modèles de régression linéaire sont les suivantes :

- **Linéarité** : La relation entre la variable dépendante $y$ et les variables explicatives $x_1, x_2, \ldots, x_k$ est linéaire. Le modèle de régression peut être écrit sous la forme :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- **Indépendance des erreurs** : Les erreurs $\varepsilon$ sont supposées indépendantes entre elles, c'est-à-dire qu'il n'y a pas de corrélation entre les termes d'erreur des différentes observations.

- **Espérance nulle des erreurs** : Les erreurs ont une espérance nulle, $\mathbb{E}(\varepsilon) = 0$, ce qui signifie que le modèle est en moyenne bien spécifié.

- **Homoscedasticité** : La variance des erreurs est constante à travers les observations, soit $\mathbb{V}(\varepsilon) = \sigma^2$, ce qui signifie que l'ampleur de l'erreur est la même pour toutes les valeurs des variables explicatives.

- **Absence de multicolinéarité parfaite** : Les variables explicatives ne doivent pas être parfaitement corrélées entre elles, ce qui garantirait l'inversibilité de la matrice des variables explicatives $X$.

- **Normalité des erreurs (facultative pour l'estimation mais nécessaire pour les tests statistiques)** : Les erreurs suivent une distribution normale, $\varepsilon \sim N(0, \sigma^2 I)$, ce qui est essentiel pour effectuer des tests de significativité.

# 7 Interprétation des sorties économétriques de `statsmodels`

Lors de l'estimation d'un modèle de régression avec `statsmodels`, les principales sorties sont les suivantes :

- **Coefficients ($\hat{\beta}$)** : Ils représentent l'effet estimé de chaque variable explicative sur la variable dépendante. Un coefficient $\hat{\beta}_1 = 2.5$ pour $x_1$ signifie qu'une augmentation de 1 unité de $x_1$ entraîne une augmentation de 2.5 unités de $y$, toutes choses égales par ailleurs.

- **Erreur standard des coefficients** : L'erreur standard mesure la précision de l'estimation des coefficients. Une petite erreur standard indique une estimation plus précise.

- **Valeur p ($p$-value)** : Elle permet de tester l'hypothèse nulle selon laquelle le coefficient associé est égal à zéro. Si $p$ est inférieur à 0.05, on rejette l'hypothèse nulle et conclut que le coefficient est significatif.

- **$R^2$ et $R^2$ ajusté** : $R^2$ mesure la proportion de la variance expliquée par le modèle. Un $R^2$ élevé indique un modèle qui explique bien les données. Le $R^2$ ajusté corrige cette mesure en fonction du nombre de variables explicatives.

- **Test F** : Il teste la significativité globale du modèle, c'est-à-dire si l'ensemble des variables explicatives est significatif pour expliquer la variable dépendante. La valeur $p$ du Test F permet de conclure sur cette significativité.

- **Intervalle de confiance des coefficients** : L'intervalle de confiance indique la plage dans laquelle la véritable valeur des coefficients se trouve avec un certain niveau de confiance, typiquement 95%.

- **Durbin-Watson** : Ce test évalue l'autocorrélation des erreurs. Une valeur proche de 2 indique qu'il n'y a pas d'autocorrélation significative des erreurs.