

Définitions

- **La régression linéaire** (analyse bivarié) concerne les variables quantitatives. C'est une étude statistique qui vise à modéliser la relation entre une variable dépendante quantitative et une ou plusieurs variables indépendantes quantitative(s). Elle est utilisée pour traiter les données afin de prendre une décision.(on doit avoir au moins une corrélation pour l'appliquer).
- **régression simple vs régression multiple** : la régression simple utilise une seule variable indépendante (revenue par exemple), tandis que la régression multiple utilise plusieurs variables indépendantes (revenue, age, situation familiale...etc) pour prédire une variable dépendante(dépenses). Le choix entre les deux dépend des objectifs de l'analyse, de la complexité de la relation étudiée et des données disponibles.
- **Analyse bivarié:** méthode statistique pour étudier la relation entre deux variables, notamment comment les variations d'une variable sont associés aux variations d'une autre variable, afin d'explorer leur corrélation, leur dépendance, ou leur association.
- **La statistique:** discipline mathématique qui utilise des méthodes statistiques pour collecter, analyser et interpréter des données en vue d'étudier des phénomènes.
- **Les statistiques:** Ensemble de mesures numériques qui décrivent les caractéristiques d'un ensemble de données, telles que la moyenne, l'écart-type, la variance, le coefficient de corrélation, etc.
- **Graphe:** interprétation / visualisation de données pour données une première idée sur le comportement de la variable.
- **Degré de concordance:** Est une mesure qui permet d'évaluer la similitude entre deux ensembles de données ou entre deux variables. Il permet de déterminer si les observations d'un ensemble de données correspondent (ou cohérents) avec les observations d'un autre ensemble de données.
- **Le résidu** : est la différence entre les valeurs réelles de la variable dépendante et les valeurs prédites par le modèle de régression. Ils mesurent l'erreur ou la variation non expliquée par le modèle.

- **Les tests d'hypothèse:** sont des outils statistiques utilisés pour évaluer l'importance statistique des coefficients de régression du modèle.
- **Le coefficient de détermination (R^2) :** est une mesure qui fournit une indication de la qualité d'ajustement d'un modèle de régression linéaire. Il représente la proportion de la variation totale de la variable dépendante qui est expliquée par le modèle.
- **La corrélation linéaire:** C'est une statistique qui synthétise l'importance de la relation entre deux variables métriques quantitatives.

⇒ Une corrélation proche de 1 ou de -1 en valeur absolue signifie que deux variables sont liées entre elles et peuvent s'expliquer mutuellement. Lorsque r est proche de 0, il y a une faible corrélation. Si r est proche de $+1$, cela veut dire que les deux variables varient dans le même sens. Si r est proche de -1 , cela signifie que les deux variables varient en sens inverse l'une de l'autre.

- **Autocorrélation :** (corrélation temporelle ou corrélation sérielle), est une mesure de la corrélation entre les observations d'une même variable dans une série temporelle ou une séquence ordonnée. Elle examine si les observations successives d'une variable sont linéairement dépendantes les unes des autres.
- **Outlier :** (observation atypique, également appelée valeur aberrante) est une observation qui diffère considérablement des autres observations de l'échantillon. Elle peut être située loin de la tendance générale des données et avoir un effet disproportionné sur l'ajustement du modèle de régression.
- **Homoscédasticité :** L'homoscédasticité signifie que la dispersion des résidus d'un modèle statistique est constante à travers l'ensemble des valeurs prédites de la variable dépendante. Cela indique que la variance des erreurs est la même quel que soit le niveau des variables indépendantes.
- **Modalité :** Une modalité est une valeur spécifique ou une catégorie d'une variable qualitative. Elle représente une option ou une observation particulière au sein d'une variable.
- **Variable muette** (ou variable factice) : Une variable muette, également appelée variable factice, est utilisée dans l'analyse statistique pour représenter des variables catégorielles avec plusieurs modalités. Elle est codée en utilisant des valeurs binaires (0 ou 1) pour chaque modalité, avec

une modalité de référence définie. Les variables muettes sont utilisées pour capturer l'effet des différentes modalités sur une variable dépendante dans les modèles statistiques.

- **Loi de Student** : La loi de Student est une distribution statistique utilisée pour effectuer des tests d'hypothèses et construire des intervalles de confiance lorsque l'écart type de la population est inconnu et doit être estimé à partir d'un échantillon.
- **Loi de Fisher** : La loi de Fisher est une distribution statistique utilisée dans l'analyse de la variance (ANOVA) et les tests de comparaison de plusieurs moyennes. Elle est utilisée pour comparer la variabilité entre différents échantillons en se basant sur le rapport des variances. La loi de Fisher est également utilisée pour évaluer la significativité globale d'un modèle de régression.
- **Le coefficient de variation CV** : est une mesure utile pour évaluer l'homogénéité ou l'hétérogénéité d'une distribution en prenant en compte à la fois la moyenne et l'écart-type des données.