Annales

Table des matières

1 2024

1.1 Comment représenter graphiquement la distribution de la variable « niveau socio-économique » codée de la façon suivante : agriculteur, cadre, profession intermédiaire, commerçant, employé, ouvrier, sans emploi, autre.

- Le plus efficace = diagramme en bâtons, en prenant bien soin de présenter au mieux les différentes modalité pour que la lisibilité soit optimale et si besoin d'inclure une modalité « données manquantes »
- Camembert : possible mais peu recommandé, car moins lisible et moins précis qu'un diagramme en bâtons

1.2 Dans une étude épidémiologique vous mettez en évidence une corrélation entre l'IMC (indice de masse corporelle) et la tension artérielle moyenne égale à r = 0,22 avec un « p » égal à 0,0012. Un reviewer vous fait remarquer que le pourcentage de variance partagé par ces deux variables, égal à 0,222 = 0,0484, est très faible et donc que cette relation est négligeable. Que lui répondez-vous ?

- La corrélation de Pearson (r) mesure la force et la direction d'une relation linéaire entre deux variables continues. Un r de 0,22 indique une corrélation positive faible entre l'IMC et la tension artérielle moyenne.
- Dans le cas d'une régression linéaire simple (donc à une seule variable explicative), le carré du coefficient de corrélation (r^2) représente le pourcentage de variance dans la variable dépendante (tension artérielle moyenne) qui peut être expliqué par la variable indépendante (IMC). Dans ce cas, un r^2 de 0,0484 signifie que seulement 4,84 % de la variance dans la tension artérielle moyenne peut être expliquée par l'IMC.
- Le p-value de 0,0012 indique que cette corrélation est statistiquement significative, ce qui signifie qu'il y a une faible probabilité que cette relation soit due au hasard.
- Le pourcentage de variance partagé (r^2) de 4,84 % indique que seulement une petite partie de la variance dans la tension artérielle moyenne peut être expliquée par l'IMC. Cependant, cela ne signifie pas nécessairement que la relation est négligeable. Même une faible corrélation peut être cliniquement significative, surtout si elle est cohérente avec d'autres recherches ou si elle a des implications pratiques importantes.
- Il est important de considérer le contexte clinique et les implications pratiques de cette relation, plutôt que de se concentrer uniquement sur la force de la corrélation ou le pourcentage de variance partagé.

Vraie correction :

L'interprétation de la force d'une association quand celle-ci est représentée par un coefficient de corrélation (de Pearson).

Il n'y a pas de consensus sur ce point dans la littérature. Dans certaines disciplines, comme en économétrie, voire en génétique dans le domaine biomédical, il est effectivement habituel de discuter en termes de pourcentage de variance expliquée ou partagée.

C'est cependant critiqué, notamment du fait qu'un pourcentage de variance expliqué dépend fortement :

- 1/ de l'échantillonnage de l'étude (un échantillon homogène conduira à une faible variance phénotypique et donc à des pourcentages de variance expliquée qui seront également faibles),
- 2/ de l'importance des erreurs de mesure et du bruit (quand ce dernier est important, puisque par définition il ne peut pas être expliqué les pourcentages de variance partagées seront faibles).

Le coefficient r lui-même n'est pas simple à interpréter, dans la littérature il est souvent considéré qu'un $r < 0,2$ est plutôt faible, mais il s'agit d'un point de vue purement subjectif.

Notamment parce que l'importance clinique et de santé publique de la relation va influer sur le caractère négligeable ou pas de cette dernière.

1.3 A quoi servent les tests statistiques ?

- En médecine (et en recherche biomédicale en général), les tests statistiques sont utilisés pour analyser des données et tirer des conclusions sur des populations à partir d'échantillons. Ils permettent de déterminer si les observations faites dans un échantillon sont suffisamment fortes pour être généralisées à une population plus large.
- Il s'agit d'un processus inférentiel, c'est à dire qu'on utilise les données d'un échantillon pour faire des inférences sur une population.
- Tout résultat tiré d'un échantillon présente une incertitude quant à sa généralisation à la population. Les tests statistiques aident à quantifier cette incertitude en fournissant des mesures telles que les p-values et les intervalles de confiance.

1.4 Comment vérifier les conditions de validité d'une régression linéaire ?

- Linéarité : La relation entre les variables indépendantes et dépendantes doit être linéaire. Cela peut être vérifié en examinant les graphiques de dispersion des résidus.
- Homoscédasticité : La variance des résidus doit être constante à travers toutes les valeurs des variables indépendantes. Cela peut être vérifié en traçant les résidus.
- Normalité des résidus : Les résidus doivent suivre une distribution normale. Cela peut être vérifié en utilisant des tests de normalité (comme le test de Shapiro-Wilk) ou en examinant les graphiques Q-Q des résidus.
- Indépendance des résidus : Les résidus doivent être indépendants les uns des autres, mais c'est beaucoup plus difficile à vérifier.

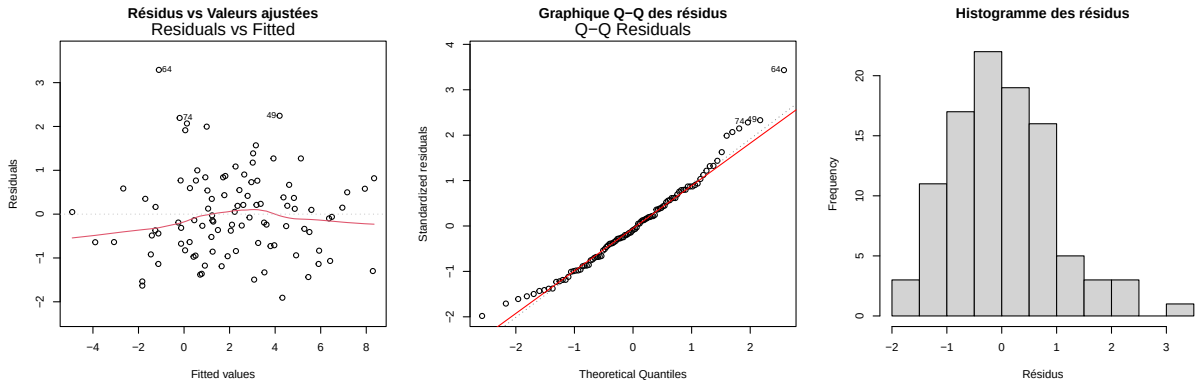

Figure 1: Vérification des conditions de validité d’une régression linéaire