

Chapitre 08 – Cours

Statistiques bivariées, ajustement et causalité

Dans ce chapitre, tu vas apprendre à :

- Calculer et interpréter un ajustement
- Distinguer corrélation, prédiction et causalité

Comprendre les notions

Deux variables sur les mêmes individus

Définition – Notion de départ

Une série bivariée est un ensemble de couples $(x_i; y_i)$. Le nuage de points permet d'observer la forme de la relation et la présence de valeurs atypiques. Le point moyen $G(\bar{x}; \bar{y})$ résume le centre du nuage mais ne décrit pas à lui seul sa dispersion. Les axes, les unités et l'échelle doivent être identifiés avant toute interprétation.

Ajustement affine

Propriété – Lecture et conséquences

Quand le nuage est approximativement rectiligne, on peut chercher $y = ax + b$. La droite des moindres carrés minimise la somme des carrés des écarts verticaux. Pour une variance des x non nulle,

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad b = \bar{y} - a\bar{x}.$$

Elle passe par le point moyen. Le coefficient de corrélation linéaire mesure une association linéaire entre moins un et un ; il ne mesure pas toutes les formes de dépendance.

Prédire avec prudence

Propriété – Lecture et conséquences

Une interpolation concerne une valeur située dans le domaine des abscisses observées. Une extrapolation le dépasse : elle suppose que le modèle continue d'être pertinent. Pour un ajustement exponentiel $y = Ce^{ax}$, le changement de variable $z = \ln y$ donne $z = ax + \ln C$, à condition que les y soient positifs.

Première lecture : application et causalité

Propriété – Lecture et conséquences

Pour les points $(0; 1)$, $(1; 3)$, $(2; 5)$, G vaut $(1; 3)$ et la droite exacte est $y = 2x + 1$. Avec des données réelles, un ajustement comporte en général des résidus et ne crée pas une égalité exacte. Une association entre deux grandeurs ne suffit jamais à établir que l'une cause l'autre ; une variable cachée ou le choix de l'échantillon peut expliquer le lien apparent.

Résidu et qualité du modèle

Définition – Résidu et qualité du modèle

Pour un ajustement $\hat{y}_i = ax_i + b$, le résidu est l'écart vertical entre la donnée et sa valeur prédite. Les moindres carrés choisissent la droite minimisant la somme des carrés de ces résidus.

$$e_i = y_i - \hat{y}_i, \quad Q(a, b) = \sum_i e_i^2.$$

Une suite de résidus présentant une forme courbe révèle qu'un ajustement affine peut être insuffisant. Une corrélation proche de zéro n'exclut pas une relation non linéaire, comme une relation quadratique symétrique.

Deux exemples résolus

Exemple résolu – Calculer une droite de régression

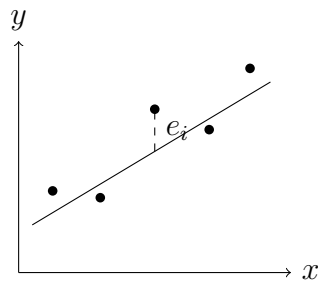
1. Pour les données $(0; 1)$, $(1; 2)$, $(2; 4)$, on obtient $\bar{x} = 1$ et $\bar{y} = 7/3$. La somme des carrés centrés des x vaut deux.
2. La somme des produits centrés vaut trois : $a = 3/2$, puis $b = 7/3 - 3/2 = 5/6$.

3. La droite prédit $7/3$ pour $x = 1$. Les résidus sont $1/6, -1/3, 1/6$; leur somme est nulle, mais aucun argument n'en fait des erreurs de mesure nécessairement aléatoires.
-

Exemple résolu – Linéariser une évolution

1. Les valeurs observées sont $(0; 2), (1; 4), (2; 8)$. Les logarithmes des ordonnées valent $\ln 2, 2 \ln 2$ et $3 \ln 2$.
 2. Les points $(x; \ln y)$ sont alignés sur $z = (\ln 2)x + \ln 2$. En exponentiant, on obtient $y = 2e^{(\ln 2)x} = 2 \times 2^x$.
 3. Pour $x = 1,5$, le modèle donne $4\sqrt{2}$: c'est une interpolation. Pour $x = 5$, il donne soixante-quatre : c'est une extrapolation, qui exige une hypothèse supplémentaire de prolongement du modèle.
-

Une représentation pour comprendre



Un résidu est un écart vertical entre observation et prédiction.

Le nuage et sa droite sont représentés avec un segment vertical matérialisant un résidu. La distance minimisée par les moindres carrés est verticale, car x joue le rôle de variable explicative ; ce n'est pas la distance perpendiculaire à la droite. Inverser les rôles des variables change généralement l'ajustement. Une association statistique peut provenir d'un facteur commun.

Par exemple, les ventes de glaces et les baignades augmentent ensemble en été ; la température peut expliquer ce lien sans que l'achat d'une glace provoque une baignade. Pour parler de causalité, la forme du nuage et le coefficient de corrélation ne suffisent pas.

À retenir – Mini-bilan

Observer le nuage, choisir un modèle et examiner ses résidus. Distinguer interpolation et extrapolation. Une régression décrit une association sur des données ; une relation causale exige d'autres arguments et un contexte adapté.
