Préparation à l'agrégation externe de Sciences Sociales

Statistique inférentielle - Modèle linéaire

2023-2024

Exercice 1

On souhaite mettre en évidence une corrélation entre le temps passé chaque jour devant la télévision (`time_tv`, en heures) et le taux de cholestérol (`cholesterol`, en mmol par litre de sang).

1. On suppose qu'il existe des paramètres $a, b, \sigma \in \mathbb{R}$ tels que les n observations $(\texttt{time\_tv}_i, \texttt{cholesterol}_i)$ réalisées sont des observations du couple (X, Y), où

   $$Y = aX + b + \varepsilon$$

   avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ indépendante de X.
   Tous les résultats qui vont suivre, y compris la pertinence des grandeurs données dans la sortie logicielle, sont conditionnés à la validité de ce modèle.

2. Les estimateurs des moindres carrés

   $$\hat{a} = \frac{\sum_{i=1}^{n}(X_i - \overline{X_n})(Y_i - \overline{Y_n})}{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}$$

   et

   $$\hat{b} = \overline{Y_n} - a\overline{X_n}$$

   sont des estimateurs normaux, sans biais et convergents des paramètres respectifs a et b.

3. Les estimations de a et b par $\hat{a}$ et $\hat{b}$ se trouvent dans la première colonne du tableau inférieur.
   Le coefficient qui nous intéresse particulièrement est a : en effet, c'est lui qui permet de dire si le nombre d'heures passées devant la télévision chaque jour a ou non un effet sur le taux de cholestérol, et le cas échéant d'en donner le signe. En examinant l'écart-type associé à l'estimation, on voit que celui-ci est quatre fois plus petit que ladite estimation : même en tenant compte de l'imprécision de l'estimation, le véritable paramètre a a de grandes chances d'être strictement positif. Plus précisément, l'intervalle de confiance pour a donné dans la dernière colonne (double) du tableau inférieur permet de dire avec 95% de certitude que a est positif, et donne une idée des valeurs possibles de a.
   Si l'on s'intéresse uniquement à l'existence d'un effet de `time_tv` sur `cholesterol` sans chercher à quantifier cet effet, l'outil pertinent est la p-value du test de l'hypothèse $H_0 : a = 0$ contre $H_1 : a \neq 0$. C'est le nombre représenté à la première ligne de la quatrième colonne du tableau inférieur : le fait qu'il soit très faible indique que l'on peut affirmer avec un niveau de certitude proche de 1 que $a \neq 0$ et, pourtant, que `time_tv` a bien un effet sur `cholesterol`.

4. Le R^2 de la régression est relativement faible. Cela peut être dû à l'existence de variables explicatives pertinentes autres que `time_tv` (par exemple l'alimentation, la pratique sportive, l'âge, le patrimoine génétique...) ou au fait que l'effet de `time_tv` sur `cholesterol` n'est pas véritablement linéaire. Une piste privilégiée dans un premier temps est donc d'ajouter des variables explicatives, ce qui augmente mécaniquement la valeur du R^2 de la régression.

Exercice 2 (2009)

1. Étude des séries statistiques

   (a)

$$\begin{aligned} \bar{y} &= \frac{1}{13}\sum_{i=1}^{13} y_i = 183,42 \\ \mathrm{Var}(y) &= \frac{1}{13}\sum_{i=1}^{13} y_i^2 - \bar{y}^2 = 48602 \end{aligned}$$

   (b)

$$\begin{aligned} \bar{x} &= \frac{1}{13}\sum_{i=1}^{13} x_i = 286,85 \\ \mathrm{Var}(x) &= \frac{1}{13}\sum_{i=1}^{13} x_i^2 - \bar{x}^2 = 27638 \end{aligned}$$

2. Le coefficient de corrélation de x et y est :

$$\begin{aligned} r_{x,y} &= \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(y)}\sqrt{\mathrm{Var}(x)}} \\ &= \frac{1}{\sqrt{\mathrm{Var}(y)}\sqrt{\mathrm{Var}(x)}}\left(\frac{1}{13}\sum_{i=1}^{13} x_i y_i - \bar{x}\bar{y}\right) \\ &= 0,853 \end{aligned}$$

   Le coefficient de corrélation de x et y étant proche de 1, il est raisonnable de penser que le nombre de salles dans lesquelles le film est projeté est corrélé linéairement au nombre d'entrées réalisé par le film en première semaine d'exploitation.

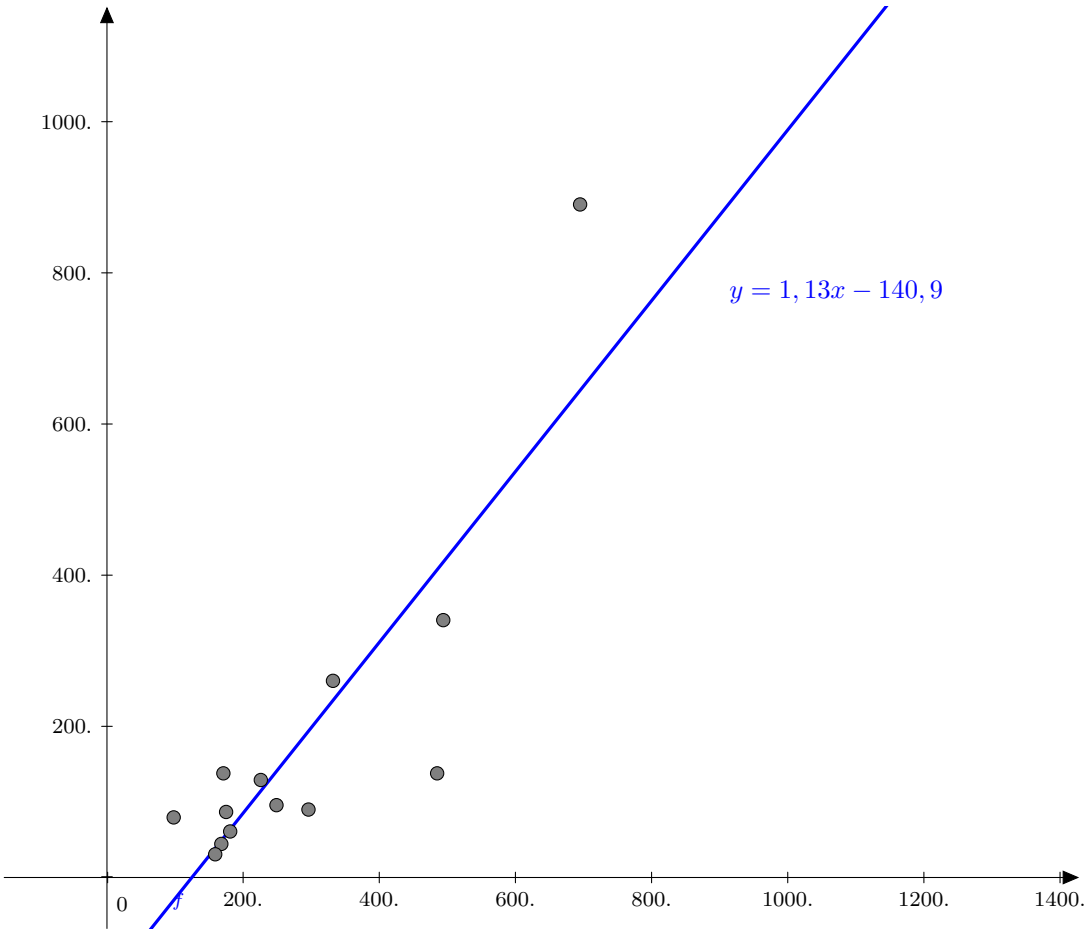
3. Modélisation du lien entre les séries statistiques.

   (a) La méthode des moindres carrés donne les coefficients a et b de la droite de régression $y = \hat{a}x + \hat{b}$.

$$\begin{aligned} \hat{a} &= \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = 1,13 \\ \hat{b} &= \bar{y} - \hat{a}\bar{x} = -140,9 \end{aligned}$$

   Cette droite passe par les points moyens ($\bar{y} = \hat{a}\bar{x} + \hat{b}$) et a et b sont choisis pour minimiser

$$\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2.$$



(b)

$$\begin{aligned} \hat{y}_i &= \hat{a}x_i + \bar{b} \\ &= 407,48 \\ &\gg 138,7 = y_i \end{aligned}$$

Le score du film 5 n'est donc pas conforme aux attentes de son producteur.

4. Analyse de la variance.

(a)

$$\begin{aligned} \text{Var}(\hat{y}) &= \frac{1}{13}\sum_{i=1}^{13}(\hat{a}x_i + \hat{b})^2 - (\hat{a}\bar{x} + \hat{b})^2 \\ &= 35336 \end{aligned}$$

(b) Variance totale = Variance expliquée + Variance résiduelle.
La variance résiduelle est donc égale à $48602 - 35336 = 13266$.

(c) R^2 = Variance expliquée/Variance totale $= 0,727$
Lors de l'établissement d'une équation de régression, le coefficient de détermination R^2 détermine à quel point l'équation de régression est adaptée pour décrire la distribution des points. Si le R^2 est nul, cela signifie que l'équation de la droite de régression détermine 0% de la distribution des points. Cela signifie que le modèle mathématique utilisé n'explique absolument pas la distribution des points. Si le R^2 vaut 1, cela signifie que l'équation de la droite de régression est capable de déterminer 100% de la distribution des points. Cela signifie que le modèle mathématique utilisé, ainsi que les paramètres $\hat{a}$ et $\hat{b}$ calculés sont ceux qui déterminent la distribution des points.
Ici R^2 est suffisamment élevé pour pouvoir considérer que la régression est bien significative.