Nantes Université

Master 1 Ingénierie Mathématique

RÉGRESSION LINÉAIRE ET LOGISTIQUE
FEUILLE 1

Exercice 1. Coefficients de corrélation de Bravais-Pearson et de Spearman

On considère deux échantillons de n variables $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$.

(1) Rappeler la définition de la corrélation linéaire empirique r entre les deux échantillons précédents (corrélation de Bravais-Pearson).

(2) En utilisant la fonction f définie par $f(z) = \sum_{i=1}^n (X_i - zY_i)^2$, montrer que $-1 \leq r \leq 1$.
On note $(r_1, \dots, r_n)$ (resp. $(s_1, \dots, s_n)$) les rangs des variables X_i (resp. Y_i) dans chaque échantillon. On suppose qu'il n'y a pas d'ex-aequo, de telle sorte que les rangs vont de 1 à n. La corrélation de Spearman ρ_S entre les échantillons $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ correspond à la corrélation linéaire entre leurs rangs.

(3) Calculer la moyenne et la variance empirique de l'échantillon $(r_1, \dots, r_n)$.

(4) En déduire que
$$\rho_S = \frac{12 \sum_{i=1}^n r_i s_i - 3n(n+1)^2}{n(n^2-1)}.$$

(5) Soit $d_i = r_i - s_i$. Montrer que $12 \sum_{i=1}^n r_i s_i = 2n(n+1)(2n+1) - 6 \sum_{i=1}^n d_i^2$.

(6) En déduire la valeur de ρ_S en fonction de l'échantillon des différences d_i.

Exercice 2. Application

On considère les données suivantes :

x	1000	800	600	450	300	200	100
y	573	534	495	451	395	337	253

(1) Représenter sur deux graphiques différents y en fonction de x et les rangs de y en fonction des rangs de x. Commenter.

(2) Écrire des fonction `R` permettant de calculer le coefficient de Bravais-Pearson et celui de Spearman. Calculer leur valeur sur les données considérées.

(3) Cet exemple permet d'illustrer une différence entre les deux coefficients précédents, laquelle ? Commenter.

Exercice 3. Régression simple par MCO

On souhaite exprimer la hauteur y d'un arbre en fonction de son diamètre x à 1m30 du sol. Pour cela, on a mesuré 20 couples diamètre-hauteur et les résultats ci-dessous sont disponibles :

$$\overline{x} = 34.9,\ \overline{y} = 18.34,\ \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x})^2 = 28.29,\ \frac{1}{20}\sum_{i=1}^{20}(y_i - \overline{y})^2 = 2.85,\ \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x})(y_i - \overline{y}) = 6.26$$

(1) On note $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ l'estimation de la droite de régression par la méthode des moindres carrés ordinaires. Donner l'expression de $\hat{\beta}_0$ et $\hat{\beta}_1$ en fonction des statistiques élémentaires ci-dessus. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.

(2) Donner une mesure de qualité d'ajustement des données au modèle. Exprimer cette mesure à l'aide des statistiques élémentaires. Calculer et commenter.

Exercice 4. Comparaison de la méthodes MCO et la méthode inverse pour la régression simple

On va illustrer par des simulations les propriétés des estimateurs des coefficients de la régression simple par la méthode MCO et par la méthode inverse.

Partie 1 : On considère deux échantillons de n variables $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$. On suppose que les $(X_1, \ldots, X_n)$ sont connus.
On considère le modèle de régression simple

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

où les ε_i sont i.i.d de moyenne nulle, non corrélées et de variance σ^2.

(1) Rappeler les expressions de $\widehat{\beta}_{0MCO}, \widehat{\beta}_{1MCO}$.

(2) Écrire le modèle de régression inverse, exprimer ses coefficients en fonction de β_0, β_1. En déduire $\widehat{\beta}_{0RI}, \widehat{\beta}_{1RI}$.

(3) Écrire une fonction `R` permettant de calculer les paramètres $\widehat{\beta}_{0MCO}$, $\widehat{\beta}_{1MCO}$ de la régression simple par MCO.

(4) Écrire une fonction `R` permettant de calculer les paramètres $\widehat{\beta}_{0RI}$, $\widehat{\beta}_{1RI}$ de la régression inversée.

(5) Récupérer le vecteur X dans le fichier X.txt. Calculer les estimateurs précédents sur $B = 10000$ échantillons $(Y_i)_i$ définis par $Y_i = 10 + 2X_i + \varepsilon_i$, ε_i sont gaussiennes i.i.d de moyenne nulle, non corrélées et de variance $\sigma^2 = 1$. Tracer les boxplot des estimateurs obtenus (on tracera les vraies valeurs des paramètres). Donner une estimation du biais et de la variance des estimateurs MCO et RI.

(6) Reprendre la question précédente avec $\sigma^2 = 16$.

(7) Commenter les résultats précédents.

Partie 2 : Utilisation de la fonction. `lm` Charger dans `R` le jeu de données `ozone.txt`. Le but est d'expliquer la concentration en ozone (variable `O3` dans le fichier) en fonction de la température (variable `T12` dans le fichier).

(1) Représenter le nuage de points.

(2) Á l'aide de la fonction `lm` ajouter la droite de régression par MCO sur le graphe précédent.

(3) Á l'aide de la fonction `lm` donner la valeur de $\widehat{\beta}_{1RI}$. Ajouter la droite de régression inversée sur le graphe précédent.

EXERCICE 5. UTILISATION DE LA MÉTHODE DES MOINDRES CARRÉS PAR CHANGEMENT DE VARIABLES

Charger les données `Animals` de la librairie `MASS` On cherche à modéliser le poids du cerveau `brain` en fonction du poids des animaux `body`.

(1) Représenter le nuage de points.
(2) Appliquer une transformation logarithmique aux variables et représenter le nouveau nuage de points.
(3) Sur quel nuage de points le modèle de régression simple vous parait-il adapté ?
(4) Estimer le modèle avec la fonction `lm`
(5) Représenter la droite estimée sur le nuage de points.
(6) Pourquoi ce modèle n'est pas adapté aux gros animaux ? Justifier à l'aide du graphique.
(7) Reprendre la construction du modèle de régression après avoir supprimé les trois plus gros animaux.
(8) En déduire une relation entre le poids du cerveau et le poids de l'animal.