Université Claude Bernard Lyon 1
L2 MASS UE 42
Année 2010-2011

Proba-Stat : fiche n° 2

Exercice 1. On considère $n \geq 2$ points (x_i, y_i) et on utilise les notations usuelles pour les moyennes et variances empiriques :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
$$\bar{v}_x = \frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}^2 \qquad \bar{v}_y = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{y}^2$$

On définit également la covariance empirique du nuage de points par

$$\overline{c_{xy}} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})$$

1. Pour deux vecteurs $u = (u_1, \ldots, u_n)$ et $v = (v_1, \ldots, v_n)$ de $\mathbf{R}^n$, prouver l'inégalité de Cauchy-Schwarz : $\langle u, v\rangle \leq \|u\|\|v\|$.
2. Montrer que l'on a
$$\overline{c_{x,y}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
3. À l'aide de la question 1), vérifier que $|\overline{c_{xy}}| \leq \sqrt{\overline{v_x}\,\overline{v_y}}$.
4. Dans quel cas y a-t-il égalité ?

Exercice 2. On considère n points $(X_i, Y_i)_{i\leq n}$, où les Y_i sont supposées de la forme $Y_i = aX_i + b + \epsilon_i$, avec les hypothèses usuelles sur les X_i et les ϵ_i, et on leur associe par la méthode des moindres carrés la droite de régression linéaire d'équation $y = A_n x + B_n$. On définit le résidu du point (X_i, Y_i) par $R_i = Y_i - (A_n X_i + B_n)$.

Quelle est la différence entre les termes d'erreur et les résidus ?

Exercice 3. Dans un modèle de régression linéaire par la méthode des moindres carrés, montrer que la somme des résidus est nulle.

Exercice 4. Le tableau suivant représente le poids et la taille de 10 filles âgées d'environ 6 ans :

Enfant	i	1	2	3	4	5	6	7	8	9	10
Taille (cm)	x_i	101	111	107	106	119	112	105	115	118	114
Poids (kg)	y_i	13	16	15	15	19	17	15	18	19	18

On donne $\overline{x} = 110,8$; $\overline{y} = 16,5$; $\sum x_i^2 = 123082$; $\sum y_i^2 = 2759$ et $\sum x_i y_i = 18388$.

1. Représenter graphiquement les données. Un modèle linéaire vous semble-t-il raisonnable ?
2. Déterminer les coefficients $\hat{a}_n$ et $\hat{b}_n$ de la droite de régression linéaire par la méthode des moindres carrés ordinaires.

3. Calculer et représenter les résidus.
4. Calculer l'erreur quadratique moyenne

$$EQ_n = \frac{1}{n}\sum (y_i - \hat{a}_n x_i - \hat{b}_n)^2.$$

5. Calculer le coefficient $R^2 = \overline{c_{xy}}^2/(\bar{v}_x \bar{v}_y)$.

Exercice 5. Vérifier que, si X et Y sont deux variables aléatoires de loi normale $\mathcal{N}(0,1)$ et indépendantes, alors pour tous a et b, la variable aléatoire $aX + bY$ suit une loi normale. Quels sont ses paramètres ? Généraliser au cas de variables normales indépendantes et de paramètres quelconques.

Exercice 6. *Un autre modèle linéaire.*

On observe des réalisations de couples (X_i, Y_i), où les X_i sont déterministes et les Y_i vérifient $Y_i = aX_i + \epsilon_i$, les variables ϵ_i étant normales, centrées et indépendantes. Autrement dit, on impose à la droite de régression de passer par le point (0,0). On construit un estimateur des moindres carrés pour a, c'est-à-dire que l'on cherche $\tilde{A}$ tel que

$$\sum_{i=1}^{n} (Y_i - \tilde{A} X_i)^2$$

soit minimal. Expliciter $\tilde{A}$.