Examen d'Analyse des données

L'examen se compose de deux parties: la première de 30 mn est personnelle, la seconde de 45 mn s'effectue en binôme et en salle des machines. Le cours, les feuilles de T.P. ainsi que tout autres documentations sur Splus sont autorisés.

Première partie

On considère des données présentant deux caractères X et Y de n individus de même poids. On notera $X = {}^t(x_1, .., x_n)$ et $Y = {}^t(y_1, .., y_n)$, où les x_i et y_i sont des réels. On note encore I_n la matrice identité de $\mathbb{R}^n$ et $U = {}^t(1, 1, .., 1)$. On suppose que X et Y ne sont pas centrées.

1. Donner une écriture matricielle de X_0 et Y_0, obtenus en centrant X et Y, à l'aide uniquement de X, Y, I_n et U.
2. Interpréter la régression par moindres carrés de Y par X comme une projection orthogonale pour la distance euclidienne dans un espace de dimension n. En déduire l'expression matricielle du coefficient linéaire a de la droite de régression à l'aide uniquement de X, Y, I_n et U (on utilisera l'expression matricielle d'un projecteur orthogonal).
3. Donner les coordonnées d'un vecteur du premier axe principal de l'ACP réduite du tableau composé par X et Y.

Seconde partie

1. On a interrogé 4 femmes et 4 hommes. On a obtenu le tableau de données suivant:

$$X = \begin{array}{c} \\ I1 \\ I2 \\ I3 \\ I4 \\ I5 \\ I6 \\ I7 \\ I8 \end{array} \begin{pmatrix} Taille & Poids & Sexe & Etudes & Age \\ 160 & 48 & F & L & 26 \\ 172 & 52 & F & S & 29 \\ 178 & 59 & F & L & 41 \\ 180 & 74 & M & S & 61 \\ 174 & 69 & M & SH & 23 \\ 162 & 51 & F & L & 51 \\ 182 & 72 & M & S & 30 \\ 187 & 82 & M & SH & 36 \end{pmatrix}$$

   (M: Masculin, F: Féminin, S: Scientifique, L: Littéraire, SH: Sciences Humaines). Après avoir implanté ce tableau dans votre *.Data* (le rentrer sous forme d'un *data.frame* en faisant attention à la nature (*factor* ou *vector*) des variables), indiquer les variables vous semblant fortement corrélées entre elles, puis justifier le numériquement.

2. On travaille sur le data.frame **consoprote**, dont une explication est donnée par **?consoprote**. Effectuer une ACP sur ce jeu de données (on précisera en justifiant les options et l'ordre de l'ACP, les contributions aux nouvelles variables, le nombre pertinent et la composition de groupes d'individus pouvant être mis en évidence).