Faculté

des **sciences économiques** et de **gestion**

Université de Strasbourg

Année universitaire 2018/2019

Licence économie-gestion 3ème année

Parcours gestion

Parcours économie

Parcours Economie quantitative + Math Eco + DUAS

Semestre 2 – Session 1 / Contrôle terminal

Avril / Mai 2019

Matière : Analyse de Données
Enseignant : S. Bianchini
Durée : 120 minutes
Document : Non autorisé
Calculatrice : Autorisé

Question 1 [5 points] Différents algorithmes de sélection de variables (*feature selection*) peuvent vous aider à choisir un ensemble de predicteurs pour un modèle de prévision.

1.1 Décrire les méthodes suivantes : (i) recherce exhaustive, (ii) sélection en avant (*forward selection*), (iii) élimination en arrière (*backward selection*). [3 pt]

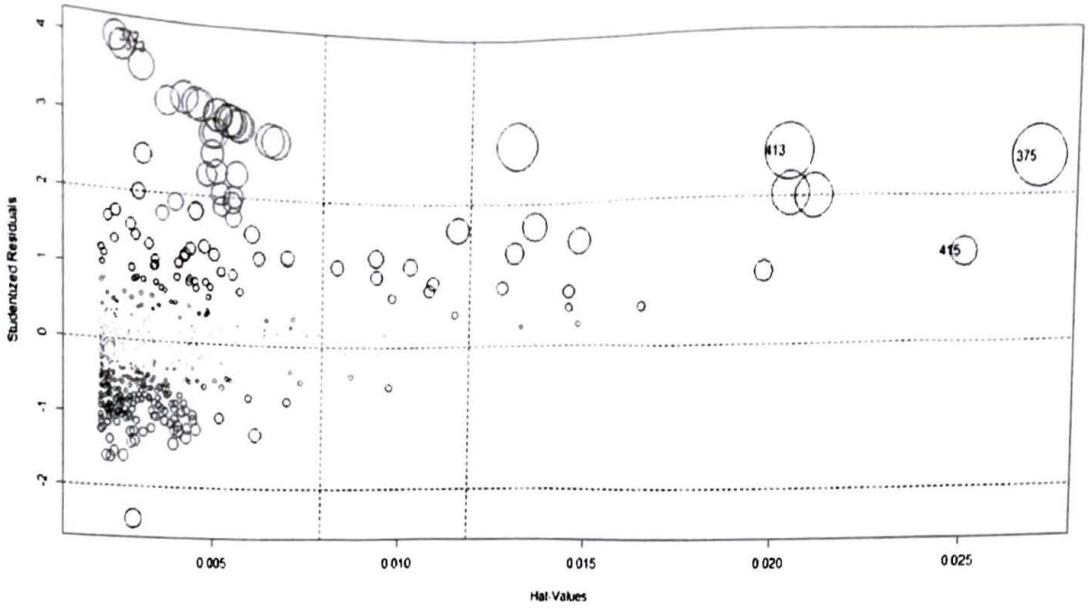
1.2 Laquelle des trois méthodes donnera le R^2 la plus élevée ? Justifier votre réponse. [2 pt]

Question 2 [5 points] Dans le context d'un modèle de regression simple (n = 1,000) :

2.1 Eablir l'intervalle de confiance à 95 % pour le coefficient $\beta_1 = 0.10$ dont l'erreur-type est égale à 0.20. [1 pt]

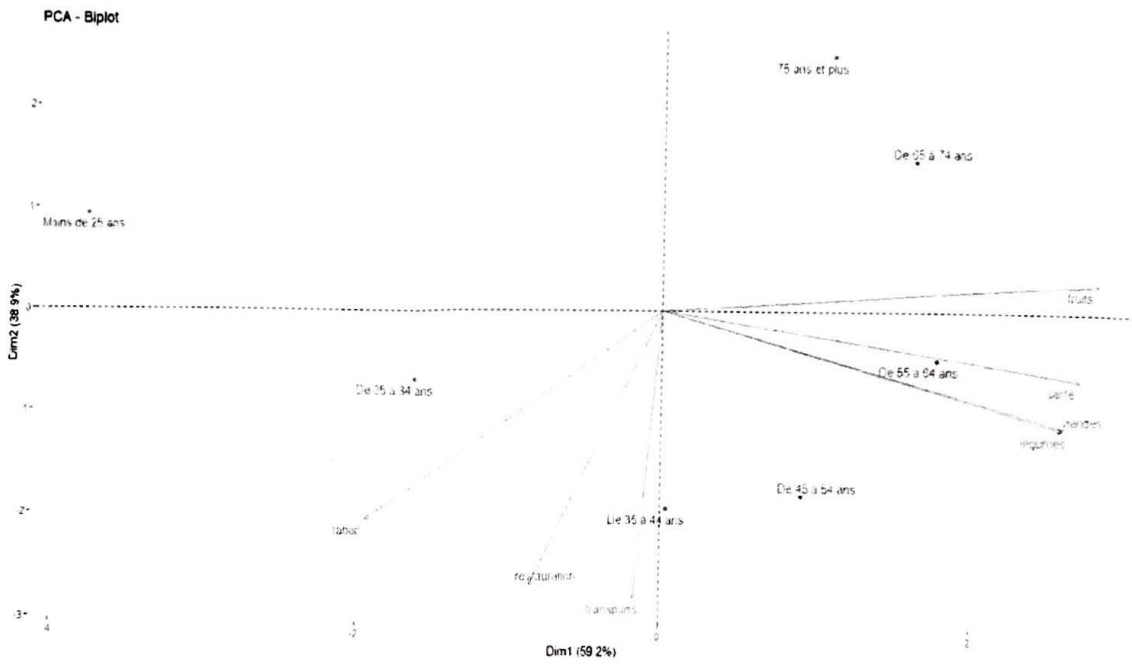
2.2 Pouvez-vous conclure que le prédicteur est positivement et statistiquement lié à la variable de réponse ? [1 pt]

2.3 Décrire la formule permettant de trouver le *score de levier* (hat values) pour la *i*-ème observation, puis commenter le graphique ci-dessous par rapport à la position des observations le long de l'axe horizontal. [3 pt]



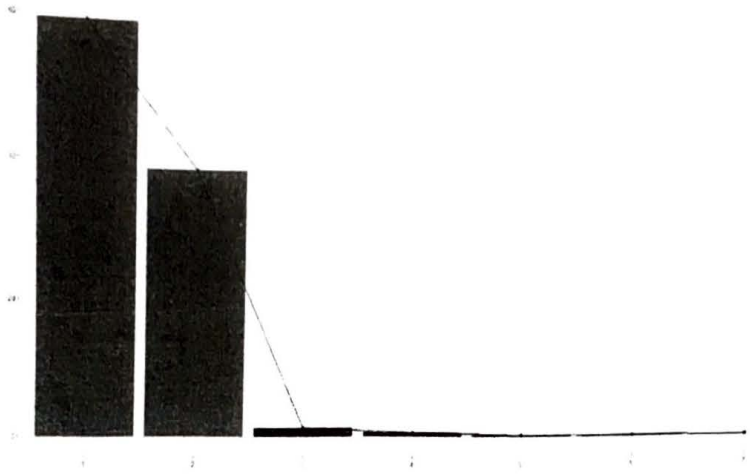
Question 3 [5 points] Vous appliquez une *Analyse en Composantes Principales* (ACP) sur un jeu de données concernant les dépenses des ménages (7 categories) classés par tranches d'âges (8 groupes).

3.1 En regardant le graphique ci-dessous, des tendance se dégagent par rapport aux composantes principales. Quelle signification pouvez-vous attribuer aux première et deuxième composantes principales ? [2 pt]



3.2 Quel comportement caractérise la tranche d'âge « 75 ans et plus » ? Et la tranche d' âge « moins de 25 ans » ? [2 pt]

3.2 Décrivez le *scree plot* ci-dessous. Que représentent les axes des abscisses et des ordonnées? Que pouvez-vous conclure en regardant le graphique ? [1 pt]



Question 4 [5 points] Vous avez quatre observations, pour lesquelles vous avez calculé la matrice de dissimilarité suivante :

$$\begin{bmatrix} & 0.1 & 0.4 & 0.7 \\ 0.1 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.6 \\ 0.7 & 0.8 & 0.6 & \end{bmatrix}$$

Remarque : la dissimilarité entre la première et la deuxième observation est de 0,1, la dissimilarité entre le deuxième et le quatrième est de 0,8, et ainsi de suite.

Sur la base de cette matrice de dissimilarité, esquisser le *dendrogramme* qui résulte du clustering hiérarchique de ces quatre observations en utilisant la methodé *complete linkage* (c.-à-d. dissimilarité inter-clusters maximale). Note : expliquer chaque étape et indiquer sur le graphique la hauteur à laquelle chaque fusion a lieu.