Code Module : 4056 **Intitulé du Module : Analyse de données**

Date : 16 mai 2014 **Durée : 1 heure 30**

Professeur : Mme Bertrand Myriam **Nombre de pages: 9**

Examen: X **Contrôle: ☐** **Classe: Info CFA**

Documents autorisés : Oui ☒ Non ☐
Calculatrice autorisée : Oui ☒ Non ☐
Ordinateur autorisé : Oui ☐ Non ☒

Précision sur le barème si QCM :

Commentaires :

NOM de l'étudiant:

Prénom de l'étudiant:

Code étudiant :

Sujet d'analyse des données

Le sujet comporte trois exercices indépendants. **Il vous est demandé de traiter obligatoirement l'exercice 3 et de faire un choix entre l'exercice 1 et l'exercice 2.**

— Les calculatrices sont autorisées.
— Le cours, les exercices de travaux dirigés, leurs corrigés ainsi que les notes de cours sont autorisés. Tout autre document est interdit
— Afin de pouvoir traiter les questions, plusieurs résultats numériques et graphiques ont été intégrés au document.
— Vous prendrez un soin particulier à préciser quelles sont les hypothèses testées.
— Tous les tests seront effectués au seuil de signification $\alpha = 5\%$.

Exercice 1. Le classement de 40 collèges de la région parisienne.

Une étude est menée dans le but de comparer les scores moyens obtenus à un test de mathématiques, par les élèves de sixième dans quarante collèges de la région Parisienne. Des différences importantes apparaissent d'une classe à une autre. Afin d'améliorer les scores, nous cherchons à déterminer des facteurs influents, concernant dans un premier temps, les enseignants. Nous disposons pour chacun des collèges, de quatre variables :

— Score : score moyen obtenu par les élèves de sixième d'un collège, variable qui est notée $V1$.
— Licence : pourcentage d'enseignants qui ont au moins une licence de mathématiques, variable qui est notée $V2$.
— Age : âge moyen des enseignants, variable qui est notée $V3$.
— Salaire : salaire moyen des enseignants, variable qui est notée $V4$.

1. Nous construisons un modèle linéaire multiple expliquant le score moyen en fonction des trois variables Licence, Age et Salaire. Ce modèle permet-il d'expliquer les variations d'un collège à un autre ? Commenter. Interpréter les coefficients du modèle.
2. Les conditions d'application du modèle de régression linéaire multiple sont-elles vérifiées ?
3. Un problème de colinéarité entre deux variables apparaît. Lequel ? Expliquer. Quel est le signe de l'estimation du coefficient de corrélation entre les estimateurs des coefficients de ces deux variables et pourquoi ce signe ?
4. Supprimer du modèle à trois variables celle qui vous paraît la moins utile en indiquant quelle est cette variable et pourquoi vous choisissez celle-ci. Quel est le nom de cette procédure de choix de modèle ? Construire le nouveau modèle. Les conditions d'application du modèle de régression linéaire multiple sont-elles vérifiées ?
5. Donner, pour le modèle à deux variables, les coefficient de détermination et ajusté. Que pouvez-vous conclure ? Réaliser le test de Fisher pour ce modèle. Pourquoi ne construisez-vous pas un modèle à une variable ?

Voici les sorties réalisées avec le logiciel `R` qui pourront vous aider à répondre aux différentes questions.

```
> exo1<-read.table(file.choose())
> head(exo1)
    V1 V2 V3    V4
1 73.9 77 52 26.10
2 59.4 48 32 28.82
3 64.6 33 50 30.94
4 59.8 25 43 23.29
5 58.8 25 40 23.94
6 58.7 39 33 22.35

> tail(exo1)
     V1 V2 V3    V4
35 66.1 56 36 13.00
36 71.7 57 59 32.41
37 68.8 41 40 19.82
38 45.0 27 40 11.64
39 61.9 37 44 28.35
40 56.0 36 56 32.88

> str(exo1)
'data.frame':   40 obs. of  4 variables:
 $ V1: num  73.9 59.4 64.6 59.8 58.8 58.7 68.5 54.7 61.7 81.6 ...
 $ V2: int  77 48 33 25 25 39 71 24 25 49 ...
 $ V3: int  52 32 50 43 40 33 37 48 47 50 ...
 $ V4: num  26.1 28.8 30.9 23.3 23.9 ...

> summary(exo1)
       V1               V2              V3                V4
 Min.   :44.00   Min.   :23.0   Min.   :23.00   Min.   :11.64
 1st Qu.:56.88   1st Qu.:32.5   1st Qu.:36.75   1st Qu.:20.13
 Median :63.60   Median :46.0   Median :40.00   Median :23.79
 Mean   :63.22   Mean   :47.7   Mean   :42.02   Mean   :24.02
 3rd Qu.:69.83   3rd Qu.:63.0   3rd Qu.:49.00   3rd Qu.:28.04
 Max.   :83.70   Max.   :79.0   Max.   :59.00   Max.   :32.88

> cor(exo1)
          V1         V2         V3         V4
V1 1.0000000 0.50662598 0.33249518 0.31199028
V2 0.5066260 1.00000000 0.07659737 0.09930797
V3 0.3324952 0.07659737 1.00000000 0.56977018
V4 0.3119903 0.09930797 0.56977018 1.00000000

> modele1<-lm(V1~V2+V3+V4,data=exo1)
```

```
> residus<-residuals(modele1)

> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.9841, p-value = 0.8348

> summary(modele1)

Call:
lm(formula = V1 ~ V2 + V3 + V4, data = exo1)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7769  -4.4512   0.3697   3.5837  15.6096

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.67802    7.27861   4.902 2.03e-05 ***
V2           0.24748    0.06985   3.543  0.00112 **
V3           0.24480    0.18521   1.322  0.19460
V4           0.22667    0.25980   0.872  0.38872
---

Residual standard error: 7.724 on 36 degrees of freedom
Multiple R-squared: 0.357,      Adjusted R-squared: 0.3034
F-statistic: 6.663 on 3 and 36 DF,  p-value: 0.001077

> modele2<-lm(V1~V2+V3,data=exo1)

> residus<-residuals(modele2)

> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.9773, p-value = 0.5914

> summary(modele2)

Call:
lm(formula = V1 ~ V2 + V3, data = exo1)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-17.4334  -4.6922   0.7114   3.3378  15.3754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.07675    7.07692   5.239 6.72e-06 ***
V2           0.25162    0.06946   3.623  0.00087 ***
V3           0.33637    0.15212   2.211  0.03328 *
---

Residual standard error: 7.7 on 37 degrees of freedom
Multiple R-squared: 0.3434,     Adjusted R-squared: 0.3079
F-statistic: 9.677 on 2 and 37 DF,  p-value: 0.0004166

> modele3<-lm(V1~V2+V4,data=exo1)
> residus<-residuals(modele3)
> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.9762, p-value = 0.5524

> summary(modele3)

Call:
lm(formula = V1 ~ V2 + V4, data = exo1)

Residuals:
     Min       1Q   Median       3Q      Max
-22.4476  -4.1501   0.2186   3.3822  17.1286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.18492    6.02807   6.832 4.75e-08 ***
V2           0.24974    0.07053   3.541   0.0011 **
V4           0.42124    0.21622   1.948   0.0590 .
---

Residual standard error: 7.802 on 37 degrees of freedom
Multiple R-squared: 0.3258,     Adjusted R-squared: 0.2894
F-statistic: 8.941 on 2 and 37 DF,  p-value: 0.0006796

> modele4<-lm(V1~V3+V4,data=exo1)
> residus<-residuals(modele4)
> shapiro.test(residus)
```

```
        Shapiro-Wilk normality test

data:  residus
W = 0.9758, p-value = 0.5385

> summary(modele4)

Call:
lm(formula = V1 ~ V3 + V4, data = exo1)

Residuals:
     Min       1Q   Median       3Q      Max
-20.6759  -4.9742   0.6824   7.0274  15.6649

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.3059     7.7353   5.857 9.81e-07 ***
V3            0.2609     0.2121   1.230    0.227
V4            0.2892     0.2969   0.974    0.336
---

Residual standard error: 8.849 on 37 degrees of freedom
Multiple R-squared: 0.1328,     Adjusted R-squared: 0.08591
F-statistic: 2.833 on 2 and 37 DF,  p-value: 0.07167
```

Exercice 2. Le classement de logements dans une grande ville française.

Cette étude est réalisée à partir d'un échantillon aléatoire de trente logements situés dans une grande ville française. Pour chaque logement nous disposons de deux variables :

— Prix : prix de vente exprimé en milliers de dollars.

— Situation : variable à trois modalités :

1 : logements situés dans des quartiers du centre ville peu côtés.

2 : logements situés dans les faubourgs.

3 : logements situés dans les banlieues résidentielles.

1. Proposer un modèle statistique qui permet d'étudier une relation (préciser le type de relation) entre le prix de vente et la situation géographique. Préciser la nature de chacune des variables présentes dans le modèle statistique proposé.
2. Les conditions d'application du modèle linéaire sont-elles vérifiées ? Si oui, expliquer votre réponse.
3. Donner le tableau de l'analyse de la variance.
4. D'après les sorties statistiques réalisées avec le logiciel `R` qui se trouvent ci-dessous, pouvez-vous conclure à une éventuelle significativité de la situation géographique sur le prix de vente ? Pour répondre à cette question, utiliser un test. Vous citerez le nom du test, les hypothèses, la statistique du test et donnerez la conclusion du test (vous préciserez quelle règle vous utilisez).
5. Pouvez-vous séparer les situations géographiques en groupes ne présentant pas de différence significative au seuil de 5% ? Si oui, expliquer comment vous procédez.
6. Dans le cas où vous avez répondu dans l'affirmative à la question précédente, faire cette répartition en groupes homogènes, en indiquant les situations géographiques et les moyennes correspondantes au prix de vente.

```
> exo3
   prix situation
1   198         3
2   185         3
3   165         2
4   170         2
5   170         2
6   183         2
7   158         2
8   146         2
9   168         2
10  162         2
11  184         3
12  154         1
13  170         1
14  122         1
15  175         1
```

```
16  168         1
17  181         1
18  162         1
19  173         3
20  178         1
21  167         3
22  158         1
23  151         1
24  157         3
25  175         3
26  181         3
27  184         3
28  175         3
29  181         2
30  183         2
> modele1<-aov(prix~situation)
> residus<-residuals(modele1)
> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.9444, p-value = 0.1198
> bartlett.test(residus~situation)

        Bartlett test of homogeneity of variances

data:  residus by situation
Bartlett's K-squared = 2.0298, df = 2, p-value = 0.3624
> anova(modele1)
Analysis of Variance Table
Response: prix
          Df Sum Sq Mean Sq F value  Pr(>F)
situation  2 1291.3  645.63  3.4354 0.04685 *
Residuals 27 5074.2  187.93
---
> TukeyHSD(modele1)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = prix ~ situation)
$situation
    diff        lwr      upr     p adj
2-1  6.7 -8.5008055 21.90081 0.5266048
3-1 16.0  0.7991945 31.20081 0.0376200
3-2  9.3 -5.9008055 24.50081 0.2990454
```

Exercice 3. La vigne se traite.
Nous nous proposons de comparer l'efficacité de deux traitements T_1 et T_2 destinés à combattre une certaine maladie de la vigne. Dans un vignoble atteint de cette maladie, nous choisissons au hasard deux échantillons, l'un de 110 pieds de vigne l'autre de 90 pieds de vigne, auquels nous appliquons respectivement les traitements T_1 et T_2. Quelques mois après la fin des traitements, nous observons les résultats obtenus. À cet effet, nous partageons chacun des échantillons obtenus en trois catégories :

a) A : disparition totale de la maladie

b) B : présence de quelques séquelles

c) C : persistence de la maladie.

Les résultats obtenus figurent dans le tableau suivant :

	A	B	C
T_1	80	25	5
T_2	60	18	12

Les effets des deux traitements sont-ils significativement différents au seuil $\alpha = 5\%$? Pour répondre à la question, vous effectuerez un test dont vous donnerez le nom, puis vous énoncerez les deux hyptohèses associées à ce test ainsi que la valeur de la statistique de ce test. Enfin, il manque deux valeurs dans la sortie de `R`, retrouvez ces valeurs.

```
> vigne<-matrix(c(80,25,5,60,18,12),byrow=T,nrow=2,
dimnames=list(c("T1","T2"),c("A","B","C")))
> vigne
```

	A	B	C
T1	80	25	5
T2	60	18	12

```
> chisq.test(vigne,correct=FALSE)
Pearson's Chi-squared test
data: vigne
X-squared = 4.9283, df = 2, p-value = 0.08508
> chisq.test(vigne,correct=FALSE)$expected
```

	A	B	C
T1	?	23.65	9.35
T2	?	19.35	7.65