Modèle Linéaire Général
Examen

Durée : 2h.
Les slides des cours et TD et les notes personnelles manuscrites sont autorisées.
Calculette autorisée.

Exercice 1
Une étude est menée pour étudier la dépendance entre l'âge à la mort de patients (variable *AgeAtDeath*) et les facteurs *Weight-Status* (3 modalités : Normal, Overweight, Underweight), *Chol-Status* (statut Cholestérol, 3 modalités : Borderline, Desirable, High) et *Sex* (2 modalités : Female, Male).

I) Une approche par modèle linéaire est réalisée et les résultats obtenus sont donnés dans le listing suivant :

```
Informations sur les niveaux de classe
Classe         Niveaux   Valeurs
Weight-Status  3         Normal Overweight Underweight
Chol-Status    3         Borderline Desirable High
Sex            2         Female Male

Nombre d'observations lues : 5209
Nombre d'observations utilisées : 1919

Variable dépendante : AgeAtDeath

Source  DDL  Somme des carrés  Carré moyen  Valeur F  Pr > F
Modèle  x     11461.6880             x       15.68      <.0001
Erreur  x    199588.4673             x
Total   x     211050.1553

R-carré  Coef de var  Racine MSE  AgeAtDeath Moyenne
0.054308  14.46302            x     70.66076

Source         DDL  Type I SS   Carré moyen  Valeur F  Pr > F
Weight-Status   2  1334.867893  667.433946   6.39      0.0017
Chol-Status     2  2516.232261  1258.116131  12.05     <.0001
Sex             1  1350.202936  1350.202936  12.93     0.0003
Chol-Status*Sex 2  6260.384931  3130.192465  29.97     <.0001

Source         DDL  Type III SS  Carré moyen  Valeur F  Pr > F
Weight-Status   2  1016.213944   508.106972     4.86    0.0078
Chol-Status     2  3338.595983   1669.297991    15.98   <.0001
Sex             1  115.164482    115.164482     1.10    0.2938
Chol-Status*Sex 2  6260.384931   3130.192465    29.97   <.0001

Paramètre                          Estimation    Erreur type  Valeur du test t  Pr > |t|
Constante                          68.07273148 B  1.36801906     49.76           <.0001
Weight-Status Normal               -0.34351994  B  1.36138491    -0.25           0.8008
Weight-Status Overweight            1.35904828  B  1.30537999     1.04           0.2980
Weight-Status Underweight           0.00000000  B  .               .             .
Chol-Status Borderline              1.41997020  B  0.71816793     1.98           0.0482
Chol-Status Desirable               1.34648798  B  0.82722500     1.63           0.1037
Chol-Status High                    0.00000000  B  .               .             .
Sex Female                          5.00755488  B  0.70686823     7.08           <.0001
Sex Male                            0.00000000  B  .               .             .
Chol-Status*Sex Borderline Female  -3.77786000  B  1.06962161    -3.53           0.0004
Chol-Status*Sex Borderline Male     0.00000000  B  .               .             .
Chol-Status*Sex Desirable Female   -9.68123249  B  1.25559534    -7.71           <.0001
```

```
Chol-Status*Sex Desirable Male      0.00000000  B  .                .              .
Chol-Status*Sex High Female         0.00000000  B  .                .              .
Chol-Status*Sex High Male           0.00000000  B  .                .              .
```

I.1) Ecrire de façon claire et explicite le modèle statistique utilisé.

I.2) Remplacer les $\times$ de la colonne DDL du premier tableau du listing par les valeurs des DDL.

I.3) Comment est calculée la probabilité donnée dans le premier tableau du listing ? Quelle interprétation en faites-vous ?

I.4) Les paramètres du modèle sont-ils estimés sous contrainte(s) ? si oui, lesquelles ? si non, pourquoi ?

I.5) Quelle est ici l'estimation de la variance des résidus du modèle ?

I.6) Et quelle est l'estimation de la variance de la variable *AgeAtDeath* ?

I.7) Que peut-on dire de l'effet de la variable *Sex* ?

I.8) En supposant les postulats vérifiés, le modèle proposé vous semble-t-il d'intérêt ?

I.9) En supposant les postulats vérifiés, donner une interprétation des sorties du listing.

I.10) En exploitant le modèle, quelle serait l'équation permettant de prédire une réponse moyenne pour une personne homme ayant les modalités *Borderline* pour *Chol-Status* et *Overweight* pour *Weight-Status* ? et pour une personne femme ayant les mêmes modalités de co-variables ?

I.11) Que pourriez vous proposer pour continuer l'analyse statistique ?

II) Le responsable de l'étude décide de créer une nouvelle variable *New* croisant les modalités des facteurs *Chol-Status* et *Sex* et de l'inclure dans l'étude. La variable *New* a les modalités ordonnées suivantes :
Borderline-Female, Borderline-Male, Desirable-Female, Desirable-Male, High-Female, High-Male.

II.1) Cette initiative vous paraît-elle d'intérêt ? si oui, pourquoi ? si non, pourquoi ?

II.2) Une comparaison multiple des résultats moyens obtenus selon les modalités de *New* est menée, selon les approches Bonferroni et Student-Newman-Keuls et les résultats obtenus sont donnés ci-dessous :

Approche Bonferroni :

New	$Estimation$	
$High - Female$	74.0285	a
$Borderline - Female$	71.6931	ab
$Borderline - Male$	70.3929	ab
$Desirable - Male$	70.1498	ab
$High - Male$	67.1199	bc
$Desirable - Female$	65.4813	c

Approche Student-Newman-Keuls :

New	$Estimation$	
$High - Female$	74.0285	a
$Borderline - Female$	71.6931	ab
$Borderline - Male$	70.3929	b
$Desirable - Male$	70.1498	b
$High - Male$	67.1199	b
$Desirable - Female$	65.4813	c

II.2.a) Interpréter les résultats obtenus.

II.2.b) Comment expliquez-vous que les différences que l'on peut voir dans les sorties ?

II.2.c) Quels résultats vous semblent les plus intéressants et pourquoi ?

II.2.d) Si l'on voulait tester l'effet de la variable *Sex* pour les patients dont le taux de Cholestérol est *High*, quel contraste devrait-on utiliser ? Donner les coefficients à considérer pour le construire.

Exercice 2

Un biologiste mesure la résistivité de membranes cellulaires (variable Y) soumis à des courants électriques de très faible intensité (valeurs x).

I) Les données sont les suivantes :

	x	y
1		1.48
2		1.02
3		2.89
4		2.59
5		4.09

Il décide de faire une régression simple de Y sur x. Dans la suite, ce modèle est appelé Mod1.

(a) Donner la matrice de design $\mathbf{X}$ du modèle Mod1.

(b) Le listing des sorties SAS relatives à Mod1 est le suivant :

Mod1 Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.61041	4.61041	10.88	0.0458
Erreur	3	1.27171	0.42390		
Corrected total	4	5.88212			

Root MSE	0.65108	R square	0.7838
Mean y	2.41400	R adj. squar.	0.7117
Coeff Var	26.97094		

Variable	DF	Valeur estimee	Erreur type	Test T	Pr > |t|
Intercept	1	0.37700	0.68286	0.55	0.6194
x	1	0.67900	0.20589	3.30	0.0458

Le modèle proposé est-il significatif ?

(c) Quelle est l'estimation de la pente de régression ? Donner un I.C. à 95% pour ce paramètre.

(d) Donner une interprétation de ce listing. Eventuellement, quelle(s) sortie(s) SAS supplémentaire(s) souhaiteriez avoir et pourquoi ?

II) Le biologiste se rend compte que les valeurs précédentes de Y correspondent en fait à des valeurs moyennes : 5 réalisations de Y avaient été mesurées pour chaque x fixé. Disposant de ce nouveau jeu de données, le biologiste hésite à refaire son étude de la dépendance entre Y et x. Que lui conseillez-vous et pourquoi ?

III) Le biologiste décide finalement de faire une nouvelle régression simple avec ce nouveau jeu de données. Soit Mod2 le modèle considéré.

(a) Pour obtenir Mod2, comment est modifiée l'écriture de Mod1 ? Donner l'écriture explicite de Mod2 (*écriture pour une réalisation de Y*).

(b) Les sorties SAS pour Mod2 sont les suivantes :

Mod2 Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	23.14241	x	x	0.0082
Erreur	23	63.50786	x		
Corrected Total	24	86.65027			

Root MSE	x	R square	x
Mean y	2.41400	R adj. squar.	0.2352

Coeff Var	68.73297

Variable	DF	Valeur estimee	Erreur type	Test T	Pr > |t|
Intercept	1	0.37700	0.77940	0.48	0.6335
x	1	0.67900	0.23500	2.90	0.0082

Complétez les valeurs pour Mean Squares, F-Value, Root-MSE et R square.

(c) Les résultats vous semblent-ils différents de ceux de Mod1 ? Si oui, comment expliquez vous ces différences ? Quelle approche vous semble la plus intéressante pour rendre compte de l'expérience du biologiste ?

IV) Le biologiste décide de faire une approche par analyse de variance en considérant x comme un facteur. Soit Mod3 le modèle correspondant ; le listing SAS obtenu pour Mod3 le suivant :

Mod3 Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	29.54234885	7.38558721	2.59	0.0682
Error	20	57.10791889	2.85539594		
Corrected Total	24	86.65027			

R-Square	Coeff Var	Root MSE	y Mean
0.340938	69.89538	1.689792	2.417602

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	4	29.542334885	7.38558721	2.59	0.0682

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	4	29.542334885	7.38558721	2.59	0.0682

(a) Ecrire de façon claire et explicite explicite le modèle Mod3.

(b) Interprétez de façon concise les résultats du listing.

(c) Pourquoi une même probabilité apparaît 3 fois ?

V) Le biologiste se demande si finalement le choix d'une dépendance linéaire entre Y et x conduit à une bonne représentation du phénomène étudié.

(a) Quel test statistique permettrait de répondre à cette question ?

(b) Comment pourrait-on calculer la p-value associée à ce test ?