Exercice 1

On s'intéresse à la taille de crevettes ayant grandies dans des eaux de salinité différente. La salinité est représentée par la variable *sal*, qui est une variable quantitative continue. Les crevettes considérées peuvent appartenir à cinq groupes génétiques représentés par la variable qualitative *crts* (cinq modalités : A, B, C, D et E). L'objectif est d'expliquer la taille des crevettes selon les variables salinité et groupe génétique. Il y a au total 42 individus.
Un modèle linéaire faisant intervenir la salinité (*sal*) et le groupe génétique (*crts*) est considéré. Le listing obtenu est donné ci-dessous.

The GLM Procedure

Informations sur le niveau de classe

Classe	Niveaux	Valeurs
crts	5	A B C D E

Number of Observations Read	42
Number of Observations Used	42

Dependent Variable: taille

Source	DDL	Somme des carres	Moyenne quadratique	Valeur F	Pr > F
Model	x	533.5089523	x	12.79	<.0001
Error	x	148.2967620	x		
Corrected Total	x	681.8057143			

R-carre	Coef de Var	Racine MSE	taille Moyenne
x	9.709507	2.152736	22.17143

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
crts	4	184.3020833	46.0755208	9.94	<.0001
sal*crts	5	349.2068690	69.8413738	15.07	<.0001

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
crts	4	271.9584900	67.9896225	14.67	<.0001
sal*crts	5	349.2068690	69.8413738	15.07	<.0001

Parametre	Valeur estimee	Erreur type	Valeur du test t	Pr > |t|
Intercept	19.43745434 B	4.11008182	4.73	<.0001

crts	A	22.82133857 B	4.95629217	4.60	<.0001
crts	B	17.78270238 B	5.32789102	3.34	0.0022
crts	C	13.65507899 B	4.89994616	2.79	0.0089
crts	D	0.25368183 B	4.56157115	0.06	0.9560
crts	E	0.00000000 B	.	.	.
sal*crts	A	-0.34882376	0.05272691	-6.62	<.0001
sal*crts	B	-0.27166982	0.06620112	-4.10	0.0003
sal*crts	C	-0.20853333	0.05558341	-3.75	0.0007
sal*crts	D	0.03992542	0.04998373	0.80	0.4303
sal*crts	E	-0.01969316	0.10971806	-0.18	0.8587

1. Quel type d'approche par modèle linéaire est considéré ici ?
2. Donner une écriture explicite du modèle considéré.
3. Donner une écriture concise de la formulation matricielle de ce modèle.
4. Retrouver les valeurs des DDL remplacées ici par des 'x'.
5. Pourquoi le DDL associé à $sal * crts$ est il de 5 ?
6. Quelle est la valeur de R-carre ? Comment s'interprète ce coefficient ?
7. Quelle est la valeur estimée de l'écart-type résiduel ? même question pour l'écart-type de la taille des crevettes ?
8. Dans le tableau des estimations des paramètres, pourquoi la ligne *crts E* comporte-t-elle des points ?
9. L'expérience considérée est-elle équilibrée ?
10. Donner une interprétation des sorties du listing obtenu.
11. En utilisant ce modèle, quelle est l'équation qui permet de prévoir la taille moyenne d'une crevette du groupe *crs* A se développant dans une salinité donnée x_s ?

Exercice 2

On s'intéresse à des données concernant la rapidité de lecture suivant le caractère (*concret, abstrait*) des mots, le sexe (*fille, garçon*) et l'âge des enfants (*petit ou grand*). Voici ci-dessous les mesures (en secondes) obtenues pour les garçons puis pour les filles lors de la lecture de deux textes (l'un avec des mots concrets et l'autre avec des mots abstraits) :

garçons :

	concret	abstrait
petit	1450 1495 1668	1445 1433 1702
grand	966 1168 590	1104 1027 788

filles :

	concret	abstrait
petit	1224 1286 1098	1344 1182 1524
grand	765 840 1183	1092 893 1150

1. Quelle(s) question(s) vous semble(nt) naturelle(s) ici et quelle(s) approche(s) vous semble(nt) appropriée(s) pour y répondre ?
2. On décide de faire une analyse de variance à trois facteurs. Un premier modèle, appelé Model1, est ajusté. La sortie SAS correpondante est donnée ci-dessous.

Model1

Analysis of Variance Procedure

Dependent Variable: RAPIDITE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	x	1421487.25000000	x	7.81	0.0004
Error	x	515975.70833334	x		
Corrected Total	x	1937462.95833333			

R-Square	C.V.	Root MSE	RAPIDITE Mean

x	14.71374	174.21685411	1184.04166667

Source	DF	SSI	Mean Square	F Value	Pr > F	
SEXE	x	65626.04166666	65626.04166666	2.16	0.1597	
AGE	x	1163801.04166666	1163801.04166666	38.34	0.0001	
CARACTER	x	37683.37500000	37683.37500000	1.24	0.2807	
SEXE*AGE	x	137259.37500000	137259.37500000	4.52	0.0484	
SEXE*CARACTER	x	16380.37500000	16380.37500000	0.54	0.4726	(*)
AGE*CARACTER	x	737.04166666	737.04166666	0.02	0.8780	

Source	DF	SSIII	Mean Square	F Value	Pr > F	
SEXE	x	65626.04166666	65626.04166666	2.16	0.1597	
AGE	x	1163801.04166666	1163801.04166666	38.34	0.0001	
CARACTER	x	37683.37500000	37683.37500000	1.24	0.2807	
SEXE*AGE	x	137259.37500000	137259.37500000	4.52	0.0484	
SEXE*CARACTER	x	16380.37500000	16380.37500000	0.54	0.4726	(**)
AGE*CARACTER	x	737.04166666	737.04166666	0.02	0.8780	

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

(b) Quels sont les postulats qui s'y rattachent ?

(c) Dans le premier tableau, complétez la colonne DF en remplaçant les $\times$ par les valeurs correspondantes.

(d) Même question pour les deux tableaux suivants.

(e) Quel est le nombre de paramètres non-liés (*indépendants*) dans Model1 ?

(f) Comment interprétez-vous la probabilité 0.0004 du premier tableau ?

(g) Quelle est l'estimation de l'écart-type résiduel ?

(h) Quelle est la valeur de *R-square* ? Comment l'interprétez-vous ?

(i) Ecrire les hypothèses H_0 et H_1 testées dans les lignes (*) et (**). Que concluez-vous pour ces tests ?

(j) Le modèle Model1 vous paraît-il satisfaisant ? Pourquoi ?

3. Le modèle Model1 est peu à peu affiné et le modèle Model2, dont les sorties SAS sont listées ci-dessous, est au final obtenu.

Model2

Analysis of Variance Procedure

Dependent Variable: RAPIDITE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1366686.45833333	455562.15277778	15.96	0.0001
Error	20	570776.50000000	28538.82500000		
Corrected Total	23	1937462.95833333			

R-Square	C.V.	Root MSE	RAPIDITE Mean
0.705400	14.26760	168.93438075	1184.04166667

Source	DF	SSI	Mean Square	F Value	Pr > F
SEXE	1	65626.04166666	65626.04166666	2.30	0.1451
AGE	1	1163801.04166666	1163801.04166666	40.78	0.0001
SEXE*AGE	1	137259.37500000	137259.37500000	4.81	0.0403

Source	DF	SSIII	Mean Square	F Value	Pr > F
SEXE	1	65626.04166666	65626.04166666	2.30	0.1451
AGE	1	1163801.04166666	1163801.04166666	40.78	0.0001
SEXE*AGE	1	137259.37500000	137259.37500000	4.81	0.0403

(a) Selon vous, quelle démarche a été suivie pour aboutir au modèle Model2 ?

(b) Donner l'écriture matricielle de Model2.

(c) Comment pourrait-on voir si le modèle Model2 est significativement meilleur que le modèle Model1 ? Comment pourrait-on en obtenir la p-value associée ?

(d) Quelle est l'estimation de la variance de la rapidité de lecture ?

(e) Que peut-on dire de l'effet de la variable *Sexe* ?

(f) En supposant les postulats vérifiés, interprétez le listing obtenu.

4. L'option *Solution* ayant été utilisée dans le programme SAS relatif à Model2, le tableau suivant est obtenu :

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		1532.166667 B	22.22	0.0001	68.9671721
SEXE	fille	-255.833333 B	-2.62	0.0163	97.5343102
	garcon	0.000000 B	.	.	.
AGE	grand	-591.666667 B	-6.07	0.0001	97.5343102
	petit	0.000000 B	.	.	.
SEXE*AGE	fille grand	302.500000 B	2.19	0.0403	137.9343443
	fille petit	0.000000 B	.	.	.
	garcon grand	0.000000 B	.	.	.
	garcon petit	0.000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations.
Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

(a) Que représentent les valeurs de la colonne *Estimate* ?

(b) Que signifient les '.' et comment peut-on les expliquer ?

(c) Quels paramètres d'interaction ont été estimés ?

(d) Compte tenu du problème initialement considéré, que pouvez-vous dire du modèle Model2 finalement retenu ? Quelles conclusions pouvez-vous tirer en exploitant Model2 ?

(e) Pour un enfant petit et de sexe féminin, quelle prédiction de rapidité moyenne de lecture d'un texte concret obtient-on avec Model2 ? et pour un texte abstrait ?

5. Afin de détecter d'éventuelles différences significatives, un test de comparaisons multiples est mené et les résultats suivants sont obtenus :

Test de Newman-Keuls pour la variable RAPIDITE

Alpha 0.05

Nombre de moyennes	2	3	4
Etendue critique	5.9914986	7.3365405	8.1641879

SNK Groupement	Moyenne	N	interact
A	1532.17	4	petit-garcon
B	1276.33	4	petit-fille
C	987.16	4	grand-fille
D	940.5	4	grand-garcon

(a) Quel est le principe général de l'approche utilisée ?

(b) Que peut-on dire de l'effet *Sexe* ?

(c) Donner une interprétation des résultats obtenus.

Exercice 3

Lors de décès de patients atteints d'une maladie M, une variable $V1$ *Temps de survie* (en mois) est calculée et les valeurs de 2 régresseurs $V2$ et $V3$ liées à des caractéristiques cliniques au moment du décès sont enregistrées.

I) On décide de faire une régression de $V1$ sur $V2$ et $V3$ et le listing de l'approche est donné ci-dessous :

```
Variable dépendante : V1
Nb d'observations lues 200
Nb d'obs. utilisées 200

Analyse de variance
Source  DDL Somme des    Moyenne    Valeur F   Pr > F
               carrés   quadratique
Modèle  x    455.62849  x    82.43    <.0001
Erreur  x   544.47300     x
Total    x    1000.10149



Root MSE x R carré x
Moyenne dépendante 5.05175  R car. ajust. 0.4501
Coeff Var 32.90889


Variable DDL  Valeur estimée      Erreur   Valeur  Pr > |t|
               des paramètres      type     du test t
Intercept 1      4.99165        0.11786    42.35    <.0001
V2         1      1.16887       0.12096    9.66     <.0001
V3        1      1.04154        0.11527    9.04     <.0001
```

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

I.2) Remplacer les $\times$ du listing par les valeurs prévues.

I.3) En supposant les postulats du modèle linéaire vérifiés, donner une interprétation du listing obtenu

II) On décide de tester si l'interaction entre les deux régresseurs $V2$ et $V3$ est significative. Une régression faisant intervenir les régresseurs $V2$ et $V3$ ainsi que leur interaction $V4$ est menée et les résultats sont les suivants ;

```
Nb d'observations lues 200
Nb d'obs. utilisées 200

Analyse de variance
Source  DDL Somme des    Moyenne    Valeur F   Pr > F
                carrés   quadratique
Modèle 3      455.82685    151.94228   54.72     <.0001
Erreur 196    544.27464   2.77691
Total  199    1000.10149

Root MSE 1.66641 R carré 0.4558
Moyenne dépendante 5.05175 R car. ajust. 0.4475
Coeff Var 32.98672

Variable DDL  Valeur estimée      Erreur   Valeur  Pr > |t|
               des paramètres      type     du test t
Intercept 1 4.99326              0.11830 42.21 <.0001
V2         1 1.17234              0.12194 9.61 <.0001
V3         1 1.03875              0.11602 8.95 <.0001
V4         1 0.03030              0.11338 0.27 0.7895
```

II.1) Comment s'exprime la matrice de design X du modèle considéré ? (donner sa forme générale permettant de la caractériser) ;

II.2) Quelle est l'estimation de la variance résiduelle ?

II.3) Comment interprétez vous les valeurs de R^2 et $R^2 adjust$? Que pouvez-vous en déduire ?

II.4) Comment est calculée la variable $V3$? Que pouvez-vous die de l'interaction ?

III) Dans le fichier de données, on se rend compte qu'il existe en fait 3 autres régresseurs disponibles, $W1$, $W2$ et $W3$. On cherche à tester **globalement** l'utilité de prendre en compte ces 3 regresseurs après prise en compte des régresseurs $V1$ et $V2$.

III.1) Comment proposez vous de procéder ?

III.2) Comment pourrait-on obtenir la p-value du test concerné ?