Examen de Statistique Approfondie II

****

• *Le cours, les exercices de travaux dirigés, leurs corrigés ainsi que les notes de cours sont autorisés. Tout autre document est interdit.*

• *Les téléphones portables sont formellement interdits.*

• *Les calculatrices sont autorisées.*

• *Tous les tests seront effectués au seuil de signification $\alpha = 5$ %.*

• *Les deux exercices sont indépendants.*

• *Afin de pouvoir traiter les questions, plusieurs résultats numériques et graphiques ont été intégrés au document.*

• *On prendra un soin particulier à préciser quelles sont les hypothèses testées.*

Durée de l'épreuve 2 heures

****

Exercice 1. Mucoviscidose

O'Neill *et al.* (1983)[1] a recueilli des données sur les capacités pulmonaires de patients atteints de mucoviscidose. Les valeurs des variables suivantes ont été reportées dans le tableau ci-après.

- **age** une valeur numérique : l'âge exprimé en années.
- **sexe** une valeur numérique associant la valeur 0 à un homme et la valeur 1 à une femme.
- **taille** une valeur numérique : la taille exprimée en centimètres.
- **masse** une valeur numérique : la masse exprimée en kilogrammes.
- **bmp** une valeur numérique : l'indice de masse corporelle exprimé en %.
- **fev1** une valeur numérique : volume expiré maximal.
- **rv** une valeur numérique : volume résiduel.

[1] **Altman, D.G.**, *Practical Statistics for Medical Research*, Table 12.11, Chapman & Hall, 1991.

- **frc** une valeur numérique : capacité fonctionnelle résiduelle.
- **tlc** une valeur numérique : capacité pulmonaire totale.
- **pemax** une valeur numérique : pression d'expiration maximale.

age	*sexe*	*taille*	*masse*	*bmp*	*fev1*	*rv*	*frc*	*tlc*	*pemax*
7	0	109	13,1	68	32	258	183	137	95
7	1	112	12,9	65	19	449	245	134	85
8	0	124	14,1	64	22	441	268	147	100
8	1	125	16,2	67	41	234	146	124	85
8	0	127	21,5	93	52	202	131	104	95
9	0	130	17,5	68	44	308	155	118	80
11	1	139	30,7	89	28	305	179	119	65
12	1	150	28,4	69	18	369	198	103	110
12	0	146	25,1	67	24	312	194	128	70
13	1	155	31,5	68	23	413	225	136	95
13	0	156	39,9	89	39	206	142	95	110
14	1	153	42,1	90	26	253	191	121	90
14	0	160	45,6	93	45	174	139	108	100
15	1	158	51,2	93	45	158	124	90	80
16	1	160	35,9	66	31	302	133	101	134
17	1	153	34,8	70	29	204	118	120	134
17	0	174	44,7	70	49	187	104	103	165
17	1	176	60,1	92	29	188	129	130	120
17	0	171	42,6	69	38	172	130	103	130
19	1	156	37,2	72	21	216	119	81	85
19	0	174	54,6	86	37	184	118	101	85
20	0	178	64,0	86	34	225	148	135	160
23	0	180	73,8	97	57	171	108	98	165
23	0	175	51,1	71	33	224	131	113	95
23	0	179	71,5	95	52	225	127	101	195

1. Écrire le modèle de régression linéaire multiple permettant d'expliquer la réponse *pemax* à l'aide de toutes les variables explicatives proposées. On précisera la nature des facteurs explicatifs ainsi que les hypothèses faites.

2. Les hypothèses du modèle sont-elles vérifiées ? Calculer les estimations de tous les paramètres du modèle.

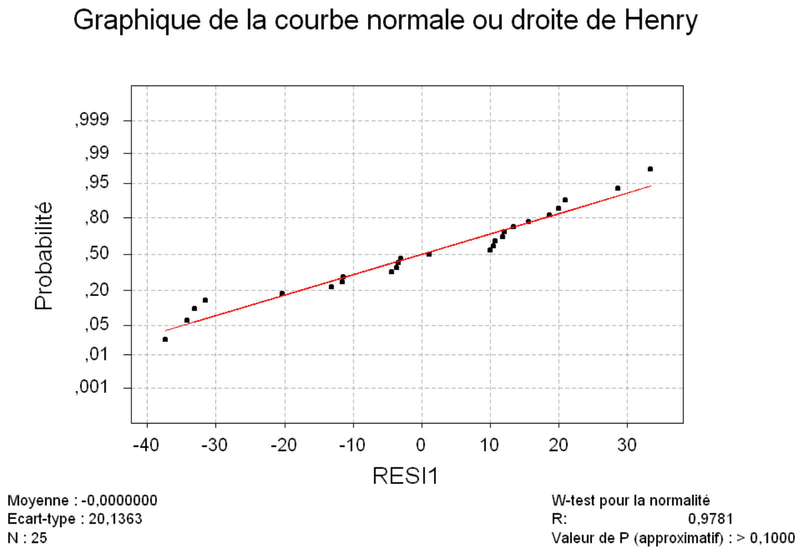
3. Ce modèle est-il intéressant ?

4. Quelles sont les variables explicatives qui influent significativement sur la réponse *pemax* ?

5. Les résultats de la question **4.** sont-ils en accord avec l'étude des corrélations non paramétriques de Kendall du *pemax* et des autres variables ?

6. Calculer la corrélation de Kendall partielle de la masse et de la taille ajustée par l'influence de l'âge. Cette corrélation est-elle significative ?

7. Quel modèle choisir pour expliquer la réponse *pemax* ? Que faudrait-il également vérifier ?



`Modèle linéaire généralisé : pemax en fonction de`

`Facteur      Type Niveaux  Valeurs`

`Analyse de la variance pour pemax, en utilisant la SC ajustée pour les tests`

Source	DL	SC séq	SC ajust	CM ajust	F	P
age	1	10098,5	181,8	181,8	0,28	0,604
sexe	1	955,4	37,9	37,9	0,06	0,812
taille	1	155,0	158,3	158,3	0,24	0,628
masse	1	632,3	1441,2	1441,2	2,22	0,157
bmp	1	2862,2	1480,1	1480,1	2,28	0,152
fev1	1	1549,1	648,4	648,4	1,00	0,333
rv	1	561,9	653,8	653,8	1,01	0,331
frc	1	194,6	254,6	254,6	0,39	0,540
tlc	1	92,4	92,4	92,4	0,14	0,711
Erreur	15	9731,2	9731,2	648,7		
Total	24	26832,6				

Terme	Coef	Er-T coef	T	P
Constante	176,1	225,9	0,78	0,448
age	-2,542	4,802	-0,53	0,604
sexe	-3,74	15,46	-0,24	0,812
taille	-0,4463	0,9034	-0,49	0,628
masse	2,993	2,008	1,49	0,157

bmp	-1,745	1,155	-1,51	0,152
fev1	1,081	1,081	1,00	0,333
rv	0,1970	0,1962	1,00	0,331
frc	-0,3084	0,4924	-0,63	0,540
tlc	0,1886	0,4997	0,38	0,711

Régression sur les meilleurs sous-ensembles : pemax en fonction de age; sexe;

La réponse est pemax

Vars	R-carré	R-carré(ajust)	C-p	S	age	sexe	taille	masse	bmp	fev1	rv	frc	tlc
1	40,4	37,8	3,7	26,380				X					
1	37,6	34,9	4,8	26,974	X								
2	47,5	42,7	2,7	25,308				X	X				
2	45,8	40,8	3,4	25,716	X					X			
3	57,0	50,9	0,8	23,440				X	X	X			
3	53,0	46,2	2,5	24,514	X					X	X		
4	61,4	53,7	1,0	22,754				X	X	X	X		
4	59,9	51,9	1,6	23,197				X	X	X		X	
5	62,1	52,2	2,7	23,122				X	X	X	X		X
5	62,1	52,2	2,7	23,128			X	X	X	X	X		
6	62,7	50,2	4,4	23,592	X		X	X	X	X	X		
6	62,6	50,2	4,5	23,601				X	X	X	X	X	X
7	63,2	48,0	6,2	24,114	X		X	X	X	X	X	X	
7	63,1	47,9	6,3	24,132	X			X	X	X	X	X	X
8	63,6	45,4	8,1	24,710	X		X	X	X	X	X	X	X
8	63,4	45,1	8,1	24,779	X	X	X	X	X	X	X	X	
9	63,7	42,0	10,0	25,471	X	X	X	X	X	X	X	X	X

Kendall's Correlations

		age	sexe	taille	masse	bmp	fev1	rv	frc	tlc	pemax
age	Cor.	1,00	-0,14	0,85	0,78	0,40	0,22	-0,39	-0,53	-0,36	0,40
	Sig.		0,43	0,00	0,00	0,01	0,14	0,01	0,00	0,01	0,01
sexe	Cor.	-0,14	1,00	-0,20	-0,17	-0,12	-0,45	0,21	0,13	0,05	-0,22
	Sig.	0,43		0,25	0,32	0,48	0,01	0,21	0,46	0,78	0,21
taille	Cor.	0,85	-0,20	1,00	0,87	0,45	0,30	-0,40	-0,46	-0,32	0,43
	Sig.	0,00	0,25		0,00	0,00	0,04	0,01	0,00	0,03	0,00
masse	Cor.	0,78	-0,17	0,87	1,00	0,58	0,35	-0,49	-0,47	-0,34	0,35
	Sig.	0,00	0,32	0,00		0,00	0,01	0,00	0,00	0,02	0,02
bmp	Cor.	0,40	-0,12	0,45	0,58	1,00	0,43	-0,50	-0,40	-0,36	0,16
	Sig.	0,01	0,48	0,00	0,00		0,00	0,00	0,01	0,01	0,27
fev1	Cor.	0,22	-0,45	0,30	0,35	0,43	1,00	-0,52	-0,47	-0,36	0,21
	Sig.	0,14	0,01	0,04	0,01	0,00		0,00	0,00	0,01	0,16
rv	Cor.	-0,39	0,21	-0,40	-0,49	-0,50	-0,52	1,00	0,69	0,44	-0,22
	Sig.	0,01	0,21	0,01	0,00	0,00	0,00		0,00	0,00	0,13
frc	Cor.	-0,53	0,13	-0,46	-0,47	-0,40	-0,47	0,69	1,00	0,51	-0,28
	Sig.	0,00	0,46	0,00	0,00	0,01	0,00	0,00		0,00	0,05
tlc	Cor.	-0,36	0,05	-0,32	-0,34	-0,36	-0,36	0,44	0,51	1,00	-0,12
	Sig.	0,01	0,78	0,03	0,02	0,01	0,01	0,00	0,00		0,41
pemax	Cor.	0,40	-0,22	0,43	0,35	0,16	0,21	-0,22	-0,28	-0,12	1,00
	Sig.	0,01	0,21	0,00	0,02	0,27	0,16	0,13	0,05	0,41	

....

Exercice 2. Sécurité routière[2]

On a réalisé une étude sur les relations entre la taille des voitures et la gravité des accidents. Plusieurs informations complémentaires sur la nature de l'accident ont été également été recueillies. Il s'agit du type de sinistre, collision ou tonneau, de savoir si le conducteur a été ejecté au cours de l'accident. Les données ont été reproduites dans le tableau ci-dessous.

Masse	Ejecté	Type	Accident grave	Accident léger
Léger	Non	Collision	150	350
Léger	Non	Tonneau	112	60
Léger	Oui	Collision	23	26
Léger	Oui	Tonneau	80	19
Lourd	Non	Collision	1022	1878
Lourd	Non	Tonneau	404	148
Lourd	Oui	Collision	161	111
Lourd	Oui	Tonneau	265	22

1. Écrire le modèle de régression permettant d'expliquer la gravité de l'accident à l'aide de toutes les variables explicatives proposées. On précisera la nature des facteurs explicatifs ainsi que les hypothèses faites.
2. Les hypothèses du modèle sont-elles vérifiées ? Calculer les estimations de tous les paramètres du modèle.
3. Ce modèle est-il adéquat ?
4. Ce modèle est-il intéressant ?
5. Quelles sont les variables explicatives qui influent significativement sur la gravité de l'accident ?

```
Régression logistique binaire : Accident gra; Accident lég en fonction de Masse;


Fonction de liaison :  Logit

Informations de réponse

Variable  Valeur  Dénombrement
Accident  Succès          2217
Accident  Echec           2614
          Total           4831

Informations de facteur
```

[2]Cet exercice est tiré du livre de B.S. Everitt, *Making Sense of Statistics in Psychology : A second level course.*

Facteur	Niveaux	Valeurs
Masse	2	Léger Lourd
Ejecté	2	Non Oui
Type	2	Collision Tonneau

Tableau de régression logistique

					Ratio de	IC à 95%	
Prédicteur	Coef	Er-T coef	Z	P	probab.	Infér	Supér
Constante	-0,94006	0,08284	-11,35	0,000			
Masse							
Lourd	0,33667	0,08612	3,91	0,000	1,40	1,18	1,66
Ejecté							
Oui	1,03036	0,09891	10,42	0,000	2,80	2,31	3,40
Type							
Tonneau	1,63859	0,08281	19,79	0,000	5,15	4,38	6,06

Log de vraisemblance = -2966,971
Test de toutes les pentes sont zéro : G = 730,585; DL = 3; P = 0,000

Tests d'adéquation de l'ajustement

Méthode	Khi deux	DL	P
Pearson	7,000	4	0,136
Deviance	7,309	4	0,120
Hosmer-Lemeshow	2,664	3	0,446
Brown:			
Alternative générale	4,066	2	0,131
Alternative symétrique	3,012	1	0,083

Tableau des effectifs observés et de leur espérance mathématique :
(voir le test de Hosmer-Lemeshow pour la statistique du Khi deux de Pearson)

Valeur	Succès		Echec		
Groupe	Observé	Espérance	Observé	Espérance	Total
1	150	140,4	350	359,6	500
2	1022	1025,3	1878	1874,7	2900
3	296	305,1	197	187,9	493
4	404	407,3	148	144,7	552
5	345	338,8	41	47,2	386

					4831

Mesures d'association :
(entre la variable de réponse et les prévisions de probabilité)

Paires	Nombre	Pourcentage	Mesures récapitulatives	
Concordant	2909138	50,2%	D de Somers	0,36

Discordant	821953	14,2%	Gamma de Goodman-Kruskal	0,56
Ex aequo	2064147	35,6%	Tau-a de Kendall	0,18
Total	5795238	100,0%		