Exercices de biostatistique

Rappel: pour visualiser la formule associée aux résultats obtenus, il vous suffit d'aller cliquer sur la case concernée(uniquement dans excel et non avec "Adobe Acrobat") !!

Régression multiple

Exercice 1

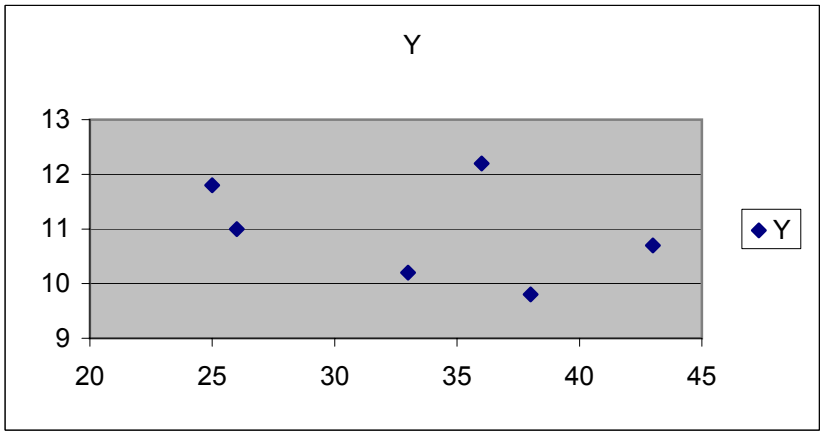
Tester l'hypothèse de curvilinéarité revient à tester l'apport du modèle curvilinéaire (quadratique par exemple) par rapport au modèle linéaire. On commence donc par effectuer un régression linéaire, puis une régression curviligne, et on teste l'augmentation de la somme des carrés due au passage d'un modèle à l'autre. La différence est toujours positive, mais n'est pas nécessairement significative.

1) Régression linéaire

X	Y	x	y	xx	yy	xy
43	10,7	9,5	-0,25	90,25	0,0625	-2,375
38	9,8	4,5	-1,15	20,25	1,3225	-5,175
36	12,2	2,5	1,25	6,25	1,5625	3,125
33	10,2	-0,5	-0,75	0,25	0,5625	0,375
25	11,8	-8,5	0,85	72,25	0,7225	-7,225
26	11	-7,5	0,05	56,25	0,0025	-0,375
33,5	10,95	Moyennes				
			Sommes	245,5	4,235	-11,65

b = -0,04745418

	SC	DF	VAR	F	pF
Reg	0,55284114	1	0,55284114	0,600561966	0,48162566
Err	3,68215886	4	0,92053971		
Tot	4,235	5	0,847		



La régression linéaire n'est pas significative (ce qui était prévisible sur le graphique).

2) Régression quadratique

X1	X1*X1=X2	Y	x1	x2	y
43	1849	10,7	9,5	685,8333333	-0,25
38	1444	9,8	4,5	280,8333333	-1,15
36	1296	12,2	2,5	132,8333333	1,25
33	1089	10,2	-0,5	-74,16666667	-0,75

25	625	11,8	-8,5	-538,1666667	0,85
26	676	11	-7,5	-487,1666667	0,05
33,5	1163,16667	10,95	Moyennes		

x1^2	x2^2	y^2	x1^x2	x1y	x2y
90,25	470367,361	0,0625	6515,41667	-2,375	-171,458333
20,25	78867,3611	1,3225	1263,75	-5,175	-322,958333
6,25	17644,6944	1,5625	332,083333	3,125	166,041667
0,25	5500,69444	0,5625	37,0833333	0,375	55,625
72,25	289623,361	0,7225	4574,41667	-7,225	-457,441667
56,25	237331,361	0,0025	3653,75	-0,375	-24,3583333
Sommes					
245,5	1099334,83	4,235	16376,5	-11,65	-754,55

D = Sx1^2*Sx2^2-Sx1x2^2	1696949,33	b1	-0,26539551
D1 = Sx1y*Sx2^2-Sx1x2*Sx2y	-450362,733	b2	0,00326716
D2 = Sx1^2*Sx2y-Sx1x2*Sx1y	5544,2		

	SC	DF	VAR	F	pF
Reg	0,62662433	2	0,31331216	0,260487426	0,78647966
Err	3,60837567	3	1,20279189		
Tot	4,235	5	0,847		

La régression curviligne n'est pas significative non plus, et il n'y a évidemment pas lieu dans ce cas de tester l'apport de la régression curviligne. Néanmoins, pour montrer le principe, il suffirait de calculer la valeur de F obtenue par:

F = [(SCR quadra - SCR lin)/1] / Var_Err_Quadra
avec 1 et 3 degrés de liberté

On obtient: F = 0,06134327
qui n'est, comme attendu, pas significative.

Exercice 2

On pourrait aborder ce problème de deux manières: soit on considère les variables 'Age' et Catégorie' comme deux variables de classification (discrètes), et on effectue une analyse de variance à deux critères sur le poids. Soit on les considère comme des variables continues, et on effectue une étude de régression. On effectuera ici les deux types d'études pour l'exemple.

Age	Catégorie	Poids
1	1	4,24951157
1	2	6,61071057
1	3	9,11051429
2	1	3,62519041
2	2	7,54958343
2	3	9,31315574
3	1	5,91150923
3	2	4,85003638
3	3	10,1962766

1) Régression

X1	X2	Y	x1	x2	y
1	1	4,24951157	-1	-1	-2,57454268

1	2	6,61071057	-1	0	-0,21334368
1	3	9,11051429	-1	1	2,28646004
2	1	3,62519041	0	-1	-3,19886384
2	2	7,54958343	0	0	0,72552918
2	3	9,31315574	0	1	2,48910149
3	1	5,91150923	1	-1	-0,91254502
3	2	4,85003638	1	0	-1,97401787
3	3	10,1962766	1	1	3,37222235
Moyennes					
2	2	6,82405425			

x1^2	x2^2	y^2	x1^x2	x1y	x2y
1	1	6,62826999	1	2,574542677	2,57454268
1	0	0,04551552	0	0,213343677	0
1	1	5,22789953	-1	-2,286460043	2,28646004
0	1	10,2327298	0	0	3,19886384
0	0	0,5263926	0	0	0
0	1	6,19562624	0	0	2,48910149
1	1	0,83273841	-1	-0,912545017	0,91254502
1	0	3,89674654	0	-1,974017867	0
1	1	11,3718836	1	3,372222353	3,37222235
Sommes					
6	6	44,9578023	0	0,98708578	14,8337354

D = Sx1^2*Sx2^2-Sx1x2^2	36	b1 =	0,1645143
D1 = Sx1y*Sx2^2-Sx1x2*Sx2y	5,92251468	b2 =	2,47228924
D2 = Sx1^2*Sx2y-Sx1x2*Sx1y	89,0024125		

	SC	DF	VAR	F	pF
Reg	36,8356741	2	18,4178371	13,60567336	0,0058965
Err	8,12212814	6	1,35368802		
Tot	44,9578023	8	5,61972528		

Globalement, la régression est significative. On pourrait tester si les deux variables sont individuellement significatives. La meilleure manière de procéder serait de tester l'apport de X2 quand on a déjà corrigé pour X1, et inversément, l'apport de X2 quand on a déjà corrigé pour X1.

Régression sur X1

b1 = Σ x1*y/Σ x1² =	0,1645143
SCR1 = b1 * Σ x1y =	0,16238972
SCR - SCR1 =	36,6732844
F(1;6) =	27,0913858
p(>F) =	0,00200521

L'amélioration due à la variable X2 par rapport au modèle où seul X1 est utilisé est très substantielle.

Régression sur X2

b2 = Σ x2*y/Σ x2² =	2,47228924
SCR2= b2 * Σ x2y =	36,6732844
SCR - SCR1 =	0,16238972
F(1;6) =	0,11996097
p(>F) =	0,7408996

L'amélioration due à la variable X1 par rapport au modèle où seul X2 est utilisé n'est pas significative.

A remarquer que les valeurs obtenues pour les coefficients de régression linéaire sont identiques à celles obtenues pour la régression multiple, une situation qui n'est pas le cas général, et qui correspond au cas où Σx1*x2 = 0 (voir les formules des coefficients de régression multiple)

2) ANOVA II

	Catégorie			
Age	1	2	3	Moyennes
1	4,24951157	6,61071057	9,11051429	6,656912143
2	3,62519041	7,54958343	9,31315574	6,82930986
3	5,91150923	4,85003638	10,1962766	6,985940737
Moyennes	4,59540374	6,33677679	9,53998221	6,824054247

SCT

6,62826999	0,04551552	5,22789953	
10,2327298	0,5263926	6,19562624	
0,83273841	3,89674654	11,3718836	
			44,95780228

SCAge

0,02793648	0,02793648	0,02793648	
2,7621E-05	2,7621E-05	2,7621E-05	
0,02620724	0,02620724	0,02620724	
			0,162514019

SCCategorie

4,9668831	0,23743932	7,3762647	
4,9668831	0,23743932	7,3762647	
4,9668831	0,23743932	7,3762647	
			37,74176134

SCErreur

0,03195159	0,19454793	0,06881483	
0,95153965	1,45817947	0,05386209	
1,33222151	2,71797067	0,24443917	
			7,053526917

Vérification: SCT -SCAge - SCCategorie - SCErreur = 1,95399E-14

Table d'analyse de la variance

Source	SC	DDL	CM	F	p(>F)
Age	0,16251402	2	0,08125701	0,046080215	0,95546477
Categorie	37,7417613	2	18,8708807	10,70152898	0,02479408
Erreur	7,05352692	4	1,76338173		
Total	44,9578023	8	5,61972528		

Encore une fois, le facteur 'Categorie' est significatif, et le facteur 'Age' ne l'est pas.

La différence de résultats entre les deux analyses vient évidemment de la différence entre les deux modèles. Dans le modèle de régression, on suppose que pour une augmentation d'une unité de la variable indépendante, la variable dépendante augmente d'une quantité égale à la

valeur du coefficient de régression sur cette variable.Par contre, dans l'ANOVA, on suppose qu'à chaque niveau correspond une solution, sans relation d'ordre entre les différents niveaux.

Exercice 3

3) Lors de la croissance de veaux, un paramètre physiologique est mesuré, et on a obtenu les résultats suivants:

Temps X1	Paramètre Y
1	10
2	15
3	13
4	16
5	21

a) Testez une régression linéaire sur ces données.
b) Mesurez ensuite l'intérêt d'adjoindre une seconde variable, à savoir $X2=X1^3$.

a)

X1	Y	x1	y	x1y	$x1^2$	y^2
1	10	-2	-5	10	4	25
2	15	-1	0	0	1	0
3	13	0	-2	0	0	4
4	16	1	1	1	1	1
5	21	2	6	12	4	36
15	75	0	0	23	10	66
3	15					

Le coéfficient de régression linéaire est donc: 2,3

La signification est vérifiée (par exemple) par un test de F:

Source	SC	Ddl	Variance	F	P(>F)
Régression	52,9	1	52,9	12,11450382	0,04003818
Erreur	13,1	3	4,36666667		
Total	66	4	16,5		

Au seuil de signification 5 %, nous pourrons dire que la régression linéaire est significative, puisque la probabilité d'avoir un F aussi grand ou plus grand que 12,1145038, avec 1 ddl régression et 3 ddl erreur, est de 4,003818% et donc plus petite que 5 %, ce qui conduit à rejeter l'hypothèse nulle selon laquelle la régression linéaire nétait pas significative.

b) Au tableau ci-dessus, il suffit de rajouter les colonnes suivantes:

$X2=X1^3$	x2	x2y	$x2^2$	x1x2
1	-44	220	1936	88
8	-37	0	1369	37
27	-18	36	324	0
64	19	19	361	19
125	80	480	6400	160
225		755	10390	304

45

$$b1 * \Sigma x^2 1 + b2 * \Sigma x1x2 = \Sigma x1y$$
$$b1 * \Sigma x1x2 + b2 * \Sigma x^2 2 = \Sigma x2y$$

D=	11484
D1=	9450
D2=	558

b1=	0,82288401
b2=	0,04858934

Pour vérifier la signification de cette régression multiple, nous utiliserons un test de F:

SCRégressior	55,6112853
SCErreur	10,3887147
Variance err.	5,19435737
Delta	2,71128527
F	0,52196741

Vu que la valeur de F est inférieure à 1, on peut directement dire que cette régression multiple n'est pas significative et donc il n'y avait pas d'intérêt d'adjoindre cette seconde variable X2=X1³.