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 linéaire

Exercice 1

On a mesuré le poids (en kg) et le périmètre thoracique (en cm) de 8 taureaux à l'engrais:

	Périmètre thoracique (X)	Poids corporel (Y)
	158	540
	164	544
	167	553
	170	549
	171	560
	176	557
	179	556
	183	565
Somme:	1368	4424
Moyenne:	171	553

a) On demande de mesurer la relation qui existe entre ces deux variables et d'en tester la signification par deux tests différents. Calculez l'intervalle de confiance (95%) du coéfficient de régression.

b) Deux animaux ont un périmètre thoracique de 172 cm et 186 cm, respectivement.
Pouvez-vous prédire leur poids avec 95 % de chances de ne pas vous tromper?

a) Il s'agit d'établir une relation entre une variable continue (variable dépendante Y) et une autre variable continue (variable indépendante X). On utilise une régression linéaire, comme suit:

X	Y	x	y	xx	yy	xy	
	158	540	-13	-13	169	169	169
	164	544	-7	-9	49	81	63
	167	553	-4	0	16	0	0
	170	549	-1	-4	1	16	4
	171	560	0	7	0	49	0
	176	557	5	4	25	16	20
	179	556	8	3	64	9	24
	183	565	12	12	144	144	144
					468	484	424

b= 0,90598291

La relation calculée entre les deux variables est donc: Y = 553 + 0,90598291 * (X - 171)
soit: Y =0,90598291 * X + 398.076923

Encore une fois, il faut tester si cette relation a un sens, ce que l'on peut réaliser par un test de signification de la régression. Deux méthodes sont possibles:

1) Analyse de la variance

source	sc	ddl	var	F	p
regression	384,136752	1	384,136752	23,0797672	0,00298807
erreur	99,8632479	6	16,6438746		
total	484	7	69,1428571		

Le résultat de l'analyse indique qu'il est très peu vraisemblable que l'hypothèse nulle de la régression (H0: $\beta = 0$), soit vraie. On en conclut qu'il y a régression, et donc que $\beta > 0$.

2) Test de t de la régression

Encore une fois, l'hypothèse nulle d'absence de régression s'exprime par H0: $\beta = 0$.

S^2b = var_err/Sx^2 =	0,03556383
t = (b-β)/Sb = b/Sb =	4,80414063
P(> t) =	0,00134838

On remarquera que les deux tests sont équivalents, et que $t^2 = F$.

23,0797672 23,0797672

3) Limites de l'intervalle de confiance

L'intervalle de confiance de β au seuil α donne la plage dans laquelle le coéfficient de régression a (1 - α) chances de tomber. Il vaut: b +/- t(α/2)*Sb, soit:

t(6 ddl)	2,44691364
lim inf	0,44453473
lim sup	1,36743108

b)

$$S_Y^2 = S_{yx}^2 \left(\frac{1}{n} + \frac{x^2}{\Sigma x^2} + 1\right)$$

1) Prédiction de Y pour X= 172.

Y=Yb+β*(X-Xb)	553,905983

2) Intervalle de confiance 95 % de Y en fonction de X=172.

Ypredit +/- t*Serr*racine(1+1/n + x^2/Sx^2)

Ecart	10,5982501
Lim. Inférieure	543,307733
Lim. Supérieure	564,504233

Il y a donc 95 chances sur 100 que la valeur prédite pour Y correspondant à une valeur de X = 172 soit située entre 543,31 et 564,5.

3) Prédiction de Y pour X= 186.

Y=Yb+β*(X-Xb)	566,589744

4) Intervalle de confiance 95 % de Y en fonction de X=186.

Ypredit +/- t*Serr*racine(1+1/n + x^2/Sx^2)

Ecart	12,6499097

Lim. Inférieure	553,939834
Lim. Supérieure	579,239653

Il y a donc 95 chances sur 100 que la valeur prédite pour Y correspondant à une valeur de X = 186 soit située entre 553,94 et 579,24.

Exercice 2

Au terme d'une expérience comportant l'administration d'un même régime à 10 souches de White Leghorn, on a mesuré le poids moyen de 50 poules de chaque souche après 350 jours et la consommation alimentaire moyenne de chacune de ces souches. Les résultats ont été les suivants (exprimés en livres):

	Poids corporel (X).	Consommation alimentaire (Y).
	4,6	87,1
	5,1	93,1
	4,8	89,8
	4,4	91,4
	5,9	99,5
	4,7	92,1
	5,1	95,5
	5,2	99,3
	4,9	93,4
	5,1	94,4
Somme:	49,8	935,6
Moyenne:	4,98	93,56

a) On demande de mesurer la relation qui existe entre ces deux variables et d'en tester la signification au moyen de deux tests. Montrez que ces deux tests conduisent au même résultat. Calculez l'intervalle de confiance 95 % du coéfficient de régression.

b) Soit un lot de 50 poules ayant atteint, au terme de l'éxpérience, le poids moyen de 5,3 livres: on demande de calculer l'intervalle de confiance 95 % de la prédiction de la consommation moyenne en 350 jours, pour ce lot particulier.

c) Quel serait l'intervalle de confiance 95% de la prédiction de la consommation moyenne en 350 jours pour des lots ayant atteint un poids moyen de 5,3 livres.

a)

X	Y	x	y	xx	yy	xy	
	4,6	87,1	-0,38	-6,46	0,1444	41,7316	2,4548
	5,1	93,1	0,12	-0,46	0,0144	0,2116	-0,0552
	4,8	89,8	-0,18	-3,76	0,0324	14,1376	0,6768
	4,4	91,4	-0,58	-2,16	0,3364	4,6656	1,2528
	5,9	99,5	0,92	5,94	0,8464	35,2836	5,4648
	4,7	92,1	-0,28	-1,46	0,0784	2,1316	0,4088
	5,1	95,5	0,12	1,94	0,0144	3,7636	0,2328
	5,2	99,3	0,22	5,74	0,0484	32,9476	1,2628
	4,9	93,4	-0,08	-0,16	0,0064	0,0256	0,0128
	5,1	94,4	0,12	0,84	0,0144	0,7056	0,1008
	49,8	935,6	4,4409E-15	1,1369E-13	1,536	135,604	11,812
	4,98	93,56					

b= 7,69010417

La relation calculée entre les deux variables est donc: Y = 93,56 + 7,69010417 * (X - 4,98)
soit: Y =7,69010417 * X + 55,2632813

Encore une fois, il faut tester si cette relation a un sens, ce que l'on peut réaliser par un test de signification de la régression. Deux méthodes sont possibles:

1) Analyse de la variance

source	sc	ddl	var	F	p
regression	90,8355104	1	90,8355104	16,2320438	0,00260058
erreur	44,7684896	8	5,5960612		
total	135,604	9	15,0671111		

Le résultat de l'analyse indique qu'il est très peu vraisemblable que l'hypothèse nulle de la régression (H0: $\beta = 0$), soit vraie. On en conclut qu'il y a régression, et donc que $\beta > 0$.

2) Test de t de la régression

Encore une fois, l'hypothèse nulle d'absence de régression s'exprime par H0: $\beta = 0$.

S^2b = var_err/Sx^2 =	3,64326901
t = (b-β)/Sb = b/Sb =	4,02890106
P(> t) =	0,00379385

On remarquera que les deux tests sont équivalents, et que $t^2 = F$.
16,2320438 16,2320438

3) Limites de l'intervalle de confiance

L'intervalle de confiance de β au seuil α donne la plage dans laquelle le coéfficient de régression a (1 - α) chances de tomber. Il vaut: b +/- t(α/2)*Sb, soit:

t(8 ddl)	2,30600563
lim inf	3,28855069
lim sup	12,0916576

b) 1) Prédiction d'une valeur particulière de Y pour X= 5,3.

Y=Yb+β*(X-Xb) 96,0208333

2) Intervalle de confiance 95 % de Y en fonction de X=5,3.

Ypredit +/- t*Serr*racine(1+1/n + x^2/Sx^2)

Ecart	5,89216615
Lim. Inférieure	90,1286672
Lim. Supérieure	101,912999

Il y a donc 95 chances sur 100 que la valeur prédite pour Y correspondant à une valeur de X = 5,3 soit située entre 90,13 et 101,91.

c) prédiction d'une moyenne.

La valeur prédite pour une moyenne peut varier en fonction de la variation sur la prédiction de Ym

et sur la prédiction du coefficient de régression. On obtient:

L'intervalle de confiance 95 % est égal à: Yprédit+/-t0,05*Serr*racine(1/n+x²/Sx²)

Ecart	2,22702948
Lim. Inférieure	93,7938039
Lim. Supérieure	98,2478628

Exercice 3

En race de Moyenne et Haute Belgique, dans la province de Luxembourg, on a suivi l'évolution du pourcentage de césariennes chez les génisses entre 1966 et 1972:

Dates (X)	% de césariennes (Y)
1966	22,22
1967	23,32
1968	30,81
1969	34,78
1970	36,52
1971	37,17
1972	37,71

En supposant une augmentation linéaire du nombre de césariennes, que sera ce pourcentage en 1980? En quelle année atteindra-t-on 100 % de césariennes chez les génisses?

X	Y	x	y	xx	yy	xy
1966	22,22	-3	-9,57	9	91,5849	28,71
1967	23,32	-2	-8,47	4	71,7409	16,94
1968	30,81	-1	-0,98	1	0,9604	0,98
1969	34,78	0	2,99	0	8,9401	0
1970	36,52	1	4,73	1	22,3729	4,73
1971	37,17	2	5,38	4	28,9444	10,76
1972	37,71	3	5,92	9	35,0464	17,76
13783	222,53	0	1,0658E-14	28	259,59	79,88
1969	31,79					

b= 2,85285714

Pourcentage de césariennes prévu pour 1980.

Y prédit= 63,1714286

Année prévue où 100 % des génisses à terme nécessiteront une césarienne pour la mise-bas.

100=31,79+2,85285714*(X-1969)
X= 1992,90936
L'année prévue où 100 % des génisses nécessiteront une césarienne est donc 1993!

Exercice 4

On a constitué deux lots de 5 porcs, auxquels on a attribué, au hasard, deux régimes différents

a et b. On a mesuré le gain quotidien moyen (en grammes) entre le moment du sevrage (+/- 20 kgs) et celui de l'abattage :

Lot A	Lot B
885 (86.0)	840 (79.9)
861 (84.0)	842 (79.2)
845 (80.5)	765 (74.5)
830 (81.0)	750 (71.0)
859 (83.8)	688 (69.0)

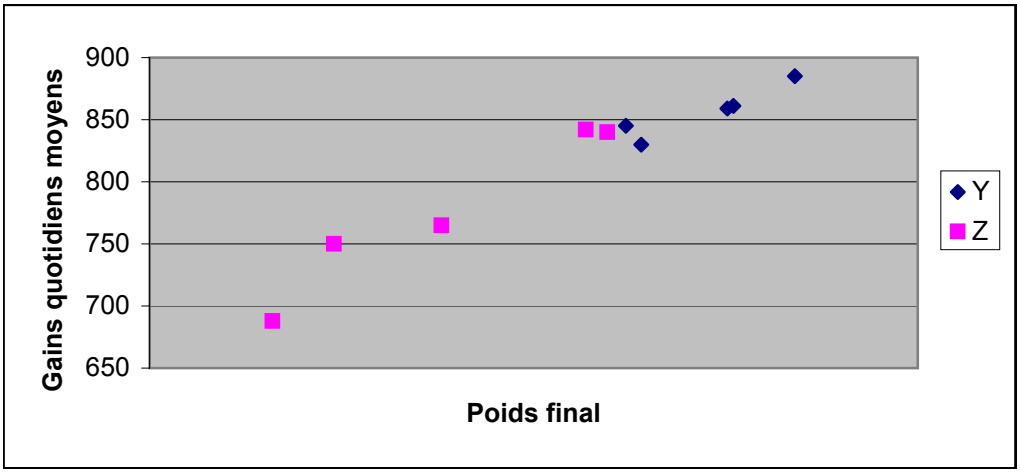
La valeur indiquée entre parenthèses est le poids final. On demande:

de comparer les moyennes des deux lots (au moyen de 2 tests).
de réaliser un ajustement sur le poids (pourquoi ?)
de recomparer les moyennes après ajustement, et de justifier les résultats.

Dans cet exercice, on désire comparer les GQM dans deux lots, mais on s'aperçoit que les poids finaux dans les deux lots sont assez différents, ce qui pourrait vouloir dire que les abattages n'ont pas eu lieu aux mêmes âges dans les deux groupes. Si le GQM varie avec l'âge, on pourrait donc avoir des différences de GQM dues à l'âge et non pas au lot. La solution est de corriger les données pour l'âge en effectuant une régression du GQM sur l'âge.

X	Y	Z
86	885	
84	861	
80,5	845	
81	830	
83,8	859	
79,9		840
79,2		842
74,5		765
71		750
69		688

Comme on le voit sur le graphique, les GQM élevés correspondent aux animaux pour lesquels le poids final est le plus élevé: les données sont mal balancées entre les deux groupes (les données ne sont pas réparties de manière aléatoire). Il risque alors d'y avoir un biais, qui se traduirait dans cet exemple par une confusion entre effet du lot et effet de l'âge.



1) Test de l'effet du lot avant ajustement

Lot A	Lot B
885	840
861	842

Somme des carrés "erreur"	
841	3969
25	4225

	845	765		121	144
	830	750		676	729
	859	688		9	7921
Moyennes	856	777	816,5	SCT	18660

Le calcul des sommes de carré se fait comme d'habitude: la somme des carrés due au modèle se calcule par la somme des carrés des écarts entre moyenne du groupe et moyenne générale, soit 5*(856-816,5)² + 5*(777-816,5)² =15602,5 . La somme des carrés "erreurs" se calcule par la somme des carrés entre données individuelles et moyenne du groupe (voir tableau)

Anova	SC	DDL	VAR	F	pF
Modèle	15602,5	1	15602,5	6,68917471	0,03229616
Erreur	18660	8	2332,5		
Total	34262,5	9	3806,94444		

Test t

La variance Sc² est la même que la variance erreur de l'Anova (Pourquoi ?). On peut donc facilement calculer la valeur de t correspondant à ce problème (deux échantillons, données non pairées, variance inconnue) :

t = (Xb1-Xb2)/racine(S2c*(1/n1+1/n2)/)

	2,58634389
p(>t)	0,03229616

Les deux tests (équivalents) montrent qu'avant ajustement, l'effet du lot est significatif.

2) Ajustement pour l'âge

On va donc utiliser une régression linéaire (voir graphique) pour ramener tous les GQM à ce qu'ils auraient été si on les avait mesurés sur chaque animal au même âge (choisi comme l'age moyen de tous les animaux). On commence par calculer le coefficient de régression à travers tous les points:

Régression

X	Y	x	y	xx	yy	xy
86	885	7,11	68,5	50,5521	4692,25	487,035
84	861	5,11	44,5	26,1121	1980,25	227,395
80,5	845	1,61	28,5	2,5921	812,25	45,885
81	830	2,11	13,5	4,4521	182,25	28,485
83,8	859	4,91	42,5	24,1081	1806,25	208,675
79,9	840	1,01	23,5	1,0201	552,25	23,735
79,2	842	0,31	25,5	0,0961	650,25	7,905
74,5	765	-4,39	-51,5	19,2721	2652,25	226,085
71	750	-7,89	-66,5	62,2521	4422,25	524,685
69	688	-9,89	-128,5	97,8121	16512,25	1270,865
78,89	816,5	Moyennes				
			Sommes	288,269	34262,5	3050,75

b= 10,5829971

Ajustement en X=Xb => Yd-Ya=b*(Xd-Xa)

	Yd	Ya	Ecarts²
	885	809,75489	4,4690449
	861	806,920885	24,4828752
Lot A	845	827,961375	258,967676

	830	807,669876	Moyenne		17,6318186
	859	807,037484	811,868902		23,3425987
	840	829,311173			66,9136247
	842	838,719271			309,343824
Lot B	765	811,459357			93,5425668
	750	833,499847	Moyenne		152,98596
	688	792,665842	821,131098		810,270822
				Somme	1761,95081

t =	0,98680667
p(>t) =	0,35263896

Après ajustement, l'effet de l'âge a disparu, mais il en est de même de l'effet du lot.