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 logistique

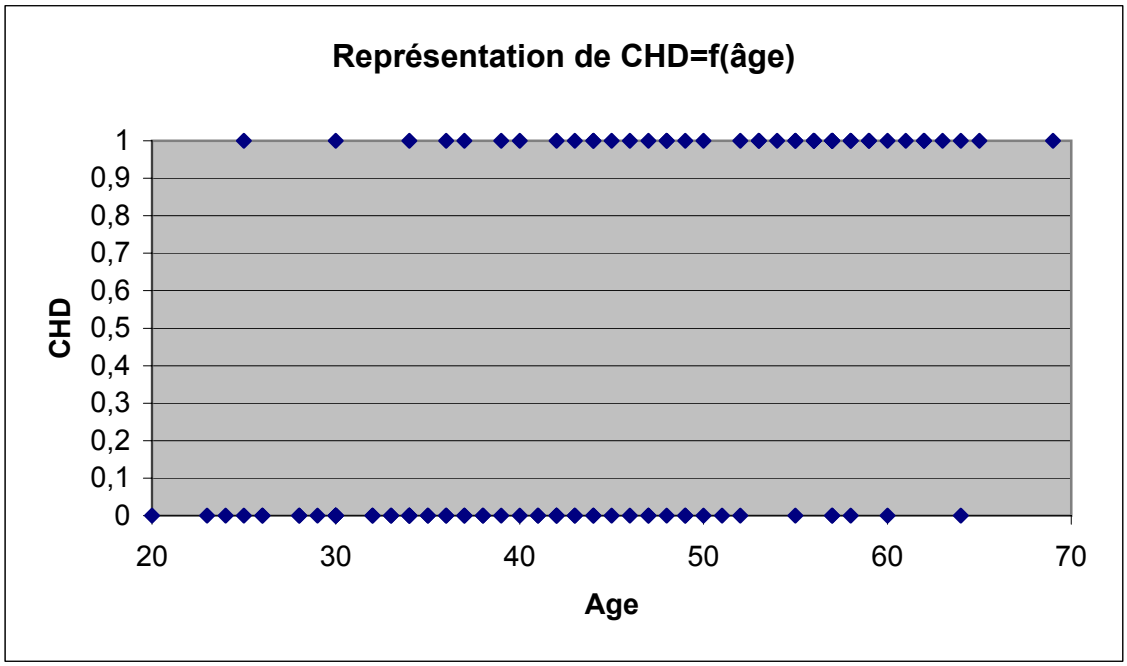
Exercice 1

Les données sont donc:

Id	Groupe âge	Age	CHD
1	1	20	0
2	1	23	0
3	1	24	0
4	1	25	0
5	1	25	1
6	1	26	0
7	1	26	0
8	1	28	0
9	1	28	0
10	1	29	0
11	2	30	0
12	2	30	0
13	2	30	0
14	2	30	0
15	2	30	0
16	2	30	1
17	2	32	0
18	2	32	0
19	2	33	0
20	2	33	0
21	2	34	0
22	2	34	0
23	2	34	1
24	2	34	0
25	2	34	0
26	3	35	0
27	3	35	0
28	3	36	0
29	3	36	1
30	3	36	0
31	3	37	0
32	3	37	1
33	3	37	0
34	3	38	0
35	3	38	0
36	3	39	0
37	3	39	1
38	4	40	0
39	4	40	1
40	4	41	0
41	4	41	0
42	4	42	0
43	4	42	0
44	4	42	0
45	4	42	1

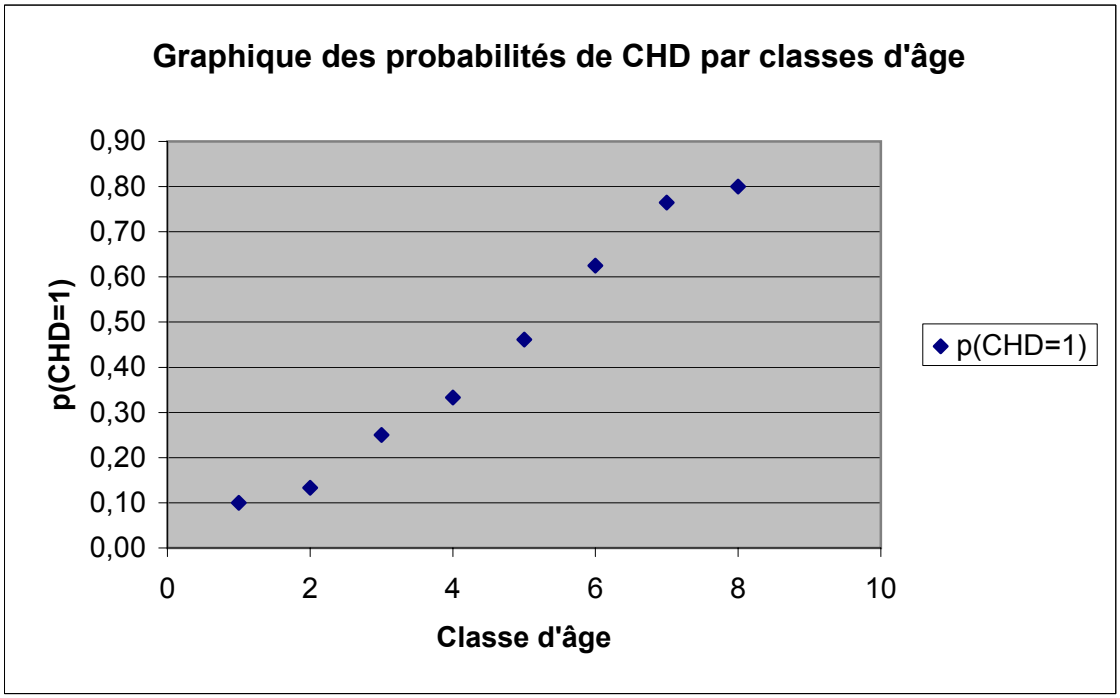
46	4	43	0
47	4	43	0
48	4	43	1
49	4	44	0
50	4	44	0
51	4	44	1
52	4	44	1
53	5	45	0
54	5	45	1
55	5	46	0
56	5	46	1
57	5	47	0
58	5	47	0
59	5	47	1
60	5	48	0
61	5	48	1
62	5	48	1
63	5	49	0
64	5	49	0
65	5	49	1
66	6	50	0
67	6	50	1
68	6	51	0
69	6	52	0
70	6	52	1
71	6	53	1
72	6	53	1
73	6	54	1
74	7	55	0
75	7	55	1
76	7	55	1
77	7	56	1
78	7	56	1
79	7	56	1
80	7	57	0
81	7	57	0
82	7	57	1
83	7	57	1
84	7	57	1
85	7	57	1
86	7	58	0
87	7	58	1
88	7	58	1
89	7	59	1
90	7	59	1
91	8	60	0
92	8	60	1
93	8	61	1
94	8	62	1
95	8	62	1
96	8	63	1
97	8	64	0
98	8	64	1
99	8	65	1
100	8	69	1

a) Graphique du CHD en fonction de l'âge



b) Graphique de p(CHD=1) en fonction de la classe d'âge

Classe	Ages	Effectif	CHD=1	p(CHD=1)
1	A < 30	10	1	0,10
2	30 <= A < 35	15	2	0,13
3	35 <= A < 40	12	3	0,25
4	40 <= A < 45	15	5	0,33
5	45 <= A < 50	13	6	0,46
6	50 <= A < 55	8	5	0,63
7	55 <= A < 60	17	13	0,76
8	60 <= A	10	8	0,80
	Total	100	43	0,43



c) Calcul des $\pi(x)$

Pour calculer les valeurs de $\pi(x)$, qui sont sensées représenter la probabilité associée à une valeur de x, il faut connaître les valeurs de $\beta 0$ et $\beta 1$, qu'il faudra estimer. On utilisera, à titre provisoire, les valeurs suivantes:

$\beta 0 = -4$
$\beta 1 = 0,2$

Id	Groupe âge	Age (X)	CHD	$\pi(X)$	logit(X)
1	1	20	0	0,5	0
2	1	23	0	0,645656306	0,6
3	1	24	0	0,689974481	0,8
4	1	25	0	0,731058579	1
5	1	25	1	0,731058579	1
6	1	26	0	0,768524783	1,2
7	1	26	0	0,768524783	1,2
8	1	28	0	0,832018385	1,6
9	1	28	0	0,832018385	1,6
10	1	29	0	0,858148935	1,8
11	2	30	0	0,880797078	2
12	2	30	0	0,880797078	2
13	2	30	0	0,880797078	2
14	2	30	0	0,880797078	2
15	2	30	0	0,880797078	2
16	2	30	1	0,880797078	2
17	2	32	0	0,916827304	2,4
18	2	32	0	0,916827304	2,4
19	2	33	0	0,93086158	2,6
20	2	33	0	0,93086158	2,6
21	2	34	0	0,942675824	2,8
22	2	34	0	0,942675824	2,8
23	2	34	1	0,942675824	2,8
24	2	34	0	0,942675824	2,8
25	2	34	0	0,942675824	2,8
26	3	35	0	0,952574127	3
27	3	35	0	0,952574127	3
28	3	36	0	0,960834277	3,2
29	3	36	1	0,960834277	3,2
30	3	36	0	0,960834277	3,2
31	3	37	0	0,967704535	3,4
32	3	37	1	0,967704535	3,4
33	3	37	0	0,967704535	3,4
34	3	38	0	0,973403006	3,6
35	3	38	0	0,973403006	3,6
36	3	39	0	0,978118729	3,8
37	3	39	1	0,978118729	3,8
38	4	40	0	0,98201379	4
39	4	40	1	0,98201379	4
40	4	41	0	0,985225968	4,2
41	4	41	0	0,985225968	4,2
42	4	42	0	0,987871565	4,4
43	4	42	0	0,987871565	4,4
44	4	42	0	0,987871565	4,4

45	4	42	1	0,987871565	4,4
46	4	43	0	0,990048198	4,6
47	4	43	0	0,990048198	4,6
48	4	43	1	0,990048198	4,6
49	4	44	0	0,991837429	4,8
50	4	44	0	0,991837429	4,8
51	4	44	1	0,991837429	4,8
52	4	44	1	0,991837429	4,8
53	5	45	0	0,993307149	5
54	5	45	1	0,993307149	5
55	5	46	0	0,994513701	5,2
56	5	46	1	0,994513701	5,2
57	5	47	0	0,995503727	5,4
58	5	47	0	0,995503727	5,4
59	5	47	1	0,995503727	5,4
60	5	48	0	0,99631576	5,6
61	5	48	1	0,99631576	5,6
62	5	48	1	0,99631576	5,6
63	5	49	0	0,996981584	5,8
64	5	49	0	0,996981584	5,8
65	5	49	1	0,996981584	5,8
66	6	50	0	0,997527377	6
67	6	50	1	0,997527377	6
68	6	51	0	0,99797468	6,2
69	6	52	0	0,998341199	6,4
70	6	52	1	0,998341199	6,4
71	6	53	1	0,99864148	6,6
72	6	53	1	0,99864148	6,6
73	6	54	1	0,998887464	6,8
74	7	55	0	0,999088949	7
75	7	55	1	0,999088949	7
76	7	55	1	0,999088949	7
77	7	56	1	0,999253971	7,2
78	7	56	1	0,999253971	7,2
79	7	56	1	0,999253971	7,2
80	7	57	0	0,999389121	7,4
81	7	57	0	0,999389121	7,4
82	7	57	1	0,999389121	7,4
83	7	57	1	0,999389121	7,4
84	7	57	1	0,999389121	7,4
85	7	57	1	0,999389121	7,4
86	7	58	0	0,999499799	7,6
87	7	58	1	0,999499799	7,6
88	7	58	1	0,999499799	7,6
89	7	59	1	0,999590433	7,8
90	7	59	1	0,999590433	7,8
91	8	60	0	0,99966465	8
92	8	60	1	0,99966465	8
93	8	61	1	0,999725422	8,2
94	8	62	1	0,999775183	8,4
95	8	62	1	0,999775183	8,4
96	8	63	1	0,999815928	8,6
97	8	64	0	0,99984929	8,8
98	8	64	1	0,99984929	8,8
99	8	65	1	0,999876605	9
100	8	69	1	0,999944551	9,8

d) Calcul des logit(X)

La fonction logit(X) peut se calculer au départ des valeurs de $\pi(X)$, ou plus simplement au départ de l'équation:

$$\text{logit}(X) = g(X) = \beta 0 + \beta 1 * X$$

Les valeurs correspondantes sont données dans la table ci-dessus (colonne F)

e) Calcul des erreurs d'estimation

Pour chaque valeur observée, l'erreur peut se mesurer par:

(valeur prédite - valeur observée)

Il n'y a donc que deux valeurs possibles d'erreur:

Y observé	Y prédit	Erreur	Probabilité
0	$\pi(X)$	$-\pi(X)$	$1-\pi(X)$
1	$\pi(X)$	$1-\pi(X)$	$\pi(X)$

La moyenne associée à l'erreur est donc:
$E[\text{Erreur}] = -\pi(X)*(1-\pi(X))+(1-\pi(X))*\pi(X) = 0$

La variance associée à l'erreur est donc:
$E[\text{Erreur}^2] = \pi^2(X)*(1-\pi(X))+(1-\pi(X))^2*\pi(X) = (1-\pi(X))*\pi(X)$

Les erreurs on donc une distribution binomiale (par opposition à la régression linéaire, où la distribution était supposée normale, de moyenne 0 et de variance σ^2)

f) Calcul de la vraissemblance (log-vraissemblance)

La vraissemblance peut se définir comme la probabilité associée à l'événement qui consiste à avoir rencontré les données observées. Cette vraissemblance est associée à un modèle que l'on postule (le modèle logistique ici), et devient alors "la probabilité des observations si le modèle postulé est vrai". Il est clair que cette probabilité dépendra des paramètres $\beta 0$ et $\beta 1$ dans notre cas, ce qui conduit à une méthode (générale) d'estimation de ces derniers: on choisit les valeurs de $\beta 0$ et $\beta 1$ qui maximisent la vraissemblance, c'est-à-dire la probabilité d'avoir rencontré ces observations.
Deux remarques s'imposent:
1) la vraissemblance est toujours positive, et est maximale en même temps que le log de la vraissemblance. Il est souvent plus simple, pour des raisons techniques, de maximiser le log que la vraissemblance elle-même, les solutions (valeurs de $\beta 0$ et $\beta 1$) étant de toutes manières identiques.
2) l'utilisation des méthodes "moindres carrés" utilisées pour estimer les paramètres dans le cadre de la régression linéaire ne sont pas applicables ici parce que l'erreur n'a pas une distribution normale. La méthode du "maximum de vraissemblance" est plus générale.

La seule chose qui est stochastique est l'erreur associée aux observations. On va donc maximiser l'erreur conjointe.

La probabilité associée à l'erreur de l'observation i peut s'écrire:

$$\pi(Xi)^{Yi} * (1-\pi(Xi))^{(1-Yi)}$$

Comme les observations sont supposées indépendantes, la probabilité conjointe est le produit des probabilités individuelles. En prenant le logarithme, ce produit devient la somme suivante:

$$L(\beta 0,\beta 1) = \Sigma \{ Yi*\ln[\pi(Xi)] + (1-Yi)*\ln[1-\pi(Xi)] \}$$

Les valeurs correspondantes sont dans le tableau ci dessous, ainsi que leur somme:

-0,00493244	-0,00493244	-0,00493244	-5,31438725	-0,004932437	
-0,00493244	-0,00493244	-0,00493244	-5,31438725	-5,314387254	
-0,00493244	-5,31438725	-0,00493244	-0,00493244	-5,314387254	
-0,00493244	-0,00493244	-0,00493244	-0,00493244	-5,314387254	
-5,31438725	-0,00493244	-5,31438725	-5,31438725	-5,314387254	
-0,00493244	-0,00493244	-0,00493244	-0,00493244	-0,004932437	
-0,00493244	-0,00493244	-0,00493244	-5,31438725	-5,314387254	
-0,00493244	-0,00493244	-5,31438725	-0,00493244	-5,314387254	
-0,00493244	-5,31438725	-0,00493244	-0,00493244	-5,314387254	
-0,00493244	-0,00493244	-0,00493244	-5,31438725	-5,314387254	
-0,00493244	-0,00493244	-5,31438725	-5,31438725	-0,004932437	
-0,00493244	-5,31438725	-5,31438725	-5,31438725	-5,314387254	
-0,00493244	-0,00493244	-0,00493244	-5,31438725	-5,314387254	
-0,00493244	-0,00493244	-5,31438725	-0,00493244	-5,314387254	
-0,00493244	-0,00493244	-0,00493244	-5,31438725	-5,314387254	
-5,31438725	-0,00493244	-5,31438725	-5,31438725	-5,314387254	
-0,00493244	-5,31438725	-0,00493244	-5,31438725	-0,004932437	
-0,00493244	-0,00493244	-0,00493244	-5,31438725	-5,314387254	
-0,00493244	-5,31438725	-5,31438725	-5,31438725	-5,314387254	Somme
-0,00493244	-0,00493244	-0,00493244	-0,00493244	-5,314387254	-228,799801

avec b0 = -5,30945482 et b1 = 0

Les valeurs de b0 et b1 ont été obtenues en maximisant la somme. La procédure utilisée pour cette maximisation consiste à faire appel à une fonctionnalité d'Excel appelée "Solveur" (dans le menu "Outils"): la cellule cible est la somme (F356 dans cette feuille), l'option est de maximiser (cocher "Max"). Les cellules variables sont celle de b0 et b1
(soit: C358;E358 dans cette feuille). On demande alors au solveur de résoudre, ce qui conduit aux valeurs fournies (on part évidemment avec des valeurs qui sont différentes, et peuvent être choisies de manière arbitraire, à condition que ces valeurs ne soient pas trop éloignées de la solution finales).
Ces valeurs de b0 et b1 sont les estimateurs au maximum de vraissemblance des vraies valeurs de $\beta 0$ et $\beta 1$, qui sont comme d'habitude, inaccessibles.

g) Calcul de la valeur estimée de la probabilité de CHD à 38 ans.

On peut calculer p(38), l'estimateur de $\pi(38)$, par:

$$p(38) = \exp(b0 + b1*38)/(1+\exp(b0 + b1*38)$$

ce qui donne:

p(38) = 0,00492029

Il y a donc à peu près 1 chance sur 200 de souffrir de CHD à 38 ans.

h) Test du modèle

Tester le modèle consiste à tester si l'inclusion de l'âge permet d'augmenter la vraissemblance.

La procédure standard consiste à tester le rapport suivant:

G = -2ln[vraiss. **avec** la variable / vraiss. **sans** la variable]

On montre que ce rapport est distribué comme χ^2 (1 ddl) sous H0: pas d'effet de la variable.

On peut écrire:

G = -2{ln[vraiss. **avec** la variable] - ln[vraiss. **sans** la variable]}

soit, G = -2*(-53.677-ln[vraiss. **sans** la variable])

Le log de la vraissemblance sans la variable s'obtient en maximisant à nouveau la vraisemblance mais cette fois en fixant b1=0 et en utilisant que b0 pour maximiser (essayez avec le solveur).

ln[vraiss. **sans** la variable] = -68.331

G = 2*(-53.677 - (-68.331)) = 29,31

valeur très significative car: $p(\chi^2 > 29.31) =$ 6,16766E-08

L'âge a un effet très significatif sur le CHD.