Exercice : Régression logistique simple

Considérons le jeu de données sur le ronflement.

AGE	*POIDS*	*TAILLE*	*ALCOOL*	*SEXE*	*TABAC*	*RONFLE*
47	71	158	A	1	1	0
56	58	164	I	0	0	1
46	116	208	M	1	1	0
…						

Tab 1. Extrait du jeu de données

L'objectif est d'expliquer le ronflement d'une personne en fonction des autres caractéristiques.

- RONFLE est la variable cible Y. C'est une variable binaire telle que RONFLE=1 si la personne ronfle et RONFLE=0 sinon
- AGE∈ [20,80], POIDS∈ [42,120], TAILLE∈ [158,208] sont des variables explicatives quantitatives
- SEXE (H=1/F=0), TABAC (oui=1/non=0) sont des variables qualitatives à deux modalités

Si dessous les résultats obtenus pour les différents modèles de régression logistique à une variable explicative.

	Variable explicative	$\hat{a}_0$	$\hat{\sigma}_0$	$\hat{a}_1$	$\hat{\sigma}_1$	Déviance
1	SEXE (H=1)	0.20000	0.09476	0.20000	0.10942	22.00
2	AGE	-0.180638	0.220530	0.010152	0.004123	21.42

1) Modèle 1
   a. Explicitez le modèle 1 avec Y=RONFLE et X=SEXE.
   b. Calculez l'odds des hommes et l'odds des femmes. Faites une phrase pour expliquer ce que signifient les résultats.
   c. Comment interprétez-vous $\hat{a}_1$?
   d. Quelle est l'augmentation du risque des hommes par rapport aux femmes ?
   e. Calculez un intervalle de confiance pour l'odds-ratio.

a. $\text{logit}(\hat{p}(x)) = \hat{a}_0 + \hat{a}_1 x = 0.2 + 0.2x$ ou bien $\log it(\hat{p}(x)) = \hat{a}_0 + \hat{a}_1 \mathbf{1}_{\{x=1\}} = 0.2 + 0.2 \times \mathbf{1}_{\{x=1\}}$

Remarque : pour une variable binaire, mettre x ou $\mathbf{1}_{\{x=1\}}$ dans le modèle revient au même

$$\Rightarrow \quad \hat{p}(x) = \frac{e^{0.2(1+x)}}{1+e^{0.2(1+x)}} = \frac{1}{1+e^{-0.2(1+x)}}$$

b. $odds(x) = \dfrac{\hat{p}(x)}{1-\hat{p}(x)} = e^{0.2(1+x)}$

Homme : $odds(1) = 0.6 = \dfrac{P(Y=1|SEXE=1)}{P(Y=0|SEXE=1)} \Rightarrow 1/odds(1) = 1{,}67 = \dfrac{P(Y=0|SEXE=1)}{P(Y=1|SEXE=1)}$

Femme : $odds(0) = 0.55 = \dfrac{P(Y=1|SEXE=0)}{P(Y=0|SEXE=0)} \Rightarrow 1/odds(0) = 1{,}82 = \dfrac{P(Y=0|SEXE=0)}{P(Y=1|SEXE=0)}$

Un homme a 1,67 plus de chance de ne pas ronfler que de ronfler alors que qu'une femme a 1,82 plus de chance de ne pas ronfler

c. L'odds-ratio entre les hommes et les femmes

$$OR_{SEXE} = \frac{odds(1)}{odds(0)} = \frac{e^{\hat{a}_0+\hat{a}_1\times 1}}{e^{\hat{a}_0+\hat{a}_1\times 0}} = e^{\hat{a}_1} = 1.22 \quad \Leftrightarrow \quad \ln(OR_{SEXE}) = \hat{a}_1 = 0.2$$

Dans le cas d'un variable binaire le coefficient a_1 correspond au logarithme de l'odds-ratio.

d. Ici on peut dire que le risque pour les hommes augmente de 1,22 par rapport à celui des femmes.

e. L'intervalle de confiance pour a_1 est $\left[\hat{a}_1 - z_{1-\alpha/2}\hat{\sigma}_1 ; \hat{a}_1 - z_{1-\alpha/2}\hat{\sigma}_1\right]$. Pour α=5%, on a $z_{1-\alpha/2}$=1.96, d'où l'IDC [-0.01 ;0.41]. Cela signifie que $P(-0.01<a_1<0.41)=0.95 \Rightarrow P(e^{-0.01}<e^{a1}<e^{0.41})=0.95 \Rightarrow P(0.99<OR_{SEXE}<1.51)=0.95$

2) Modèle 2 : AGE
   a. Explicitez le modèle 2 avec Y=RONFLE et X=AGE.
   b. Quelle est la probabilité de ronfler pour une personne de 40 ans.
   c. Calculez l'odds en fonction de l'âge. En déduire l'odds pour une personne de 80 ans. Qu'est-ce cela signifie (Faites une phrase) ?
   d. Comment interprétez-vous $\hat{a}_1$?
   e. Quel est le rapport de chance de ronfler quand on passe de 20 à 30 ans ou de 20 à 60 ans ? Pensez-vous qu'il y a une différence importante ?
   f. Afin de confirmer ou infirmer votre conclusion, effectuez un test pour déterminer si a_1 est significativement non nul.

a. $\text{logit}(\hat{p}(x)) = \hat{a}_0 + \hat{a}_1 x = -0.1806 + 0.0101x \quad \Rightarrow \quad \hat{p}(x) = \dfrac{e^{-0.1806+0.0101x}}{1+e^{-0.1806+0.0101x}} = \dfrac{1}{1+e^{-(-0.1806+0.0101x)}}$

où x est l'age $\in [20,80]$.

b. $\hat{p}(40) = \dfrac{1}{1+e^{-(-0.1806+0.0101*40)}} = 0.5953$

Il y 59,5% de risque qu'une personne de 40 ans ronfle.

c. $odds(x) = \dfrac{\hat{p}(x)}{1-\hat{p}(x)} = e^{-0.1806+0.0101x}$

$odds(80) = 1{,}87 = \dfrac{P(Y=1|AGE=80)}{P(Y=0|AGE=80)}$

Cela signifie qu'une personne de 80 ans a 1,87 fois plus de risque de ronfler que de ne pas ronfler.

d. Si on augmente l'âge de x_1 à x_2, on a l'odds-ratio

$$OR_{AGE} = \frac{odds(x_2)}{odds(x_1)} = \frac{e^{\hat{a}_0+\hat{a}_1 x_2}}{e^{\hat{a}_0+\hat{a}_1 x_1}} = e^{\hat{a}_1(x_2-x_1)} \quad \Leftrightarrow \quad \ln(OR_{AGE}) = \hat{a}_1(x_2-x_1) = 0.0101(x_2-x_1)$$

Si on augmente l'âge de 20 à 30 ans alors OR_{AGE}=1,11, c'est-à-dire que le rapport de chance est 1,11 fois plus élevé.

Si on augmente l'âge de 20 à 60 ans alors OR_{AGE}=1,5 c'est-à-dire que le rapport de chance est 1,5 fois plus élevé.
On constate que le rapport de force augmente avec l'âge et cette augmentation est une fonction du coefficient a_1.
Remarque : Qu'on augmente l'âge de 20 à 30 ans ou de 60 à 70 ans, l'odds-ration reste le même !

e. On effectue le test de Wald (équivalent du test de Student dans le cas de la régression multilinéaire)

$H_0 : a_1=0$ contre $H_1 : a_1 \neq 0$

La statistique du test est le rapport a_1/σ_1. Sous l'hypothèse H_0 elle suit une loi N(0,1), donc pour α=5%, on obtient un seuil $z_{1-\alpha/2}$=1.96. Sur l'échantillon, $\hat{a}_1 / \hat{\sigma}_1 = 2.46 > 1.96$. Donc l'âge est une variable significativement influente dans le modèle 2.
Si on effectue ce test avec le modèle 3, on obtient $\hat{a}_1 / \hat{\sigma}_1 = 0.15 < 1.96$. Donc la variable taille est non influente dans le modèle 3.