TD1, sur la Régression Logistique
(STA 2211)

Exercice 1 :

Un sondage international cité dans un article de presse (le 14 décembre 2004) rapportait le faible taux d'approbation de la politique du Président des États-Unis d'Amérique, George W. Bush, dans les pays traditionnellement alliés des États-Unis : 32% au Canada, 30% au Royaume-Uni, 19% en Espagne et 17% en Allemagne.

Soient Y la variable aléatoire binaire indiquant l'approbation d'une personne à la politique de G. W. Bush (1= approuve, 0= désapprouve), X_1 la variable indiquant si la personne est Canadienne (1=Canadien, 0=autre), X_2 la variable indiquant si la personne est Britannique (1=Britannique, 0=autre), X_3 la variable indiquant si la personne est Espagnol (1=Espagnol, 0=autre).

Montrez qu'en terme de modèle logistique les résultats du sondage se traduisent par

$$\text{logit}\Big[\widehat{\mathbb{P}}(Y=1|X_1=x_1,X_2=x_2,X_3=x_3)\Big] = -1.59 + 0.83x_1 + 0.74x_2 + 0.14x_3.$$

Solution :

$$-1.59 = \log\Big(0.17/(1-0.17)\Big)$$

$$0.83 = \log\left(\frac{0.32/(1-0.32)}{0.17/(1-0.17)}\right)$$

$$0.74 = \log\left(\frac{0.30/(1-0.30)}{0.17/(1-0.17)}\right)$$

$$0.14 = \log\left(\frac{0.19/(1-0.19)}{0.17/(1-0.17)}\right)$$

Exercice 2 :

Soit Y une variable aléatoire à valeur dans $\{0,1\}$ et X une variable aléatoire réelle. Supposons que la distribution de X sachant $Y=j$ soit $\mathcal{N}\big(\mu_j,\sigma^2\big)$ pour $j=0,1$.
Montrez que la distribution de Y sachant X suit un modèle logistique et exprimez les paramètres du modèle en fonction de μ_1, μ_2, σ et $\rho=\mathbb{P}(Y=1)$.

Solution :

D'après le théorème de Bayes,

$$\mathbb{P}(Y=1|X=x) = \frac{f(x|Y=1)\mathbb{P}(Y=1)}{f(x|Y=1)\mathbb{P}(Y=1) + f(x|Y=0)\mathbb{P}(Y=0)},$$

avec

$$f(x|Y=j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_j)^2}{2\sigma^2}\right], \quad j=0,1.$$

D'où

$$\begin{aligned}
\mathbb{P}(Y=1|X=x) &= \frac{\rho \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right]}{\rho \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right] + (1-\rho)\exp\left[-\frac{(x-\mu_0)^2}{2\sigma^2}\right]} \\
&= \frac{1}{1+\frac{(1-\rho)}{\rho}\exp\left[\frac{(x-\mu_1)^2}{2\sigma^2} - \frac{(x-\mu_0)^2}{2\sigma^2}\right]} \\
&= \frac{1}{1+\frac{(1-\rho)}{\rho}\exp\left[\frac{(2x-\mu_1-\mu_0)(\mu_0-\mu_1)}{2\sigma^2}\right]} \\
&= \frac{1}{1+\frac{(1-\rho)}{\rho}\exp\left[\frac{2x(\mu_0-\mu_1)-(\mu_1^2-\mu_0^2)}{2\sigma^2}\right]}
\end{aligned}$$

et ainsi,

$$\begin{aligned}
\mathbb{P}(Y=1|X=x) &= \frac{1}{1+\exp[-(\alpha+\beta x)]} \\
&= \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}
\end{aligned}$$

avec

$$\alpha = -\log\left[\frac{(1-\rho)}{\rho}\right] + \frac{\mu_1^2-\mu_0^2}{2\sigma^2}$$

$$\beta = \frac{\mu_1-\mu_0}{\sigma^2}$$

Exercice 3 :

En épidémiologie, il est fréquent de s'intéresser à l'étude de maladies rares. Lorsqu'on sélectionne aléatoirement un échantillon de personnes dans la population (par exemple par tirage au sort sur les listes électorales), les données contiennent alors souvent très peu de personnes atteintes de cette maladie.

Question 1

Soit π la prévalence de la maladie rare dans la population, c'est à dire la probabilité qu'un individu tiré au hasard soit malade. Soit n la taille d'un échantillon de la population tiré au hasard. Soit Z_n le nombre de personnes malades observé dans un échantillon aléatoire de taille n.

1.1 Proposez un intervalle de fluctuation à 95% de Z_n pour n grand, c'est à dire un intervalle dans lequel Z_n sera inclut avec une probabilité de 95% pour n grand.

Solution :

D'après le Théorème Central Limite (TCL), pour n grand alors on a approximativement

$$\sqrt{n}\left(\frac{Z_n}{n} - \pi\right) \sim \mathcal{N}\Big(0, \pi(1-\pi)\Big).$$

D'où, pour n grand, l'intervalle

$$\Big[n\pi \pm z_{1-\alpha/2}\sqrt{n\pi(1-\pi)}\Big], \tag{1}$$

où $z_{1-\alpha/2}$ est le quantile à $100(1-\alpha/2)\%$ d'un loi normale centrée réduite, contient Z_n avec une probabilité approximativement égale à α. Pour $= 5\%$, on connaît $z_{1-\alpha/2} = 1.96$.

1.2 En déduire un intervalle de fluctuation de Z_n pour $n = 1000$ et $\pi = 0.006$.

Solution :

Numériquement,

$$\Big[1000 \times 0.006 \pm 1.96\sqrt{1000 \times 0.006(1-0.006)}\Big] \approx [1; 11]$$

1.2 Supposons que **(i)** $\pi = 0.006$, **(ii)** que les épidémiologies n'ont d'argent que pour interroger $n = 1000$ personnes et **(iii)** qu'ils récoltent leurs données en interrogeant des personnes tirées aléatoirement sur les listes électorales.
Pensez vous que les épidémiologistes auront alors probablement assez de données pour estimer un modèle logistique pour étudier cette maladie rare ?

Question 2

Les épidémiologistes souhaitent collecter leurs données retrospectivement, c'est à dire qu'ils choisissent d'interroger n_1 patients malades (récemment diagnostiqués à l'hôpital) et n_0 personnes non malades, ces derniers étant tirés au sort sur les listes électorales, pour faire des comparaisons.

Soit Y la variable aléatoire binaire à expliquer ($Y = 1$ signifiant malade, $Y = 0$ non malade) et $X = (X_1, \ldots, X_p)$ le vecteurs des variables explicatives qui modélise les informations qu'ils récoltent sur chaque personne. On cherche à modéliser la probabilité de Y sachant X dans la **population générale**.

2.1 Rappelez la définition du modèle logistique.

Solution :

Pour tout $X = (X_1, \ldots, X_p) = (x_1, \ldots, x_p) = x$:

$$\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_j}}$$

On définit par S la variable aléatoire binaire qui indique si une personne (dans la population générale) est inclue dans l'échantillon ($S = 1$) ou non ($S = 0$).

2.2 Quelles sont les valeurs des probabilités $\tau_1 = \mathbb{P}(Y = 1|S = 1)$ et $\tau_0 = \mathbb{P}(Y = 0|S = 1)$?

Solution :

$$\tau_1 = n_1/(n_1 + n_0)$$
$$\tau_0 = n_0/(n_1 + n_0)$$

2.3 Pourquoi est-il raisonnable de supposer que quelque soit $j \in \{0, 1\}$, pour tout $x \neq x'$ on a la relation suivante ?

$$\mathbb{P}(S = 1|Y = j, X = x) = \mathbb{P}(S = 1|Y = j, X = x') \tag{2}$$

Solution :
Ici on suppose que les informations (X) sont récoltés après le tirage au sort. Les probabilités d'être tiré au sort (sachant le statut malade ou non) n'ont donc aucune raison de dépendre de X (i.e. des réponses des personnes aux questions de l'enquête).

Dans la suite on suppose (2) et on pose $\tau_1 = \mathbb{P}(S = 1|Y = 1)$, $\tau_0 = \mathbb{P}(S = 1|Y = 0)$, $p(x) = \mathbb{P}(Y = 1|X = x, S = 1)$ et $\pi(x) = \mathbb{P}(Y = 1|X = x)$.

2.3 Montrez que

$$\text{logit}\{p(x)\} = \log\left(\frac{\tau_1}{\tau_0}\right) + \text{logit}\{\pi(x)\} \tag{3}$$

Solution :

$$
\begin{aligned}
p(x) &= \mathbb{P}(Y=1|X=x,S=1) \\
&= \frac{\mathbb{P}(S=1|Y=1,X=x)\mathbb{P}(Y=1|X=x)}{\mathbb{P}(S=1|Y=1,X=x)\mathbb{P}(Y=1|X=x)+\mathbb{P}(S=1|Y=0,X=x)\mathbb{P}(Y=0|X=x)} \\
&= \frac{\tau_1\mathbb{P}(Y=1|X=x)}{\tau_1\mathbb{P}(Y=1|X=x)+\tau_0\mathbb{P}(Y=0|X=x)} \\
&= \frac{\tau_1\pi(x)}{\tau_1\pi(x)+\tau_0\big[1-\pi(x)\big]} \\
&= \left(\frac{\tau_1}{\tau_0}\times\frac{\pi(x)}{1-\pi(x)}\right)\Big/\left(\frac{\tau_1}{\tau_0}\times\frac{\pi(x)}{1-\pi(x)}+1\right)
\end{aligned}
$$

d'où

$$
\begin{aligned}
\operatorname{logit}\big\{p(x)\big\} &= \log\left(\frac{\tau_1}{\tau_0}\times\frac{\pi(x)}{1-\pi(x)}\right) \\
&= \log\left(\frac{\tau_1}{\tau_0}\right)+\operatorname{logit}\{\pi(x)\}
\end{aligned}
$$

Question 3

En déduire si la collecte rétrospective des données permet d'estimer simplement certains paramètres du modèle logistique, et si oui, lesquels ?

Conclure : une approche consistant à collecter rétrospectivement des données pour ensuite les analyser avec un modèle logistique vous parait-elle puissante pour l'étude des facteurs de risques d'une maladie rare ?

Oui, tous les paramètres sauf β_0.

La collecte rétrospective de données et l'utilisation du modèle logistique est une méthode puissante pour l'étude des facteurs de risque d'une maladie rare