Exercice

Régression logistique avec une seule variable explicative

On considère le cas d'une variable aléatoire réelle qualitative Y à deux modalités (0/1) que l'on souhaite expliquer à l'aide d'une variable aléatoire réelle quantitative X. Pour cela, on dispose des valeurs $(x_1, y_1), \ldots, (x_n, y_n)$ qui sont les valeurs observées d'un n-échantillon aléatoire $(X_1, Y_1), \ldots, (X_n, Y_n)$. Par la suite, on utilise le modèle de la régression logistique pour expliquer la variable Y à l'aide de la variable X. Dans cette régression logistique, la constante et le coefficient de X seront notés respectivement β_0 et β_1. Pour tout réel x, on notera $\pi(x) = \mathbb{E}(Y \mid X = x)$.

PARTIE 1 - *Quelques résultats théoriques*

1. Donner l'expression de $\pi(x)$ en fonction de β_0, β_1 et x.

2. Montrer que la log-vraisemblance du paramètre $\beta = (\beta_0, \beta_1)$ associée aux valeurs observées $y_1, \ldots, y_n$ s'écrit :

$$\ln L(\beta) = \sum_{i=1}^{n} \big[y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \big].$$

3. Écrire les deux équations du maximum de vraisemblance (MV), c.-à-d.

$$\frac{\partial}{\partial \beta_0} \ln L(\beta) = \frac{\partial}{\partial \beta_1} \ln L(\beta) = 0,$$

en fonction de β_0, β_1 et des x_i, y_i ($i \in \{1, \ldots, n\}$). Par la suite, on notera $\widehat{\beta_0}$ et $\widehat{\beta_1}$ les solutions de ces équations qui sont donc les estimateurs du MV respectifs de β_0 et β_1.

4. Quel estimateur $\widehat{\pi}(x)$ de $\pi(x)$ peut-on déduire de ce qui précède ?

5. Calculer la matrice d'information de Fisher du paramètre β, que l'on notera $\mathbb{I}(\beta)$.

6. Expliquer comment l'on peut déduire des résultats précédents une estimation des variances de $\widehat{\beta_0}$, de $\widehat{\beta_1}$ et de la covariance entre $\widehat{\beta_0}$ et $\widehat{\beta_1}$,

PARTIE 2 - *Application à des données réelles*

Lors d'une enquête de santé publique, on a observé que 133 individus souffraient d'une maladie chronique parmi 307 personnes âgées entre 18 et 85 ans. Étant donné que la proportion de personnes atteintes d'une maladie chronique augmente avec l'âge, on applique le modèle logistique pour estimer la probabilité d'être atteint d'une maladie chronique en fonction de l'âge. Á l'aide d'un logiciel statistique, on a obtenu les résultats suivants :

$$\widehat{\beta_0} = -2,284 \quad \text{et} \quad \widehat{\beta_1} = 0,04468.$$

$$\widehat{\mathbb{V}}(\widehat{\beta}) = \begin{pmatrix} 0,1349 & -0.2639 \times 10^{-2} \\ -0.2639 \times 10^{-2} & 0,5814 \times 10^{-4} \end{pmatrix}$$

1. Donner l'expression de $\widehat{\pi}(x)$ en fonction de x. Quelle est la probabilité (estimée) d'être atteint d'une maladie chronique à l'âge de 50 ans ?

2. En utilisant un test de Student, montrer que l'âge est une variable explicative significative.

3. Déterminer un intervalle de confiance, au seuil de confiance 95%, de la probabilité d'être atteint d'une maladie chronique à l'âge de 50 ans.