Chapitre 1

La régression logistique et son application en biologie

Introduction

1.1 Cadre théorique de la régression logistique

La régression logistique est l'une des méthodes d'analyse multivariée les plus couramment utilisées en épidémiologie. Elle a pour but de mesurer la survenue d'un évènement (variable à expliquer) et les variables susceptibles de l'influencer (variables explicatives). Il s'agit donc d'un modèle permettant de relier la variable dépendante (Y) à des variables explicatives (X_1, X_2, ..., X_n).

À la différence de la régression linéaire (où la variable à expliquer est quantitative), la régression logistique s'applique lorsque la variable à expliquer Y est qualitative de type binaire (oui/non), (présence/absence)....

Dans le cas d'une variable explicative qualitative, une propriété très intéressante de la régression logistique est qu'elle permet d'estimer un odds ratio (OR) qui fournit une information sur la force et le sens de l'association entre la variable explicative (X_i) et la variable à expliquer Y.

1.2 Régression logistique binaire simple

On parle de la régression logistique binaire lorsque la variable à expliquer Y est de type binaire :

- 0 en cas de non occurrence de l'évènement.
- 1 si occurrence.

1.2.1 Définition mathématique du modèle de régression logistique

ou bien : Spécification du modèle

L'astuce de la régression logistique consiste non pas à modéliser la variable qualitative Y mais la probabilité que celle-ci se réalisé. Y aléatoire et X non aléatoire.

- On cherche à expliquer la survenue d'un évènement.
- On cherche la probabilité de succès.
- On travaille en terme d'espérance.

La fonction de régression de Y par rapport à X à estimer est l'espérance de Y conditionnelle à X :

$$E(Y|X = x) = 1.p + 0(1 - p) = p(x)$$

où $p(x) = P(Y = 1|X = x)$

Les variables Y_i sont indépendantes entre elles et suivant la loi de Bernoulli de paramètre p_i

$$p_i = E(Y_i|X_i) = P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Le problème est de déterminer comment la probabilité de "succès" évolue en fonction du niveau de la variable X.

Le modèle logistique stipule que la probabilité conditionnelle de succès est de la forme :

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

La fonction $g(u) = \frac{e^u}{1+e^u}$ est appelée **fot logistique**, elle est strictement croissante et prend ses valeurs dans l'intervalle [0,1]. Sa fonction inverse est :

$$g^{-1}(u) = \ln \frac{u}{1 - u}$$

est s'appelle **fot logit**.

• Pour une loi de Bernoulli le rapport $\frac{p}{1-p}$ a une certaine signification. On l'appelle parfois la chance ou la cote de succès. Dans le modèle logistique, le logarithme de ce rapport est donc

une fonction ln de la variable explicative :

$$\ln \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

Le modèle comporte donc deux paramètre β_0 et β_1. On notera par β le couple (β_0 , β_1).

• Contrairement à la régression classique, il n'y a pas de variance de l'erreur à estimer.

1.2.2 L'odds Ratio

L'odds ratio noté OR est l'outil qui nous permet une interprétation facile et logique des résultats d'une régression logistique.

Définition 1.

L'odds ratio (Rapport des chances) est le rapport de deux odds associés à deux valeurs différentes de X respectivement x et t avec x=t+1.

$$odds(x) = \frac{p(x)}{1-p(x)}$$
$$odds(t) = \frac{p(t)}{1-p(t)}$$

avec

$$p(x) = P(Y=1|X=x) = \frac{e^{\beta x}}{1-e^{\beta x}} \qquad \text{(prob de succès)}$$
$$p(t) = P(Y=1|X=t) = \frac{e^{\beta t}}{1-e^{\beta t}}$$

Ainsi

$$OR(x,t) = \frac{odds(x)}{odds(t)} = \frac{p(x)}{1-p(x)}.\frac{1-p(t)}{p(t)}$$

• Si la variable explicative X est quantitative, on obtient en posant X=t+1 et en fixant les autres variables :

$$OR_{x|t} = \frac{P(Y_i=1|X_i=x)}{1-P(Y_i=1|X_i=x)}.\frac{1-P(Y_i=1|X_i=t)}{P(Y_i=1|X_i=t)}$$

on trouve $OR = e^{\beta}$

• Si la variable explicative X est qualitative binaire, l'odds ratio permet de comparer les odds

de deux modalités de cette variable. Pour l'individu statistique i, on a :

$$OR_{1|0} = \frac{P(Y_i = 1|X_i = 1)}{1 - P(Y_i = 1|X_i = 1)} . \frac{1 - P(Y_i = 1|X_i = 0)}{P(Y_i = 1|X_i = 0)}$$

$OR_{1|0}$ représente le rapport de cote du risque de la survenue de l'évènement chez les individus exposés ($X_i = 1$) par rapport aux individus statistiques non exposés ($X_i = 0$).

Interprétation de l'OR

a. Pour la variable explicative quantitative

Le logarithme népérien de la probabilité de survenue d'un évènement augmente de $\hat{\beta}$ pour chaque unité supplémentaire de X.

- Si OR > 1, X est considérée comme facteur favorisant du risque.
- Si OR < 1, X est considérée comme facteur handicapant la survenue de l'évènement.
- Si OR $= 1$, X est considérée sans effet sur la survenue de l'évènement.

b. Pour la variable explicative qualitative

- Si OR > 1, le risque de la survenue de l'évènement ($Y_i = 1$) chez les individus exposés ($X_{ij} = 1$) est plus élevé que celui chez les individus non exposés ($X_{ij} = 0$).

$\rightarrow$ Dans ce cas, la variable explicative X est appelée *facteur favorisant.*

- Si OR < 1, le risque de la survenue de l'évènement ($Y_i = 1$) chez les individus exposés ($X_{ij} = 1$) est plus faible que celui chez les individus non exposés ($X_{ij} = 0$).

$\rightarrow$ X est appelée *facteur handicapant* ou *freinant.*

- Si OR $= 1$, le risque de la survenue de l'évènement ($Y_i = 1$) chez les individus exposés ($X_{ij} = 1$) est égale à celui chez les individus non exposés ($X_{ij} = 0$).

$\rightarrow$ X est considérée sans effet sur la survenue de l'évènement.

1.2.3 Estimation des paramètres

Pour la régression logistique, la méthode adéquate pour estimer les paramètres β est la méthode de maximum de vraisemblance. Supposons que nous observons indépendamment les variables aléatoires binaires $Y_1, Y_2, ..., Y_n$ au points $x_1, x_2, ..., x_n$ de la variable explicative.
Pour tout i, $Y_i \sim B(p(x_i))$, et la fonction de probabilité de Y_i est :

$$p(y) = p(x_i)^y (1 - p(x_i))^{1-y}, \qquad y \in \{0, 1\}$$

-Déterminons l'estimateur de maximum de vraisemblance de β
La fonction de vraisemblance de β associée à une réalisation $(y_1, y_2, ..., y_n)$ de $(Y_1, Y_2, ..., Y_n)$ est

donnée par :

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} P(Y = y_i | X = x_i) \\ &= \prod_{i=1}^{n} [p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}] \end{aligned}$$

avec $p(x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1-\exp(\beta_0+\beta_1 x_i)}$

La log-vraisemblance est égale à :

$$\begin{aligned} \log L(\beta) &= \log\{\prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}\} \\ &= \sum_{i=1}^{n} [\log(p(x_i)^{y_i}(1 - p(x_i))^{1-y_i})] \\ &= \sum_{i=1}^{n} \{y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))\} \end{aligned}$$

on obtient $\hat{\beta}$ en annulant $\frac{\partial \log L(\beta)}{\partial \beta}$

Ainsi on doit dériver cette fonction par rapport à β_0 et β_1.

la fonction logistique :

$$g(u) = \frac{e^u}{1 + e^u}$$

La dérivée de la fonction logistique est :

$$\begin{aligned} g^{'}(u) &= \frac{e^u}{(1 + e^u)^2} = \frac{e^u}{1 + e^u} . \frac{1}{1 + e^u} \\ &= g(u)(1 - g(u)) \end{aligned}$$

Ainsi :

$$\begin{cases} \frac{\partial}{\partial \beta_0} p(x_i) = p(x_i)[1 - p(x_i)] \\ \frac{\partial}{\partial \beta_1} p(x_i) = x_i p(x_i)[1 - p(x_i)] \end{cases}$$

La dérivée du i$^{\text{ème}}$ terme de la log-vraisemblance par rapport à β_0 est donc :

$y_i \frac{p(x_i)(1-p(x_i))}{p(x_i)} - (1 - y_i) \frac{p(x_i)(1-p(x_i))}{1-p(x_i)}$

$= y_i(1 - p(x_i)) - p(x_i) + y_i p(x_i)$

$= y_i - p(x_i)$

D'où les équations de vraisemblance sont :

$$\begin{cases} \frac{\partial}{\partial\beta_0}\log L(\beta) = \sum_{i=1}^{n}\{y_i - p(x_i)\} = 0 \\ \frac{\partial}{\partial\beta_1}\log L(\beta) = \sum_{i=1}^{n}\{x_i[y_i - p(x_i)]\} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n}\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Ces équations n'ont pas de solution explicite et paraissent complexes, mais elles peuvent être résolues de façon itérative pour trouver l'EMV de β qui est $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$, et par la suite, l'estimateur de la fonction de régression en x quelconque est : ‘

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}}$$

Il existe des algorithmes numériques itératifs permettant d'obtenir cette solution :

— Algorithme de Newton-Raphson.

— Algorithme du score de Fisher.

L'algorithme de Newton-Raphson

L'algorithme de Newton-Raphson est une des méthodes numériques les plus utilisées pour optimiser la log-vraisemblance. Il démarre avec une initialisation arbitraire du paramètre β_0, pour passer de l'étape (i) à l'étape (i+1).

Il se rapproche de la solution finale $\hat{\beta}$ en utilisant la formule suivante :

$$\beta^{i+1} = \beta^i - [\frac{\partial^2 \log L}{\partial\beta\partial\beta'}]^{-1}.\frac{\partial \log L}{\partial\beta}$$

$$\beta_{i+1} = \beta_i - H^{-1}(\beta_i).\nabla \log L(\beta_i)$$

où H(β) désigne la matrice hessienne avec $H(\beta_0) = \nabla^2 \log L(\beta_0) = \frac{\partial^2 \log L(\beta_0)}{\partial\beta_k\partial\beta_l}$, $\quad 0 < k < l < n$ et $\nabla \log L(\beta_i) = \frac{\partial}{\partial\beta}\log L$ est le gradient au point β_i.

$*$ Sur R, le modèle logistique s'ajuste avec la fonction glm :

>model ← glm(Y∼X, data= données, Family=binomial)

>summary(model)

↩

les résultats.

Quelques remarques

Plusieurs règles d'arrêt sont possibles pour stopper le processus de recherche :

1. On fixe à l'avance le nombres maximum d'itérations pour limiter le temps de calcul. C'est utile pour éviter les boucles infinis faute de convergence.
2. On stoppe les itérations lorsque l'évolution de la log-vraisemblance d'une étape à l'autre n'est pas significative. Pour cela, on fixe souvent un seuil ε, on arrêtte les processus si l'écart d'une étape à l'autre est plus petit que le seuil.
3. On stoppe les itérations lorsque l'écart entre les vecteurs solutions $\hat{\beta}$ est faible d'une étape à l'autre. Ici également, souvent il s'agit de fixer un seuil à l'avance auquel on compare la somme des écarts aux carrés où la somme des écarts absolus entre les composantes des vecteurs solutions.

1.2.4 Signification statistique des paramètres : (test $H_0 : \beta_1 = 0$)

Une fois les paramètres sont estimés, s'interroge sur leur signification statistique au seuil α. Pour répondre à cette question, il existe plusieurs tests, les plus utilisés sont :

— Test de Wald

— Test du rapport de vraisemblance

⋆ **Test de Wald**

Les hypothèses du test sont :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

En raison de la normalité asymptotique du maximum de vraisemblance, sous H_0 la statistique $\frac{\hat{\beta}_1}{S(\hat{\beta}_1)}$ où $S(\hat{\beta}_1) = \sqrt{s^2(\hat{\beta}_1)}$ (avec $s^2(\hat{\beta}_1)$ est l'estimation de la variance de $\hat{\beta}_1$) suit approximativement une loi N(0,1), et on rejettera H_0 au niveau 0.05 si sa réalisation n'est pas comprise dans l'intervalle ± 1.96.

Parfois ce test est présenté avec le carre de la statistique ci-dessus telle que $w = \frac{\hat{\beta}_1^2}{S^2(\hat{\beta}_1)} \sim \chi^2(1)$ dont la valeur critique doit être lue sur une loi $\chi^2(1)$.

⋆ **Test du rapport de vraisemblance**

Ce test est fondé sur la déviance : $-2\,[\log L(\hat{\beta}_{H_0}) - \log L(\hat{\beta})\,]$ où $\hat{\beta}_{H_0}$ est la valeur de β maximisent la log vraisemblance $\log L(\beta)$ sous l'hypothèse H_0 c-à-d avec : $p(x_i) = \frac{e^{\beta_0}}{1+e^{\beta_0}} = p_0$

Alors la première équation de vraisemblance est $\sum_{i=1}^{n}(y_i - p_0) = 0$, dont la solution est

$\hat{p}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ (la proportion de succès observée).

Cette solution est naturelle puisque sous H_0, les Y_i sont des variables aléatoire de même moyenne et donc indépendant identiquement distribués (i.i.d), en déduit : $\hat{\beta}_{H_0} = \log\frac{\hat{p}_0}{1-\hat{p}_0}$, qui permet de calculer la déviance ci-dessus qui suit approximativement une loi $\chi^2(1)$ car H_0 ne spécifie qu'un seul paramètre. Ce test donne des décisions généralement en accord avec celles du test de Wald.

Exemple

Lors d'une enquête de santé publique 307 individus d'âge variant entre 18 et 35 ans ont été étudiés. Permis ceux-ci 133 souffraient d'une maladie chronique. Sachant que la proportion de personnes ayant une maladie chronique augmente avec l'âge, on envisage un modèle logistique pour estimer la probabilité d'un tel type d'affectation en fonction de l'âge.

Écrivons le modèle logistique :

Soit Y la variable désignant la maladie chronique.

X l'âge de la personne ayant une maladie chronique.

La probabilité d'avoir une maladie chronique à l'âge x est :

$$p(x) = P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 - e^{\beta_0+\beta_1 x}}$$

Le modèle logistique est donnée par :

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

une estimateur de p(x) par la méthode de maximum de vraisemblance est :

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}}$$

La solution des deux équations de vraisemblance par R a donnée :

$$\hat{\beta}_0 = -2.284$$
$$\hat{\beta}_1 = 0.04468$$

X=c(18 : 35) , $Y = \begin{cases} 1 & \text{si maladie} \\ 0 & \text{sinon} \end{cases}$

$p = \frac{133}{307} = 0.433 \rightarrow$ proportion de personne ayant une maladie chronique

$\hat{p}(x) = \frac{e^{-2.284+0.04468x}}{1+e^{-2.284+0.04468x}}$