

LA RÉGRESSION LOGISTIQUE

Spécialisé dans le traitement statistique des études cliniques,
le Département Biométrie de FOVEA vous propose
au travers de ce fascicule quelques notions simples
concernant la régression logistique.



Traitement statistique

DES ETUDES CLINIQUES

LA RÉGRESSION LOGISTIQUE EN PRATIQUE

L'objectif principal de l'épidémiologie est de rechercher les causes des maladies, ces causes étant le plus souvent la résultante de plusieurs facteurs de risques associés. Dans la recherche clinique, on s'intéresse également aux facteurs pronostiques : ce sont les facteurs qui influent sur le devenir de la maladie étudiée. Les principaux modèles multivariés permettant d'identifier les facteurs de risque (ou les facteurs pronostiques) sont au nombre de 3 :

- **La régression linéaire multiple** : la variable à expliquer est quantitative (PAS).
- **Le modèle de Cox** : la variable à expliquer est dichotomique (Malade / Non malade) et on s'intéresse à la date de survenue de la maladie.
- **La régression logistique** : la variable à expliquer est dichotomique (Malade/Non malade) et la maladie est caractérisée par un risque.

1

Introduction

En épidémiologie comme en recherche clinique, lorsque l'on veut évaluer l'influence d'un ou de plusieurs facteurs sur la survenue ou le devenir d'une maladie, une des premières solutions qui vient à l'esprit est de constituer des groupes en fonction de la présence de 1, 2, ..., n facteurs et d'estimer le risque lié à chaque situation, par la fréquence de la maladie ou du résultat observé dans chacun des groupes.

L'inconvénient de cette méthode est de multiplier le nombre de groupes parallèlement au nombre de facteurs et de se heurter très rapidement à un

problème d'effectif. En effet, en appliquant la formule :

$$\text{Nombre de groupes} = 2^n$$

(n étant le nombre de facteurs étudiés)

on arrive très rapidement à un nombre de groupes important :

3 facteurs → 8 (2^3) groupes
4 facteurs → 16 (2^4) groupes
5 facteurs → 32 (2^5) groupes
...

Il faut alors faire appel à un **modèle permettant de combiner les risques liés à plusieurs facteurs**. Plusieurs modèles sont utilisables, parmi ceux-ci la régression (ou modèle) logistique.

2

Principes généraux

Le modèle de régression logistique ou modèle logistique est un modèle multivarié qui permet d'exprimer sous forme de **risque** (ou de **probabilité**) la **relation** entre **une variable Y dichotomique** et **une ou plusieurs variables X_i**, qui peuvent être **qualitatives** ou **quantitatives**.

- Y caractérise **la maladie** (Présence/Absence, Guérison/Echec, ...);
- Les X_i caractérisent **les i facteurs de risque** (ou **facteurs pronostiques**) **de la maladie**.



Ce modèle permet de calculer le **risque de survenue** ou le **probabilité de guérison** (ou **d'échec**) de la **maladie** lorsque les valeurs des variables X_i sont connues.

D'une façon générale, la **formule de la régression** (ou **modèle**) **logistique** s'écrit :

$$P(M^+|X_i) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i X_i)}}$$

$P(M^+|X_i)$ correspond à la probabilité ou risque de survenue de la maladie en fonction de l'exposition ou non à un ou plusieurs facteurs (X_1, \dots, X_p).

Ces facteurs sont les **caractéristiques des patients** : âge, poids, sexe, tabagisme, consommation d'alcool, ...

Comme on le constate, les facteurs X_i peuvent être **qualitatifs à 2 classes**, prenant alors la valeur 0 ou 1 selon que le facteur soit absent ou présent. Ils peuvent être également **quantitatifs**.

Dans ce cas, le modèle suppose que les distributions des facteurs quantitatifs soient **normales** et que les **relations** entre **ces facteurs** et **la maladie** soient **linéaires**. Les conclusions que l'on peut tirer des résultats d'un modèle logistique sont conditionnées par la vérification de ces hypothèses (Tests d'adéquation).

Ce modèle est appelé fonction logistique car sa transformation logistique en fait une fonction linéaire :

$$\text{Logit } P = \beta_0 + \sum \beta_i X_i$$

dépend de la fréquence de la maladie ;

β_i mesure l'association entre le facteur i et la maladie :

- si $\beta_i = 0$ l'association entre la maladie et le facteur i est nulle,
- si β_i est grand l'association est forte.

et β_i sont estimés par la **méthode du maximum de vraisemblance** ou par la **méthode des moindres carrés**.

3

Méthodologie

Le modèle logistique peut être à un ou plusieurs facteurs.

n Un seul facteur X

Dans le cas d'un seul facteur X , peut se poser l'intérêt d'utiliser le modèle logistique.

Deux raisons amènent à faire ce choix :

- Le modèle logistique permet d'exprimer l'association entre la maladie et l'exposition au facteur étudié au moyen de l'**odds ratio**, indicateur très fréquemment employé en épidémiologie ;

- La courbe de la fonction logistique a une forme sigmoïde qui correspond à une forme souvent observée dans la relation dose-effet.

Si le facteur X est dichotomique (Présence/Absence, Exposition/Non exposition, ...), le modèle logistique s'écrit :

$$P(M^+|X) = \frac{1}{1 + e^{-(\beta_0 + \beta X)}}$$

avec $X = 1$ quand le facteur est présent (sujet exposé), $X = 0$ quand le facteur est absent (sujet non exposé).

Pour les sujets exposés ($X = 1$) :

- La **probabilité de survenue de la maladie** quand le **facteur** est **présent** [$P(M^+|X = 1)$] s'écrit :

$$P_1 = \frac{1}{1 + e^{-(\beta_0 + \beta)}}$$

- Alors que la **probabilité de non survenue de la maladie** quand le **facteur** est **présent** [$P(M^-|X = 1)$] s'écrit :

$$1 - P_1 = \frac{e^{-(\beta_0 + \beta)}}{1 + e^{-(\beta_0 + \beta)}}$$

Pour les sujets non exposés ($X = 0$) :

- La **probabilité de survenue de la maladie** quand le **facteur** est **absent** [$P(M^+|X = 0)$] s'écrit :

$$P_0 = \frac{1}{1 + e^{-\beta_0}}$$

($\beta = 0$ car il n'existe pas dans ce cas de liaison entre le facteur de risque et la maladie)

LA RÉGRESSION LOGISTIQUE EN PRATIQUE

- La probabilité de non survenue de la maladie quand le facteur est absent [$P(M|X = 0)$] s'écrit :

$$1 - P_0 = \frac{e^{-\beta_0}}{1 + e^{-\beta_0}}$$

D'où la valeur de l'odds ratio :

$$OR = \frac{P_1 (1 - P_0)}{P_0 (1 - P_1) = e^{\beta_1}}$$

(P_1 est la fréquence de la maladie chez les patients exposés
 P_0 est la fréquence de la maladie chez les patients non exposés)

Le coefficient du facteur X dans le modèle logistique est donc le logarithme de l'odds ratio mesurant l'association entre X et la maladie.

n Plusieurs facteurs X_i

Lorsque plusieurs facteurs X_i ($i = 1, \dots, p$) interviennent, la formule du modèle logistique devient :

$$P(M^+|X_1, \dots, X_p) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i X_i)}}$$

d'où

$$\text{Logit } P = \beta_0 + \sum \beta_i X_i$$

Ce modèle permet d'exprimer la probabilité de survenue de la maladie (ou de la guérison ou de l'échec) en fonction des valeurs prises par les facteurs X_i .

La valeur du coefficient β_i du facteur X_i dépend de la présence des autres facteurs. Un même facteur X_i n'a pas nécessairement le même coefficient β_i dans un modèle où il est seul et dans un modèle où figurent d'autres facteurs. Lorsque l'on veut donner **plus de poids à l'un des facteurs (E)** auquel on s'intéresse plus particulièrement, on l'isole en utilisant la formule suivante :

$$\text{Logit } P = \beta_0 + E + \beta_i X_i$$

e^{β_i} est alors l'odds ratio lié à l'exposition E ajustée sur les X_i .

4

Conclusion

La régression logistique est l'un des 3 modèles multivariés les plus utilisés en épidémiologie.

Elle est particulièrement indiquée lorsque l'on veut exprimer une variable dichotomique (Malade/Non malade) sous forme d'un risque (ou probabilité) en fonction des valeurs prises par un certain nombre de facteurs, appelés facteurs de risque ou facteurs pronostiques.

Ces facteurs caractérisent les sujets analysés (sexe, âge, poids, présence ou non d'une anomalie clinique, ...).

Le modèle de régression logistique peut fonctionner avec **un seul facteur de risque** (ou **pronostique**) : même dans ce cas et plus particulièrement lorsque le facteur est **quantitatif**, son **utilisation** est **préférable** à la **segmentation de l'échantillon en plusieurs sous-groupes**.

