« Cours Statistique et logiciel R »

Rémy Drouilhet [1], Adeline Leclercq-Samson [1],
Frédérique Letué [1], Laurence Viry [2]

[1] Laboratoire Jean Kuntzmann, Dép. Probabilites et Statistique,

[2] Laboratoire Jean Kuntzmann, Dép. Modele et Algorithmes Déterministes

mars-mai 2016

Plan de la présentation

Introduction

La régression logistique permet de relier

- une variable qualitative (souvent binaire)
- à des variables $X_1, \cdots, X_p$

Exemples

- En épidémiologie, survenue d'une maladie à un groupe de facteurs de risque, avec des poids spécifiques pour chaque facteur de risque
- En fiabilité : survenu d'un accident sur une ligne de production

Exemple

Le but est d'expliquer une pathologie coronarienne par la présence de deux facteurs de risque bien connus :

- la consommation de tabac
- la cholestérolémie.

Chaque patient est représenté par 3 variables. La variable *pathologie coronarienne* est binaire (1 pour oui, 0 pour non), la variable *tabagisme* et la variable *cholestérolémie*.

Idée basique

Une première idée est de modéliser la probabilité d'être atteint d'une coronopathie par une régression linéaire avec une équation du type :

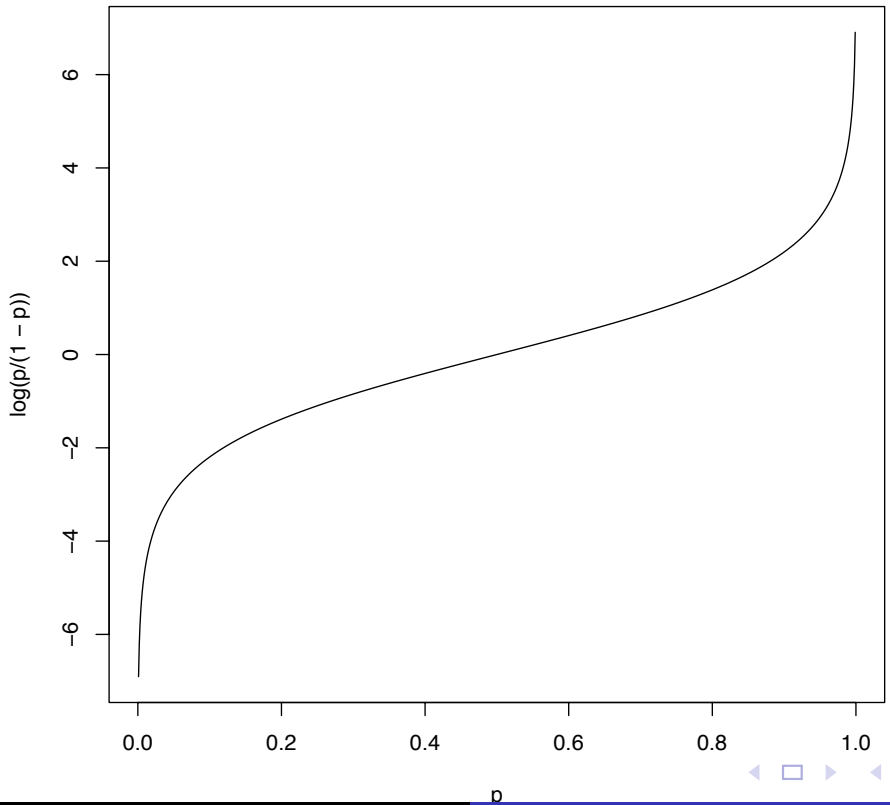
$$P(coronopathie = 1) = a_0 + a_1 \times tabac + a_2 \times cholesterol$$

Mais ce modèle peut conduire à des probabilités prédites négatives ou supérieures à 1.

La fonction logit

On définit la fonction logit sur l'intervalle $[0, 1]$ par la relation

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$



Le modèle

On note

- Y_i la variable qui vaut 1 si le patient i est atteint de la maladie et 0 sinon
- $p_i = P(Y_i = 1)$ la probabilité du patient i d' être atteint d'une maladie
- $X_1, \cdots, X_p$ des facteurs de risque,

Régression logistique

Le modèle logistique ou logit correspond à la relation :

$$\text{logit}(p_i) = a_0 + a_1 X_{1,i} + a_2 X_{2,i} + \cdots + a_p X_{p,i}$$

Exemple

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p_i) = a_0 + a_1 \times tabac + a_2 \times cholesterol$$

où Y_i désigne la présence d'une coronopathie pour le patient i.

Estimation des paramètres

- L'estimation des paramètres du modèle de régression logistique se fait par la méthode du maximum de vraisemblance.
- On maximise la vraisemblance par rapport aux paramètres $a_0, a_1, \cdots, a_p$ au moyen d'un algorithme numérique (par exemple, la méthode du gradient).
- Dans certains cas, l'algorithme ne peut proposer une estimation stable des paramètres.
- En pratique, il faut s'assurer que la procédure numérique a convergé vers des valeurs stables des paramètres.

Tests de significativité des paramètres

Pour tester la significativité d'un coefficient $H_0 : a_p = 0$ on utilise un test de Wald. Sous H_0,

$$\frac{\hat{a}_p}{\sqrt{\hat{V}(\hat{a}_p)}} \sim \chi^2(1)$$

Une absence d'association entre la variable binaire Y et la variable X_p se traduit par l'hypothèse $H_0 : a_p = 0$.

Odds-Ratio

Exemple

- variable Y qui vaut 1 si le patient est malade et 0 s'il est sain
- variable `Sexe` codée par 0 pour les femmes et 1 pour les hommes

Tableau de contingence

`Sexe` / `Maladie`	0	1
1	p_0	p_1
0	$1-p_0$	$1-p_1$

L'Odds-Ratio permet de mesurer l'association entre deux variables binaires.

$$OR = \frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}$$

Odds-Ratio et régression logistique

Le modèle de régression logistique est très utilisé à cause de son lien avec l'Odds-Ratio.

Dans le modèle $\text{logit}\left[P(Y=1|\texttt{Sexe})\right] = a_0 + a_1\texttt{Sexe}$, on a

$$OR = \exp(a_1)$$

Interprétation : Si $OR = 1$, dans notre exemple, la probabilité d'être malade est la même chez les Hommes que chez les Femmes, donc le risque de maladie n'est pas associé à la variable Sexe.

Odds-Ratio Ajusté

- Calculer un Odds-Ratio entre les variables binaires Y et X_1 en tenant compte des variables $(X_j, 2 \leq j \leq p)$
- Estimer un modèle

$$\text{logit}\left[P(Y=1|X_1, X_2, \cdots, X_p)\right] = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$

- Calculer l'Odds-Ratio ajusté :

$$OR_{(X_j, 2 \leq j \leq p)} = \exp(a_1)$$

Avec R

```
data=read.table("usi.txt", header=TRUE)
res=glm(DECEDE~SEXE, data=data, family = "binomial")
summary(res)

resC=glm(DECEDE~AGE+SERV+INF_JO+TAS+SEXE+FC+TYP_AD+PH+BICAR
 +CONSC+GLASGOW, data=data, family = "binomial")
summary(resC)
step(resC)

resF=glm(DECEDE~AGE+TAS+TYP_AD+PH+CONSC, data=data, family
= "binomial")
summary(resF)
boxplot(resF$residuals~data$TYP_AD); abline(h=0)
```