Université de Lille

Régression logistique

M. Fourcot (marie.fourcot@univ-lille.fr)

2021-2022

Notations et objectifs

Notations

- Y : la variable cible qualitative, le plus souvent à deux modalités
- X_j : prédicteurs qualitatifs ou quantitatifs

Objectifs

- mesurer le pouvoir prédictif des X_j par rapport à Y
- construire une règle de décision pour la prédiction de Y à partir des X_j

Comparaison à l'analyse discriminante

Objectif identique entre analyse discriminante et régression logistique mais approches différentes :

- analyse discriminante probabiliste : modélisation de X conditionnellement à la classe.
- régression logistique : modélisation directe de $P(Y = i/X = x)$

Surtout on peut utiliser la régression logistique avec plusieurs variables explicatives.
Et on peut inclure dans le modèle des variables explicatives quantitatives.

Types de régression logistique

3 types de régression logistique selon le type de variable à expliquer (VAE)

- binaire : VAE binaire (ex : vivant / décés)
- ordinale : VAE ordinale (ex : stades de cancer)
- multinomiale : VAE qualitative (ex : types de cancer)

Suite du cours basée sur la régression logistique binaire car :

- *Reg. Ordinale* : hypothèses complémentaires fortes (proportionnalité entre les modalités de Y)
- *Reg. Multinomiale* : peut être vue comme plusieurs régressions logistiques binaires. L'interprétation des coefficients est plus difficile.

Problématique

En régression linéaire multiple, le modèle est linéaire :

$$Y = f(X_1, X_2, \dots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

Question : Qu'en est-il de la régression logistique ? ?

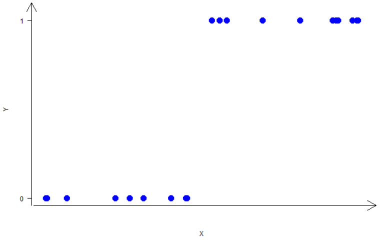
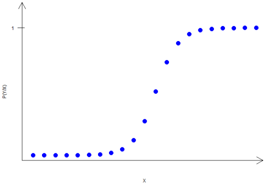


Figure – Variable binaire en fonction d'une variable quantitative

Introduction



Transformation logit :

$$\text{Logit}[\pi(X)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

$$X \to +\infty \text{ alors } \pi(X) \to 1$$
$$X \to -\infty \text{ alors } \pi(X) \to 0$$
$$\pi(X) \in [0, 1]$$

Introduction

Si :

$$P(Y = 1|X) = \pi(X) = \frac{e^{\beta_0+\beta_1 X_1}}{1 + e^{\beta_0+\beta_1 X_1}}$$

Alors :

$$\text{Logit}[\pi(X)] = \beta_0 + \beta_1 X$$

$\Rightarrow$ on revient au modèle linéaire classique !

Introduction

Si :

$$P(Y=1|X)=\pi(X)=\frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$$

Alors :

$$\text{Logit}[\pi(X)]=\beta_0+\beta_1 X$$

$\Rightarrow$ on revient au modèle linéaire classique !

En multivarié :

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

$$\text{Logit}[\pi(\{X_j\})]=\beta_0+\sum_{j=1}^{p}\beta_j X_j$$

Notion d'odds-ratio

Mesure d'association entre exposition et maladie

	M	$\bar{M}$
E^+	a	b
E^-	c	d

$$\left.\begin{array}{l} a = P(M/E^+) \\ b = P(\bar{M}/E^+) \end{array}\right\} \frac{a}{b}$$: cote (odd) d'être malade pour le groupe exposé

$$\left.\begin{array}{l} c = P(M/E^-) \\ d = P(\bar{M}/E^-) \end{array}\right\} \frac{c}{d}$$: cote (odd) d'être malade pour le groupe non exposé

Notion d'odds-ratio

Rapport des odds $\rightarrow$ odds-ratio

$$OR = \frac{\frac{P(M/E^+)}{P(\bar{M}/E^+)}}{\frac{P(M/E^-)}{P(\bar{M}/E^-)}} = \frac{P(M/E^+)}{P(M/E^-)} \times \frac{P(\bar{M}/E^-)}{P(\bar{M}/E^+)} = \frac{ad}{bc}$$

Note : si la prévalence est faible ($P(M) < 10\%$),
alors OR $\approx$ RR ($RR = \frac{P(M/E^+)}{P(M/E^-)}$)

Interprétation de l'odds-ratio :

- $OR = 1$: pas d'association
- $OR > 1$: E^+ est un facteur de risque de M
- $OR < 1$: E^+ est un facteur protecteur de M

Notion d'odds-ratio

De manière générale,

$$\text{odds}(x) = \frac{\pi(x)}{1-\pi(x)}$$

combien de fois on a plus de chance d'avoir $Y = 1$ au lieu d'avoir $Y = 0$ lorsque $X = x$

odds-ratio : facteur par lequel la cote est multipliée quand X change (ou plus souvent, quand une variable change, toutes causes inchangées par ailleurs).

Lien entre OR, modèle Logit et coefficient de la régression

$$\text{logit}\left(\frac{\pi(1)}{\pi(0)}\right) = \log \underbrace{\left[\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}\right]}_{OR} = \beta_1$$

On en déduit que :

$$OR = e^{\beta_1}$$

L'exponentiel du coefficient peut être interprété comme un odds-ratio.

Interprétation des coefficients

Supposons que X soit quantitative :

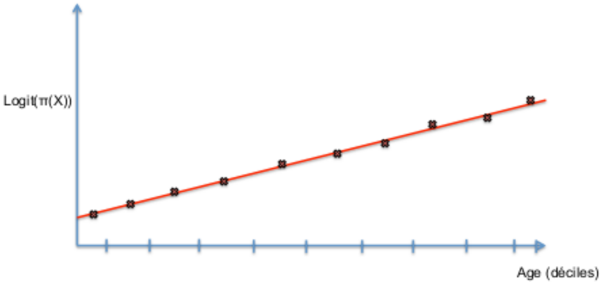
$$OR = e^{\beta_1} = OR^{X=x_0+1/X=x_0} \ \forall x_0$$

Cela sous-tend une hypothèse forte : log-linéarité de X qui est à vérifier.

Principe :

- Découper X en déciles
- Pour chaque intervalle calculer $P(Y = 1/X = c_l)$ (proportion de malades)
- Représenter graphiquement Logit$(\pi(X))$ en fonction des déciles de X

Interprétation des coefficients



Objectif : vérification de la présence d'une relation linéaire entre X et Logit$(\pi(X))$

Sinon :

- Transformations mathématiques ($\log(X)$, $\sqrt{X}$,...)
- Discrétisation de X en classes appropriées

Interprétation des coefficients

Cas des variables nominales :

Exemple : niveau de conscience $\begin{cases} \text{normal} \\ \text{Coma leger} \\ \text{Coma profond} \end{cases}$

1. On choisit une modalité de référence (normal)
2. On construit 2 variables binaires $\begin{cases} \text{Coma leger}(0/1) \\ \text{Coma profond}(0/1) \end{cases}$
3. Introduction dans le modèle
   - Test de la variable dans sa totalité
   - Test des variables binaires une par une (test individuel)

L'OR s'interprète relativement à la modalité de référence, modalité dont l'effet est fixé à 0.

Estimation des coefficients

En régression linéaire multiple $\Rightarrow$ Méthode des moindres carrés (MCO)

$$\min \sum_{i=1}^{n} (e_i)^2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_j))$$

En régression logistique, la MCO ne permet pas d'obtenir une estimation des coefficients

On utilise la **méthode du maximum de vraisemblance**

Estimation des coefficients

Méthode du maximum de vraisemblance

Objectifs : trouver $\hat{\beta}_0$ et $\hat{\beta}_1$ qui maximisent la probabilité d'observer l'échantillon (*i.e.* maximisation de la vraisemblance)

$$\begin{aligned} L(\beta) &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n; \beta) \\ &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | X_i = x_i; \beta) \end{aligned}$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, L(\beta)$$

On passe par le log (plus simple à calculer)

$$\operatorname{argmax} L(\beta) = \operatorname{argmax} \log(L(\beta))$$

Estimation des coefficients

Par exemple, dans le cas univarié, pour trouver $\hat{\beta}_0$ et $\hat{\beta}_1$ qui maximisent L, on a recours aux dérivées partielles :

$$\frac{\partial L}{\partial \beta_0} = 0 \text{ et } \frac{\partial L}{\partial \beta_1} = 0$$

Une fois $\hat{\beta}_0$ et $\hat{\beta}_1$ estimés, on peut calculer pour tout i $\hat{\pi}(x_i)$:

$$\hat{\pi}(x_i) = \mathbb{P}(Y_i = 1/X = x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1}}$$

Ou avec le modèle Logit

$$\text{logit}(\hat{\pi}(x_i)) = \ln\left(\frac{\hat{\pi}(x_i)}{1 - \hat{\pi}(x_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Tests dans le modèle

Tests dans le modèle

Test global de significativité basé sur le Test du Rapport de Vraisemblance (TRV) :

- $\mathcal{H}_0$: Pas de liaison entre Y et les $X_j \Leftrightarrow$ $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
- $\mathcal{H}_1$: Le modèle a du sens $\Leftrightarrow$ Au moins 1 $\beta_j \neq 0$

Considérons le cas avec une seule variable explicative X

Principe du TRV : Comparer la vraisemblance L_X (avec variable explicative) avec la vraisemblance L_0 sans variable explicative

Intuitivement :
si $L_X > L_0$ alors la variable X apporte à l'estimation de $P(Y)$

Test global de significativité

Test avec p variables explicatives X_j

- $\mathcal{H}_0$: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ ($\emptyset$ liaison)
- $\mathcal{H}_1$: $\exists$ au moins un $\beta_j \neq 0$ (liaison)

La statistique de test a pour expression :

$$D = -2 \ln \left(\frac{L_0}{L_X} \right) \sim \chi^2{}_{p\ d.l.l.}$$

Test global / Tests individuels

Dans le cas multiple :

Si on ne rejette pas $\mathcal{H}_0$ alors STOP

Si on rejette $\mathcal{H}_0$ alors test individuel de chaque coefficient :

- $\mathcal{H}_0 : \beta_j = 0$ (la variable n'est pas significative dans le modèle)
- $\mathcal{H}_1 : \beta_j \neq 0$ (la variable est significative dans le modèle)

On peut montrer que si $\mathcal{H}_0$ est vraie alors :

$$\frac{\hat{\beta}_j^2}{s^2_{\hat{\beta}_j}} \sim \chi^2_1 \text{ ddl} \quad \text{(Test de Wald)}$$

Dans les études réelles, beaucoup de variables disponibles, plus ou moins pertinentes, concurrentes...
Trop de variables tue l'interprétation, il y a le danger du sur-apprentissage aussi.

Principe du Rasoir d'Occam : à performances égales, plus un modèle sera simple, plus il sera robuste ; plus aisée sera son interprétation également.

2 approches :

- sélection de variables par optimisation d'un critère
- sélection sur la significativité des variables

Sélection par optimisation

Définitions :

- critère d'Akaike (AIC) : $AIC = -2ln(L) + 2k$
- critère BIC : $BIC = -2ln(L) + ln(n)k$

avec L le maximum de la fonction de vraissemblance du modèle, k le nombre de paramètres à estimer du modèle et n le nombre d'observations.
Dans les deux cas, on veut ce critère petit.
On évalue des successions de modèles :

- en ajoutant les variables au fur et à mesure $\rightarrow$ forward
- en enlevant des variables au fur et à mesure $\rightarrow$ backward

Sélection basée sur des critères statistiques

Utilisation du test de Wald

On fait un test de Wald sur chaque coefficient

- $\mathcal{H}_0 : \beta_j = 0$ (la variable n'est pas significative dans le modèle)
- $\mathcal{H}_1 : \beta_j \neq 0$ (la variable est significative dans le modèle)

On peut montrer que si $\mathcal{H}_0$ est vraie alors :

$$\frac{\hat{\beta}_j^2}{s^2_{\hat{\beta}_j}} \sim \chi^2_1 \text{ ddl}$$

$\Rightarrow$ en backward pur, on aurait que J régressions à effectuer

Sélection basée sur des critères statistiques

Utilisation du test de Wald

On fait un test de Wald sur chaque coefficient

- $\mathcal{H}_0 : \beta_j = 0$ (la variable n'est pas significative dans le modèle)
- $\mathcal{H}_1 : \beta_j \neq 0$ (la variable est significative dans le modèle)

On peut montrer que si $\mathcal{H}_0$ est vraie alors :

$$\frac{\hat{\beta}_j^2}{s^2_{\hat{\beta}_j}} \sim \chi^2_1 \text{ ddl}$$

$\Rightarrow$ en backward pur, on aurait que J régressions à effectuer

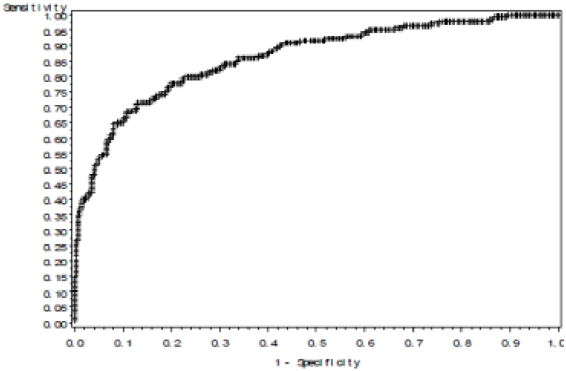
Utilisation du test du score

Utiliser les résultats de la régression à p variables pour calculer les SCORES de chaque (J-p) variable restantes, choisir celle qui a le meilleur score

$\Rightarrow$ en forward pur, on aurait au pire J régressions à effectuer

Pouvoir discriminant du modèle

Rappel : courbe ROC S_e en fonction de $1 - S_p$



Pouvoir discriminant du modèle

Quelques repères pour l'évaluation de l'aire sous la courbe :

AUC	Discrimination
0.5	Nulle
0.7 - 0.8	Acceptable
0.8 - 0.9	Excellente
> 0.9	Exceptionnelle

Remarques :

- Si $AUC = 0.5$ alors le modèle classe de manière complètement aléatoire les observations
- Si $AUC > 0.9$ le modèle est très bon, voire trop bon, il faut évaluer s'il y a sur-ajustement.

Calibration du modèle

Calibration : comparaison des probabilités prédites par le modèle $\hat{\pi}_i(X_j)$ à celles observées dans l'échantillon.

$\Rightarrow$ Mesure d'adéquation

Idée : on cherche à avoir un modèle qui minimise la distance entre les probabilités observées et celles prédites par le modèle

Test de Hosmer - Lemeshow

Principe :

On calcule pour chaque observation la probabilité prédite par le modèle $\hat{\pi}(X)$. On classe les observations par déciles de probabilités prédites.
On compare dans chaque classe les effectifs observés et les effectifs théoriques.

- Si dans chaque classe ces deux effectifs sont proches alors le modèle est calibré
- S'il existe des classes dans lesquelles les effectifs sont trop différents, alors le modèle est mal calibré

Test de Hosmer - Lemeshow

Construction :

1. Calculer les $\hat{\pi}(X)$ prédites par le modèle
2. Classer les données (observations + $\hat{\pi}(X)$) par ordre croissant de $\hat{\pi}(X)$
3. Regrouper les données par déciles de $\hat{\pi}(X)$
4. Construire le tableau suivant

Test de Hosmer - Lemeshow

	Malade (Y=1)		Non-Malade (Y=0)	
	Observés	*Prédits*	*Observés*	*Prédits*
G1 : 0 - 10%	#M	#prédits	#NM	#G1 - #prédits
G2 : 10% - 20%	.	.	.	.
G3 : 20% à 30%	.	.	.	.
G4 : 30 à 40%	.	.	.	.
G5 : 40% à 50%	.	.	.	.
G6 : 50% à 60%	.	.	.	.
G7 : 60% à 70%	.	.	.	.
G8 : 70 à 80%	.	.	.	.
G9 : 80% à 90%	.	.	.	.
G10 : 90 à 100%	.	.	.	.

- #M : le nombre de malades dans la classe (#NM : nb de non-malades)
- $\#\text{prédits} = \sum_{G1} \hat{\pi}(X)$

Test de Hosmer - Lemeshow

Hypothèses du test :

- $\mathcal{H}_0$: les probabilités théoriques sont proches de celles observées (modèle calibré)
- $\mathcal{H}_1$: les probabilités théoriques sont différentes des observées (modèle non calibré)

Sous $\mathcal{H}_0$,

$$\hat{C} = \underbrace{\sum_G \frac{(\#M - \#predits)^2}{\#predits}}_{\text{Malades}} + \underbrace{\sum_G \frac{(\#NM - \#G + \#predits)^2}{\#G - \#predits}}_{\text{Non Malades}}$$

$$\sim \chi^2_{G-2} \text{ ddl}$$

Test de Hosmer - Lemeshow

Le modèle est calibré si **on ne rejette pas** $\mathcal{H}_0$

En pratique, on ne rejette pas $\mathcal{H}_0$ si $p > 0.2$