Classification supervisée
Régression logistique

V. Vandewalle (vincent.vandewalle@univ-lille.fr)
C. Preda (cristian.preda@univ-lille.fr)

Polytech'Lille GIS2A4

Année universitaire 2019-2020

Sources : G. Marot (Univ. Lille), M. Genin (Univ. Lille), J. Jacques (Univ Lyon 2)

Rappel des notations et objectifs

Notations

- Y : la variable cible (**qualitative**) : $Y \in \{1, 2, \ldots, K\}$, $K \geq 2$
- $\{X_1, X_2, \ldots, X_p\}$: p prédicteurs quantitatifs des groupes

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

Introduction

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

- analyse discriminante probabiliste :
  1. Modélisation de X conditionnellement à la classe Y
  2. Estimation des paramètres de la loi de X pour chaque Y
  3. On déduit une estimation de $P(Y = i|X = x)$ via Bayes
- régression logistique :
  1. Modélisation directe de $P(Y = i|X = x)$
  2. Estimation des paramètres par un algorithme itératif
  3. Calcul de $P(Y = i|X = x)$ via la formule directe

Remarque : dans le cas où on étend l'analyse à des variables explicatives qualitatives, on considère les indicatrices correspondantes.
Par souci d'identifiabilité, on ne considère que J-1 indicatrices pour une variable à J modalités (cf. codage binaire en enlevant la colonne correspondant à la première modalité (modalité de référence))

Introduction à la régression logistique

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

- binaire $\Rightarrow$ VAE binaire (ex : vivant / décés)
- ordinale $\Rightarrow$ VAE ordinale (ex : stades de cancer)
- multinomiale $\Rightarrow$ 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.

Introduction à la régression logistique

Rappel :

En régression linéaire multiple, le modèle est linéaire :

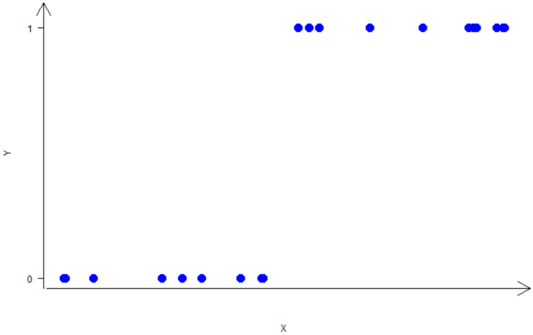
$$Y = f(X_1, X_2, \ldots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

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

$$Y = f(X_1, X_2, \ldots, X_p) = \ ?$$

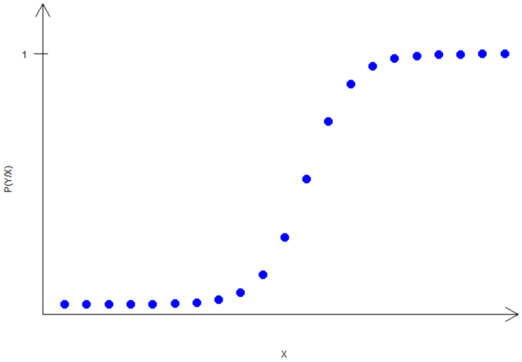
Régression logistique

Variable binaire en fonction d'une variable quantitative



Le modèle de régression linéaire ne s'applique pas.

Régression logistique



$$\pi(x) = P(Y = 1|X = x) = F(\tilde{x}^T\beta) \Leftrightarrow F^{-1}(\pi(x)) = \tilde{x}^T\beta$$

avec $x = (x_1, \ldots, x_p)$, $\tilde{x} = (1, x^T)^T$ et $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$
et F^{-1} **la fonction de lien** dans le modèle linéaire généralisé

Introduction

Pour F, n'importe quelle fonction de répartition peut convenir. Si F est la fonction de répartition de la loi normale centrée réduite, on parle de régression probit.

En régression logistique, transformation logit :

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

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

Régression logistique

Remarque :

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

Si $\beta > 0$:

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

En multivarié $x = (x_1, \ldots, x_p)$:

$$\mathbb{P}(Y = 1/X = x) = \pi(x) = \frac{e^{\beta_0+\sum_{j=1}^{p} \beta_j x_j}}{1 + e^{\beta_0+\sum_{j=1}^{p} \beta_j x_j}}$$

Modèle logistique

Exercice :

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

Réponse :

$$\text{logit}[\pi(x)] = \beta_0 + \beta_1 x$$

Intérêts du modèle logistique :

- Permet de revenir au modèle linéaire classique
- Interprétation des coefficients du modèle comme une mesure d'association de X par rapport à Y (Notion d'odds-ratio)

Régression logistique

Notion d'odds-ratio

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

Odds-ratio : mesure d'association entre exposition et maladie

$$\text{OR} = \frac{ad}{bc} = \frac{\frac{a}{b}}{\frac{c}{d}}$$

$$\left.\begin{array}{l} a = P(M/E^+) \\ b = P(\bar{M}/E^+) \end{array}\right\} \frac{a}{b}$$: 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

Régression logistique

Mesure de l'association entre maladie et exposition ? ?

Rapport des odds $\rightarrow$ odds-ratio

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

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

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

Régression logistique

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

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).

Régression logistique

Lien entre OR, modèle Logit et coefficient de la régression :

Exercice :

Considérons une seule variable explicative X binaire $\begin{cases} 1 : E^+ \\ 0 : E^- \end{cases}$

Calculer logit $\left(\dfrac{\pi(1)}{\pi(0)}\right)$

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x$$
$$\text{logit}(\pi(1)) = \beta_0 + \beta_1$$
$$\text{logit}(\pi(0)) = \beta_0$$
$$\text{logit}(\pi(1)) - \text{logit}(\pi(0)) = \beta_1$$

Régression logistique

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 :

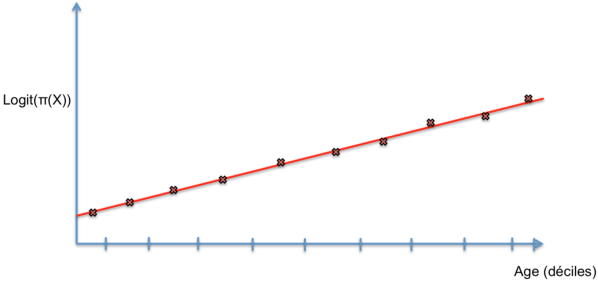
$$OR = e^{\beta_1} = OR^{X=x_0+1/X=x_0} \; \forall x_0$$

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 :

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

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)

LOR 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)

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

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

$$
\begin{aligned}
L(\beta) &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \beta) \\
&= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | X_i = x_i; \beta).
\end{aligned}
$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}}\, L(\beta)$$

On passe par le log (plus simple à calculer)

$$\underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}}\, L(\beta) = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}}\, \underbrace{\log(L(\beta))}_{\ell(\beta)}$$

Estimation des coefficients

Par exemple dans le cas univarié, pour trouver $\hat{\beta}_0$ et $\hat{\beta}_1$ qui maximisent $\ell(\beta_0, \beta_1)$, on a recours aux dérivées partielles :

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

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

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

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 (Likelihood Ration Test : LRT) :

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

Considérons le cas avec une seule variable explicative X

Principe : 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

- $\mathcal{H}_\mathbf{0}$: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ ($\emptyset$ liaison)
- $\mathcal{H}_\mathbf{1}$: $\exists$ au moins un $\beta_j \neq 0$ (liaison)

La statistique de test a pour expression :

$$D = -2\ln\left(\frac{L_0}{L_X}\right) \sim {\chi^2}_{p\ d.l.l.} = -2\ln L_0 - (-2\ln L_X) = D_0 - D_X$$

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 :

- $\mathcal{H}_\mathbf{0}$: $\beta_j = 0$ (la variable n'est pas significative dans le modèle)
- $\mathcal{H}_\mathbf{1}$: $\beta_j \neq 0$ (la variable est significative dans le modèle)

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

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

Remarque : d'autres tests sont envisageable tels que le test du score ou le test du rapport de vraisemblance.

Test de modèles emboîtés

Considérons deux modèles sur logit$[\pi(x)]$:

- M_1 : $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- M_2 : $\beta_0 + \beta_1 x_1$

- $M_2 \subset M_1$: si $\beta_2 = \beta_3 = 0$ on se ramène à M_1.
- Test de modèles emboîtés : test de l'apport des variables X_2 et X_3 par rapport à la variable X_1 déjà présente dans le modèle.

Ainsi on peut tester :
$H_0 : \{\beta_2 = \beta_3 = 0\}$ (M_2 vrai), contre $H_1 : \{\beta_2 \neq 0 \text{ ou } \beta_3 \neq 0\}$ (M_1 vrai)
A l'aide du tet du rapport de vraisemblance :

$$D = -2 \ln \frac{L_{M_2}}{L_{M_1}} = D_{M_2} - D_{M_1}$$

avec L_{M_1} et L_{M_2} les log-vraisemblances de modèle des modèle M_1 et M_2 et D_{M_1} et D_{M_2} les déviances associées.
Sous H_0 : $D \sim \chi^2_{\nu_{M_1} - \nu_{M_2}}$.
avec ν_{M_1} et ν_{M_2} les nombres de paramètres des modèle M_1 et M_2.

Test de l'apport d'une variable qualitative

- Pour une variable explicative X_j qualititave à j modalités on doit estimer $m_j - 1$ coefficients.
- Donc tester l'effet de X_j revient donc à tester la nullité de ces $m_j - 1$ coefficients.
- On se ramène à la question du test de modèle emboîté.
- Cela se rapproche de la question de l'ANOVA en régression linéaire

Choix de modèle par critère AIC ou BIC

Outre les tests d'hypothèses on peut utiliser des critères de **choix de modèle** pour sélectionner le meilleur modèle :

- $AIC(M) = -2 \ln L_M + 2\nu_M$
- $BIC(M) = -2 \ln L_M + \nu_M \ln n$

Dans ces deux cas on cherche le modèle parmis une collection de modèle qui **minimise** ce critère.
Ces critères peuvent entre autre être utilisés pour faire de la **sélection de variables** à l'aide d'un procédure **pas à pas** :

- Ascendante (forward) : ajout d'une variable à chaque itération (si diminution du critère)
- Descendante (backward) : suppression d'une variable à chaque itération (si diminution du critère)
- Dans le deux sens (forward-backward : both).

Pouvoir discriminant du modèle

Pouvoir discriminant via la courbe ROC
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 *overfitting*

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 $\hat{\pi}_i(X_j)$. 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 $\hat{\pi}_i(X_j)$ prédites par le modèle
2. Classer les données (observations + $\hat{\pi}_i(X_j)$) par ordre croissant de $\hat{\pi}_i(X_j)$
3. Regrouper les données par déciles de $\hat{\pi}_i(X_j)$
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)
- #prédits $= \sum_{G1} \hat{\pi}_i(X_j)$

Test de Hosmer - Lemeshow

Hypothèses du test :

- $\mathcal{H}_{\mathbf{0}}$: les probabilités théoriques sont proches de celles observées (modèle calibré)
- $\mathcal{H}_{\mathbf{1}}$: 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 $p > 0.2$