MA431 : Mathématiques appliquées à la sécurité

Regression Logistique

D. Barcelo

Grenoble INP ESISAR

2022/2023

Objectifs

- Régression logistique ou régression binomiale
- Objectif : prévoir une variable binaire Y
- A l'aide de p **variables explicatives continues** (notamment) : $X_1, \ldots, X_p$.
- On observe $x = (x_1, \ldots, x_p)$. On veut donc pouvoir déterminer $\mathbb{P}_{X=x}(Y = 1)$.
- Affectation de la valeur 1 à Y :
- Chercher une régression, c'est donc obtenir une relation de la forme :

$$\mathbb{P}_{X=x}(Y = 1) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Mais cela pose des problèmes ...

Objectifs

- Régression logistique ou régression binomiale
- Objectif : prévoir une variable binaire Y
- A l'aide de p **variables explicatives continues** (notamment) : $X_1, \ldots, X_p$.
- On observe $x = (x_1, \ldots, x_p)$. On veut donc pouvoir déterminer $\mathbb{P}_{X=x}(Y = 1)$.
- Affectation de la valeur 1 à Y : si la prévision de $\mathbb{P}_{X=x}(Y = 1)$ est supérieure à 0,5.
- Chercher une régression, c'est donc obtenir une relation de la forme :

$$\mathbb{P}_{X=x}(Y = 1) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Mais cela pose des problèmes ...

Résoudre les problèmes : odds

- Pour prendre une décision, on peut aussi prévoir un "odds" (côte ou rapport de chances).
- Affectation de la valeur 1 à Y :
- On peut donc chercher une relation de la forme :

$$\frac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=O)} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Mais cela pose des problèmes ...

Résoudre les problèmes : odds

- Pour prendre une décision, on peut aussi prévoir un "odds" (côte ou rapport de chances). Par exemple calculer $\frac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=O)}$.
- Affectation de la valeur 1 à Y :
- On peut donc chercher une relation de la forme :

$$\frac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=O)} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Mais cela pose des problèmes ...

Résoudre les problèmes : odds

- Pour prendre une décision, on peut aussi prévoir un "odds" (côte ou rapport de chances). Par exemple calculer $\dfrac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=O)}$.
- Affectation de la valeur 1 à Y : si la côte est supérieure à 1.
- On peut donc chercher une relation de la forme :

$$\frac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=O)} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Mais cela pose des problèmes ...

Résoudre les problèmes : fonction logit

- Une solution est de chercher une relation linéaire avec le logarithme de l'odds :

- Affectation de la valeur 1 à Y :

- La relation linéaire avec le logarithme de l'odds peut s'écrire :

$$\ln\left(\frac{\mathbb{P}_{X=x}(Y=1)}{1-\mathbb{P}_{X=x}(Y=1)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Mais il y a d'autres modélisations ...

Résoudre les problèmes : fonction logit

- Une solution est de chercher une relation linéaire avec le logarithme de l'odds :

$$\ln\left(\frac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=O)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Affectation de la valeur 1 à Y :

- La relation linéaire avec le logarithme de l'odds peut s'écrire :

$$\ln\left(\frac{\mathbb{P}_{X=x}(Y=1)}{1-\mathbb{P}_{X=x}(Y=1)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- Mais il y a d'autres modélisations ...

Résoudre les problèmes : fonction logit

- Une solution est de chercher une relation linéaire avec le logarithme de l'odds :

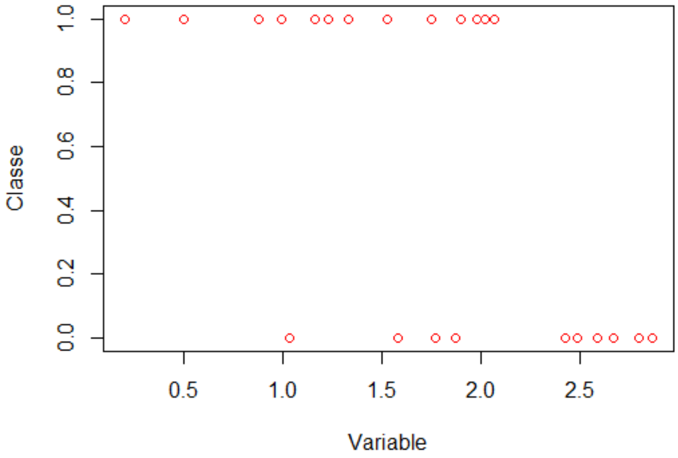
$$\ln\left(\frac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=O)}\right)=\beta_0+\sum_{i=1}^{p}\beta_i x_i$$

- Affectation de la valeur 1 à Y : si la côte prévision du logarithme de l'odds est supérieure à 0.
- La relation linéaire avec le logarithme de l'odds peut s'écrire :

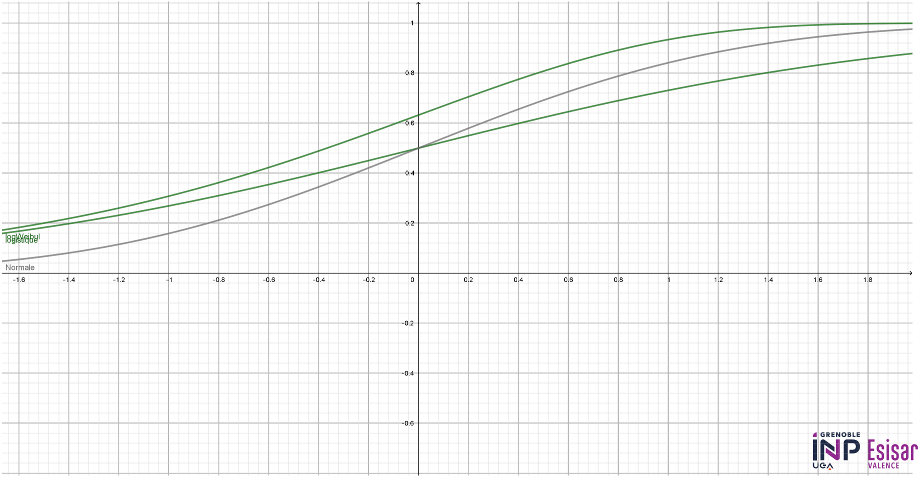
$$\ln\left(\frac{\mathbb{P}_{X=x}(Y=1)}{1-\mathbb{P}_{X=x}(Y=1)}\right)=\beta_0+\sum_{i=1}^{p}\beta_i x_i$$

- Mais il y a d'autres modélisations ...

Modélisation logistique



Modélisation logistique



Modélisation logistique

On note $\pi(x) = \mathbb{P}_{X=x}(Y = 1)$.
On cherche à déterminer une relation de la forme :
$\pi(x) = \dfrac{f\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}{1 + f\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}$.

Modèle Logit :

$$\pi(x) = \frac{e^{\left(\beta_0+\sum_{i=1}^{p} \beta_i x_i\right)}}{1 + e^{\left(\beta_0+\sum_{i=1}^{p} \beta_i x_i\right)}}$$

et

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

On note aussi $Logit\left(\pi(x)\right) = \dfrac{\pi(x)}{1-\pi(x)}$.

Il existe d'autres modèles possibles (Probit et log-log notamment).

Autres modèles

On note $\pi(x) = \mathbb{P}_{X=x}(Y = 1)$.
On cherche à déterminer une relation de la forme :
$\pi(x) = \dfrac{f\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}{1 + f\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}$.

Modèle Log-Log :

$$\pi(x) = 1 - e^{\left(-e^{\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}\right)}$$

et

$$\ln\left(-\ln\left(1 - \pi(x)\right)\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

Courbe dissymétrique mais proche du logit.

Autres modèles

On note $\pi(x) = \mathbb{P}_{X=x}(Y = 1)$. On cherche à déterminer une relation de la forme : $\pi(x) = \dfrac{f\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}{1 + f\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}$.

Modèle Probit :

$$\pi(x) = \Phi\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)$$

et

$$\Phi^{-1}(\pi(x)) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

Avec Φ fonction de répartition de la loi normale centrée réduite.

Moins pratique et moins utilisé que le logit, minimise les cas extrèmes.

Odds

La modélisation $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$ permet de calculer :

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\mathbb{P}_{X=x}(Y=1)}{1-\mathbb{P}_{X=x}(Y=1)} = \frac{\mathbb{P}_{X=x}(Y=1)}{\mathbb{P}_{X=x}(Y=0)}$$

Il s'agit d'un odds (côte) : rapport mesurant la côte d'une modalité par rapport à une autre.

Règle d'affectation

- Le modèle logistique permet d'obtenir une estimation de $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$.
- Il faut pouvoir, pour un x donné, proposer une valeur à Y (0 ou 1).

Règle d'affectation

- Si $\pi(x) > 0,5$, on affecte la valeur 1 à Y.
- Si $\dfrac{\pi(x)}{1-\pi(x)} > 1$, on affecte la valeur 1 à Y.
- Si $\beta x = \beta_0 + \displaystyle\sum_{i=1}^{p} \beta_i x_i > 0$, on affecte la valeur 1 à Y.

Estimation des coefficients

Pour estimer les coefficients β du modèle, on va utiliser la méthode du maximum de vraisemblance.
Le modèle est binomial, la vraisemblance va donc s'exprimer :

$$L(x_1, ..., x_n, \beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

Rappels

Estimation des coefficients

Et la log-vraisemblance :

$$\ln L(x_1, \ldots, x_p, \beta) = \sum_{i=1}^{p} y_i \ln\left(\pi(x_i)\right) + (1 - y_i) \ln(1 - \pi(x_i))$$

$$\ln L(x_1, \ldots, x_p, \beta) = \sum_{i=1}^{p} y_i \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i}\right) + \ln(1 - \pi(x_i))$$

$$\ln L(x_1, \ldots, x_p, \beta) = \sum_{i=1}^{p} y_i \beta x + \sum_{i=1}^{p} \ln(1 - \pi(x_i))$$

On peut donc, en maximisant la log-vraisemblance, obtenir une estimation $\hat{\beta}$ de β.

L'optimisation de la log-vraisemblance se fait à l'aide d'approximation numérique (algorithme de Newton-Raphson ou Fisher)

Qualité d'un modèle

Déviance

- Si on note $L(\beta_i)$ la vraisemblance du modèle avec i variables alors on définit la déviance : $D(\beta_i) = -2 \ln L(\beta_i)$.
- But de la régression logistique : maximiser $L(\beta_i)$ donc minimiser $D(\beta_i)$.
- La déviance permet de comparer des modèles emboités en ajoutant des variables explicatives supplémentaires à un modèle existant.
- La déviance est l'analogue de la somme des carrés résiduels en regression linéaire.
- La déviance suit asymptotiquement une loi du χ^2.

Qualité d'un modèle

AIC

Critère d'Akaïké : $AIC = -2 \ln L(\beta_i) + 2(i+1)$.

- permet de comparer des modèles, emboités ou non.
- L'AIC doit être le plus petit possible.
- peut aussi aider à déterminer le nombre de variables à sélectionner.

Exemple

Exemple de régression logistique. ◂ Exemple

Courbe ROC

Principe de la courbe ROC

Le modèle donne un résultat numérique avec un seuil s tel que la prédiction est positive si $x > s$ et négative si $x < s$.
Au fur et à mesure que s augmente alors :

- le taux de vrais négatifs augmente (spécifité).
- le taux de vrais positifs diminue (sensibilité).

La courbe ROC représente l'évolution de la spécificité (taux de vrais positifs) en fonction de 1-sensibilité (taux de faux négatifs).

Courbe ROC

Courbe ROC

- Visualiser le pouvoir discriminant d'un modèle.
- Courbe Receiver Operating Characteristic
- Représente $\alpha(s)$ en fonction de $1 - \beta(s)$

$\alpha(s) \approx$ Proportion de vrais positifs au score supérieur à s

$\beta(s) \approx$ Proportion de vrais négatifs au score supérieur à s

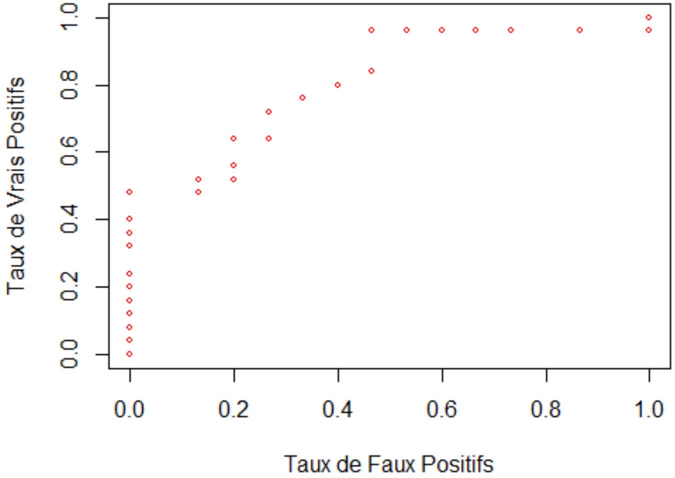
$1 - \beta(s) \approx$ Proportion de faux positifs au score supérieur à s

Courbe ROC

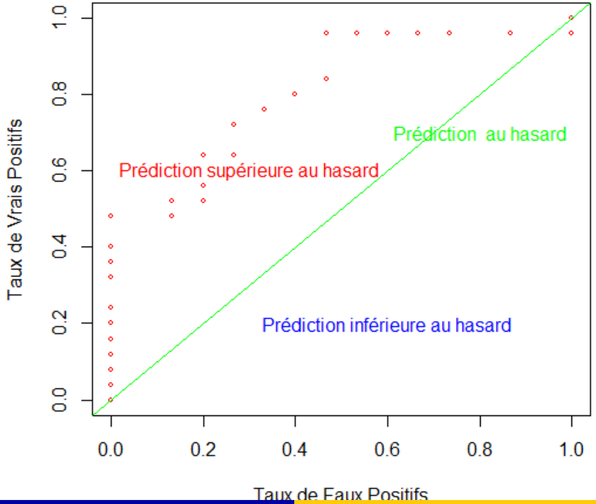
Courbe ROC

- Pour un seuil $s = 1$: Ni vrais positifs ni faux positifs donc point (0 ;0)
- Pour un seuil $s = 0$: Tous les vrais positifs et tous les faux positifs donc point (1 ;1)
- Modèle parfait : Tous les vrais positifs et aucun faux positifs
- Modèle aléatoire : Autant de vrais positifs que de faux positifs

Courbe ROC



Courbe ROC


Taux de Vrais Positifs
0.0
0.2
0.4
0.6
0.8
1.0
Prédiction au hasard
Prédiction supérieure au hasard
Prédiction inférieure au hasard
0.0
0.2
0.4
0.6
0.8
1.0
Taux de Faux Positifs

Courbe ROC

Critère AUC

Pour comparer différentes modélisations logistiques : critère AUC

- Area Under the Curve
- Modèle performant qui sépare les vrais positifs des faux positifs : AUC proche de 1
- AUC=0,5 : autant tirer à pile ou face les affectations
- AUC<0,5 : autant tirer à pile ou face les affectations
- Comparaison de deux modèles : comparaison des AUC

La régression logistique :

- permet de traiter des variables explicatives de tout type,
- permet de traiter des variables à expliquer qualitatives,
- modèle précis et facilement programmables,
- peut s'appliquer aux petits échantillons,
- modélise directement une probabilité,
- possibilité de sélectionner pas à pas les variables explicatives
- mais c'est une approximation numérique (temps et précision),
- mais traite mal les valeurs manquantes et sensibilité aux valeurs hors normes de variables continues.

Rappels

Vraisemblance

Définition :

Soit $X_1, X_2, \ldots, X_n$ un échantillon, on appelle vraisemblance de l'échantillon la fonction :

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i, \theta)$$

si les X_i sont des v.a. discrètes et

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

si les X_i sont continues.

La vraisemblance est interprétée comme l'information apportée par l'échantillon sur la population.

Maximum de vraisemblance

Définition :

On appelle estimateur du maximum de vraisemblance la valeur de $\hat{\theta}_n$ qui rend maximale la vraisemblance de l'échantillon.

$$L(x_1, ..., x_n, \hat{\theta}_n) = \max_{\theta} L(x_1, ..., x_n, \theta)$$

La vraisemblance $L(x_1, ..., x_n, \theta)$ représente la probabilité d'observer $(x_1, \ldots, x_n)$ pour une valeur fixée de θ.
Si on a observé $(x_1, \ldots, x_n)$ sur un échantillon, on peut donc penser que θ prend la valeur la plus vraisemblable, i.e. celle qui fait que la probabilité d'observer $(x_1, \ldots, x_n)$ est la plus grande possible.

Maximum de vraisemblance

Généralement, la vraisemblance s'exprime sous la forme d'un produit, on va donc chercher à maximiser $\ln L$.
Si L est deux fois dérivable, on va alors résoudre :

$$\begin{cases} \dfrac{\partial \ln L}{\partial \theta}(x_1, \dots, x_n, \theta) = 0 \\ \dfrac{\partial^2 \ln L}{\partial \theta^2}(x_1, \dots, x_n, \theta) < 0 \end{cases}$$

Cours

Exemple

On dispose d'un échantillon de 40 individus sur lesquels on a relevé une variable binaire en observant une unique variable quantitative.
On effectue une regression logistique, voici les résultats :

Deviance Residuals :

Min	1Q	Median	3Q	Max
-1.8686	-0.7764	0.3801	0.8814	2.0253

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.9462	1.9599	3.034	0.00241
x	-0.1156	0.0397	-2.912	0.00360

Null deviance : 52.925 on 39 degrees of freedom
Residual deviance : 39.617 on 38 degrees of freedom
AIC : 43.617

i) Interpréter ces résultats.
ii) Déterminer le modèle.
iii) On observe $x = 47$. Quelle classe doit-on attribuer ?

Exemple

Deviance Residuals :

Min	1Q	Median	3Q	Max
-1.8686	-0.7764	0.3801	0.8814	2.0253

Distribution de la Déviance

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.9462	1.9599	3.034	0.00241
x	-0.1156	0.0397	-2.912	0.00360

Estimation des coefficients

On a donc comme modèle : $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = 5.9462 - 0.1156x.$

Les p_valeurs permettent de tester la nullité des coefficients.

Les Std. Error sont des écarts types qui permettent de réaliser des intervalles de confiance sur les coefficients.

Exemple

Null deviance : 52.925 on 39 degrees of freedom
Deviance nulle : déviance avec seulement la constante
Residual deviance : 39.617 on 38 degrees of freedom
Deviance : permet d'effectuer un test du χ^2 pour valider le modèle
Les deux déviances suivent une loi du χ^2, leur différence également de degré de liberté la différence des degrés de libertés.
On a Deviance nulle - Déviance résiduelle $\sim \chi^2_1$.
On peut donc effectuer un test pour valider le modèle, l'hypothèse nulle est que les coefficients sont tous nuls contre une hypothèse alternative de non nullité d'au moins un des coefficients.
Un rejet de l'hypothèse nulle valide donc le modèle.
Deviance : permet de comparer des modèles emboités
AIC : 43.617 **AIC : permet de comparer des modèles différents, pas forcément emboités.** Cours