Régression logistique

Pr. Mokhtar KOUKI

Université de Carthage (ESSAI)
mokhtar.kouki@essai.ucar.tn

Octobre 2023

Contenu

1. Modèle : Définition
2. Cote et rapport des cotes : Odds and Odds Ratio
3. Effets marginaux
4. Prédiction et indicateurs de qualité de l'ajustement
5. Indicateurs de performance
6. Courbe ROC et l'Aire sous la courbe (Area Uner Curve)

On considère un échantillon d'individus pour lesquels on observe les caractéristiue X et le statut de solvabilité Y (1 si l'individu est solvable et 0 sinon). La modélisation de la variable Y en fonction de la variable X peut être formalisée comme suit :

$$\begin{aligned} Y_i^* &= \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \; i = 1, 2, \cdots, n \\ Y_i &= \begin{cases} 1 & \text{si } Y_i^* > 0 \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

avec

- Y^* : Une variable latente (inobservable) (i.e. score)
- Y : une variable binaire qui vaut 1 si l'individu est solvable et 0 sinon
- $X_1, X_2, \cdots, X_k$: k variables explicatives (exogènes)
- $\alpha, \beta_1, \cdots \beta_k$: des paramètres à estimer

Faisons remarquer que la variable Y_i suit une loi de Bernoulli ; avec

$$\begin{aligned} P(Y_i = 1) &= P(Y = 1) = P(Y_i^* > 0) \\ &= P(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i > 0) \\ &= P\left(\varepsilon_i > -\left(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}\right)\right) \\ &= 1 - F_\varepsilon(-\left(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}\right)) \end{aligned}$$

Remarque : 2 cas de figure

- si ε suit une loi normale $N(0,1)$, il s'agit d'un modèle **Probit**
- si ε suit une loi logistique, il s'agit d'un modèle **Logit**
- pour une loi normale centrée réduite

$$F_\varepsilon(-x) = 1 - F_\varepsilon(x)$$

Définition (Loi logistique)

Une variable aléatoire X suit une loi logistique si et seulement si la fonction de répartition est définie comme suit :

$$F(x) = P(X < x) = \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{1 + \exp(x)}$$

Définition (Loi logistique)

Une variable aléatoire X suit une loi logistique si et seulement si la fonction de répartition est définie comme suit :

$$F(x) = P(X < x) = \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{1 + \exp(x)}$$

Ainsi :

$$F(-x) = \frac{\exp(-x)}{1 + \exp(-x)} = \frac{1}{1 + \exp(x)} = 1 - \frac{\exp(x)}{1 + \exp(x)} = 1 - F(x)$$

Définition (Loi logistique)

Une variable aléatoire X suit une loi logistique si et seulement si la fonction de répartition est définie comme suit :

$$F(x) = P(X < x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)}$$

Ainsi :

$$F(-x) = \frac{\exp(-x)}{1+\exp(-x)} = \frac{1}{1+\exp(x)} = 1 - \frac{\exp(x)}{1+\exp(x)} = 1 - F(x)$$

Conclusion : Pour les deux modèles on a :

$$P(Y_i = 1) = F_\varepsilon \left(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}\right) = F_i$$

Et la vraisemblance d'un échantillon $(Y_i, X_i), i = 1, 2, \cdots, n$ est définie par :

Définition (cote)

On défini la cote, ou odds en anglais, par le rapport :

$$odd = \frac{P(Y_i = 1}{P(Y_i = 0} = \frac{P(Y_i = 1}{1 - P(Y_i = 1}$$

Pour un modèle logit, la cote est donnée par la relation suivante :

$$odd = \exp\left(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}\right)$$

et

$$\ln(odd) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Observation : On parle souvent de régression logistique entre autre à cause de cette relation linéaire

Exemple : odds=4, veut que l'individu a quatre chances d'être solvable contre 1 chance d'être non solvable. Et la probabilité de solvabilité est égale à 0.8 :

$$P(Y=1)=\frac{4}{4+1}=0.8$$

Considérons une modèle logistique avec une variable explicative qualitative (binaire), telle que le genre.

$$Y_i* = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \theta G_i$$

avec

$$G_i = \begin{cases} 1 & \text{si l'individu est une femme} \\ 0 & \text{sinon} \end{cases}$$

Définition (rapport des cotes)

Le rapport des cotes (ou odds ratio en anglais noté OR) entre les femmes et les hommes est défini par :

$$\begin{aligned} OR = \frac{odds/G_i = 1}{odds/D_i = 0} &= \frac{\exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \theta)}{\exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})} \\ &= \exp(\theta) \end{aligned}$$

Définition (Effet marginaux)

On définit l'effet marginal d'une variable X_l sur la probabilité que Y soit égale à 1 (individu solvable), la dérivée partielle suivante :

$$\begin{aligned}\frac{\partial P(Y_i = 1)}{\partial X_{li}} &= \frac{\partial F_\varepsilon(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}{\partial X_{li}} \\ &= \beta_l f_\varepsilon(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})\end{aligned}$$

où $f_\varepsilon()$ est la densité de probabilité de ε Et

- si $\beta_l > 0$, X_l a une inlfuence positive sur les chances que l'individu soit solvable
- si $\beta_l < 0$, X_l a une inlfuence négative sur les chances que l'individu soit solvable

Pour un modèle estimé, l'affectation de l'individu entre solvable et non solvable est définie par rapport à une seuil de probabilité (0.5 par défaut) :

$$\hat{Y}_i = \begin{cases} 1 & \text{si } F_\varepsilon(\hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}) > 0.5 \\ 0 & \text{sinon} \end{cases}$$

Pour un modèle estimé, l'affectation de l'individu entre solvable et non solvable est définie par rapport à une seuil de probabilité (0.5 par défaut) :

$$\hat{Y}_i = \begin{cases} 1 & \text{si } F_\varepsilon(\hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}) > 0.5 \\ 0 & \text{sinon} \end{cases}$$

Et on définit le tableau de contingence ou de confusion suivant :

		Estimation		
		$Y_i = 0$	$Y_i = 1$	Total
Observation	$Y_i = 0$	$VN = N_{00}$	$FP = N_{01}$	$N_{0.}$
	$Y_i = 1$	$FN = N_{10}$	$VP = N_{11}$	$N_{1.}$
	Total	$N_{.0}$	$N_{.1}$	N

Pour un modèle estimé, l'affectation de l'individu entre solvable et non solvable est définie par rapport à une seuil de probabilité (0.5 par défaut) :

$$\hat{Y}_i = \begin{cases} 1 & \text{si } F_\varepsilon(\hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}) > 0.5 \\ 0 & \text{sinon} \end{cases}$$

Et on définit le tableau de contingence ou de confusion suivant :

		Estimation		
		$Y_i = 0$	$Y_i = 1$	Total
Observation	$Y_i = 0$	$VN = N_{00}$	$FP = N_{01}$	$N_{0.}$
	$Y_i = 1$	$FN = N_{10}$	$VP = N_{11}$	$N_{1.}$
	Total	$N_{.0}$	$N_{.1}$	N

Les cellules en surbrillances jaune correspondent aux erreurs de prédiction.

Les valeurs à l'intérieur du tableau dépendent du seuil d'affectation adopté.

Avec :

- $VN = N_{00}$ le nombre de vrais négatifs ; les individus non solvables que le modèle prédit non solvables

Avec :

- $VN = N_{00}$ le nombre de vrais négatifs ; les individus non solvables que le modèle prédit non solvables
- $FP = N_{01}$ le nombre de faux positifs ; les individus non solvables que le modèle prédit solvables

Avec :

- $VN = N_{00}$ le nombre de vrais négatifs ; les individus non solvables que le modèle prédit non solvables
- $FP = N_{01}$ le nombre de faux positifs ; les individus non solvables que le modèle prédit solvables
- $FN = N_{00}$ le nombre de faux négatifs ; les individus solvables que le modèle prédit non solvable

Avec :

- $VN = N_{00}$ le nombre de vrais négatifs ; les individus non solvables que le modèle prédit non solvables
- $FP = N_{01}$ le nombre de faux positifs ; les individus non solvables que le modèle prédit solvables
- $FN = N_{00}$ le nombre de faux négatifs ; les individus solvables que le modèle prédit non solvable
- $VP = N_{11}$ le nombre de vrais positifs ; les individus solvables que le modèle prédit solvables

Avec :

- $VN = N_{00}$ le nombre de vrais négatifs ; les individus non solvables que le modèle prédit non solvables
- $FP = N_{01}$ le nombre de faux positifs ; les individus non solvables que le modèle prédit solvables
- $FN = N_{00}$ le nombre de faux négatifs ; les individus solvables que le modèle prédit non solvable
- $VP = N_{11}$ le nombre de vrais positifs ; les individus solvables que le modèle prédit solvables
- N nombre total d'individus

A partir du tableau de contingence précédent, on définit les indicateurs suivants :

$$\text{Sensibilité ou Rappel} = \frac{VP}{FN + VP}$$

$$\text{Spécificité} = \frac{VN}{VN + FP}$$

$$\text{Précision} = \frac{VP}{VP + FP}$$

$$\text{Accuracy} = \frac{VP + VN}{N}$$

$$MCC = \frac{VN * VP - FN * FP}{\sqrt{(VN + FP)(FN + VP)(VN + FP)(FP + VP)}}$$

$$F1 - Score = \frac{2 * \text{précision*Sensibilité}}{\text{précision+Sensibilité}}$$

Observation : MCC le coefficient de corrélation Mathews et $-1 < MCC < 1$. Une valeur proche de 1 constitue une classification parfaite et une valeur négative correspond à une mauvaise classification.

Observation : MCC le coefficient de corrélation Mathews et $-1 < MCC < 1$. Une valeur proche de 1 constitue une classification parfaite et une valeur négative correspond à une mauvaise classification.

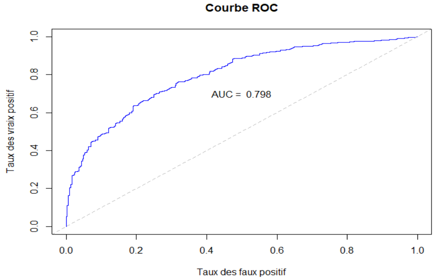
Calculer les indicateurs de performance pour le tableau de contingence / confusion suivant :

		Estimation		
		$Y_i = 0$	$Y_i = 1$	Total
Observation	$Y_i = 0$	136	14	150
	$Y_i = 1$	18	120	138
	Total	154	134	288

La courbe ROC (Receiving Operating Curve) permet de donner une idée sur la qualité de classification du modèle en représentatnt la sensibilité (taux des vais positif) en fonction de 1-spécificité (1-taux des vrais négatifs=taux des faux positifs).

La courbe ROC (Receiving Operating Curve) permet de donner une idée sur la qualité de classification du modèle en représentatnt la sensibilité (taux des vais positif) en fonction de 1-spécificité (1-taux des vrais négatifs=taux des faux positifs).
L'aire sous la courbe roc (AUC) est une mesure de la qualité du modèle.

La courbe ROC (Receiving Operating Curve) permet de donner une idée sur la qualité de classification du modèle en représentatnt la sensibilité (taux des vais positif) en fonction de 1-spécificité (1-taux des vrais négatifs=taux des faux positifs).
L'aire sous la courbe roc (AUC) est une mesure de la qualité du modèle.



MERCI POUR VOTRE ATTENTION

Pr. Mokhtar KOUKI

mokhtar.kouki@essai.ucar.tn