Régression logistique

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Professional Rodeo Cowboys Association

L'exemple suivant est inspiré de l'article

> *Daneshvary, R. et Schwer, R. K. (2000) The Association Endorsement and Consumers' Intention to Purchase. Journal of Consumer Marketing* 17, *203-213.*

Objectif: Les auteurs cherchent à voir si le fait qu'un produit soit recommandé par le *Professional Rodeo Cowboys Association* (PRCA) a un effet sur les intentions d'achats.

Données du PRCA

On dispose de 500 observations sur les variables suivantes:

- Y: seriez-vous intéressé à acheter un produit recommandé par le PRCA
  - 0: non
  - 1: oui
- X_1: quel genre d'emploi occupez-vous?
  - 1: à la maison
  - 2: employé
  - 3: ventes/services
  - 4: professionnel
  - 5: agriculture/ferme
- X_2: revenu familial annuel
  - 1: moins de 25 000
  - 2: 25 000 à 39 999
  - 3: 40 000 à 59 999
  - 4: 60 000 à 79 999
  - 5: 80 000 et plus

Données du PRCA

- X_3: sexe
  - 0: homme
  - 1: femme
- X_4: avez-vous déjà fréquenté une université?
  - 0: non
  - 1: oui
- X_5: âge (en années)
- X_6: combien de fois avez-vous assisté à un rodéo au cours de la dernière année?
  - 1: 10 fois ou plus
  - 2: entre six et neuf fois
  - 3: cinq fois ou moins

Expliquer le comportement de la **moyenne** d'une variable binaire $Y \in \{0,1\}$ en utilisant un modèle de régression avec p variables explicatives $\mathsf{X}_1, \ldots, \mathsf{X}_p$.

$$\underbrace{\mathsf{E}(Y = 1 \mid \mathbf{X})}_{\text{moyenne théorique}} = \underbrace{\mathsf{Pr}(Y = 1 \mid \mathbf{X})}_{\text{probabilité de succès}} = p$$

Objectif de la régression

1) **Inférence** : comprendre comment et dans quelles mesures les variables $\mathbf{X}$ influencent la probabilité que $Y = 1$.
2) **Prédiction** : développer un modèle pour prévoir des valeurs de Y ou la probabilité de succès à partir des $\mathbf{X}$.

Exemples

- Est-ce qu'un client potentiel va répondre favorablement à une offre promotionnelle?
- Est-ce qu'un client est satisfait du service après-vente?
- Est-ce qu'un client va faire faillite ou non au cours des trois prochaines années.

Modéliser une probabilité avec une régression linéaire?

- Aucune contrainte sur le prédicteur linéaire $\beta_0 + \beta_1 \mathsf{X}_1 + \cdots + \beta_p \mathsf{X}_p$
  - retourne des probabilités négatives ou supérieures à 1!
- les données binaires ne respectent pas le postulat d'égalité des variances
  - invalide résultat des tests d'hypothèse sur coefficients.

Illustration: linéaire vs logistique

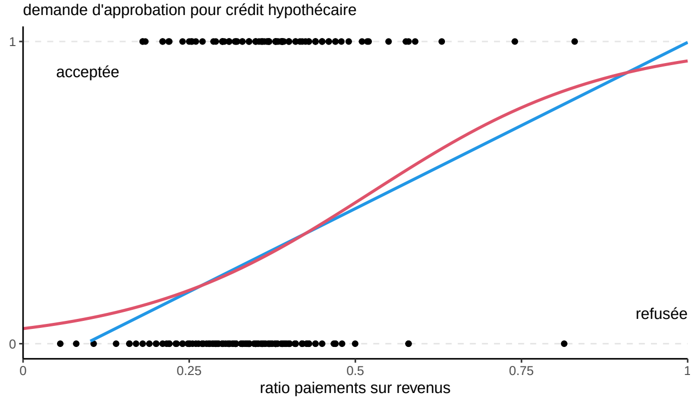


Figure 1: Données de la réserve de Boston sur l'approbation de prêts hypothécaires (1990); données tirées de Stock et Watson (2007).

Fonction de liaison

Idée: appliquer une transformation au **prédicteur linéaire**

$$\eta = \beta_0 + \beta_1 \mathsf{X}_1 + \cdots + \beta_p \mathsf{X}_p$$

pour que la prédiction soit entre zéro et un.

On considère

$$p = \mathsf{expit}(\eta) = \frac{\mathsf{exp}(\eta)}{1 + \mathsf{exp}(\eta)} = \frac{1}{1 + \mathsf{exp}(-\eta)}.$$

Courbe sigmoïde

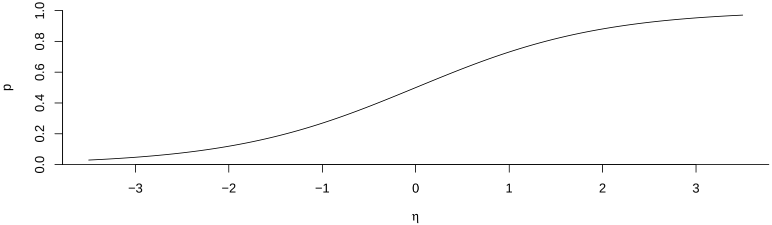


Figure 2: Valeurs ajustées du modèle de régression logistique en fonction du prédicteur linéaire η.

Cote

La cote donne le ratio de la probabilité de succès ($Y = 1$) sur la probabilité d'échec ($Y = 0$).

$$\mathsf{cote}(p) = \frac{p}{1-p} = \frac{\mathsf{Pr}(Y = 1 \mid \mathbf{X})}{\mathsf{Pr}(Y = 0 \mid \mathbf{X})}.$$

Cotes et probabilités

$\Pr(Y=1)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
cote	0.11 $\frac{1}{9}$	0.25 $\frac{1}{4}$	0.43 $\frac{3}{7}$	0.67 $\frac{2}{3}$	1 1	1.5 $\frac{3}{2}$	2.3 $\frac{7}{3}$	4 4	9 9

Ajustement du modèle

La fonction `glm` dans **R** ajuste un modèle linéaire généralisé (par défaut, Gaussien)

- L'argument `family=binomial(link="logit)` permet de spécifier que l'on ajuste un modèle logistique

Modèle avec une seule variable

```r
data(logit1, package = "hecmulti")
# Nombre d'observations par groupe
with(logit1, table(y))
# Ajustement du modèle avec une
#  seule variable explicative
modele1 <- glm(formula = y ~ x5,
               family = binomial(link = "logit"),
               data = logit1)
```

Sortie

Tableau résumé avec les coefficients (`summary`)

```
summary(modele1)
```

Cote pour une augmentation d'une unité des variables explicatives

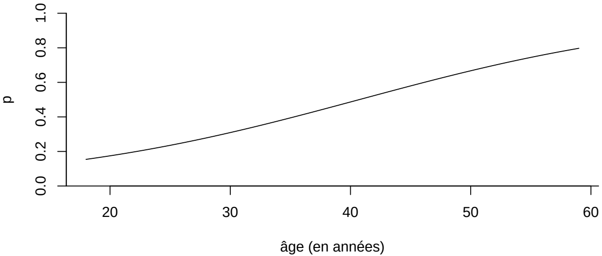
```
cote <- exp(coef(modele1))
confrcote <- exp(confint(modele1))
```

Significativité globale avec test du rapport de vraisemblance vs $\mathcal{H}_0 : p_i = \mathrm{expit}(\beta_0)$

```
anova(modele1, test = 'Chisq')
```

Interprétation

Par défaut, pour des variables $0/1$, le modèle décrit la probabilité de succès.



Si le coefficient β_j de la variable X_j est positif, alors plus la variable augmente, plus $\Pr(Y = 1)$ augmente.

Example avec données du PRCA

Le modèle ajusté en termes de cote est

$$\frac{\mathsf{Pr}(Y=1 \mid \mathsf{X}_5 = x_5)}{\mathsf{Pr}(Y=0 \mid \mathsf{X}_5 = x_5)} = \exp(-3.05)\exp(0.0749 x_5).$$

- Lorsque X_5 augmente d'une année, la cote est multipliée par $\exp(0.0749) = 1.078$ peut importe la valeur de x_5.
- Pour deux personnes dont la différence d'âge
  - est d'un an, la cote de la personne plus âgée est 7.8% plus élevée
  - est de 10 ans, la cote de la personne plus âgée est 112% plus élevées (cote est multiplié par $1.078^{10} = 2.12$)

Vraisemblance et estimation du modèle

Pour un modèle probabiliste donné, on peut calculer la « probabilité » d'avoir obtenu les données de l'échantillon.

Si on traite cette « probabilité » comme une fonction des paramètres, on l'appelle **vraisemblance**.

Maximum de vraisemblance: valeurs des paramètres qui maximisent la fonction de vraisemblance. - on cherche les valeurs des paramètres qui rendent les données les plus plausibles

Vraisemblance d'une observation

La vraisemblance d'une observation $Y_i \in \{0,1\}$ (loi Bernoulli/binomiale) est

$$L(\beta; y_i) = p_i^{y_i}(1-p_i)^{1-y_i} = \begin{cases} p_i & y_i = 1(\text{succès}) \\ 1-p_i & y_i = 0(\text{échec}) \end{cases}$$

et où

$$p_i = \mathrm{expit}(\eta_i) = \frac{\exp(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip})}{1 + \exp(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip})}.$$

Log vraisemblance

Pour des questions de stabilité numérique, on maximise le logarithme naturel $\ell(\beta) = \ln L(\beta)$ (transformation monotone croissante), qui après simplification s'écrit

$$\begin{aligned}\ell(\beta) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip}) \\ - \sum_{i=1}^{n} \ln\left\{1 + \exp(\beta_0 + \cdots + \beta_p \mathsf{X}_{ip})\right\}\end{aligned}$$

Pas de solution explicite pour $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$ dans le cas de la régression logistique.

Prédiction des probabilités de succès

Des estimés des paramètres $\widehat{\beta}$, découle une estimation de $\mathsf{Pr}(Y = 1)$ pour les valeurs $\mathsf{X}_1 = x_1, \ldots, \mathsf{X}_p = x_p$ d'un individu donné,

$$\widehat{p} = \mathrm{expit}(\widehat{\beta}_0 + \cdots + \widehat{\beta}_p \mathsf{X}_p).$$

Test du rapport de vraisemblance

Pour les modèles ajustés par maximum de vraisemblance.
Comparaison de modèles **emboîtés**

- Modèle complet (sous l'alternative) avec tous les paramètres
- Modèle restreint (sous l'hypothèse nulle) sur lequel on impose $k \leq p$ restrictions (typiquement $\beta_j = 0, j \in \{1, \ldots, p\}$).

Exemple

Comparons un modèle avec et sans X_6.

Variable catégorielle à trois niveaux (deux coefficients associés à $\mathsf{I}(\mathsf{X}_6 = 2)$ et $\mathsf{I}(\mathsf{X}_6 = 3)$.

```
modele2 <-  glm(y ~ x1 + x2 + x3 + x4 + x5 + x6,
                data = hecmulti::logit1,
                family = binomial(link = "logit"))
modele3 <-  glm(y ~ x1 + x2 + x3 + x4 + x5,
                data = hecmulti::logit1,
                family = binomial(link = "logit"))
# Restriction X6_2 = X6_3 = 0 (k=2 param)
```

Rapport de vraisemblance

Le test est basé sur la statistique

$$D = -2\{\ell(\widehat{\beta}_0) - \ell(\widehat{\beta})\}.$$

Cette différence D, lorsque l'hypothèse $\mathscr{H}_0$ est vraie, suit approximativement une loi khi-deux χ^2_k.

Exemple de test

```r
# modèle 2 (alternative), modèle 3 (nulle)
anova(modele3, modele2, test = "LR")
```

```
Analysis of Deviance Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5
Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       488     566.45
2       486     516.20  2   50.251 1.225e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Deviance = -2*log vraisemblance
rvrais <- modele3$deviance - modele2$deviance
pchisq(rvrais, df = 2, lower.tail = FALSE) # valeur-p
```

```
[1] 1.225046e-11
```

Tester la significativité des variables

Si un paramètre n'est pas significativement différent de 0, cela veut dire qu'il n'y a pas de lien significatif entre la variable et la réponse *une fois que les autres variables* sont dans le modèle.

```r
car::Anova(modele2, type = "3")
```

```
Analysis of Deviance Table (Type III tests)

Response: y
   LR Chisq Df Pr(>Chisq)
x1    4.291  4     0.3681
x2   32.912  4  1.245e-06 ***
x3   29.878  1  4.601e-08 ***
x4   42.957  1  5.597e-11 ***
x5   36.731  1  1.356e-09 ***
x6   50.251  2  1.225e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Intervalles de confiance pour coefficients

On peut fixer la valeur de β_j et maximiser la vraisemblance pour les autres paramètres.

La courbe de vraisemblance profilée résultante permet de déterminer l'intervalle de confiance pour le paramètre.

Les intervalles de confiance de vraisemblance sont invariants aux reparamétrisation.

Intervalles de confiance

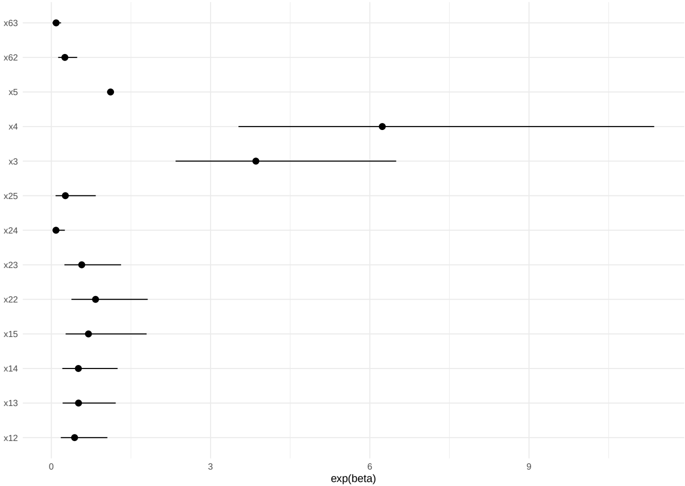


Figure 3: Intervalles de confiance profilés de niveau 95% pour les coefficients du modèle logistique (échelle exponentielle).

Tests et intervalles de confiances

Comme $\exp(\cdot)$ est une transformation monotone croissante,

$$\beta > 0 \quad \iff \quad \exp(\beta) > 1.$$

Si la valeur postulée, par exemple $\mathscr{H}_0 : \beta_j = 0$ ou $\exp(\beta_j) = 1$, est dans l'intervalle de confiance de niveau $1-\alpha$, on ne rejette pas l'hypothèse nulle.

Données complètes

Table 1: Modèle logistique avec toutes les variables catégorielles.

variables	cote[1]	IC 95%[1]	valeur-p
x1			0.4
1	—	—	
2	0.44	0.18, 1.06	
3	0.51	0.21, 1.21	
4	0.51	0.21, 1.25	
5	0.70	0.27, 1.80	
x2			<0.001
1	—	—	
2	0.83	0.38, 1.82	
3	0.57	0.25, 1.31	
4	0.09	0.03, 0.25	
5	0.26	0.08, 0.84	
x3			<0.001
0	—	—	
1	3.85	2.34, 6.50	
x4			<0.001
0	—	—	
1	6.24	3.53, 11.4	

Multicolinéarité

Il est difficile de départager l'effet individuel d'une variable explicative lorsqu'elle est fortement corrélée avec d'autres.

La multicollinéarité ne dépend pas de la variable réponse Y, mais de la matrice $\mathbf{X}$ du modèle.

Multicolinéarité pour PRCA

Mêmes diagnostics qu'en régression linéaire: considérer les facteurs d'inflation de la variance (`car::vif`).

```r
car::vif(modele2)
```

```
       GVIF Df GVIF^(1/(2*Df))
x1 1.698464  4        1.068457
x2 1.852841  4        1.080139
x3 1.450100  1        1.204201
x4 1.491202  1        1.221148
x5 1.219334  1        1.104234
x6 1.179133  2        1.042055
```

Pas d'inquiétude ici, coefficients faibles (inférieurs à 5)

Dichotomiser des variables continues