Régression logistique

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

1/1/22

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 dans la base de données `logit1`:

- 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

Régression logistique

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

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

Objectif de la régression

1) **Inférence** : comprendre comment et dans quelles mesures les variables **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 **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.

Inférence et interprétation

Ce cours est consacré à l'estimation et l'interprétation des paramètres du modèle dans le cas binaire.

Par convention, on désigne le résultat « 1 » par un succès et « 0 » par un échec.

Modéliser une probabilité avec une régression linéaire?

Mauvaise idée!

- Sans contrainte, on peut obtenir des probabilités négatives ou supérieures à 1!
- les données binaires ne respectent pas le postulat d'égalité des variances
  - invalide les résultats des tests d'hypothèse pour les coefficients.

Illustration: linéaire vs logistique

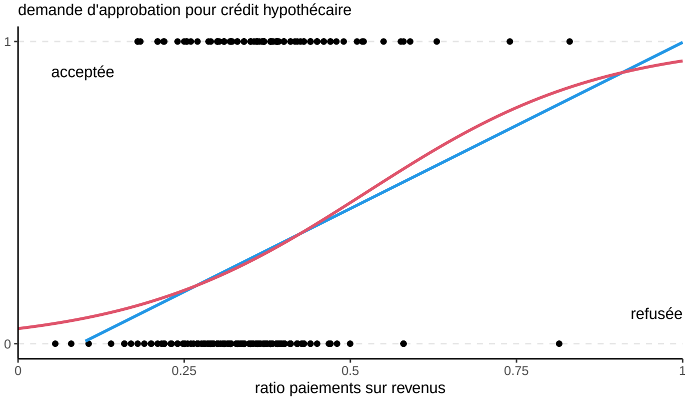


Figure 1: Données de la réserve de Boston sur l'approbation de prêts hypothécaires (1990); données tirées de Stock et Watson (2007).

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

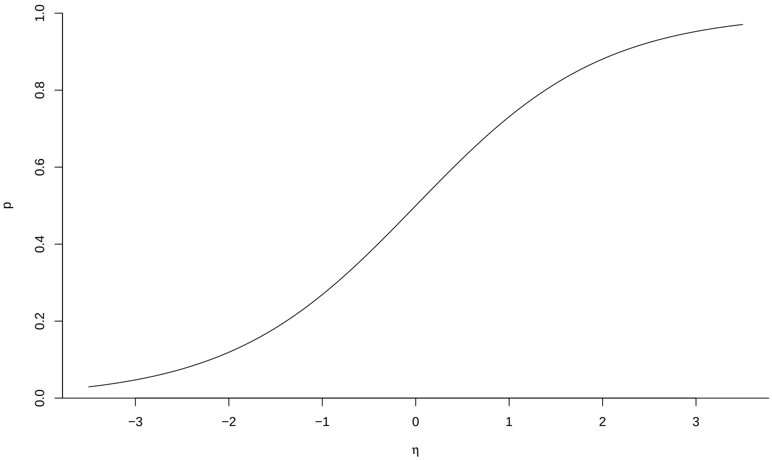


Figure 2: Valeurs ajustées du modèle de régression logistique en fonction du prédicteur linéaire η.

Ajustement du modèle

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

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

```
data(logit1, package = "hecmulti")
# Ajustement du modèle avec toutes
# les variables explicatives
modele1 <- glm(formula = y ~ .,
               family = binomial(link = "logit"),
               data = logit1)
```

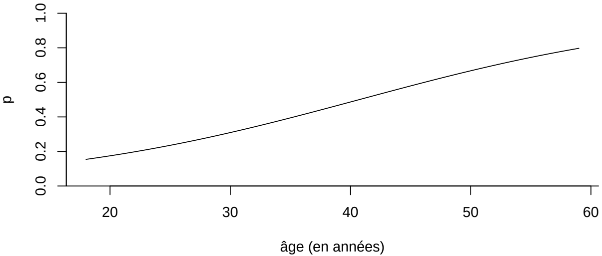
Sortie

Tableau résumé avec les coefficients (`summary`)

```
summary(modele1)
```

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 $\mathsf{Pr}(Y = 1)$ augmente.

Rappel exponentiels et logarithme

Quelques propriétés de la fonction exponentielle:

- $\exp(0) = 1$
- $\exp(a + b) = \exp(a)\exp(b)$
- $\exp(ab) = \exp(a)^b$

Quelques propriétés de la fonction logarithmique

- $\ln(1) = 0$,
- $\ln(\exp(x)) = x$ (fonction inverse)
- $\ln(ab) = \ln(a) + \ln(b)$

Cote

Si on applique la transformation inverse, on obtient

$$\ln\left(\frac{p}{1-p}\right) = \eta = \beta_0 + \beta_1 \mathsf{X}_1 + \cdots + \beta_p \mathsf{X}_p.$$

ou, en prenant l'exponentielle de chaque côté,

$$\text{cote} = \frac{p}{1-p} = \exp(\beta_0)\cdots\exp(\beta_p \mathsf{X}_p)$$

Modèle multiplicatif pour la cote.

Cote et probabilité

La cote est utilisée dans les paris sportifs

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

Table 1: Cote et probabilité de succès

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
cote	$\frac{1}{9}$	$\frac{1}{4}$	$\frac{3}{7}$	$\frac{2}{3}$	1	$\frac{3}{2}$	$\frac{7}{3}$	4	9

Interprétation avec données du PRCA

Le modèle ajusté en termes de cote est

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

- 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 élevée (cote multipliée par $\exp(10 \times 0.0749) = 1.078^{10} = 2.12$)

Modèle complet

On considère le modèle avec toutes les variables explicatives:

```r
modele2 <- glm(
  formula = y ~ .,
  data = logit1,
  family = binomial)
exp(coef(modele2))
```

```
(Intercept)         x12         x13         x14         x15
      0.062       0.438       0.512       0.509       0.699
        x23         x24         x25          x3          x4
      0.572       0.087       0.264       3.854       6.236
        x62         x63
      0.255       0.090
```

Interprétation des coefficients

Si on a plusieurs variables explicatives, les coefficients sont interprétés en modifiant une variable à la fois.

On compare deux profils identiques, sauf pour la variable en question

- toute chose étant égale par ailleurs
- *ceteris paribus*

Interprétation des coefficients

Variable continue:

- La cote de la personne plus âgée d'un an est 1.116 fois celle de la personne plus jeune, *ceteris paribus*, une augmentation de 11,6%

Le rapport de cote pour les femmes (`x3=1`) versus les hommes (`x3=0`) est de $\exp(\widehat{\beta}_{\mathrm{x}_3}) = 3.854$:

- les femmes sont plus susceptibles de suivre les recommandations d'achat toute chose étant égale par ailleurs,
- Inversement, le rapport de cote homme/femme est de 1/3.854=0.259,

On peut donc conclure que:

- la cote des femmes est 285.4% plus élevée que celle des hommes.
- la cote des hommes est 74.1% inférieure à celle des femmes.

Variable catégorielle

Toutes les comparaisons sont effectuées avec la catégorie de référence.

Pour `x1`, c'est à la maison. Le rapport de cote est

$$\frac{\mathsf{cote}\{Y \mid \mathsf{X}_1 = 2(\text{employé}), \ldots\}}{\mathsf{cote}\{Y \mid \mathsf{X}_1 = 1(\text{maison}), \ldots\}} = \exp(\widehat{\beta}_{\mathsf{X}_1=2}) = 0.438$$

Le coefficient pour $\mathsf{X}_1 = 1$ est zéro, d'où $\exp(\widehat{\beta}_{\mathsf{X}_1=1}) = 1$ (absent du tableau).

On peut ordonner les type d'emploi selon la probabilité de succès à l'aide des coefficients: *ceteris paribus* on obtient le classement

- employé < professionnel < ventes/service < agriculture < maison.

Invariance

Si on voulait le rapport de cote professionnel vs employé, inutile de réajuster le modèle.

On peut calculer

$$\frac{\dfrac{\mathsf{cote}(Y \mid \mathsf{X}_1 = 4, \ldots)}{\mathsf{cote}(Y \mid \mathsf{X}_1 = 1, \ldots)}}{\dfrac{\mathsf{cote}(Y \mid \mathsf{X}_1 = 2, \ldots)}{\mathsf{cote}(Y \mid \mathsf{X}_1 = 1, \ldots)}} = \frac{\exp(\widehat{\beta}_{\mathsf{X}_1=4})}{\exp(\widehat{\beta}_{\mathsf{X}_1=2})} = 1.162746$$

plutôt que de changer la catégorie de référence via

```r
logit2 <- logit1 |>
   mutate(x1 = relevel(x1, ref = 2))
```

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

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

Probabilité conjointe d'événements binaires

- Si les observations sont indépendantes, la probabilité conjointe d'avoir un résultat donné est le produit des probabilités pour chaque observation.
- Il n'y a pas de solution explicite pour $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$ dans le cas de la régression logistique: il faut maximiser la vraisemblance.

Log vraisemblance

- Pour des raisons de stabilité numérique, on maximise le logarithme naturel $\ell(\beta) = \ln L(\beta)$ de la log vraisemblance conjointe de l'échantillon (transformation monotone croissante).
- La log vraisemblance est simplement la somme des contributions individuelles.
- On utilise la log vraisemblance ℓ comme mesure d'ajustement et pour construire des tests d'hypothèse.

Prédiction des probabilités de succès

Des estimations des coefficients $\widehat{\beta}$ découlent 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,

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

Modèle nul

Un modèle avec uniquement l'ordonnée à l'origine retournera $\widehat{p}$, la proportion empirique de succès.

Comme pour la régression linéaire, c'est la moyenne des observations.

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 $\beta_j = 0, j \in \{1, \ldots, p\}$).

Comparons un modèle avec et sans X_6.

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

```r
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"))
```

Test d'hypothèse

On test l'hypothèse nulle $\mathscr{H}_0 : \beta_{\mathsf{X}_6=2} = \beta_{\mathsf{X}_6=3} = 0$ (soit $k = 2$ restrictions).

L'hypothèse alternative est qu'au moins un des coefficients est non-nul.

Si la valeur p est inférieure au seuil de signification, typiquement $\alpha = 0.05$, on rejette l'hypothèse nulle.

- on conclut que la variable explicative X_6 améliore significativement l'ajustement du modèle.

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 aussi considérer des intervalles de confiance pour les coefficients individuels.

Ceux obtenus par défaut dans **R** sont appelés *intervalles de confiance de vraisemblance profilée*.

```r
confint(modele2)      # IC pour beta
exp(confint(modele2)) # IC pour exp(beta)
```

Ces intervalles sont invariants aux reparamétrisation: si $[b_i, b_s]$ est l'intervalle de vraisemblance profilée pour β, l'intervalle pour $\exp(\beta)$ est simplement $[\exp(b_i), \exp(b_s)]$.

Intervalles de confiance

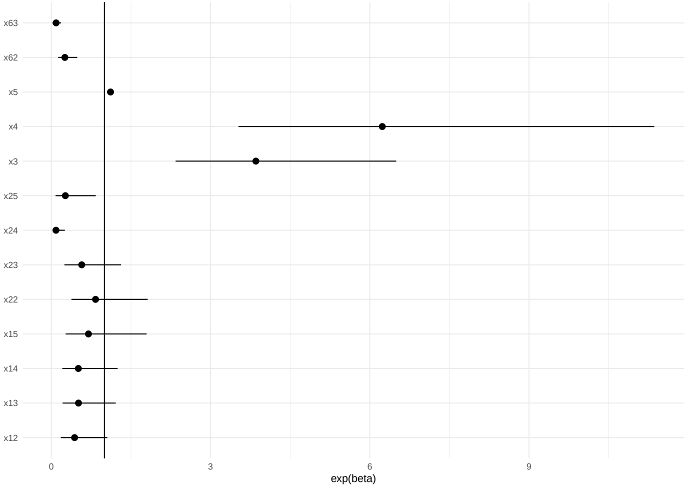


Figure 3: Intervalles de confiance profilés de niveau 95% pour les coefficients du modèle logistique (échelle exponentielle).

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.

Coefficients pour données complètes

Table 2: 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	

[1]cote = rapport de cote, IC = intervalle de confiance

Coefficients pour données complètes

Table 3: Modèle logistique avec toutes les variables catégorielles.

variables	cote[1]	IC 95%[1]	valeur-p
x4			<0.001
0	—	—	
1	6.24	3.53, 11.4	
x5	1.12	1.08, 1.16	<0.001
x6			<0.001
1	—	—	
2	0.25	0.13, 0.49	
3	0.09	0.04, 0.18	

[1]cote = rapport de cote, IC = intervalle de confiance

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`).

```
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

Si Y est continue et qu'on cherche à estimer $\Pr(Y > c \mid \mathbf{X})$ pour une valeur c donnée, il n'est **pas** recommandé de dichotomiser Y via

$$Y^* = \begin{cases} 1, & Y > c; \\ 0, & Y \leq c. \end{cases}$$

et d'ajuster une régression logistique.

Pourquoi? **On perd de l'information**.

Probabilité de dépassement

On peut estimer plutôt une régression linéaire et prendre

$$\Pr(Y > c \mid \mathbf{X}) = \Phi\left(\frac{\hat{\mu} - c}{\hat{\sigma}}\right),$$

où

- $\hat{\mu} = \hat{\beta}_0 + \cdots + \beta_p \mathrm{X}_p$ est la moyenne prédite pour le profil donné,
- $\hat{\sigma}$ est l'estimation de l'écart-type
- $\Phi(\cdot)$ est la fonction de répartition d'une loi normale standard (`pnorm`)

Récapitulatif

- Une régression logistique sert à modéliser la moyenne de **variables catégorielles**, typiquement binaires.
- C'est un cas particulier d'un modèle de régression linéaire généralisée (GLM)

Récapitulatif

Le modèle est interprétable à l'échelle de la cote

- La cote donne le rapport probabilité de réussite (1) sur probabilité d'échec (0)
- Interprétation en terme de
  - pourcentage d'augmentation si $\exp(\widehat{\beta}) > 1$, avec $\exp(\widehat{\beta}) - 1$.
  - pourcentage de diminution si $\exp(\widehat{\beta}) < 1$, avec $1 - \exp(\widehat{\beta})$

Récapitulatif

- Estimation par maximum de vraisemblance
- Tests d'hypothèse comparent modèles emboîtés
  - loi nulle asymptotique χ^2
  - degrés de liberté égal au nombre de restrictions
- Intervalles de confiance de vraisemblance profilée
  - invariants aux reparamétrisations