Régression logistique

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Rappel

La régression logistique spécifie un modèle pour la probabilité de succès

$$p = \mathsf{Pr}(Y = 1 \mid \mathbf{X}) = \frac{1}{1 + \exp(-\eta)}$$

où $\eta = \beta_0 + \cdots + \beta_p \mathsf{X}_p$.

Prédiction

En substituant l'estimation $\widehat{\beta}_0, \dots, \widehat{\beta}_p$, on calcule

- le prédicteur linéaire $\widehat{\eta}_i$ et
- la probabilité de succès $\widehat{p}_i$

pour chaque ligne de la base de données.

Classification de base

Choisir un point de coupure c:

- si $\hat{p} < c$, on assigne $\widehat{Y} = 0$.
- si $\hat{p} \geq c$, on assigne $\widehat{Y} = 1$.
- Un point de coupure de $c = 0.5$ revient à assigner l'observation à la classe (catégorie) la plus probable.
- Si $c = 0$, on catégorise toutes les observations en succès avec $\widehat{Y}_i = 1 \ (i = 1, \dots, n)$.

Qualité de l'ajustement

L'erreur quadratique pour une variable binaire est

$$(Y - \widehat{Y})^2 = \begin{cases} 1, & Y \neq \widehat{Y}; \\ 0, & Y = \widehat{Y}. \end{cases}$$

et donc on obtient le **taux de mauvaise classification** si on calcule la moyenne.

Plus le taux de mauvaise classification est petit, meilleure est la capacité prédictive du modèle.

Estimation de la performance du modèle

Utiliser les mêmes données pour l'ajustement et l'estimation de la performance n'est (toujours) pas recommandé.

Plutôt, considérer

- la validation croisée,
- la division de l'échantillon.

Base de données marketing

On considère un modèle pour `yachat`, le fait qu'une personne achète suite à l'envoi d'un catalogue.

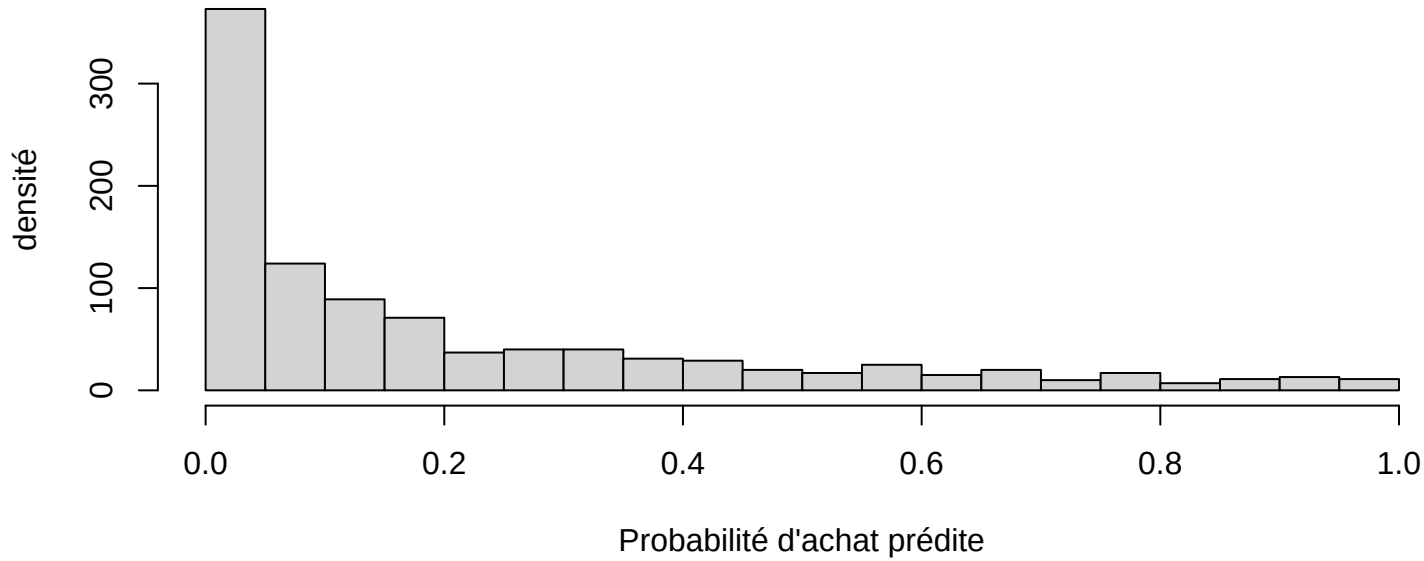
```
data(dbm, package = "hecmulti")
formule <- formula("yachat ~ x1 + x2 + x3 +
                   x4 + x5 + x6 + x7 + x8 + x9 + x10")
dbm_class <- dbm |>
  dplyr::filter(test == 0) |>
  # pour caret, convertir 0/1 en facteurs
  dplyr::mutate(yachat = factor(yachat))
```

Estimation avec validation croisée

On utilise la fonction `train` du paquet `caret`, avec le modèle linéaire généralisé.

```
set.seed(202209)
cv_glm <-
  caret::train(form = formule,
               data = dbm_class,
               method = "glm",
               family = binomial(link = "logit"),
               trControl = caret::trainControl(
                 method = "cv",
                 number = 10))
```

Prédictions



Répartition des probabilités de succès prédites par validation croisée.

Performance

On peut varier le point de coupure et regarder pour chaque valeur de c la classification résultante.

```
# predict retourne une matrice n x 2
# avec [P(Y=0), P(Y=1)]
predprob <- predict(cv_glm, type = "prob")[,2]
classif <- with(dbm, yachat[test == 0])
# Tableau de la performance
hecmulti::perfo_logistique(
  prob = predprob,
  resp = classif)
```

Matrice de confusion

On peut classer les observations dans un tableau pour un point de coupure donné.

Table 1: Matrice de confusion avec point de coupure 0.465.

	$Y = 1$	$Y = 0$
$\widehat{Y} = 1$	109	52
$\widehat{Y} = 0$	101	738

Classification et mesures de performance

Les estimés empiriques sont simplement obtenus en calculant les rapports du nombre d'observations dans chaque classe.

	$Y=1$	$Y=0$
$\widehat{Y}=1$	109	52
$\widehat{Y}=0$	101	738

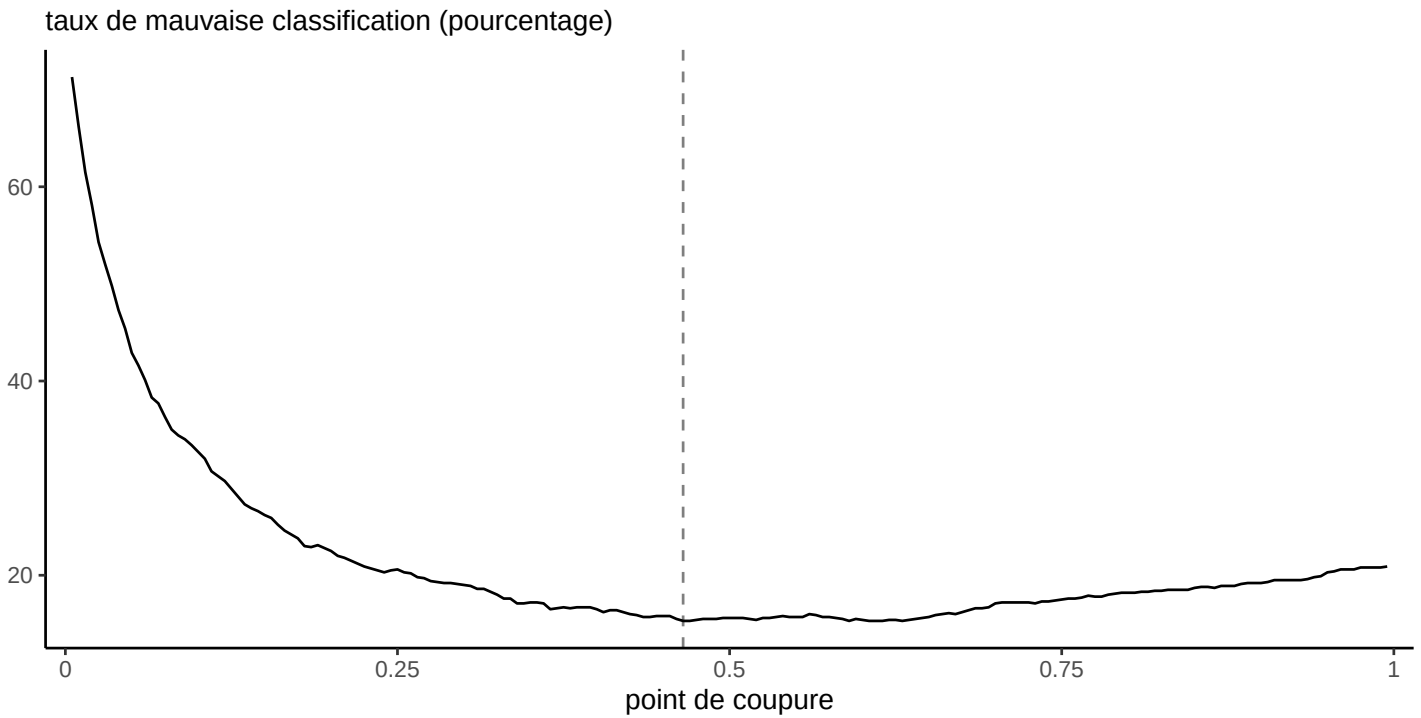
	$Y=1$	$Y=0$
$\widehat{Y}=1$	VP	FP
$\widehat{Y}=0$	FN	VN

- La **sensibilité** est le taux de succès correctement classés, $\Pr(Y=1, \widehat{Y}=1 \mid Y=1)$, soit $\mathsf{VP}/(\mathsf{VP}+\mathsf{FN})$.
- La **spécificité** est le taux d'échecs correctement classés, $\Pr(Y=0, \widehat{Y}=0 \mid Y=0)$, soit $\mathsf{VN}/(\mathsf{VN}+\mathsf{FP})$.
- Le taux de **faux positifs** est $\Pr(Y=0, \widehat{Y}=1 \mid \widehat{Y}=1)$.
- Le taux de **faux négatifs** est $\Pr(Y=1, \widehat{Y}=0 \mid \widehat{Y}=0)$.

Choix d'un point de coupure.

On peut faire varier le point de coupure et choisir celui qui minimise le taux de mauvaise classification, $(\text{FP} + \text{FN})/n$.

Ici, avec $c = 0.465$, on obtient 15.3%.



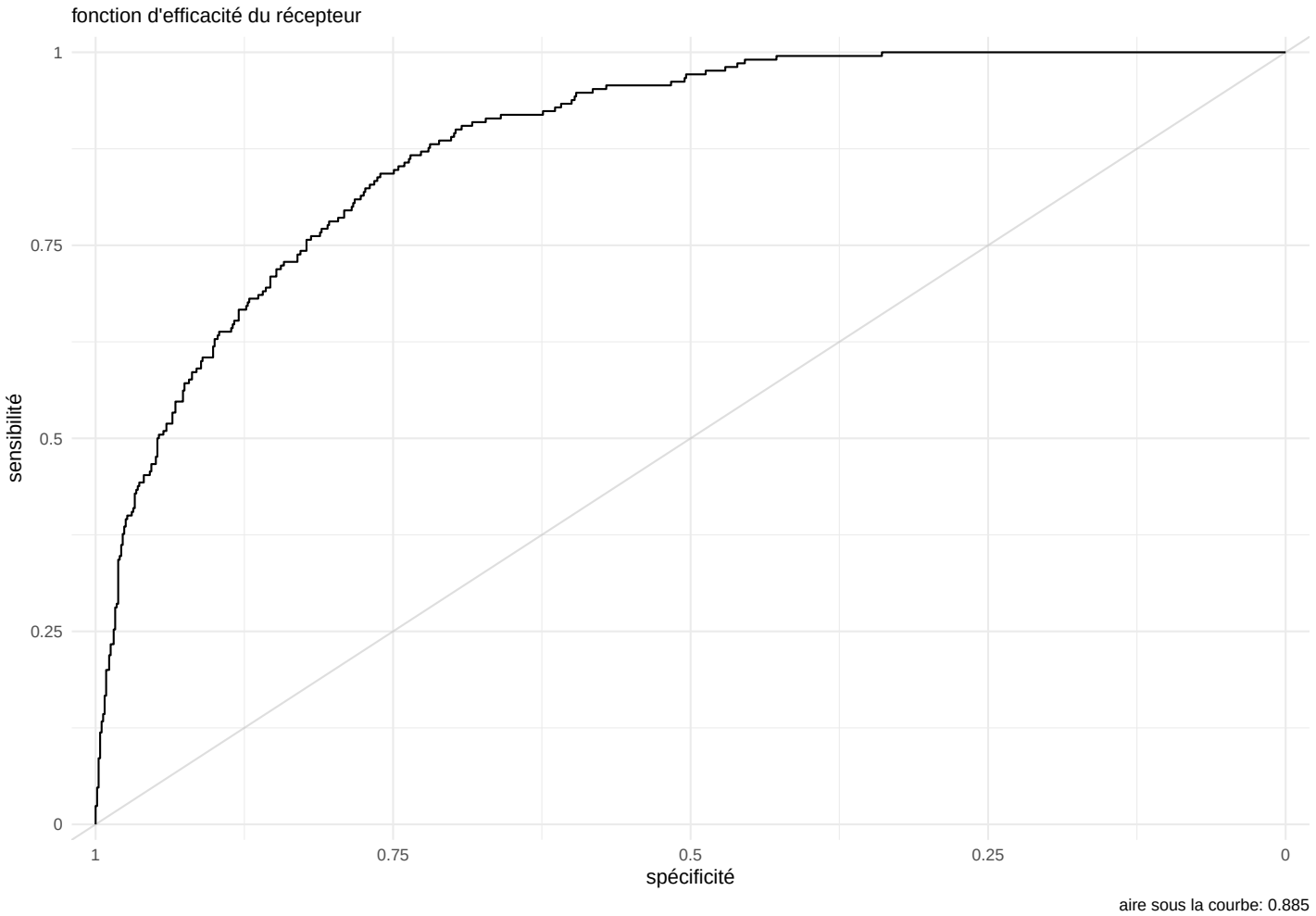
Fonction d'efficacité du récepteur

Graphique de la sensibilité en fonction de un moins la spécificité, en faisant varier le point de coupure, souvent appelé courbe ROC (de l'anglais *receiver operating characteristic*).

La fonction `hecmulti::courbe_roc` permet de tracer la courbe et de calculer l'aire sous la courbe.

```
roc <- hecmulti::courbe_roc(
  resp = classif,
  prob = predprob,
  plot = TRUE)
print(roc)
## Pour extraire l'aire sous la courbe, roc$aire
```

Courbe ROC


fonction d'efficacité du récepteur
sensibilité
1
0.75
0.5
0.25
0
1
0.75
0.5
0.25
0
spécificité
aire sous la courbe: 0.885

Aire sous la courbe

- Plus la courbe se rapproche de (0, 1) (coin supérieur gauche), meilleure est la classification.
- Autrement dit, plus l'aire sous la courbe est près de 1, mieux c'est.
- Une aire sous la courbe de 0.5 (ligne diagonale) correspond à la performance d'une allocation aléatoire.

Courbe lift

À quelle point notre modèle est meileur qu'une assignation aléatoire?

- Ordonner les probabilités de succès estimées par le modèle, $\hat{p}$, en ordre croissant.
- Regarder quelle pourcentage de ces derniers seraient bien classifiés (le nombre de vrais positifs sur le nombre de succès). La référence est la ligne diagonale, qui correspond à une détection aléatoire.

Code pour produire la courbe lift

```
tab_lift <- hecmulti::courbe_lift(
  prob = 1-predprob,
  resp = classif,
  plot = TRUE)
tab_lift
```

Courbe lift

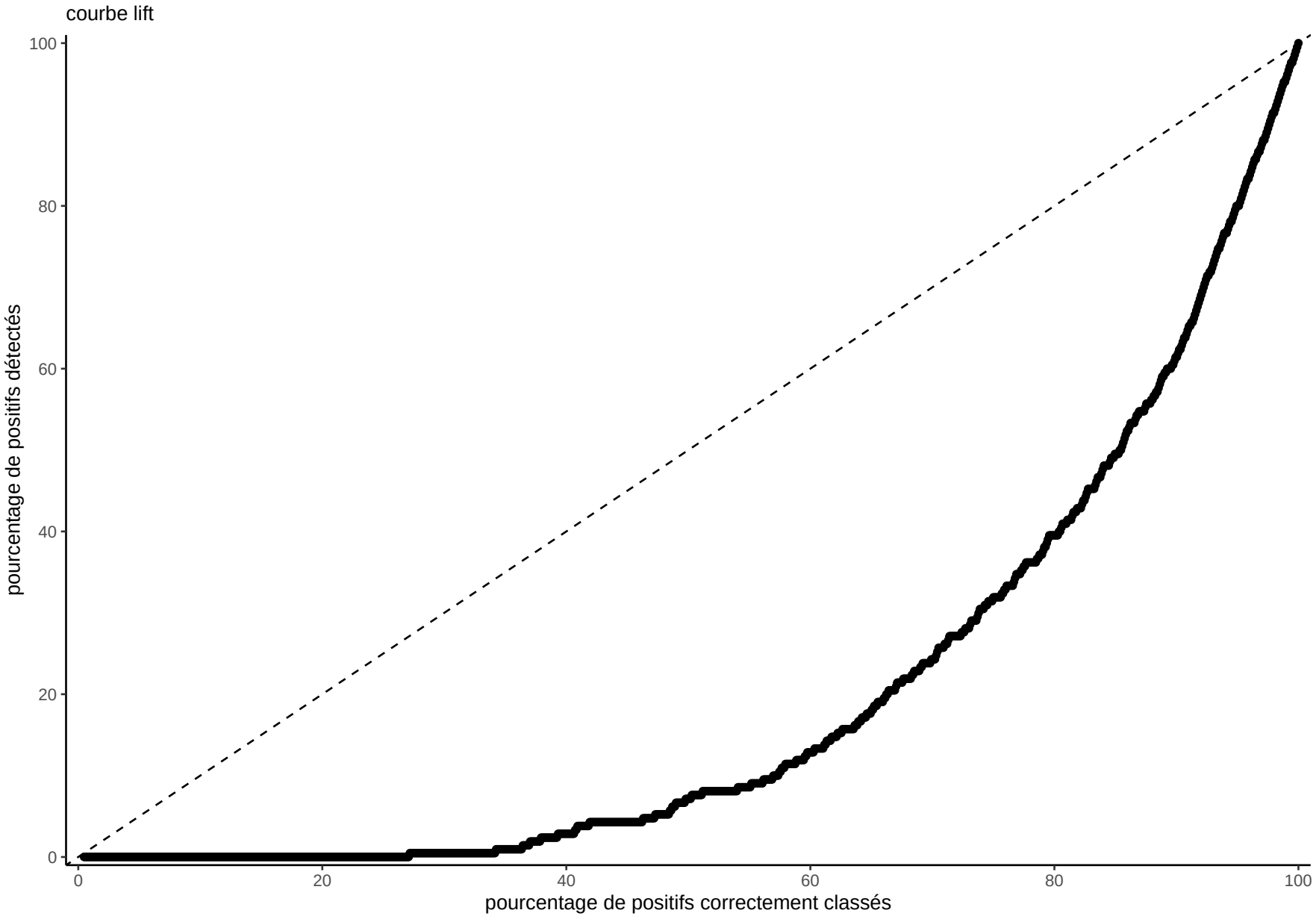

courbe lift
100
80
60
40
20
0
pourcentage de positifs détectés
0
20
40
60
80
100
pourcentage de positifs correctement classés

Tableau du lift

	pourcent	hasard	modele	lift
10%	10	21	0	0.00
20%	20	42	0	0.00
30%	30	63	1	0.02
40%	40	84	6	0.07
50%	50	105	15	0.14
60%	60	126	27	0.21
70%	70	147	51	0.35
80%	80	168	83	0.49
90%	90	189	129	0.68

Si on classifiait comme acheteurs les 10% qui ont la plus forte probabilité estimée d'achat, on détecterait 81 des 210 clients.

Le lift est le nombre détecté par le modèle sur proportion au hasard.