Régression logistique: prédiction et classification

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 $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$, 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 $\hat{Y} = 0$.
- si $\hat{p} \geq c$, on assigne $\hat{Y} = 1$.
- 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 $\hat{Y}_i = 1$ $(i = 1, \dots, n)$.

Qualité de l'ajustement

L'erreur quadratique pour un problème de classification est

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

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 ou 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.

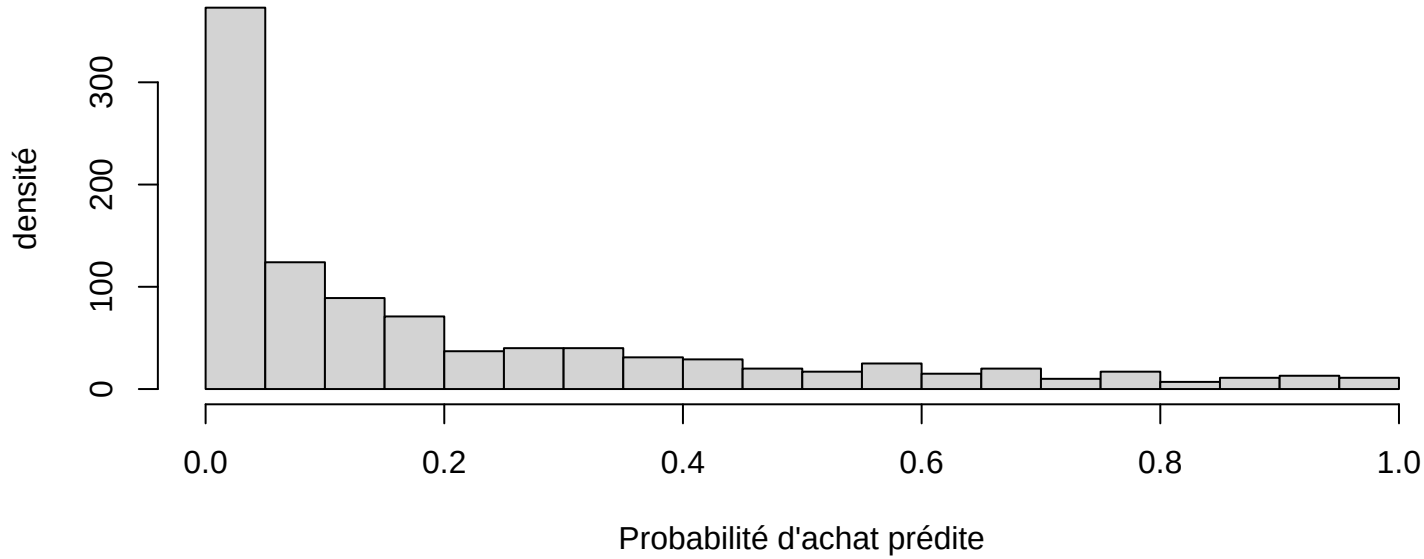
```
data(dbm, package = "hecmulti")
formule <- formula("yachat ~ x1 + x2 + x3 +
                   x4 + x5 + x6 + x7 + x8 + x9 + x10")
dbm_class <- dbm |>
  dplyr::filter(test == 0) |>
  # pour caret, convertir 0/1 en facteurs
  dplyr::mutate(yachat = factor(yachat))
```

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 classifier les observations dans un tableau pour un point de coupure donné.

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

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

Classification et mesures de performance

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

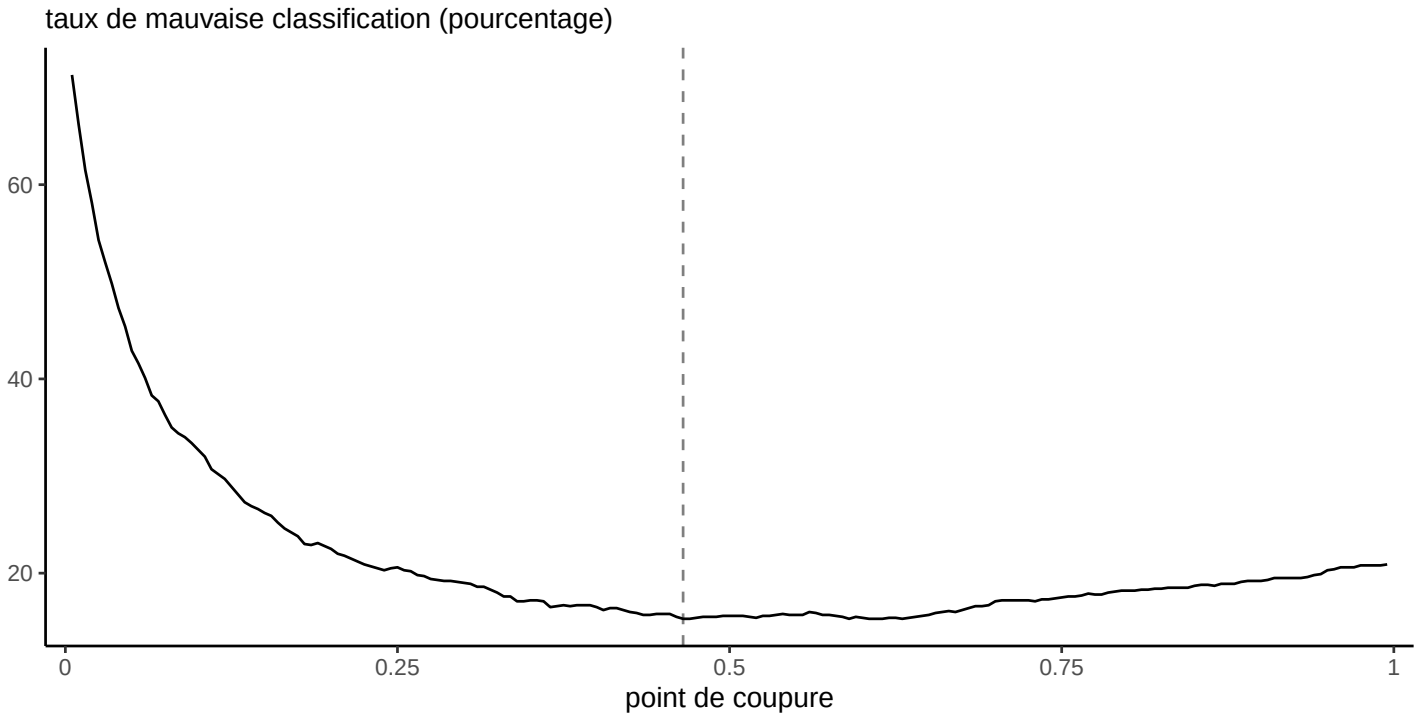
	$Y=1$	$Y=0$
$\hat{Y}=1$	109	52
$\hat{Y}=0$	101	738

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

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

Choix d'un point de coupure.

On peut faire varier le point de coupure et choisir celui qui maximise le taux de bonne classification, $\widehat{\mathsf{Pr}}(Y = \hat{Y})$.



Avec $c = 0.465$, on obtient un taux de mauvaise classification de 15.3%.

Matrice de gain

Il est également possible d'assigner un poids différent à chaque événement selon le scénario et chercher à maximiser le gain.

Table 2: Matrice de gain (cas général)

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	c_{11}	c_{10}
$\hat{Y} = 0$	c_{01}	c_{00}

On calcule le gain en faisant la somme des entrées fois les poids, soit

$$\text{gain} = c_{11}\text{VP} + c_{10}\text{FP} + c_{01}\text{FN} + c_{00}\text{VN}.$$

Gain pour taux de bonne classification

Si on cherche à maximiser le taux de bonne classification, cela revient à assigner les poids suivants.

Table 3: Matrice de gain pour le taux de bonne classification.

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	1	0
$\hat{Y} = 0$	0	1

Coûts et bénéfices du ciblage marketing

- Si on n'envoie pas de catalogue, notre gain est nul.
- Si on envoie le catalogue
  - à un client qui n'achète pas, on perd 10$ (le coût de l'envoi).
  - à un client qui achète, notre revenu net est de 57$ (revenu moyen moins coût de l'envoi).

n	moyenne	écart-type	minimum	maximum
210	67.29	13.24	25	109

Statistiques descriptives des montants d'achats pour la base de données marketing.

Table 4: Matrice de gain pour ciblage marketing.

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	57	-10
$\hat{Y} = 0$	0	0

Point de coupure avec gain

```
formule = formula(yachat ~ x1 + x2 + x3 +
                    x4 + x5 + x6 + x7 +
                    x8 + x9 + x10)
modele <- glm(formule,
              family = binomial,
              data = hecmulti::dbm)
coupe <- hecmulti::select_pcoupe(
  modele = modele,
  c00 = 0,
  c01 = 0,
  c10 = -10,
  c11 = 57,
  plot = TRUE)
```

La fonction `select_pcoupe` estime le gain pour différents points de coupures, avec probabilités estimées par validation croisée avec `ncv` groupes, répétée `nrep` fois.

Calcul du gain

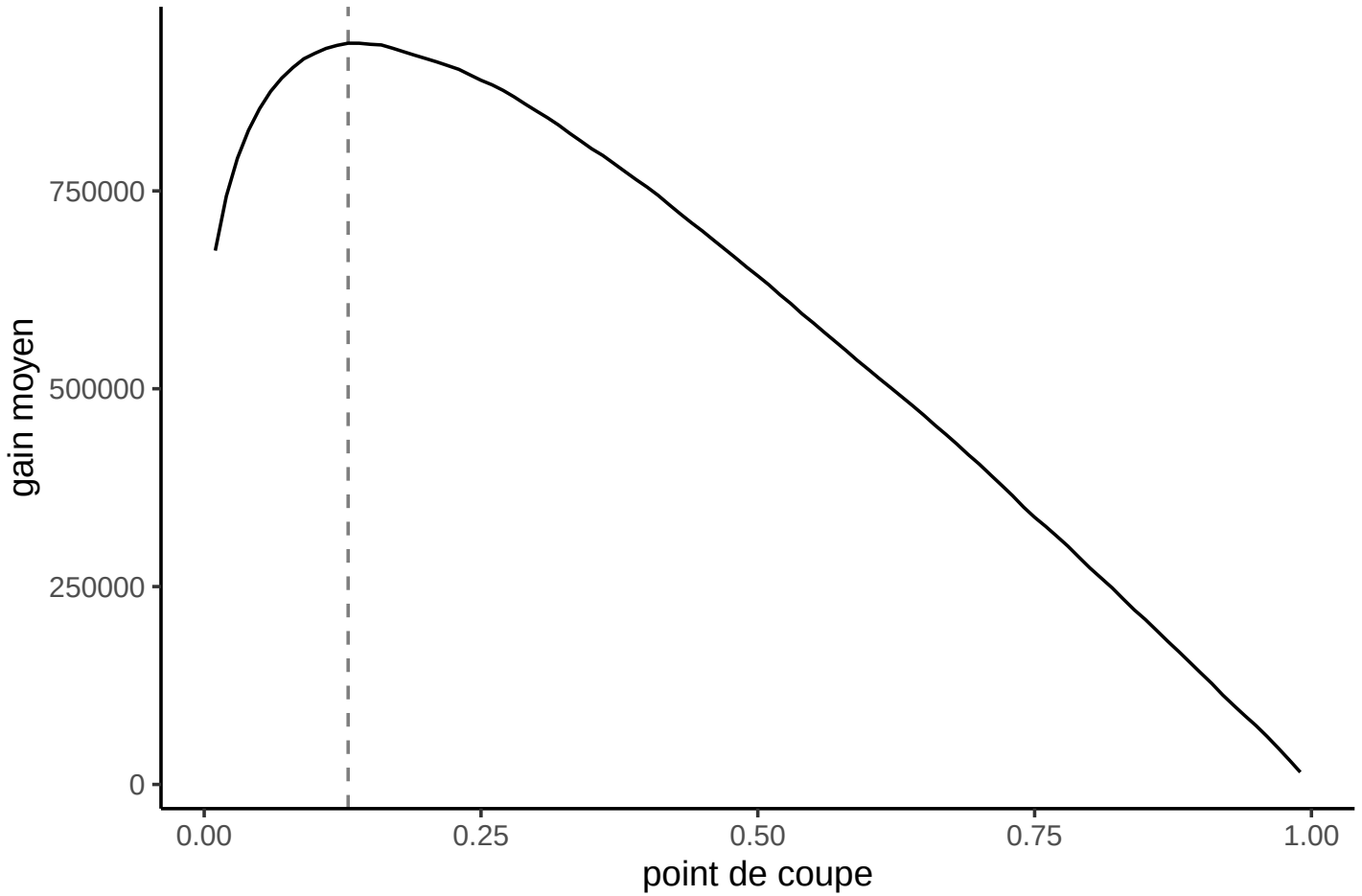


Figure 1: Estimation du gain moyen en fonction du point de coupure pour l'exemple de base de données marketing.

Point de coupure optimal et explication

Dans l'exemple, le point de coupure qui maximise le gain est 0.13. Avec ce point de coupure, on estime que

- le taux de bonne classification est de 70.3
- la sensibilité est de 90.95.

Ainsi, on va détecter environ 90.95% des clients qui achètent.

Il est coûteux de rater un client potentiel, donc la stratégie optimale est d'envoyer le catalogue à plus de clients quitte à ce que plusieurs d'entre eux n'achètent rien.

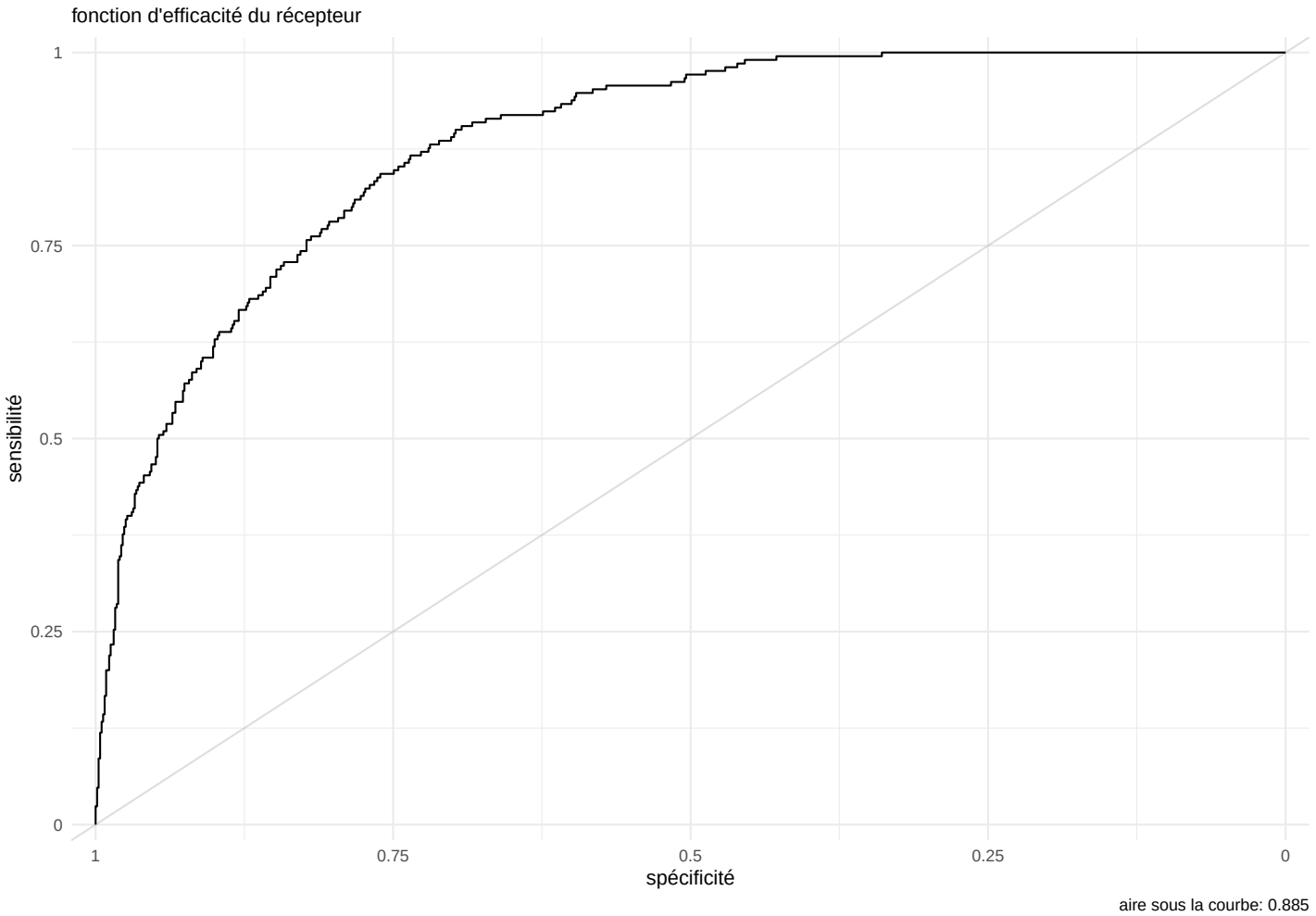
Fonction d'efficacité du récepteur

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, $\widehat{p}$, 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 = predprob,
  resp = classif,
  plot = TRUE)
tab_lift
```

Courbe lift

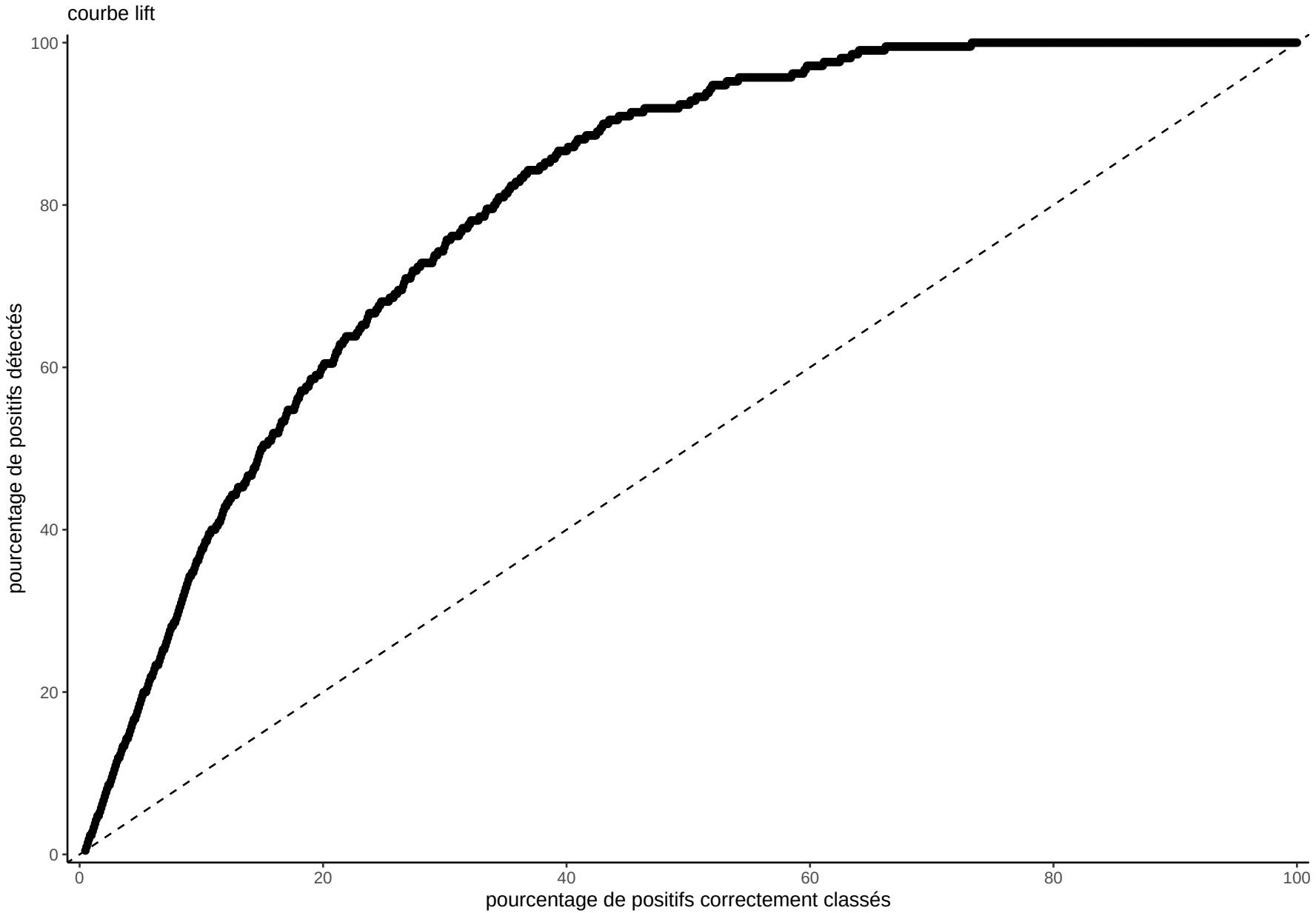

courbe lift
100
80
60
40
20
0
pourcentage de positifs détectés
0
20
40
60
80
100
pourcentage de positifs correctement classés

Tableau du lift

Table 5: Tableau du lift (déciles).

	hasard	modèle	lift
10%	21	81	3.86
20%	42	127	3.02
30%	63	159	2.52
40%	84	183	2.18
50%	105	195	1.86
60%	126	204	1.62
70%	147	209	1.42
80%	168	210	1.25
90%	189	210	1.11

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 la proportion détectée par hasard.

Calibration du modèle

Certains modèles sont trop confiants dans leurs prédictions (surajustement).

une statistique simple proposée par Spiegelhalter (1986) peut être utile à cette fin. Pour une variable Bernoulli $Y \in \{0, 1\}$, l'erreur quadratique moyenne s'écrit

$$\begin{aligned}\overline{B} = & \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - p_i)^2}_{\text{erreur quadratique moyenne}} \\ = & \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - p_i)(1-2p_i)}_{\text{manque de calibration}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}p_i(1-p_i)}_{\text{variabilité}}.\end{aligned}$$

Si notre modèle était parfaitement calibré, $\mathsf{E}_0(Y_i) = p_i$ et $\mathsf{Va}_0(Y_i) = p_i(1-p_i)$.

Test de Spiegelhalter

On peut calculer la moyenne et la variance attendue sous l'hypothèse nulle et construire une statistique de test (Spiegelhalter, 1986).

$$Z = \frac{\overline{B} - \mathsf{E}_0(\overline{B})}{\mathsf{sd}_0(\overline{B})}$$

Sous l'hypothèse nulle de calibration parfaite, $Z \sim \mathsf{No}(0, 1)$ en grand échantillon.

Calibration du modèle

```
hecmulti::calibration(
    prob = predprob,
    resp = classif)
```

```
Test de calibration de Spiegelhalter (1986)
Statistique de test: 0.28
valeur-p: 0.782
```

Il n'y a pas de preuve ici que le modèle est mal calibré.

Sélection de variables en régression logistique

Même principes que précédemment, mais les modèles de régression logistique sont plus coûteux à estimer.

On peut utiliser les critères d'information puisque le modèle est ajusté par maximum de vraisemblance.

Fonctions **R** pour la sélection de modèles

- `glmbb::glmbb` permet une recherche exhaustive de tous les sous-modèles à au plus une certaine distance (`cutoff`) du modèle avec le plus petit critère d'information (`criterion`).
- `step` permet de faire une recherche séquentielle avec un critère d'information.
- `glmulti::glmulti` permet une recherche exhaustive (`method = "h"`) ou par le biais d'un algorithme génétique (`method = "g"`).
- `glmnet::glmnet` permet d'ajuster le modèle avec pénalité LASSO.

Voir le code en ligne.

Objectif du ciblage marketing

Déterminer si le revenu prévu justifie l'envoi du catalogue

$$\mathsf{E}(\mathsf{ymontant}_i) = \mathsf{E}(\mathsf{ymontant}_i \mid \mathsf{yachat}_i = 1)\,\mathsf{Pr}(\mathsf{yachat}_i = 1).$$

On peut combiner un modèle de régression logistique avec la régression linéaire (ajustés simultanément avec un modèle Heckit).

Ou simplement ignorer le montant d'achat et envoyer un catalogue si la probabilité d'achat excède notre point de coupure optimal.

Stratégie de référence

- Parmi les 100K clients, 23 179 auraient acheté si on leur avait envoyé le catalogue
- Ces clients auraient généré des revenus de 1 601 212$.
- Si on enlève le coût des envois (100 000 $\times$ 10$), la stratégie de référence permet un revenu net de 601 212$.

Stratégie d'ajustement

En résumé, la procédure numérique à réaliser est la suivante:

- Choisir les variables à essayer (termes quadratiques, interactions, etc.)
- Choisir l'algorithme ou la méthode de sélection du modèle.
- Construire un catalogue de modèles: pour chacun, calculer les prédictions par validation croisée.
- Calculer le point de coupure optimal pour chaque modèle et le gain associé.
- Sélectionner le modèle qui **maximise le gain**.

Prédiction et envoi

- Prédire les 100 000 observations de l'échantillon test.
- Envoyer un catalogue si la probabilité d'achat excède le point de coupure.
- Calculer le revenu résultant:
  - zéro si on n'envoie pas de catalogue
  - -10 si la personne n'achète pas
  - -10 plus l'achat si la personne achète.

En pratique, on ne pourrait pas *a priori* connaître le revenu résultant de cette stratégie.

Conclusion

Si on avait fait une bête recherche séquentielle et qu'on avait pris le modèle avec le plus petit BIC (8 variables explicatives), on aurait dégagé des revenus de 978 226$.

C'est une énorme amélioration, de plus de 56%, par rapport à la stratégie de référence.

Récapitulatif

- La classification est une forme d'apprentissage supervisée.
- On peut assigner l'observation à la classe la plus plausible, ou déterminer un point de coupure.
- Si on a un objectif particulier (fonction de gain), on peut optimiser les profits en assignant une importance différente à chaque scénario.
- On peut catégoriser les observations dans une matrice de confusion.

Récapitulatif

- On s'intéresse à
  - la spécificité (proportion d'échecs correctement classifiés)
  - la sensibilité (proportion de succès correctement classifiés)
- L'aire sous la courbe de la fonction d'efficacité du récepteur (courbe ROC) et le lift donne une mesure de la qualité des prédictions.

Récapitulatif

- Les mêmes principes de sélection de variable s'appliquent (recherche exhaustive, séquentielle et LASSO).
- On peut calculer les critères d'information puisque le modèle est ajusté par maximum de vraisemblance.
- Attention au surajustement! Suspect si les probabilités estimées sont près de 0/1 (vérifier la calibration).
- Deux étapes: sélectionner le modèle (variables) et le point de coupure.
- D'autres modèles que la régression logistique (arbres de régression, etc.) sont envisageables pour la classification.