

Classification

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

La régression logistique spécifie un modèle pour la probabilité de succès

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

où $\eta = \beta_0 + \dots + \beta_p \mathbf{X}_p$.

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

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

pour chaque ligne de la base de données.

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.

Qu'arrive t'il si $c = 0$ ou $c = 1$? Exemple au tableau

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.

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

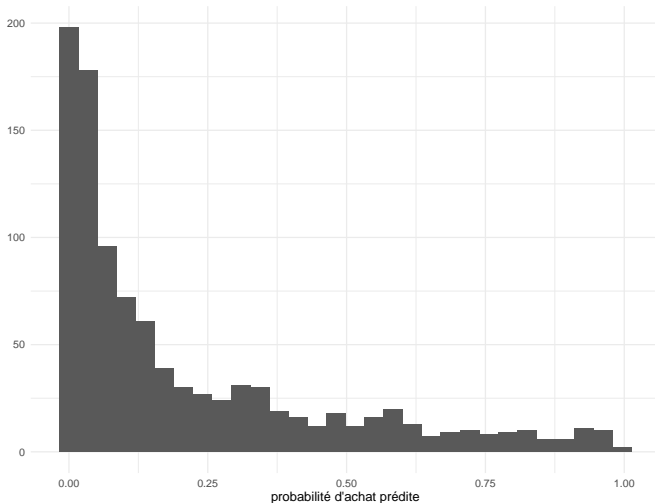
```
1 data(dbm, package = "hecmulti")
2 # Ne conserver que l'échantillon d'apprentissage
3 appr <- dbm[(dbm$test == 0),]
4 formule <- formula("yachat ~ x1 + x2 + x3 +
5                     x4 + x5 + x6 + x7 + x8 + x9 + x10")
6 modele <- glm(formule,
7               data = appr,
8               family = binomial)
```

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.

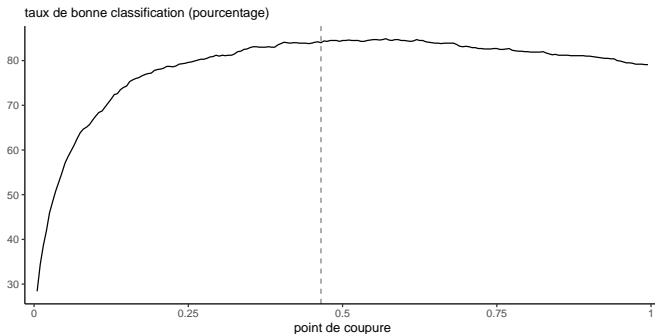
```
1 set.seed(202209)
2 # Option par défaut:
3 # validation croisée à 10 plis, 10 répétitions
4 predprob <- hecmulti::predvc(
5   modele = modele,
6   data = appr)
```



Répartition des probabilités de succès prédites par validation croisée.

Choix d'un point de coupure.

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



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

```
1 classif <- with(dbm, yachat[test == 0])  
2 # Tableau de la performance  
3 hecmulti::perfo_logistique(  
4   prob = predprob,  
5   resp = classif)
```

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.

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	89	30
$\hat{Y} = 0$	121	760

Classification et mesures de performance

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

	$Y = 1$	$Y = 0$		$Y = 1$	$Y = 0$
$\hat{Y} = 1$	89	30	$\hat{Y} = 1$	VP	FP
$\hat{Y} = 0$	121	76	$\hat{Y} = 0$	FN	VN

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

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

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

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

- 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).

Table 4: Statistiques descriptives des montants d'achats.

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

Table 5: 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

```
1 formule = formula(yachat ~ x1 + x2 + x3 +  
2               x4 + x5 + x6 + x7 +  
3               x8 + x9 + x10)  
4 modele <- glm(formule,  
5               family = binomial,  
6               data = hecmulti::dbm)  
7 coupe <- hecmulti::select_pcoupe(  
8   modele = modele,  
9   c00 = 0,  
10  c01 = 0,  
11  c10 = -10,  
12  c11 = 57,  
13  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.

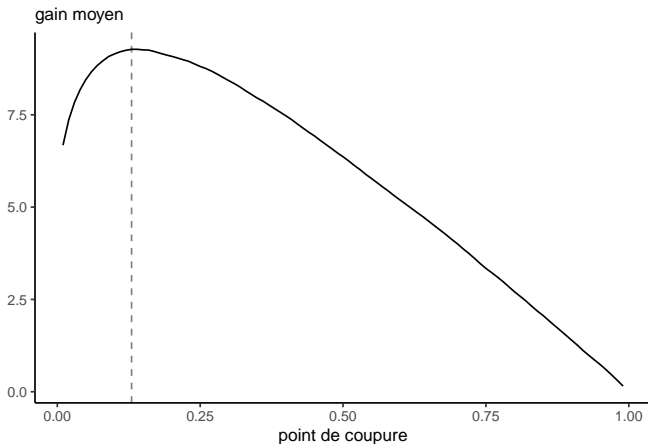


Figure 1: Estimation du gain moyen en fonction du point de coupe pour l'exemple de base de données marketing.

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 70.5
- la sensibilité est de 88.57.

Ainsi, on va détecter environ 88.57% des clients qui achètent.

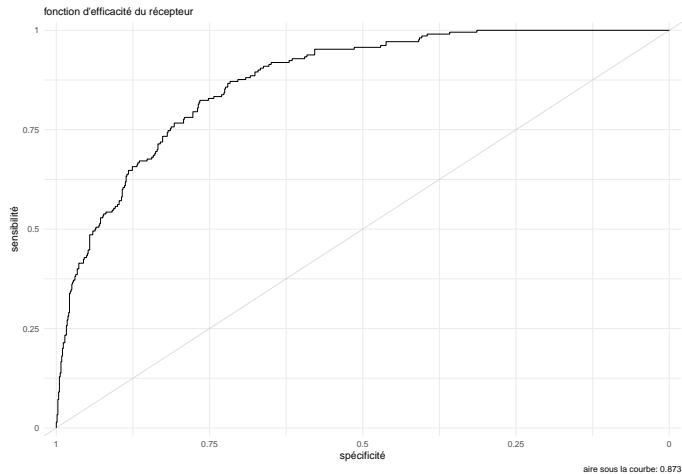
Il est coûteux de rater un client potentiel, donc la stratégie optimale est d'envoyer le catalogue à plus de clients quitte à ce que plusieurs d'entre eux n'achètent rien.

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.

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

Courbe ROC



- 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.

À quelle point notre modèle est-il meilleur 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

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


Table 6: Tableau du lift (déciles).

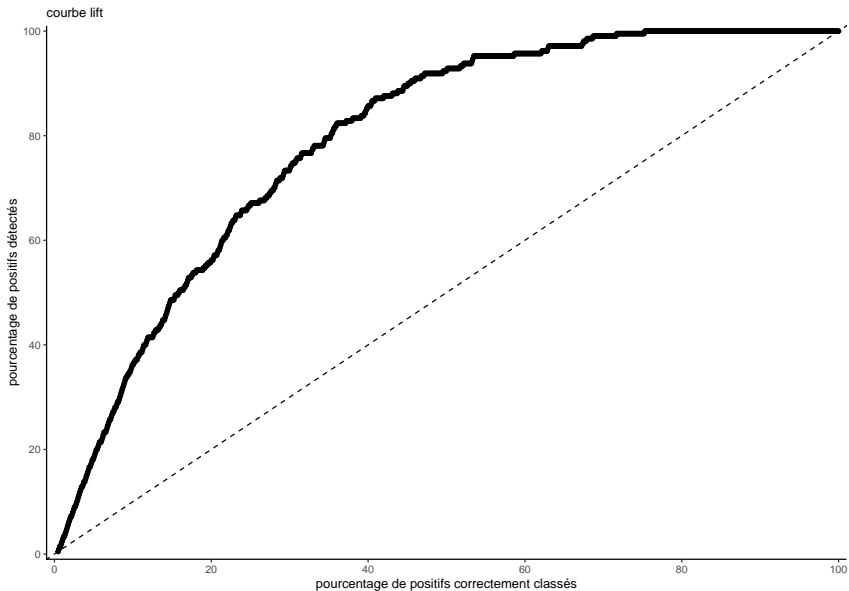
	hasard	modèle	lift
10%	21	78	3.71
20%	42	120	2.86
30%	63	157	2.49
40%	84	180	2.14
50%	105	195	1.86
60%	126	201	1.60
70%	147	208	1.41
80%	168	210	1.25
90%	189	210	1.11

Si on classifiait comme acheteurs les 10% qui ont la plus forte probabilité estimée d'achat, on détecterait 78 des 210 clients.

Si on prend 10% des clients au hasard et que 21% des observations correspondent à des achats, on détecterait en moyenne 21 clients par tranche de 100 personnes.

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

Courbe lift



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

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

Si notre modèle était parfaitement calibré, $E_0(Y_i) = p_i$ et $Va_0(Y_i) = p_i(1 - p_i)$.

Test de Spiegelhalter

On peut construire une statistique de test (Spiegelhalter, 1986) pour l'hypothèse nulle de calibration parfaite.

Sous l'hypothèse nulle \mathcal{H}_0 , le modèle est adéquat (correctement calibré).

Une petite valeur- p mène au rejet de \mathcal{H}_0 et à conclure que le modèle est surajusté.

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

Test de calibration de Spiegelhalter (1986)

Statistique de test: 1.26

valeur-p: 0.207

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

Il arrive que, lors de l'ajustement d'une régression logistique, on obtienne un message d'avertissement:

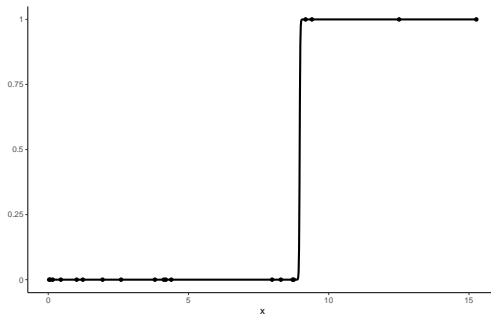
```
1 Warning messages:
2 1: glm.fit: algorithm did not converge
3 2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Le deuxième message d'erreur survient quand une combinaison linéaire de variables explicatives permet de prédire exactement la réponse: nos probabilités prédites sont 0 ou 1.

- par exemple, si on ajuste un modèle pour `yachat` en fonction de `ymontant`.

```
1 quasisep <- hecmulti::dbm |>
2   # remplacer les valeurs manquantes par des zéros
3   dplyr::mutate(ymontant = ifelse(yachat == 0, 0, ymontant))
4 # Régression logistique pour l'achat (0 ou 1) en fonction du montant
5 modele <- glm(yachat ~ ymontant,
6               family = binomial,
7               data = quasisep,
8               subset = test == 0)
```

Illustration de la séparation de variables



Quasi-séparation de variable: les estimations des paramètres sont presque infinies pour permettre une transition abrupte de la probabilité de succès de $\hat{p} = 0$ à $\hat{p} = 1$ à $x = 9$.

Les valeurs élevées des coefficients et erreurs-type sont une autre indication de quasi-séparation de variables.

Avec des variables standardisées, un coefficient $|\beta_j| > 10$ est suspect.

Table 7: Données simulées avec séparation de variable pour modèle logistique (gauche) et correction de Firth (droite).

	coef.	erreur-type		coef.	erreur-type
cst	-829.0	311290.6	cst	-5.6	2.6
x	92.5	34780.6	x	0.6	0.3

Typiquement, le problème survient parce que

- on a ajouté un dérivé de la variable réponse comme variable explicative
- on a un modèle surajusté, souvent une variable catégorielle pour laquelle toutes les observations d'un niveau donné ont la même réponse, soit 0 ou 1.

Ce n'est pas nécessairement un enjeu pour la prédiction, mais c'est souvent indicateur de problèmes plus importants. Les coefficients et interprétations ne sont plus valides...

1. Regarder quelles probabilités sont presque 0 ou 1 pour identifier les observations problématiques et s'assurer que l'échantillon que l'on emploie est adéquat.
 - Par exemple, les femmes ne peuvent pas avoir un cancer de la prostate, donc la probabilité prédite pour ces dernières est 0. On pourrait simplement enlever les femmes et prédire zéro manuellement.
2. Déterminer si une variable explicative cause du surajustement en étudiant les coefficients dont la valeur absolue est très élevée.
3. Si on a des problèmes de convergence numérique (1e message d'erreur), on peut aussi ajouter une pénalité, proposée par Firth (1993) pour empêcher les valeurs des coefficients d'être trop élevés (un peu comme le LASSO).
 - Dans R, via la fonction `logistf` du paquet éponyme.

- 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.

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

- On applique les mêmes principes que précédemment.
- Notre mesure d'ajustement (gain, taux de bonne classification, log-vraisemblance) peut différer selon l'objectif.
- Les modèles de régression logistique sont plus coûteux à estimer.
- Pour la classification, le point de coupure est à déterminer.

- `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.

Déterminer si le revenu prévu justifie l'envoi du catalogue

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

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.

- Parmi les 100K clients, 23179 auraient acheté si on leur avait envoyé le catalogue
- Ces clients auraient généré des revenus de 1601212\$.
- Si on enlève le coût des envois ($100000 \times 10\$$), la stratégie de référence permet un revenu net de 601212\$.

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 selon la fonction de gain moyen.
- Sélectionner le modèle qui maximise le gain.

- Prédire les 100000 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.

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 978226\$.

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

- Les principes de sélection de variable couverts précédemment s'appliquent toujours (recherche exhaustive, séquentielle et LASSO).
- On peut aussi 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 ou 1 (vérifier la calibration).
- Deux étapes: sélectionner le modèle (variables) et le point de coupure.
- D'autres modèles que la régression logistique (arbres de classification, etc.) sont envisageables pour la classification.