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 $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$, 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 une variable binaire 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
- 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

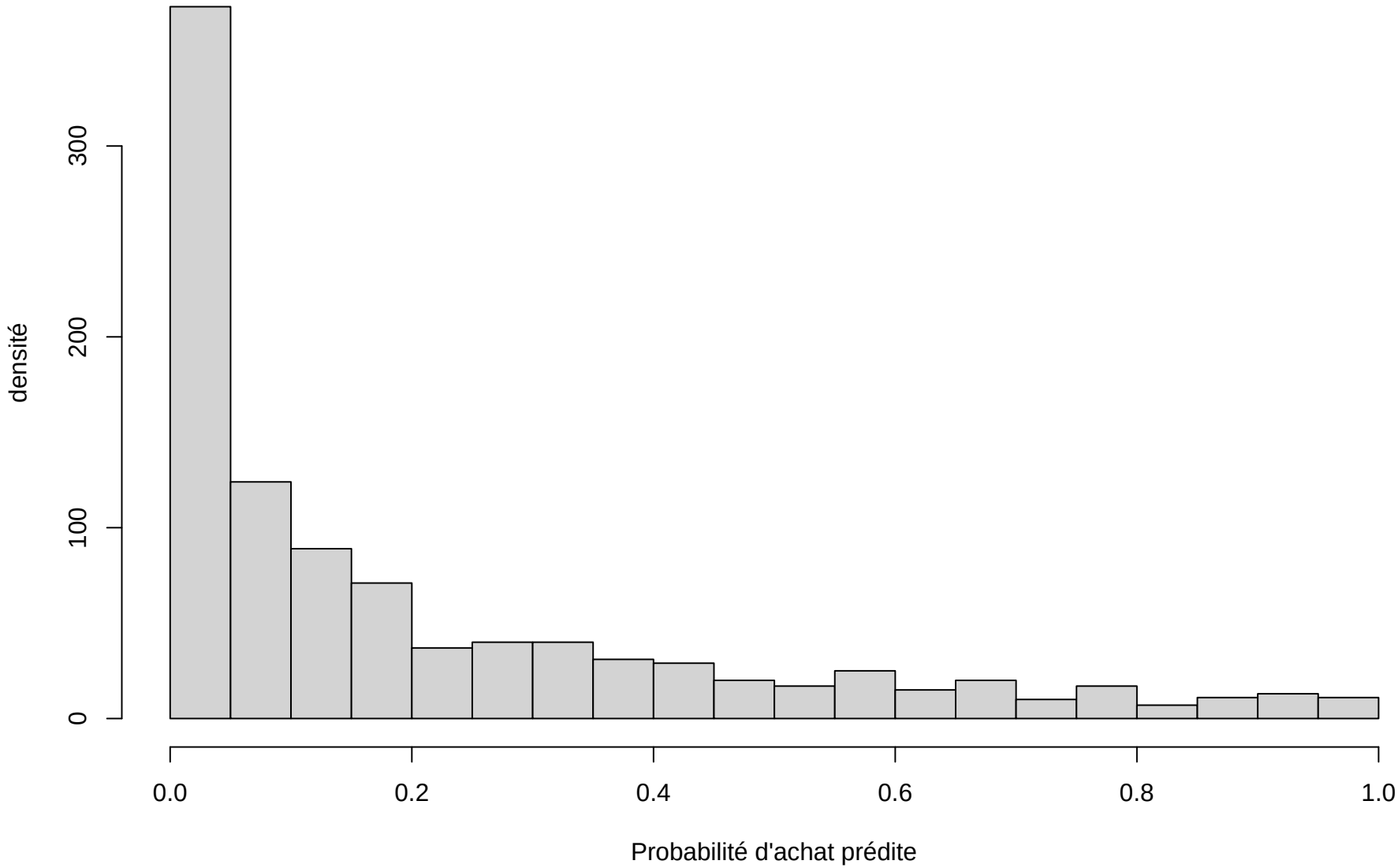

Figure 1: Répartition des probabilités de succès prédites par validation croisée.

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

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.

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