Régression logistique: prédictions et données multinomiales

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

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

- Les principes de sélection de variable couverts précédemment s'appliquent toujours (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.