Sélection de variables en régression logistique

- Plusieurs procédures empiriques sont proposées dans la littérature pour construire un modèle de régression logistique au départ d'une série de variables explicatives.
- Celle présentée dans les exemples du cours partent du modèle saturé et cherche à le simplifier en tentant de retirer certains termes du modèle en respectant la hiérarchie (: pas de retrait d'un effet principal lorsque la variable concernée apparait toujours en interaction avec une autre).
- La méthode présentée ici permet de faire face aux situations impliquant un grand nombre de variables explicatives $X_1, \ldots, X_p$.
- Dans ce cas, il n'est pas réaliste de partir du modèle saturé avec les interactions de tous les ordres. "Au pire", on pourrait imaginer démarrer avec une modèle présentant tous les effets principaux ainsi que toutes les interactions d'ordre 2,

$$(X_1 + \ldots + X_p) + \\ (X_1 * X_2 + \ldots + X_1 * X_p) + (X_2 * X_3 + \ldots + X_2 * X_p) + \ldots + (X_{p-1} * X_p)$$

  pour ensuite tenter de le simplifier comme au départ d'un modèle saturé.
- Alternativement, la procédure suivante, inspirée de Hosmer *et al.* (Section 4.2 dans *Applied Logistic Regression*, 3rd ed - Wiley 2013) pourrait être employée.

  1. Identifier isolément toutes les variables explicatives présentant un $P < 0.25$.
  2. Ajuster le modèle multivarié avec toutes les variables retenues en (1).
  3. Tenter de simplifier, variable par variable, le modèle en (2) (avec critère de $P < 0.05$ pour conserver une variable). Garder malgré tout une variable si son retrait influence fortement les estimations liées aux variables toujours présentes dans le modèle.
  4. Tenter de réintroduire une par une chacune des variables écartées à l'étape (1) au modèle issu de l'étape (3).
  5. Examiner dans quelle mesure certaines catégories d'une variable explicative retenue en fin d'étape (4) peuvent être regroupées.
  6. Déterminer quelles interactions d'ordre 2 pourraient faire sens dans le contexte de l'application et identifier (comme en (1) avec les effets principaux) celles qui sont significatives lorsqu'on les ajoute au modèle en (5) (mais cette fois avec un $P < 0.05$ comme critère d'inclusion).
  7. Ajouter toutes les interactions identifiées en (6) aux effets principaux retenus en (5) et chercher à éliminer des interactions du modèle (comme lors de l'étape (3) avec les effets principaux).

- Lorsque la réponse est ordinale, la même procédure peut être employée avec le modèle à cotes proportionnelles pour identifier les variables explicatives pertinentes (ainsi que leurs interactions).