

5.1 Les données `logistclient` contiennent des données simulées pour un cas fictif de promotion pour des clients. La base de données contient les variables suivantes :

- `promo` : variable binaire, 1 si le client s'est prévalu d'une offre promotionnelle, 0 sinon
- `sexe` : 0 pour les femmes, 1 pour les hommes
- `tclient` : variable catégorielle, soit `frequent` pour les clients réguliers ou `occasionnel` autrement
- `nachats` : nombre d'achats au magasin dans le dernier mois

Estimer le modèle logistique pour `promo=1` avec les variables explicatives `nachats`, `sexe` et `tclient` (référence `occasionnel`)

- (a) Interprétez les coefficients du modèle à l'échelle de la cote en terme de pourcentage d'augmentation ou de diminution.
 - (b) Testez si l'effet de `nachats` est statistiquement significatif à niveau $\alpha = 0.05$.
 - (c) Choisissez le point de coupure sur la base du taux de bonne classification. Pour ce faire, utilisez l'option `ctable`
 - i. Pour le point de coupure choisi, construisez une matrice de confusion
 - ii. Produisez un graphique de la fonction d'efficacité du récepteur (courbe ROC). Quelle est l'aire sous la courbe (estimée à l'aide de la validation croisée)?
- 5.2 Les données `sashelp.junkmail` de l'aide SAS contiennent 4601 corpus de courriels divisés en 59 variables colligées par Hewlett-Packard et classifiées selon que le message est un pourriel (`class=1`) ou pas (`class=0`).
- (a) Construisez un modèle de régression logistique pour classer les courriels en pourriels et messages selon leurs texte avec les 57 variables fournies telles quelles (sans transformation ni interaction).
 - Faites une partition avec la variable `test` (avec zéro pour entraînement et un pour validation).
 - Utilisez la procédure `hpgenselect` avec la méthode séquentielle et le critère BIC pour la sélection des variables `stepwise(choose=bic)`. Obtenez les probabilités estimées.
 - (b) Avec le modèle précédent et la macro `mlogisticclass` du script `logit10_macro_gainpred.sas`, comparez les observations et les prédictions pour l'échantillon de validation. Pour ce faire
 - Sélectionnez un point de coupure optimal en attribuant un poids de 1 en cas de bonne classification, de -1 pour un faux positif et de -2 en cas de classification de courriel valide en pourriel sur les données d'apprentissage.
 - Rapportez le taux de bonne classification, la sensibilité et la spécificité pour ce point de coupure.
 - (c) Commentez sur la difficulté à détecter un pourriel : est-ce que la tâche est facile?

5.3 On s'intéresse à la satisfaction de clients par rapport à un produit. Cette dernière est mesurée à l'aide d'une échelle de Likert, allant de très insatisfait (1) à très satisfait (5). Le jeu de données `multinom.sas7bdat` contient les variables suivantes :

- `y` : score de satisfaction
- `sexe` : sexe de l'individu; homme (0) ou femme (1)
- `educ` : niveau d'éducation le plus élevé complété; secondaire (`sec`), collégial (`cegep`) ou universitaire (`uni`)
- `revenu` : variable catégorielle indiquant le revenu, soit faible (1), moyen (2) ou élevé (3).
- `age` : âge de l'individu (en années).

Modélisez la satisfaction des clients, `y`, en fonction de l'âge, du niveau d'éducation, du sexe et du niveau de revenu.

- (a) Est-ce que le modèle de régression multinomiale ordinaire à cote proportionnelles est une simplification adéquate du modèle de régression multinomiale logistique? Si oui, utilisez ce modèle pour la suite. Si non, ajustez le modèle de régression multinomiale logistique avec 1 comme catégorie de référence pour `y`, 1 pour revenu et `sec` pour éducation et utilisez ce dernier pour répondre aux autres questions.

- (b) Interprétez l'effet des variables éducation et sexe pour la catégorie 2.
- (c) Est-ce que le modèle avec une probabilité constante pour chaque item est adéquat lorsque comparé au modèle qui inclut toutes les covariables?
- (d) Est-ce que l'effet de la variable âge est globalement significatif?
- (e) Fournissez un intervalle de confiance à niveau 95% pour l'effet multiplicatif d'une augmentation d'une unité de la variable âge pour chacune des cote par rapport à très insatisfait (1). Que concluez-vous sur l'effet de âge pour les réponses 2, ..., 5 par rapport à 1?
- (f) Écrivez l'équation de la cote ajustée pour satisfait (4) par rapport à très insatisfait (1).
- (g) Prédisez la probabilité qu'un homme de 30 ans qui a un diplôme collégial et qui fait partie de la classe moyenne sélectionne une catégorie donnée. Quelle modalité est la plus susceptible?