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 prévalue 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.

**Solution**

- La cote pour l'offre promotionnelle (oui versus non) des hommes est 6.4% plus faible que celle des femmes, *ceteris paribus*
- Le rapport de cotes pour `tclient` est 0.934 : les clients fréquents ont 6.6% inférieure à celle des clients occasionnels, toute chose étant égale par ailleurs.
- La cote de `nachats` augmente de 23.1% pour chaque augmentation du nombre d'achats dans le dernier mois, *ceteris paribus*

(b) Testez si l'effet de `nachats` est statistiquement significatif à niveau $\alpha = 0.05$.

**Solution**

L'intervalle de confiance à 95% pour le rapport de cote de `nachats`, basé sur la vraisemblance profilée, est de $[1.15, 1.32]$ ; comme 1 est exclu, cette différence est statistiquement significative. La statistique de Wald pour cette même variable est 35.0561, ce qui donne une valeur-$p$ négligeable (inférieure à $10^{-4}$).

(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) ?

**Solution**

Si on considère uniquement des points de coupures entre 0 et 1 par incréments de 0.2, on trouve un taux de coupure optimal à 0.44 avec un taux de bonne classification de 59.1%. Le tableau de classification 2 indique une sensibilité de 72.2% et une spécificité de 47%. L'aire sous la courbe estimée avec une approximation de la validation croisée à

| Prédiction/observation | $Y = 1$ | $Y = 0$ |
|---|---|---|
| $\widehat{Y} = 1$ | 346 | 133 |
| $\widehat{Y} = 0$ | 276 | 345 |

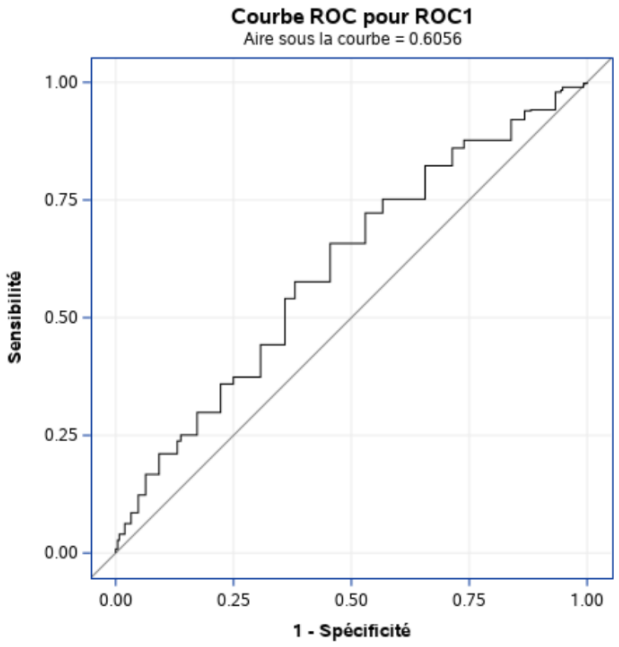
TABLEAU 1 – Matrice de confusion pour les données `tclients`

$n$ groupes est 0.6056, soit un peu mieux qu'une sélection aléatoire.

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 `stepwise(choose`
Obtenez les probabilités estimées.

**Solution**
Voir le code ; le modèle retenu contient 21 variables explicatives.

(b) Avec le modèle précédent, comparez les observations et les prédictions pour l'échantillon de validation. Sélectionnez un point de coupure optimal en attribuant un poids de 1 en cas de bonne classification, de $-1$ pour un faux positive 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.

**Solution**
Le point de coupure optimal sur la grille 0, 1 en incréments de 0.02 est 0.18 pour un gain de 0.326 ; le taux de bonne classifiacation est d'environ 75%, la spécificité de 65.6% et la sensibilité de 85.1%.

(c) Commentez sur la difficulté à détecter un pourriel : est-ce que la tâche est facile ?

**Solution**
La tâche est difficile parce qu'on a peine à bien classifier les courriels. L'échantillon d'apprentissage contenait 1847 courriels et 1218 pourriels.

| Prédiction/observation | $Y = 1$ | $Y = 0$ |
|---|---|---|
| $\widehat{Y} = 1$ | 1273 | 223 |
| $\widehat{Y} = 0$ | 540 | 1029 |

TABLEAU 2 – Matrice de confusion pour les données pourriel avec point de coupure à 0.18

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 ordinale 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.

**Solution**
On peut regarder directement la sortie du test du score pour l'hypothèse des cotes proportionnelles ; le modèle ordinal a 10 paramètres, contre 28 pour le modèle multinomial logistique. La valeur-$p$ du test du score est inférieure à $10^{-4}$. On peut également faire un test du rapport de vraisemblance en comparant la différence des log-vraisemblance des deux modèles emboîtés. La valeur-$p$ estimée est 0,00088. Le modèle à cote proportionnelle n'est pas adéquat.

(b) Interprétez l'effet des variables éducation et sexe pour la catégorie 2.

**Solution**
— La cote pour les femmes pour insatisfait par rapport à très insatisfait est 42,4% plus élevée que pour les hommes, toute chose étant égale par ailleurs.
— La cote pour les individus qui ont un diplôme collégial pour insatisfait par rapport à très insatisfait est 2,5% plus basse que pour ceux qui on un diplôme secondaire, toute chose étant égale par ailleurs.
— La cote pour les individus qui ont un diplôme universitaire pour insatisfait par rapport à très insatisfait est 16,2% plus élevée que pour ceux qui on un diplôme secondaire, toute chose étant égale par ailleurs.

(c) Est-ce que le modèle avec une probabilité constante pour chaque item est adéquat lorsque comparé au modèle qui inclut toutes les covariables ?

**Solution**
La statistique pour le test du rapport de vraisemblance que tous les coefficients associés aux covariables sont nuls (24 paramètres) est 55.41, et si le modèle sans covariable était vrai, cette statistique serait approximativement $\chi^2_{24}$. La valeur-$p$ est 0,0003 et on conclut qu'au moins une covariable est utile pour prédire une cote par rapport au modèle avec une probabilité constante.

(d) Est-ce que l'effet de la variable âge est globalement significatif ?

**Solution**
Puisqu'on modélise quatre rapport de cotes à l'aide d'un modèle logistique, on regarde ici le test de Wald (on pourrait manuellement ajuster un modèle sans âge pour faire le test du rapport de vraisemblance). La statistique de Wald vaut 21.69 et la probabilité d'obtenir un tel résultat si $\beta_{\texttt{age}_2} = \beta_{\texttt{age}_3} = \beta_{\texttt{age}_4} = \beta_{\texttt{age}_5} = 0$ est approximativement 0,0002. On conclut que l'âge impacte la probabilité des différents items de satisfaction.

(e) Fournissez un intervalle de confiance à niveau 95% pour l'effet 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 ?

**Solution**
Les intervalles de confiance sont [0,975;1,011] pour $\beta_{\texttt{age}_{2|1}}$ (pas significatif), [0,871;0,963] pour $\beta_{\texttt{age}_{3|1}}$ (significatif), [0,941;0,989] pour $\beta_{\texttt{age}_{4|1}}$ (significatif) et [0,988;1,021] pour $\beta_{\texttt{age}_{5|1}}$ (pas significatif)

(f) Écrivez l'équation de la cote ajustée pour satisfait (`4`) par rapport à très insatisfait (`1`).


Compilé le 12/11/2021 à 21:19

**Solution**

Pour obtenir l'équation ajustée, on utilise uniquement les coefficients pour `y=4` dans le tableau des coefficients.

$$\frac{\mathsf{P}(Y=4 \mid \boldsymbol{X})}{\mathsf{P}(Y=1 \mid \boldsymbol{X})} = \exp(-0{,}114 - 0{,}035\texttt{age} + 0{,}184\texttt{cegep} + 0{,}308\texttt{uni} - 0{,}028\texttt{revenu}_2 + 0{,}018\texttt{revenu}_3 + 0{,}613\texttt{sexe})$$

(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?

**Solution**

Les probabilités pour (1, 2, 3, 4, 5) sont (0,321 ; 0,222 ; 0,039 ; 0.127 ; 0.292). La modalité la plus susceptible est donc très insatisfait (1).



Compilé le 12/11/2021 à 21:19