ACT6100	Analyse de données
H2019	Série 4

Apprentissage supervisé - Introduction (suite)

1. On souhaite utiliser la régression logistique comme technique de *scoring* pour prédire si un emprunteur sera un bon ou un mauvais payeur (et donc, si on lui accordera un crédit ou non). Le fichier de données *credit.csv* disponible sur le site du cours contient les données recueillies auprès de 1 500 clients d'une institution financière.

   La variable **statut** est catégorielle et comporte deux modalités : 0 si l'emprunteur a fait défaut et 1 s'il a remboursé le prêt.

   Les variables explicatives quantitatives sont :
   — **Age** : l'âge de l'emprunteur (en années) ;
   — **Revenu** : le revenu de l'emprunteur (en euros) ;
   — **Tendett** : le taux d'endettement de l'emprunteur ;
   — **Nexp** : le nombre d'années d'expérience de l'emprunteur dans son travail actuel ; et
   — **Rabanque** : le nombre d'années depuis lesquelles l'emprunteur est un client de la banque.

   Les variables explicatives qualitatives sont :
   — **Prof** : la profession de l'emprunteur, 1 pour les professions libérales, 2 pour les chômeurs, 3 pour les employés du secteur privé et 4 pour les employés du secteur public.
   — **Genre** : le sexe de l'emprunteur, 0 pour les hommes et 1 pour les femmes.

   Répondre aux questions suivantes.

   (a) Pourquoi faut-il utiliser le modèle de régression logistique plutôt que le modèle de régression normale ?

   (b) Réaliser une régression logistique permettant d'expliquer la variable **statut**. Quel est le modèle complet ? Utiliser une méthode de sélection *Backward* avec critère BIC pour simplifier le modèle. Quel est le modèle final ?

   (c) Est-il possible de retirer **Prof_public** du modèle final ?

   (d) Est-il possible de retirer le β_0 du modèle final ?

   (e) En utilisant le modèle final déterminé à la sous-question (b) et la 10-validation croisée, déterminer le point de rupture τ du modèle qui offrira le meilleur compromis entre sensibilité et spécificité.

   (f) Un client de 35 ans fait une demande de crédit. Il a un revenu de 2 195 euros, un taux d'endettement de 25 %, 10 années d'expérience dans son emploi actuel comme employée du secteur public et est client de la banque depuis une dizaine d'années. En se basant sur le modèle final, quelle est la probabilité que le crédit lui soit accordé ?

2. La base de données *carsClaims.csv* disponible sur le site du cours est basée sur 67 856 contrats d'assurance automobile annuels souscrits en 2004 et 2005. Les variables sont
   — **clm** : la survenance (1) ou la non survenance (0) d'une réclamation ;
   — **veh_value** : la valeur du véhicule (en 10 000 euros) ;
   — **veh_age** : l'âge du véhicule (variable catégorielle : plus jeune 1, 2, 3 ou 4) ;
   — **gender** : le sexe du conducteur (M ou F) ;
   — **area** : le lieu de résidence du conducteur (A à F) ; et
   — **agecat** : la catégorie d'âge du conducteur (variable catégorielle : plus jeune 1 à 6).

   Les autres variables dans la base de données ne seront pas utilisées pour cette analyse.

   (a) Réaliser une régression logistique. Quel est le modèle complet ?

(b) Utiliser une méthode de sélection *Backward* avec critère BIC pour déterminer le modèle final.

(c) On suppose qu'en moyenne, un sinistre coute 2 000 euros à l'assureur. La prime demandée à un client est calculée selon la formule :

$$\Pi = (\text{cout moyen})(\text{prob. d'avoir une réclam.})(1+\theta),$$

où θ est une surcharge que demande l'assureur pour couvrir les frais fixes. Quelle prime sera demandée à une cliente de 60 ans (catégorie 6) dont le véhicule vaut 22 000 euros ? On suppose que l'assureur demande une surcharge de 25 %.

(d) En utilisant le modèle final déterminé à la sous-question (b) et la 16-validation croisée, déterminer le point de rupture τ du modèle qui offrira le meilleur compromis entre sensibilité et spécificité.

3. Le modèle de régression logistique fait intervenir les équations MAT8594

$$p(k) = \frac{e^k}{1+e^k} \tag{1}$$

et

$$\text{logit}(k) = \ln\left(\frac{k}{1-k}\right). \tag{2}$$

Vérifier l'exactitude des relations suivantes :

- $p(-k) = 1 - p(k)$;
- $\dfrac{dp(k)}{dk} = p(k)(1-p(k))$;
- $\dfrac{dp(SCORE)}{dx_j} = p(SCORE)(1-p(SCORE))\beta_j$, pour $j = 1, \ldots, q$ et x_j une variable continue ;
- $\text{logit}(1-k) = -\text{logit}(k)$;
- $\dfrac{d\text{logit}(k)}{dk} = \dfrac{1}{k(1-k)}$.

4. Pour la variable dépendante Y, on possède un échantillon aléatoire de taille n, $y_1, \ldots, y_n$. Pour chacun des éléments de l'échantillon (y_i), on possède un vecteur de variables explicatives $x_i = (x_{i1}, \ldots, x_{iq})$. On définit le modèle de régression logistique à l'aide de

$$p(x_i) = \Pr(Y_i = 1 | x_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}}}.$$

La fonction de vraisemblance est donnée par

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}.$$

Répondre aux questions suivantes permettant d'obtenir les estimateurs $\hat{\beta}_0, \ldots, \hat{\beta}_q$ à l'aide de la technique du maximum de vraisemblance.

(a) La fonction de log-vraisemblance est obtenue en prenant le logarithme de la fonction de vraisemblance. Déterminer la fonction de log-vraisemblance de ce modèle logistique.

(b) Déterminer la dérivée de la fonction de log-vraisemblance par rapport au paramètre β_j, $j = 0, \ldots, q$.

(c) En posant les $q+1$ dérivées égales à 0 et en résolvant pour $\hat{\beta}_0, \ldots, \hat{\beta}_q$, on trouve les estimateurs du maximum de vraisemblance. Cependant, le système n'est pas linéaire et les solutions doivent être trouvées numériquement pour $q > 0$. Trouver analytiquement la solution pour $q = 0$.