5.1 On veut estimer la probabilité qu'une personne possède la carte de crédit VISA Première à l'aide d'une régression logistique. Pour ce faire, on utilisera la base de données `visalogist` qui contient, outres des variables explicatives, la variable réponse `carvp`, le matricule du client `matric` ainsi qu'une partition aléatoire 50%/20%/30% des données en échantillons d'apprentissage, de validation et de test (`_ROLE_`). Pour la signification des différentes variables, voir les libellés SAS (procédure `format`) ou la description dans l'énoncé du Devoir 1.

(a) Faites une sélection de modèles à l'aide d'une procédure séquentielle avec le critère AIC (`select` et `stop`) sur les données d'entraînement et sélectionnez le modèle ~~à partir de l'échantillon de validation~~ avec le critère BIC (`choose`). N'incluez que les effets principaux (pas de fonctions des variables explicatives, ni de terme d'interaction). *Les données de validation ne sont employées que si on sélectionne `choose=validate` dans `hplogistic` et `hpgenselect`.*

(b) Obtenez les prédictions des probabilités de possession de la carte de crédit avec ce modèle (c'est-à-dire, avec les coefficients estimés avec les données d'entraînement) pour les données de l'échantillon test. Calculez le taux de bonne classification, la sensibilité et la spécificité avec un point de coupure de 0.6. Commentez sur la performance du modèle.

Réajustez le modèle avec les variables explicatives sélectionnées, mais cette fois avec uniquement l'échantillon test (puisque la sélection de variable invalide l'inférence statistique).

(c) Interprétez le paramètre pour la variable explicative `relat` à l'échelle de la cote (en terme de pourcentage d'augmentation ou de diminution).

(d) Fournissez un intervalle de confiance à 95% basé sur la vraisemblance profilée pour `sexe` et rapportez l'estimé ponctuel correspondant. Est-ce que la différence entre hommes et femmes est statistiquement significative une fois les autres variables prises en compte?

(e) Produisez une courbe d'efficacité du récepteur (ROC) pour les données de test à l'aide des probabilités estimées à partir des coefficients estimés (a) avec les données d'entraînement et (b) avec les données test. Rapportez l'estimé de l'aire sous la courbe et commentez sur les différences entre les deux estimés.

(f) Produisez une courbe *lift* et indiquez la valeur du *lift* si 30% des observations avec la probabilité estimée de succès la plus élevée étaient assignées à 1.

Indications :

— pour la sélection de modèle, si vous utilisez la procédure `hplogistic` ou `hpgenselect`, l'utilisation de la partition préalable requiert la ligne

```
partition rolevar=_ROLE_(train="TRAIN", validate="VALIDATE", test="TEST");
```

— utilisez l'option `output` pour obtenir les prédictions des probabilités de l'échantillon test $\widehat{p}$ avec les procédures haute performance, mais avec les coefficients du modèle estimé sur les données d'apprentissage.

— pour le modèle de l'échantillon test, sélectionnez uniquement les données avec `_ROLE_="TEST"`. Écrivez manuellement les variable explicatives du modèle sélectionné avec la procédure `logistic`.

— Avec `hpgenselect`, l'option `select=aic` n'est pas disponible, mais vous devriez obtenir le même modèle final qu'avec `hplogistic`.

Compilé le 30/03/2021 à 21:40