Ecole Polytechnique Universitaire de Lille
Département de Génie Statistique et Informatique
Université des Sciences et Technologies de Lille

TP2 Régression linéaire (GIS3)

Preda/Loingeville

Note : Il est important de lire le cours avant le TP

Simulation. Expliquer le résultat d'un test cognitif

L'objectif de ce TP est **de prédire le résultat** *d'un test cognitif* à partir des paramètres de débitmétrie. Pour évaluer les patients qui souffrent de certaines maladies neurologiques (Alzheimer, démences,...) on utilise des *tests psychologiques* et **cognitifs**. La débitmétrie cérébrale est un examen de médecine nucléaire (d'imagerie) qui permet de détecter *des anomalies* dans *la vascularisation des différentes zones du cerveau.*

Dans cette étude, on voudrait savoir si **le résultat d'un test cognitif** est **lié** à un **ensemble des paramètres** caractérisant les résultats de la débitmétrie.

Les résultats se trouvent dans le fichier **debitmetrie.xls** sur le site http ://math.univ-lille1.fr/∼preda/GIS3. Ce fichier contient des observations pour les 69 patients qui ont participé à cette étude. Pour chaque patient on dispose :

- du résultat de test cognitif : `RESUL` - **variable numérique** dont les valeurs sont comprises entre 0 et 4.
- des valeurs de 8 paramètres de débitmétrie - `fig, fid, fpg, fpd, tpg, tpd, carg, card` - des **variables numériques** exprimées en pourcentage.

A faire en R : Les résultats numériques et vos interprétations seront écrits dans un document OpenOffice.

1. Présenter des statistiques univariées pour chaque variable.
2. Quelles sont les variables de débitmétrie les plus corrélées avec `RESUL` ? présenter des statistiques bivariées (corrélations linéaires) entre la variable RESUL et les variables de débitmétrie. Fonction `cor`. Illustration graphique (nuages de points).
3. Réaliser le modèle de régression linéaire simple entre `RESUL` et `card`.
   - Inspecter les propriétés dans l'objet obtenu grace à la fonction `lm`. Fonctions utiles : `residuals`, tt coef
   - Quelle est la qualité du modèle ? Fonction `summary(`*`votre modele lineaire`*`)`.
   - Est-le modèle linéaire valide ? Représentation graphique avec la fonction `plot(`*`votre modele lineaire`*`)`. Vérifier la normalité des résidus (`shapiro.test`) + `qqnorm` et `qqline` , l'homoscédasticité (`bptest`, l'indépendance (dwtest). Package *lmtest* à charger.
   - étudier l'influence des observations sur l'estimation du modèle : la fonction `lm.influence(`*`votre modele lineaire`*`)` et la fonction `influence.measures(`*`votre modele lineaire`*`)`. Tracer le graphe des elements h_{ii} pour identifier si effet levier existe (à prendre comme limites 2/n ou 3/n).
   - Validation croisée : Calculer le PRESS (voir cours).
4. Réaliser le modèle de régression linéaire multiple pour expliquer `RESUL` à l'aide de toutes les variables de débitmétrie.
   - Avant d'estimer le modèle vérifier que la matrice X^TX est inversible. X est la matrice dite "de design" définie en cours. Fonction `det` mais aussi `solve`. Vérifier que les estimations $\hat{\beta}_i$ sont bien données par la relation $\hat{\beta} = (X^TX)^{-1}X^Ty$.
   - Qualité (R^2) et inférence sur les paramètres du modèle. Quelles sont les variables qui expliquent significativement RESUL ? Interprétation des ces coefficients.
   - Comme en 3., valider le modèle multiple. Remarque : les limites pour les h_{ii} dans le cas multiples sont soit 2p/n ou 3p/n, où p est le nombre de variables explicatives (ici p=8).

A faire en SAS : Sur le meme jeu de données, mais en SAS, réaliser les points 1-4. Vous pouvez utiliser la méthodologie (example de code et d'analyse) proposée à :

http ://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter1/sasreg1.htm