Modèles linéaires généralisés
Master 2 Pharmacologie 2016-2017

Date de retour : 21 déc. 2016

Vous devez choisir l'un des sujets proposés ci-dessous.

Vous rédigerez votre rapport sous la forme d'un article (vous pouvez cependant vous dispenser d'une introduction et de bibliographie du sujet) mais il vous faut par contre détailler votre démarche pour obtenir le meilleur modèle et votre analyse de la bonne adéquation de celui-ci.
Des analyses univariées pourront être résumées dans un tableau car elle apportent souvent des informations utiles. Dans la section « méthodes », vous donnerez les modèles et tests utilisés. Vous penserez à vérifier les hypothèses, à étudier les interactions éventuelles. Pour les variables quantitatives, la catégorisation est parfois pertinente. La qualité du modèle obtenue sera justifiée. Dans la section « résultats », vous approfondirez vos investigations et interpréterez vos résultats. Les programmes R devront figurer en annexe.

Sujet no 1 - Poids de naissance

Dans ce projet, on s'intéresse aux facteurs de risque d'un poids de nourrisson faible à la naissance. On dispose de données acquises en 1986 au centre médical Baystate. Pour chaque individu, les données disponibles sont les suivantes

- `low` 1 si le poids est inférieur à 2.5kg, 0 sinon
- `age` age de la mère
- `lwt` poids de la mère (en pounds) au moment des dernières règles
- `race` race de la mère (1 = blanche, 2 = noire, 3 = autre)
- `smoke` status de fumeuse lors de la grossesse
- `pt1` nombre de fausses couches
- `ht` la mère a fait de l'hypertension
- `ui` douleurs intra-utérines
- `ftv` nombre de visites chez le médecin pendant le 1er trimestre

On dispose aussi du poids de naissance, mais on n'utilisera pas cette variable ici.

Les données seront chargées sous R avec les commandes

```
library(MASS)
data(birthwt)
```

Sujet no 2 - Maladie coronarienne

Dans ce projet, on s'intéresse aux facteurs de risque d'apparition de maladie coronarienne. Les données proviennent de 4 centres

1. Cleveland Clinic Foundation
2. Hungarian Institute of Cardiology, Budapest
3. V.A. Medical Center, Long Beach, CA
4. University Hospital, Zurich, Switzerland

et elles portent sur 303 individus.

Deux études utilisant ces données :
- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology*, *64*,304–310.
- John Gennari – Gennari, J. H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence, 40*, 11–61.

La variable à prédire est codée en 5 modalités représentant l'absence de maladie (0) ou le niveau de gravité (1 à 4). Pour ce projet, on peut se contenter de prédire l'absence contre la présence d'une maladie cardiaque.
La table contient ples de 70 variables (voir le fichier pour la liste complète). Les plus importantes sont

— `age` age
— `sex` sexe (1 = male ; 0 = female)
— `cp` douleur thoracique (1 : typical angina, 2 : atypical angina, 3 : non-anginal pain, 4 : asymptomatic)
— `trestbps` pression sanguine au repos (au moment de l'hospitalisation)
— `chol` cholestérol sérique
— `fbs` glycémie à jeun>120 mg/dl
— `restecg` résultats d'électrocadiographie au repos (0 : normal, 1 : anormale, 2 : hypertrophie probable ou certaine du ventricule gauche)
— `thalach` fréquence cardiaque maximum
— `exang` *angina* induite par l'effort
— `oldpeak)` *ST depression* induite par l'effort
— `slope` pente du pic ST en exercice
— `ca` nombre de vaisseaux majeurs colorés en fluoroscopie
— `thal` 3 = normal ; 6 = défaut irrémédiable ; 7 = défaut réversible
— `num` diagnostique de maladie cardiaque (variable à prédire)

D'autres variables sont incluses dans les données. Elles sont décrites dans le fichier `heart-disease.names`.

Sujet no 3 - Nombre de visites chez le médecin

Deb et Trivedi (2009) étudient des données intéressantes sur le nombre de visites chez le médecin pour des adultes agés de 25 à 64 ans basées sur une étude démographique menée de 1997 à 2002. On propose de modéliser le nombre de visites.

Deb, P., & Trivedi, P. K. (2002). The structure of demand for health care : latent class versus two-part models. Journal of health economics, 21(4), 601-625.

- `educ` niveau d'éducation
- `docvis` nombre de visites chez le médecin
- `age` age/10
- `income` salaire/1000
- `female` sexe
- `black` 1 si de race noire
- `hispanic` 1 si hispanic
- `married` marié
- `noreast`, `midwest`, `south` région (north est la modalité de référence)
- `msa` =1 si le patient vit dans une zone urbaine
- `firmsize` taille/10 de l'entreprise qui emploie le chef de famille
- `famsize` taille de la famille
- `injury` a subi un accident
- `vegood, good, fairpoor` état de santé (poor est la modalité de référence)
- `physlim` 1 si la personne a une limitation physique
- `private` type d'assurance maladie
- `chronic` 1 si la personne a une maladie chronique

Les données sont disponibles en utilisant les commandes suivantes dans le fichier `docvis.Rdata`.

Sujet no 4 - Effets secondaires dans le traitement du cancer de la prostate

Important : les étudiants qui choisissent ce sujet devront s'engager à ne pas diffuser les données ni publier le travail réalisé.

Dans cette étude, on s'intéresse à la prédiction du risque d'effet secondaires dans le traitement du cancer de la prostate par radiothérapie.

Contexte - Lors du traitement par radiothérapie, la tumeur à irradier est localisée par un scanner et la dose à administrer est déterminée aussi précisément que possible. Néanmoins, les organes voisins sont souvent partiellement irradiés (en particulier le rectum et la vessie dans le cancer de la prostate). Cette irradiation peut engendrer des effets secondaires, par exemple des saignements.

Les données ont été acquises en suivant les patients jusqu'à 4 ans après le traitement. On a ici uniquement les séquelles au niveau du rectum. L'étude de la vessie est beaucoup plus difficile car c'est un organe qui se déforme.

On dispose de données multicentriques. Les variables cliniques mesurées sont les suivantes

- `age`
- `antticoagulant_i` 1 si le patient reçoit des traitements anticoagulants
- `abdomen_surg` 1 si le patient a subi une chirurgie de l'abdomen
- `diabete` 1 si le patient est diabétique
- `hormono` 1 si le patient est sous traitement hormonal
- `psa` taux de psa à baseline (biomarqueur pour le cancer de la prostate)
- `T` incateur du stade de la maladie
  (Beckendorf V, Guerif S, Le Prisé E, et al. 70 Gy versus 80 Gy in localized prostate cancer : 5-year results of GETUG 06 randomized trial. Int. J. Radiat. Oncol. Biol. Phys. 2011 ;80 :1056-1063.)
- `gleason` : score basé sur les données histologiques
  voir `http://www.cancer.ca/fr-ca/cancer-information/cancer-type/prostate/grading/gleason-classification/?region=qc`
- `X3techniques` correspond à la technique d'image utilisée pour la planification du traitement : 3D-CRT, IMRT, IGRT
  3D-RCt : technique la plus ancienne
  IMRT : technique la plus utilisée aujourd'hui
  IGRT : c'est de l'IMRt mais avant chaque séance on "recale le patient' via un scanner
- `dose_totale` dose totale planifiée pour la radiothérapie
- `Dmean` dose moyenne reçue par le rectum
- `Dmax` dose maximum reçue par le rectum
- `tox` 1 si le patient a subi des effets secondaires

Les données vous seront transmises par mail si vous voulez travailler sur ce projet. Le mail doit préciser que vous vous engagez à ne pas diffuser les données et à ne pas publier les résultats obtenus sans un accord préalable de notre part.

Lagrange, J. L., & De Crevoisier, R. (2010). La radiothérapie guidée par l'image (IGRT). Bulletin du Cancer, 97(7), 857-865.
Acosta, O., Drean, G., Ospina, J. D., Simon, A., Haigron, P., Lafond, C., & De Crevoisier, R. (2013). Voxel-based population analysis for correlating local dose and rectal toxicity in prostate cancer radiotherapy. Physics in medicine and biology, 58(8), 2581.

Sujet no 5 - Liquides articulaires, arthrite

Important : les étudiants qui choisissent ce sujet devront signer un engagement à ne pas diffuser les données ni publier le travail réalisé.

Dans cette étude, l'objectif est de construire un modèle d'aide au diagnostique de l'arthrite contre les autres pathologies articulaires. La motivation est la suivante. Le diagnostique d'arthrite est obtenu par ponction de liquide articulaire après une mise en culture qui dure environ 48h. En attendant le résultat de l'analyse, le patient est hospitalisé sous antibiotique à spectre large. L'objectif de cette étude est de développer des outils de

diagnostique rapide. L'analyse des données cliniques sera complétée par une analyse de données de spectrométrie dans le moyen infra rouge. L'analyse des spectres est plus difficile que celle des données cliniques et sort du cadre du projet.

On dispose de données multicentriques. Les variables cliniques mesurées sont décrites dans un cahier d'observation.

Les données vous seront transmises par mail si vous voulez travailler sur ce projet. Le mail doit préciser que vous vous engagez à ne pas diffuser les données et à ne pas publier les résultats obtenus sans un accord préalable de notre part.

Albert, J. D., Monbet, V., Jolivet-Gougeon, A., Fatih, N., Le Corvec, M., Seck, M., ... & Guggenbuhl, P. (2016). Une nouvelle méthode pour le diagnostic rapide d'arthrite septique utilisant la spectroscopie infrarouge. Revue du Rhumatisme, 83(4), 295-300.