Analyse Statistique Multidimensionnelle - TP
ASM-TP1 2016

Un compte-rendu est à rendre à la fin de la deuxième séance de ce TP. Suivez les règles de rédaction décrites dans https://ensiwiki.ensimag.fr/index.php/Rédaction_de_documents_écrits
En particulier, la représentation graphique est très importante en statistique. Chaque graphe doit être accompagné **au minimum d'un titre et de noms d'axes**. Il doit toujours être expliqué par quelques phrases dans le texte (**Qu'a-t-on tracé ? Qu'en déduire ?**).
Répondez aux questions dans l'ordre en vous référant au numéro des questions.
Les calculs, les graphiques et la programmation se feront avec le logiciel **R**.
Il est interdit de mettre du code R ou du copier-coller de sorties R dans le compte-rendu, mettez-le si nécessaire en annexe avec un maximum de commentaires. Pour présenter des résultats, rédigez un paragraphe en utilisant des phrases complètes et des équations. Vous enverrez le code R sur TEIDE à l'encadrant TP pour d'éventuelles vérifications.
Les questions en italiques sont celles que vous devriez avoir le réflexe de vous poser par vous-même sans qu'elles soient dans l'énoncé. À partir du TP 2 ces questions ne seront plus posées mais vous devrez y répondre tout de même !

ASM-TP1 : Cancer de la prostate.

Dans une étude de 1989, Stamey et ses collègues ont examiné la corrélation entre le taux dans le sang d'antigènes prostatiques spécifiques (PSA) et un certain nombre de mesures cliniques (âge, poids de la prostate, etc.). Le PSA est une protéine normalement sécrétée par les cellules prostatiques, mais une cellule cancéreuse en sécrète 10 fois plus qu'une cellule normale. Le taux sanguin de PSA est donc une variable de grand intérêt pour la détection du cancer de la prostate. Il peut toutefois être augmenté par d'autres facteurs (volume prostatique, inflammations, etc.). L'objectif est de poursuivre l'étude des relations entre le PSA et les 8 mesures cliniques issues des données de Stamey et al. (1989) et de construire un modèle prédictif du logarithme de PSA.

Les données contiennent le taux dans le sang d'antigènes prostatiques spécifiques (PSA) et 8 mesures cliniques chez 97 hommes qui vont subir une prostatectomie. L'objectif est de prédire le logarithme de PSA (lpsa) à partir des mesures suivantes (que l'on appellera prédicteurs) :

lcavol : log cancer volume
lweight : log prostate weight
age : age
lbph : log of benign prostatic hyperplasia amount
svi : seminal vesicle invasion
lcp : log of capsular penetration
gleason : Gleason score
pgg45 : percent of Gleason scores 4 or 5

Comme vous le verrez la corrélation entre certaines mesures cliniques et le lpsa est évidente, mais il est plus délicat de construire un bon modèle prédictif.

1. **Analyse préliminaires des données.**

   Récupérez les données sur Chamilo

   `chamilo2.grenet.fr/inp/courses/ENSIMAG4MMFDASM/document/prostate.data`

   Grâce au code qui suit, construisez un objet `pro` de classe `data.frame` contenant pour chaque individu le niveau **lpsa** et les valeurs des 8 **prédicteurs**. On choisit ici de travailler avec des variables centrées et réduites. On utilisera pour cela la fonction R `scale`.

```
prostateCancer = read.table("dir_path/prostate.data",header=T)
pro1 = prostateCancer[,-ncol(prostateCancer)]  # retire la dernière colonne du jeu

pro = as.data.frame(cbind(scale(pro1[,1:8]),pro1[,9]))
# centre et réduit les colonnes correspondant aux 8 prédicteurs
# crée un objet de type data.frame avec les 8 colonnes de prédicteurs et la colonne «output» lpsa

names(pro) = names(pro1)    # conserve les noms des colonnes
```

a) Visualisez les données de manière pertinente sur un graphique rendant compte des corrélations éventuelles entre toutes les variables, **lpsa** et mesures cliniques. Aide : `?pairs`
Joignez le graphe au compte-rendu. Numérotez-le en écrivant sous le graphe : « Figure 1 » et ajoutez un titre sur la même ligne. Citez le graphe dans le compte-rendu en écrivant « d'après la Figure 1, ... ».
Dans la légende au-dessus de « Figure 1 », explicitez la signification des graphes tracés (axes des abscisses, des ordonnées).

À partir de ce graphe, quels facteurs semblent en relation avec **lpsa** ?
Justifiez votre réponse.

b) Que dire graphiquement de la corrélation entre les **prédicteurs** eux-mêmes ? *Justifiez votre réponse.*

2. **Régression linéaire. Méthode des moindres carrés.**

   a) Faire une régression « classique » pour construire un modèle prédictif pour la variable **lpsa**. Traiter **gleason** comme une variable qualitative (utiliser `factor(gleason)`).
   Donner l'équation définissant le modèle et ses paramètres ($Y_i = \ldots + \varepsilon_i$).
   Donner la signification de la 1ère colonne de toutes le lignes nommées `gleason.`
   Résumer les résultats principaux.

   b) Au vu de ces résultats et de la partie 1., que penser du coefficient de régression de la variable **lcp** ? En donner un intervalle de confiance au seuil 1%. Que se passe-t-il si on ignore **lcavol** et **svi** ?

   c) Quelle est la loi suivie par la variable t (T statistic) ? Précisez les paramètres.
   Pour quelle valeur de t passe-t-on du seuil ** au seuil *** ie pour quelle valeur x a-t-on $P(|t|>x)=0.001$?
   Aide : `?qt`
   Vérifiez que votre valeur est cohérente avec les *** indiqués pour **lweight** et l'intercept.

   d) Observez graphiquement si les valeurs de **lpsa** prédites sont proches des valeurs réelles.
   Joindre le(s) graphe(s) au compte-rendu et les commenter.
   Aide : `?lm` La rubrique `Value` indique quel type d'objet est renvoyé par la fonction lm et la liste des attributs que vous pouvez récupérer. Ex : si `obj_lm=lm(...)` alors `obj_lm$residuals` renvoie le vecteur des résidus de la régression.
   Quelle est la valeur du RSS (Residual Sum of Squares) ?

3. **Effet des prédicteurs qualitatifs.**
   a) Y-a-t-il un effet sur **lpsa** des prédicteurs **svi** et **gleason** ? En quel sens ?
   Présenter et commenter les résultats de l'ANOVA.

4. **Sélection du meilleur sous-ensemble. Méthode « Best Subset selection »**
   a) Pour quelles raisons ne peut-on se limiter à l'analyse du modèle obtenu en partie 2 ?
   Un modèle de régression utilisant k **prédicteurs** est dit de « taille k ».
   Par exemple, $\mathbf{lpsa}=\beta_0+\beta_1*\mathbf{lcavol}+\varepsilon$ et $\mathbf{lpsa}=\beta_0+\beta_1*\mathbf{lweight}+\varepsilon$ sont tous les deux des modèles de taille 1.
   Le modèle particulier de régression sans prédicteur $\mathbf{lpsa}=\beta_0+\varepsilon$ est un modèle i.i.d. gaussien (dit « de taille 0 ».)
   Le but de cette partie est de parcourir tous les modèles possibles, de calculer leur RSS et de sélectionner les meilleurs modèles pour k variant de 0 à 8.

   Questions préliminaires :
   - Que font les commandes `lm(lpsa~1,data=pro)`, `lm(lpsa~.,data=pro[,c(1,4,9)])` et `lm(lpsa~.,data=pro[,c(2,7,9)])` ?
   - Comment récupérer le RSS d'une régression automatiquement (sans le lire vous-même à l'écran) ? (pensez à la fonction `sum`)
   - Comment effectuer automatiquement les régressions de tous les modèles possibles de taille k=2 ?
     Aide : `?combn`

b)
- Pour chaque *k* dans {0, ... , 8}, écrire le programme qui sélectionne le sous-ensemble de prédicteurs qui minimise le RSS.
- Tracez le graphe des RSS en fonction de *k*.
  Donnez explicitement (c'est-à-dire avec les noms des prédicteurs) le « meilleur »modèle pour chaque *k*.

c) Est-ce que cette méthode permet de sélectionner le meilleur modèle, toutes tailles confondues ?
Fournir une démonstration mathématique pour étayer vos arguments.

5. Split-validation

On dispose maintenant de 9 modèles, le meilleur pour chaque taille *k* de prédicteurs. On désire comparer ces modèles.

a) Rappelez brièvement le principe et l'intérêt de la split-validation.

On décide que le jeu de validation sera composé de tous les individus d'indices impairs (c'est-à-dire les lignes impaires du jeu de données). Stockez ces indices dans le vecteur `valid` en utilisant la fonction `seq` ou les fonctions modulo (ex. `1:11 %% 2 == 1`) et `which`.

b) Supposons que le meilleur modèle de taille 2 contienne les prédicteurs *i* et *j*. À quoi correspond `lm(lpsa~.,data=pro[-valid,c(i,j,9)])` ? Quelle est l'erreur d'apprentissage moyenne (*training error*) pour ce modèle ?

c) Toujours avec ce modèle de taille 2, utilisez la régression qui vient d'être faite pour prédire les valeurs de **lpsa** des individus du jeu de validation.
Aide : `?predict.lm` Il va falloir fournir à la fonction `predict` les données pour lesquelles on désire prédire le **lpsa**, c'est à dire `pro[valid,]`. Comme les colonnes de `pro` sont nommées, `predict` trouvera les valeurs correspondant aux prédicteurs utilisés dans le c) (vous pouvez vérifier en renommant `"stupidname"` la colonne n° *i*, `predict` devrait alors renvoyer un message d'erreur).
Calculez l'erreur de prédiction moyenne. *Commentez la valeur obtenue.*
Rq : la différence terme à terme de deux vecteurs de même longueur est possible avec l'opérateur "-".

d) En vous servant des étapes précédentes, appliquez la méthode de split-validation pour comparer les 9 modèles. Tracez les erreurs d'apprentissage et de prédiction. Quel modèle choisiriez-vous ? Pourquoi ? Décrire les résultats associés à ce modèle.
Aide : assurez-vous que les 2 niveaux de `gleason` sont présent dans l'ensemble d'apprentissage.

e) Quel est le principal problème de la méthode de split-validation ? Illustrez ce problème en répétant les étapes précédentes avec un nouveau jeu de validation.
Aide : utilisez `sample` pour choisir des indices au hasard.

f) Quelle méthode proposeriez vous pour l'éviter ?
La programmer et commenter ses résultats.

6. Conclusion

Quelle est votre conclusion générale quant au choix du meilleur modèle prédictif de **lpsa** ? Interprétez ce modèle.
Tout d'abord, estimez-le ! Commentez l'ensemble des prédicteurs retenus par rapport à d'autres modèles envisagés en partie 5, notamment 5d).