Master Pro M2, SITN, Université Claude Bernard, Lyon1
Outils Statistiques Avancés
Année 2013-2014

Examen du 1er avril 2014
Durée 2 heures

Avant de commencer à rédiger vos réponses, lisez avec attention ces consignes:
Pour tous les exercices, le code SAS et les sorties associées sont sur la feuilles suivantes.
Si il faut faire des tests d'hypothèse, ils sont à faire pour un seuil $\alpha = 0.05$. *Ecrire les hypothèses à tester, les statistiques de test et leur loi(si elles ont été faites en cours). Les sorties SAS sont imprimées 4 pages sur une page.*

Exercice 1.
La variable Y représente le Produit National Brut (en bilions de dollars) des Etats Unis pour la période 1969-2009, par trimestre.
1) Donnez le nombre d'observations.
2) Justifier (test d'hypothèse) pourquoi on utilise comme modèle des séries chronologiques?
3) La série chronologique considérée est-elle stationnaire (test d'hypothèse)? Appuyez votre conclusion en utilisant aussi le graphique représenté sur la page 1 des sorties SAS.
4) On a différencié la série une fois. La nouvelle variable est-elle aussi une série chronologique? Est-elle stationnaire? (Appuyez vos réponses, en dehors des tests d'hypothèse, en utilisant les graphiques de la pages 5)
5) Pour cette nouvelle série, on a considéré trois modèles de type ARMA. Justifiez le choix de ces trois modèles, notés (M1), (M2), (M3), respectivement. (voir le code SAS).
6) Donnez les détails (modèles, tests d'hypothèse, estimations, ...) pour les trois modèles. Comparaison: quel modèle choisir?

Source des données pour l'Exercice 1: http://www.economagic.com/em-cgi/data.exe/feddal/gnpcw

Exercice 2.
Il s'agit d'une étude sur le cancer de poumon des anciens combattants(vétérans) des Etats Unis.
On a mesuré les variables:
trait: le type de traitement administré: 1=standard, 2=nouveau (en test);
cell: type de cellules cancéreuses: 1=épidermoïde, 2=petites, 3=moyennes, 4=grandes;
survie: le nombre de jours en vie;
status: =1 si décès, 0=censuré
score: indice de Karnofsky ;
mois: le temps (en mois) depuis le diagnostique du cancer;
age: l'âge du patient;
therapie: traitement antérieur: 0=NON, 10=OUI.

1) D'abord on veut modéliser la probabilité de décès fonction d'autres variables. Quel type de modèle a-t-on utilisé? Sur combien d'observations a-t-il été considéré? Est-il significatif? Donner les variables qui ont une influence sur le décès. Interprétation des estimations des paramètres. Quelle est la qualité du modèle (voir page 15)?
2) On modélise ensuite, par une méthode non-paramétrique, la fonction de survie (voir la Figure de la page 22 et sorties SAS). Commentez les resultats.
3) On modélise le risque instantané que le décès survient, fonction d'autres variables. Quelles variables influent de manière significative le risque instantané? Comparez avec les variables significatives obtenues à la question 1).
4) Comment expliquez-vous la différence des résultats entre les deux modèles considérés aux questions 1) et 2)?

Source des données pour l'Exercice 2: http://lib.stat.cmu.edu/datasets/veteran

Exercice 3.
Dans cette exercice, on veut étudier l'influnce de l'ozone et des pluies acides sur les plantules de pin. Pour ceci, on a mesuré les variables:
site: codé de 1 à 6, correspond au code d'emplacement;
block: le block de chaque expérience (valeurs de 1 à 6);
rep: le numéro de la répétition;
ozone: comment on a mesuré l'ozone. Codé ainsi: 0.0 = en filtrant l'air avec charbon, 1.0 = air non filtré, "x.x"= autrement.;
rain: le pH de la pluie acide;
fam: le type de pin;
biomass: la biomasse totale (en grammes) après deux saisons de croissance (de la plantule de pin).
ppmhrs, vwpph, diam, dma, dmb, d2ha, dwhb, dmot: variables non utilisées dans l'analyse (on ne les spécifie pas).

1) Combien d'observations ont été considéré et combien d'observations ont servis effectivement aux trois modèles, notés (M4), (M5), (M6)?
2) Pour les modèles (M4) et (M5) spécifiez le type de modèle et leur forme statistique.
3) Ces deux modèles, (M4) et (M5), sont-ils significatifs? Comparez ces deux modèles (lequel est meilleur?).
4) Pourquoi on a considéré le modèle (M6)?

Source des données pour l'Exercice 3: http://lib.stat.cmu.edu/datasets/csb/ch5.txt

Exercice 4.
Les données de cet exercice concernent le dosage de la protéine *DNase* dans du sérum de rat. Les variables suivantes ont été mesurées:

obs: le numéro de l'observation;
rrun: un facteur qui indique le type de dosage;
conc: une variable numérique correspondant à la concentration de la protéine;
density: une variable numérique correspondant à la densité optique du dosage.

On modélise la variable *density* fonction de la concentration *conc* par le modèle de régression non linéaire (M7):

$$density_i = f(conc_i, \theta) + \varepsilon_i, \qquad i = 1, \cdots, n,$$

avec $f(conc_i, \theta) = c/(1 + \exp((a - \log(conc_i))/b)$, $\theta = (a, b, c)$ des paramètres inconnus, et $density_i$, $conc_i$ les valeurs de *density*, *conc* à l'observation numéro i.
1) Le modèle non linéaire (M7) est-il significatif (test d'hypothèse)? Donnez les estimations des paramètres a, b, c. Quelle est l'estimation de la variance de l'erreur ε_i?
2) On considère maintenant le modèle non linéaire mixte (M8):

$$density_i = (c + d)/(1 + \exp((a - \log(conc_i))/b) + \varepsilon_i, \qquad i = 1, \cdots, n,$$

Quelle est la loi de la variable aléatoire $density_i$?
3) Pour le modèle (M8), quel est le paramètre avec effet aléatoire et quelle est sa loi?
4) Pour le modèle (M8), donnez les estimations de **tous** les paramètres.
5) Pour les paramètres a, b, c a-t-on les mêmes estimations par les modèles (M7) et (M8)? Justification.
6) Est-il justifié, avec un risque de 0.05, de considérer un effet aléatoire, en occurrence d, dans le modèle (M8)?

Source des données pour l'Exercice 4, logiciel R. Une description des données: http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/DNase.html