Sujet n°8

*Ce devoir est fait par groupe de 2 personnes. Les réponses aux questions et le programme R permettant de répondre doivent figurer dans un seul document pdf appelé sujet**n** où vous remplacez **n** par le numéro de votre sujet. La première ligne de votre document doit contenir le nom et le prénom des deux étudiants. Ce document devra être déposé dans la zone de devoir prévue sur le site Moodle avant mardi 22 décembre minuit.*
Le nombre d'itérations sous Jags sera au minimum de 30 000. On enlèvera au moins 1000 itérations pour le temps de chauffe.

Une dizaine de machines ont été surveillées durant une année. On a relevé dans le vecteur y le nombre de pannes de chaque machine. Le vecteur x contient l'ancienneté de la machine (en année). Les données des 10 machines sont dans le tableau suivant :

x	2	14	2	9	15	7	3	14	5	2
y	3	50	7	20	44	3	1	58	8	7

Il est fréquent de modéliser le nombre de pannes y par la loi de Poisson dont le paramètre dépend éventuellement de l'ancienneté après une transformation logarithmique.

Trois différents modèles sont proposés, du plus simple au plus complexe. Le modèle M3 est optionnel pour ce devoir, les modèles M1 et M2 sont eux obligatoires.
Le modèle M1, le plus simple, est le suivant :

$$y_i \sim \text{Pois}(\lambda) \text{ avec}$$
$$\log(\lambda) = a \text{ pour } i=1 \text{ à } 10$$

Le modèle M2 est le suivant :

$$y_i \sim \text{Pois}(\lambda_i) \text{ avec}$$
$$\log(\lambda_i) = a_0 + b_0 x_i \quad \text{pour } i=1 \text{ à } 10$$

Le modèle M3 est le suivant :

$$y_i \sim \text{Pois}(\lambda_i) \text{ avec}$$
$$\log(\lambda_i) = a_0 + b_0 x_i + \varepsilon_i \quad \text{pour } i=1 \text{ à } 10$$
$$\text{où } \varepsilon_i \text{ est iid de loi } N(0,\sigma^2)$$

Questions

Rappel : si Z~Pois(m) alors E(Z)=Var(Z)=m et $P(Z=z) = m^z \exp(-m)/z!$
Grâce au format list, préparez vos données pour qu'elles soient lisibles sous Jags.

Pour M1 :

1) Donner $E(y_i|a)$ d'après ce modèle en fonction de a.
2) Mettre en place ce modèle avec, comme loi a priori sur a, une loi normale d'espérance nulle et de variance 1000. Faire 30000 itérations et enlever 1000 itérations pour le temps de chauffe. D'après l'history et les autocorrélations, voyez-vous un problème de mélangeance de l'algorithme ? Si oui, résoudre ce problème en justifiant.
3) Que vaut le nombre d'itérations pour les calculs ? Que vaut le nombre d'itérations « effectif » ?
4) Donnez la moyenne a posteriori et l'intervalle de crédibilité à 95% de a.
5) Que vaut le DIC ? Que vaut l'estimation de la complexité du modèle ? Vous semble-t-elle logique ?

6) Refaire tourner ce modèle (30000 itérations et enlever 1000 itérations pour le temps de chauffe) mais avec cette fois-ci comme loi a priori sur a, une loi normale d'espérance nulle et de variance 10000. Donnez la moyenne a posteriori et l'intervalle de crédibilité à 95% de a et commentez.

Pour M2 :

1) Donner $E(y_i|a_0,b_0,x_i)$ d'après ce modèle en fonction de a_0, b_0 et de x_i ? Si $b_0=0$, que cela signifie-t-il ? Même question si b_0 est supérieur à 0 ou si b_0 est inférieur à 0 ?
2) Mettre en place ce modèle avec, comme loi a priori sur a_0 et b_0 , une loi normale d'espérance nulle et de variance 1000. Faire 30000 itérations et enlever 1000 itérations pour le temps de chauffe. D'après l'history et les autocorrélations, voyez-vous un problème de mélangeance de l'algorithme ? Si oui, mettre un thin à 10. Cela a-t-il amélioré la mélangeance ? On considèrera que c'est suffisant.
3) Que vaut le nombre d'itérations pour les calculs ? Que vaut le nombre d'itérations « effectif » ?
4) Si ce n'est pas le cas, refaire tourner votre modèle pour que le nombre effectif d'itérations soit au moins de 10000.
5) Donnez la moyenne a posteriori et l'intervalle de crédibilité à 95% de a_0 et b_0.
6) Pensez-vous que la variable x doit être prise en compte ? Donnez rapidement une interprétation du résultat (par exemple, pour deux machines ayant une différence d'ancienneté de 1 an, que représente $\exp(b_0)$?).
7) Que vaut le DIC ? Que vaut l'estimation de la complexité du modèle ? Vous semble-t-elle logique ?
8) D'après le DIC, quel modèle choisissez-vous entre M1 et M2 ?

Question supplémentaire : selon le modèle que vous avez retenu, dire quelles machines (s'il y en a) ont une moyenne du nombre de pannes probablement (c'est à dire à 97,5%) supérieure à 40 pannes (valeur pour laquelle la machine sera remplacée) ?

Pour M3 (optionnel) :

1) Selon vous, quel est l'intérêt de l'ajout de la composante aléatoire ε ?
2) Mettre en place ce modèle avec, comme loi a priori sur a_0 et b_0 , une loi normale d'espérance nulle et de variance 1000 et sur $tau=1/\sigma^2$ une loi gamma de paramètres 0.01 et 0.01. Faire 30000 itérations et enlever 1000 itérations pour le temps de chauffe. D'après l'history et les autocorrélations, voyez-vous un problème de mélangeance de l'algorithme ? Si oui, mettre un thin à 10. Cela a-t-il amélioré la mélangeance ? Si non, rajoutez des itérations et augmentez le thin afin d'avoir des résultats raisonnables.
3) Que vaut le nombre d'itérations pour les calculs ? Que vaut le nombre d'itérations « effectif » ?
4) Donnez la moyenne a posteriori et l'intervalle de crédibilité à 95% de a_0, b_0 et tau.
5) Que vaut le DIC ? Que vaut l'estimation de la complexité du modèle ? Vous semble-t-elle logique ?
6) D'après le DIC, quel modèle choisissez-vous entre M1, M2 et M3 ?