

TD « Teneur en protéines »

Régression linéaire pour prédire la
teneur en protéines du blé

Corrections package rjags

Contexte

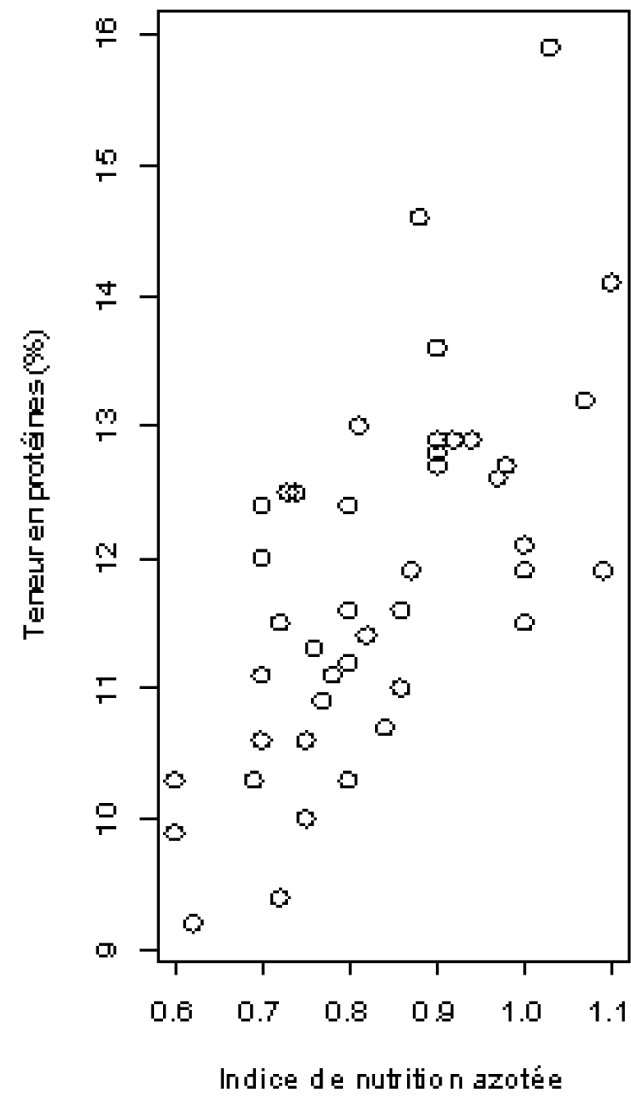
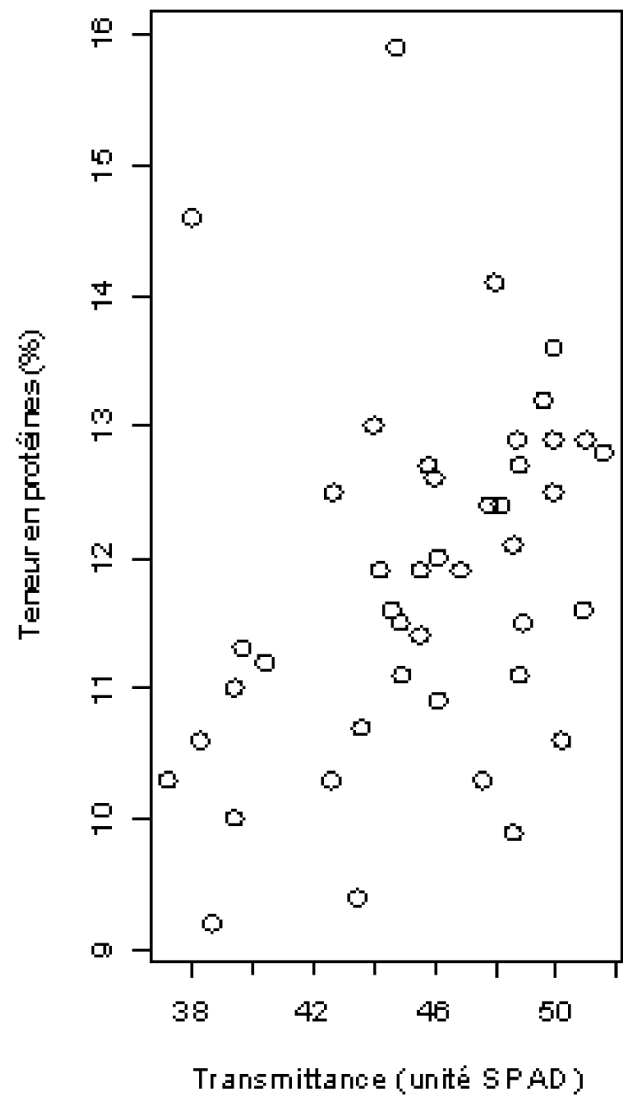
- Teneur en protéines des grains de blé :
 - Critère de qualité important pour les entreprises qui collectent et stockent les récoltes de blé;
 - Détermine le type d'utilisation industrielle d'une récolte (panification, fabrication de biscuits, alimentation animale, etc.)
- Important de pouvoir prédire, avant la récolte, la qualité du blé afin d'organiser le stockage des grains en silos et de passer des contrats.

Objectif

- Proposer un modèle linéaire pertinent pour prédire la teneur en protéines du blé
- Utiliser des variables explicatives mesurables quelques semaines avant la récolte dans les parcelles cultivées par les agriculteurs :
 - une mesure de transmittance (SPAD) réalisée sur un échantillon de feuilles de blé avec le chlorophyl meter Minolta ;
 - une mesure de l'indice de nutrition azotée du blé (INN) à floraison.

Données

- Quarante-trois expérimentations ont été réalisées en exploitations agricoles pendant trois ans (2004, 2005, 2006) sur plusieurs sites en France.
- Chaque expérimentation est constituée d'une parcelle d'agriculteur sur laquelle l'INN, le SPAD et la teneur en protéines (%) ont été mesurés.



Questions

1- Ecrire les équations de plusieurs modèles linéaires permettant de prédire la teneur en protéines à partir:

- ✓ d'une variable explicative,
- ✓ de deux variables explicatives,
- ✓ d'aucune variable explicative.

2- Représenter graphiquement ces modèles

3- Quels sont les paramètres à estimer ?

Modèle à 1 variable explicative

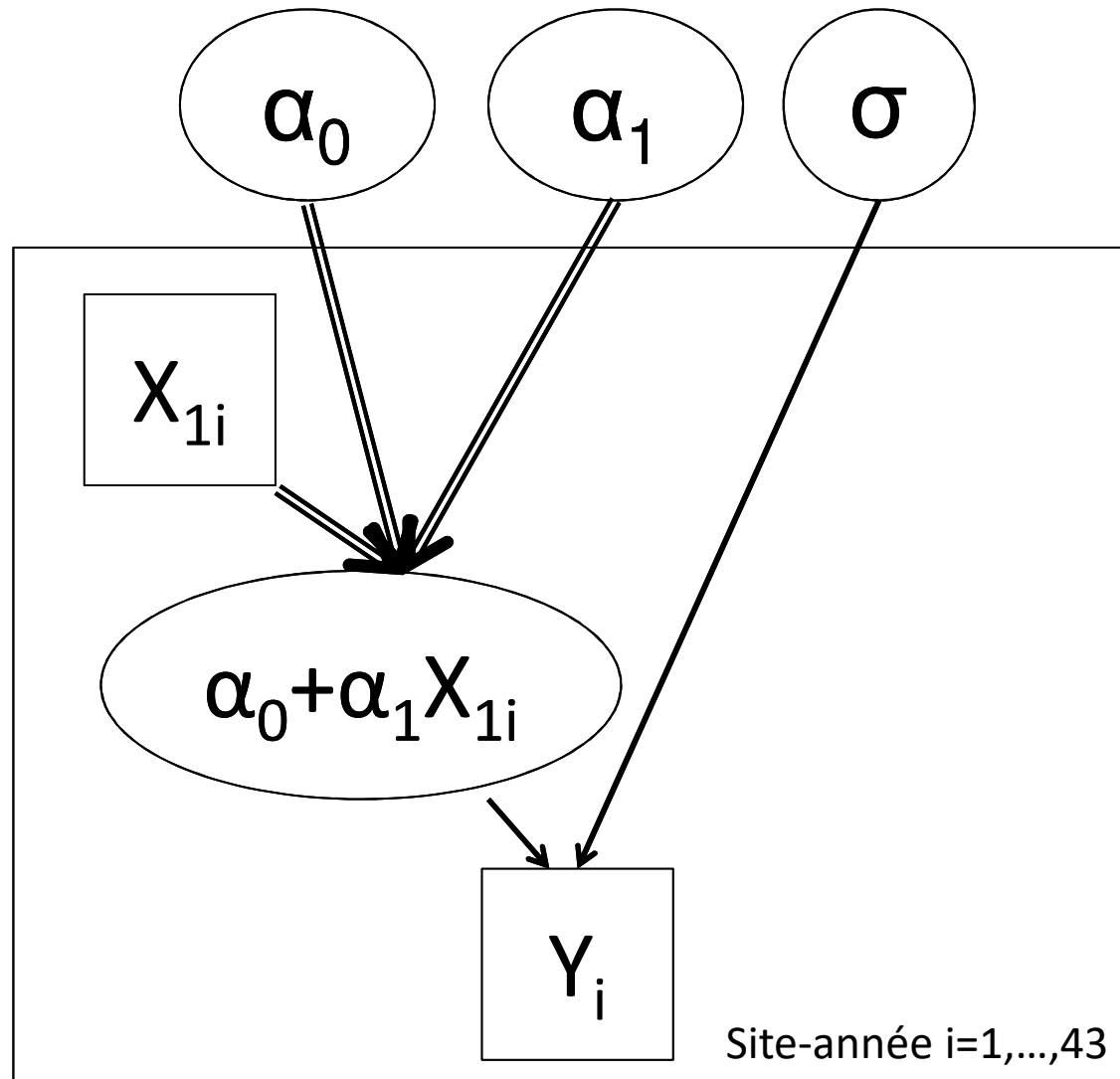
- Modèle d'observations

$$(M_1) \quad Y_i = \alpha_0 + \alpha_1 X_{1i} + \varepsilon_i$$
$$\varepsilon_i \sim^{i.i.d} N(0, \sigma^2)$$

Y_i = Teneur en protéines moyenne des grains de blé pour le site-année i

X_{i1} = Mesure de transmittance moyenne (SPAD) OU mesure de l'indice de nutrition azotée du blé (INN) pour le site-année i

ε_i = Terme d'erreur (bruit blanc) pour le site-année i



Directed Acyclic Graph (DAG) du modèle M_1 de teneur en protéines du blé.

Les variables observables sont représentées dans des rectangles et les paramètres inconnus dans des ellipses. Les double flèches correspondent à des relations déterministes. Les flèches en trait plein indiquent une relation stochastique.

Modèle à deux variables explicatives

- Modèle d'observations

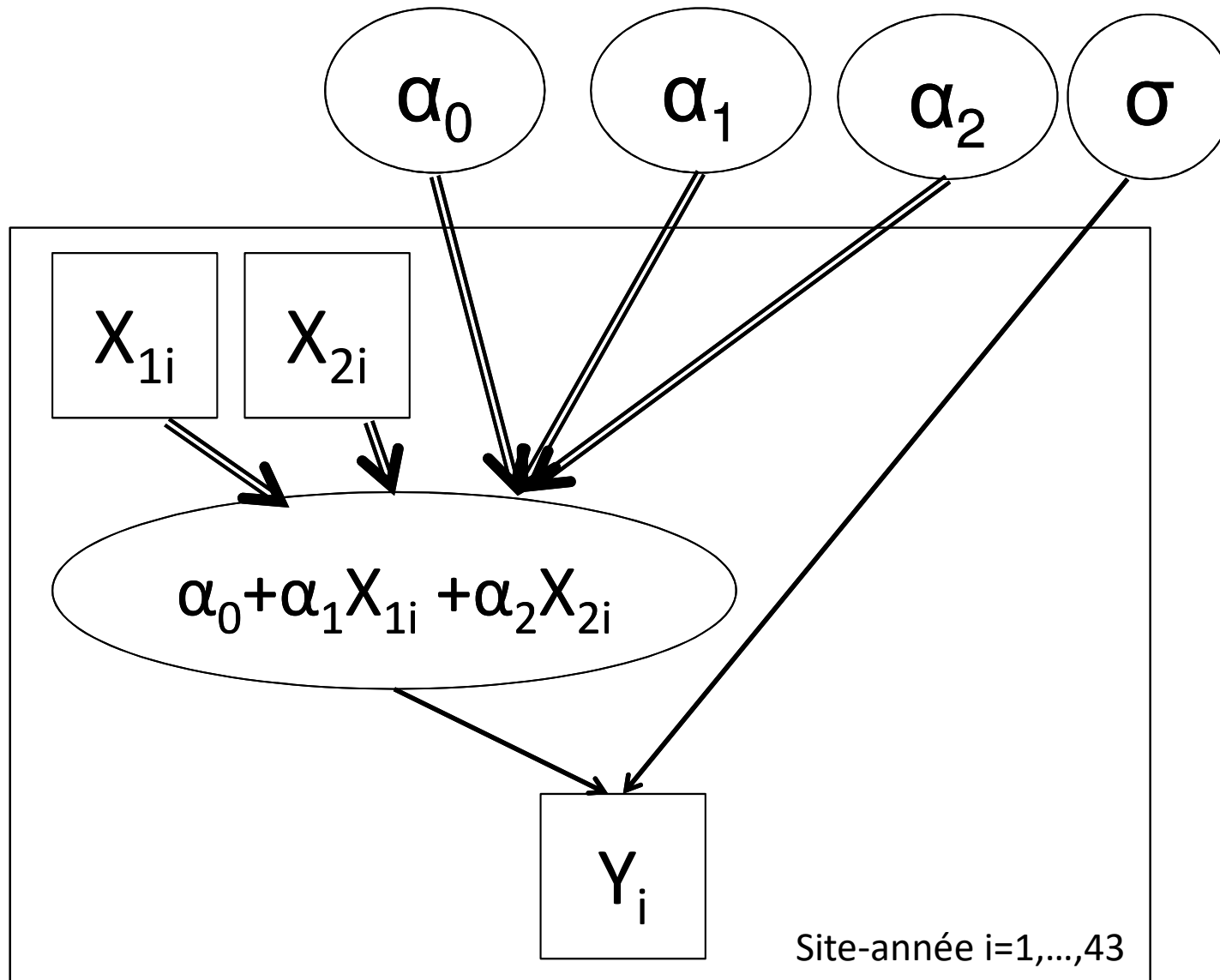
$$\begin{aligned} (M_2) \quad Y_i &= \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \varepsilon_i \\ \varepsilon_i &\sim^{i.i.d} N(0, \sigma^2) \end{aligned}$$

Y_i = Teneur en protéines moyenne des grains de blé pour le site-année i

X_{i1} = Mesure de transmittance moyenne (SPAD)

X_{i2} = Mesure de l'indice de nutrition azotée du blé (INN) pour le site-année i

ε_i = Terme d'erreur (bruit blanc) pour le site-année i



Directed Acyclic Graph (DAG) du modèle M_2 de teneur en protéines du blé.

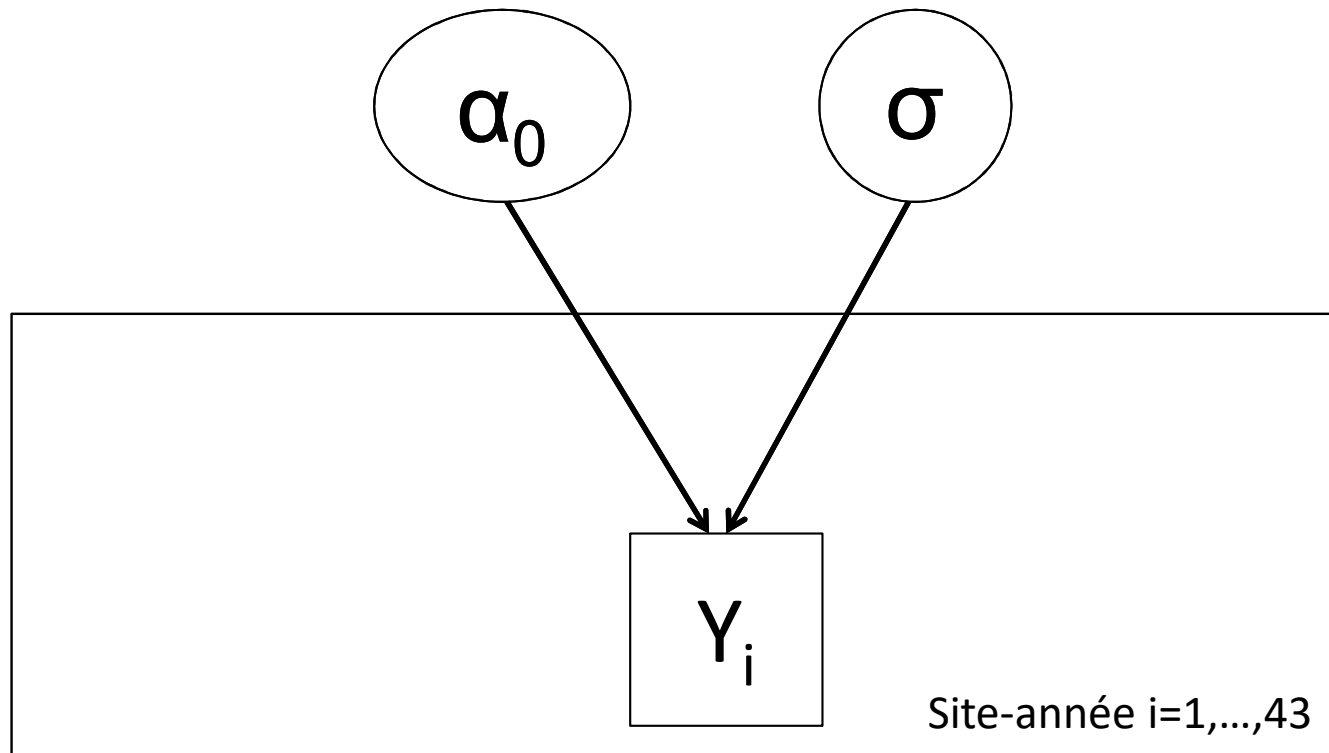
Modèle sans variable explicative

- Modèle d'observations

$$(M_0) \quad \begin{aligned} Y_i &= \alpha_0 + \varepsilon_i \\ \varepsilon_i &\sim^{i.i.d} N(0, \sigma^2) \end{aligned}$$

Y_i = Teneur en protéines moyenne des grains de blé pour le site-année i

ε_i = Terme d'erreur (bruit blanc) pour le site-année i



Directed Acyclic Graph (DAG) du modèle M_0 de teneur en protéines du blé.

Questions

4- Définir des distributions a priori pour les paramètres

5- Implémenter les modèles sous OpenBUGS ou rjags ou rstan afin d'estimer les paramètres à partir des données disponibles. Vérifier tout d'abord que les programmes fonctionnent en réalisant 5000 itérations MCMC

Choix des distributions a priori

Sur les coefficients de régression (en l'absence de connaissances a priori)

$$\alpha_0, \alpha_1, \alpha_2 \sim N(0, 10^6)$$

Sur l'écart-type des termes d'erreur

$$\sigma \sim Unif(0, 100)$$

Alternative: Sur la précision des termes d'erreur

$$\tau = \frac{1}{\sigma^2} \sim Gamma(0.001, 0.001)$$

Modèle à 1 variable explicative

```
model{
# Y= teneur en protéines du blé
# covar_cr = variable explicative (SPAD ou INN) centrée réduite
# alpha=paramètres de la régression; prec=paramètre de précision = 1/variance
# m, s= moyenne et écart-type empirique de covar
#Modèle d'observations
for (i in 1:N) {
  Y[i] ~ dnorm(mu[i],prec)
  #mu[i]<-alpha_nonc[1]+alpha_nonc[2]*covar[i] => Problème de convergence!
  mu[i]<-alpha[1]+alpha[2]*covar_cr[i]

  #Calcul de la log_vraisemblance de l'observation i
  loglik[i] <- logdensity.norm(Y[i], mu[i], prec)
}
alpha_nonc[1]<- alpha[1]-alpha[2]*(m/s)
alpha_nonc[2]<- alpha[2]/s

#Lois a priori
for (j in 1:2) {
  alpha[j]~dnorm(0.0,1.0E-6)
}
sd ~ dunif(0,100)
prec<- 1/pow(sd,2)}
```

Les données

#covar=Inn

list(
Y=c(11.9 , 10.3 , 12.4 , 12.1 , 12.4 , 12.7 , 11.5 , 13.6 , 9.9 , 11.1 , 10.6 , 12 , 12.8 , 11.6 , 11.2 , 10.3 ,
12.9 , 10.3 , 11.5 , 14.1 , 12.6 , 12.7 , 11.9 , 9.2 , 12.5 , 11.9 , 11.3 , 11 , 10 , 15.9 , 10.6 , 11.6 , 11.4 ,
10.9 , 14.6 , 10.7 , 9.4 , 13 , 12.5 , 12.9 , 11.1 , 12.9 , 13.2), covar_cr=c(1.26708817,-0.26123697, -
0.26123697,... , 1.80200197),N=43, **m=mean(INN),s=sd(INN)**)

#covar=SPAD

data=list(
Y=c(11.9 , 10.3 , 12.4 , 12.1 , 12.4 , 12.7 , 11.5 , 13.6 , 9.9 , 11.1 , 10.6 , 12 , 12.8 , 11.6 , 11.2 , 10.3 ,
12.9 , 10.3 , 11.5 , 14.1 , 12.6 , 12.7 , 11.9 , 9.2 , 12.5 , 11.9 , 11.3 , 11 , 10 , 15.9 , 10.6 , 11.6 , 11.4 ,
10.9 , 14.6 , 10.7 , 9.4 , 13 , 12.5 , 12.9 , 11.1 , 12.9 , 13.2), covar_cr=c(-0.01827560,-0.74406137,
0.55734760,...,1.00783532), N=43, **m=mean(SPAD),s=sd(SPAD)**)

Les valeurs initiales (3 chaînes)

#Chaîne 1

```
inits =list(list(alpha_c=c(0,0), sd=1),
```

#Chaîne 2

```
list(alpha_c= c(5,5), sd = 0.1),
```

#Chaîne 3

```
list(alpha_c= c(-5,-5), sd = 10))
```

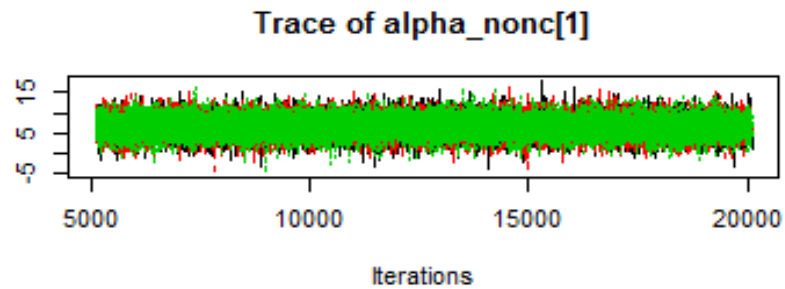
Questions

6- Pour chaque modèle : lancer trois chaînes de Markov. Utiliser le diagnostic de convergence de Gelman et Rubin et des représentations graphiques des chaînes pour déterminer le nombre d'itérations nécessaire pour espérer avoir atteint la convergence.

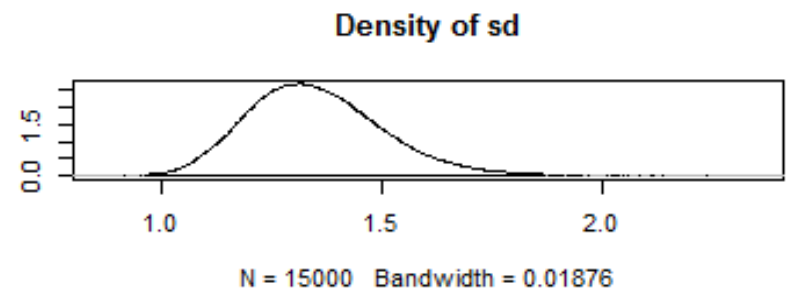
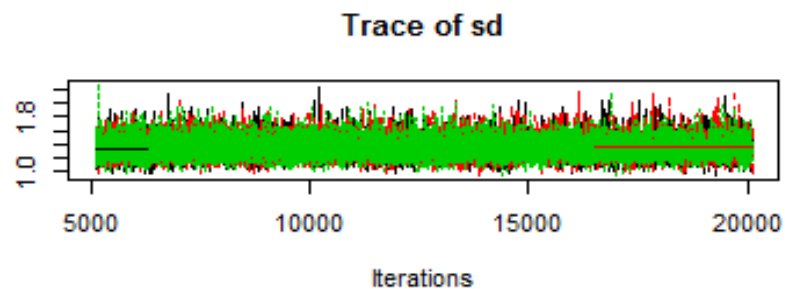
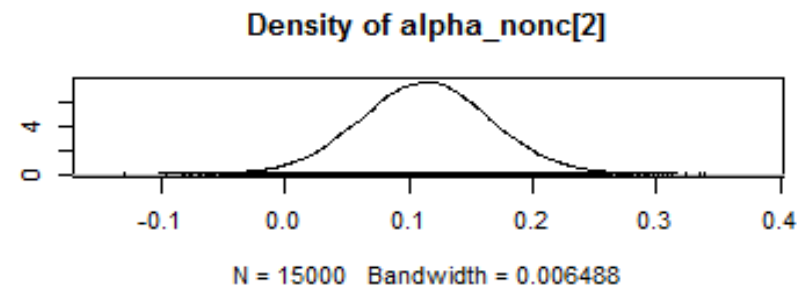
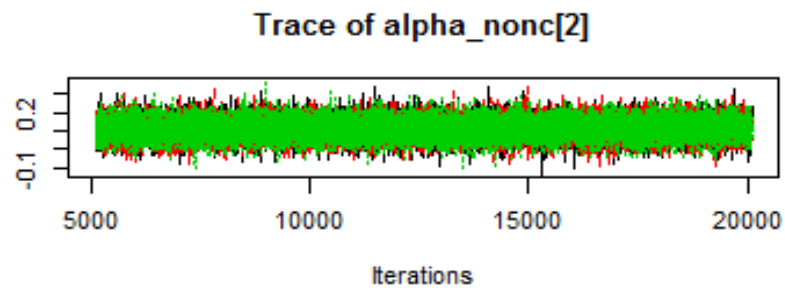
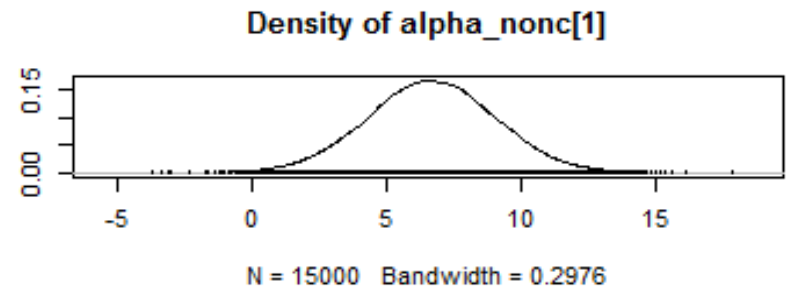
7- Analyser les auto-corrélations des chaînes. Faites de nouvelles itérations MCMC afin d'obtenir un échantillon de valeurs suffisamment représentatif de la loi a posteriori.

Résultats (SPAD)

Visualisation des 3 chaînes de Markov

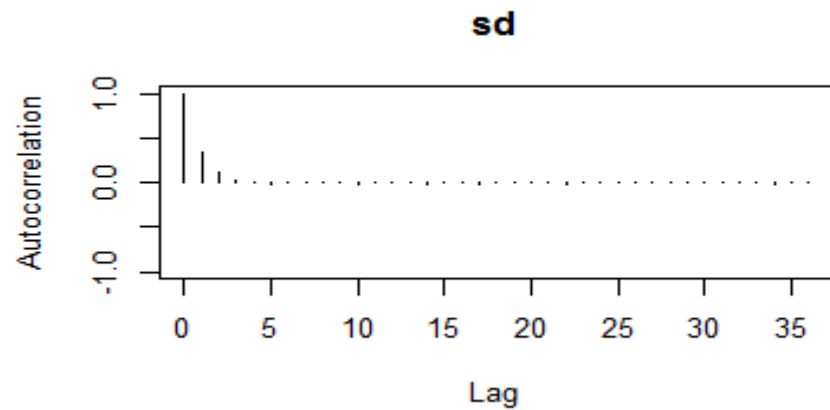
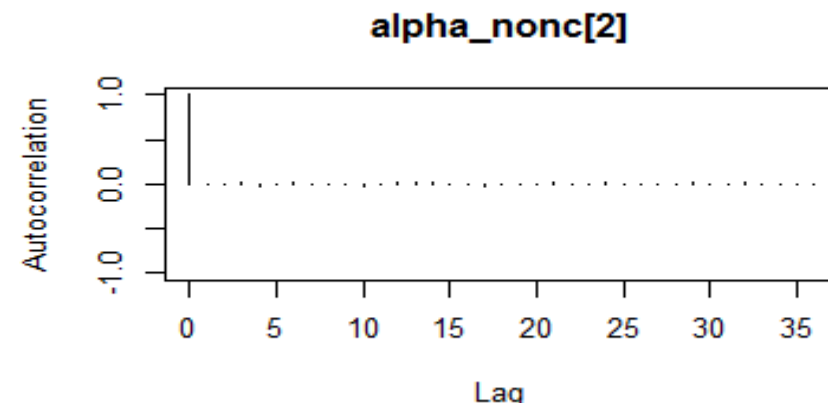
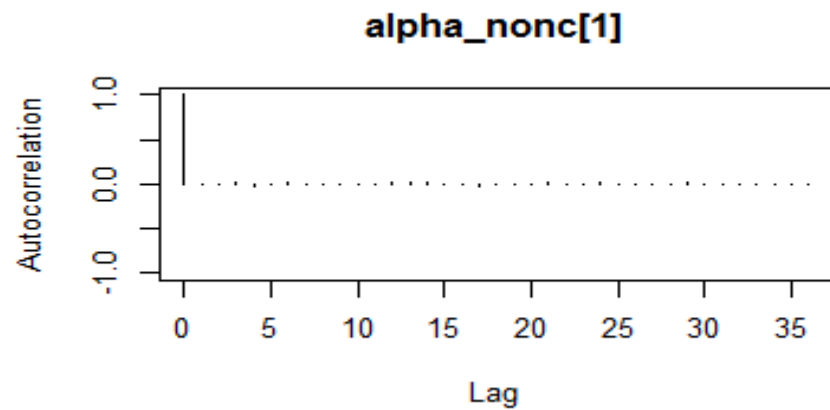


Densités a posteriori



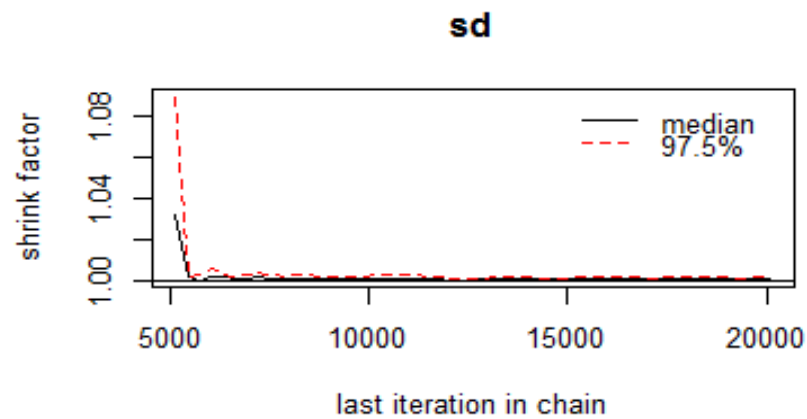
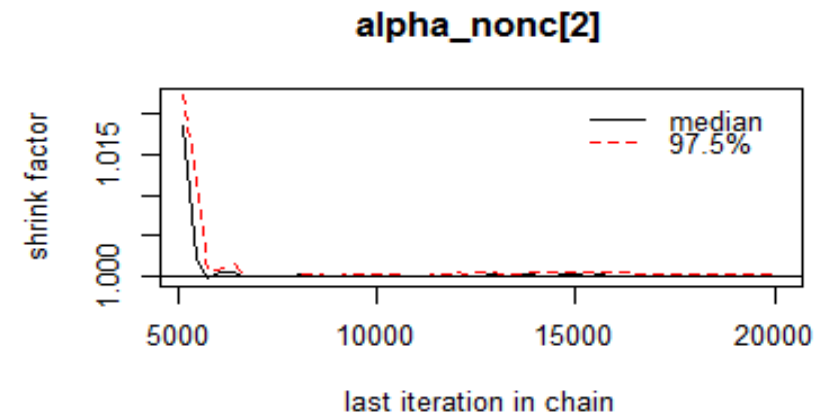
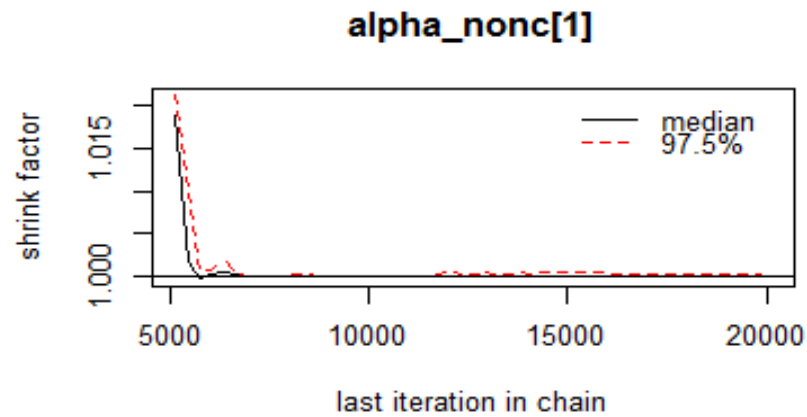
Résultats (SPAD)

Autocorrélations intra-chaîne (chaîne 1)



Résultats (SPAD)

Statistique de Gelman Rubin



Modèle à 2 variables explicatives

```
model{
# Y= teneur en protéines du blé
# INN_cr = Indice de nutrition azotée centré réduit # SPAD_cr = Mesure de transmittance centrée réduite
# alpha=paramètres de la régression # prec=paramètre de précision = 1/variance
# m_INN, s_INN , m_SPAD, s_SPAD = moyennes et écarts-type empiriques de INN et SPAD

for (i in 1:N) {
  Y[i] ~ dnorm(mu[i],prec)
  mu[i]<-alpha[1]+alpha[2]*SPAD_cr[i]+alpha[3]*INN_cr[i]
  loglik[i] <- logdensity.norm(Y[i], mu[i], prec)
}
alpha_nonc[1]<- alpha[1]-alpha[2]*(m_SPAD/s_SPAD)-alpha[3]*(m_INN/s_INN)
alpha_nonc[2]<- alpha[2]/s_SPAD
alpha_nonc[3]<- alpha[3]/s_INN

for (j in 1:3) {
  alpha[j]~dnorm(0.0,1.0E-6)
}
sd ~ dunif(0,100)
prec<- 1/pow(sd,2)
}
```

Modèle sans variable explicative

```
model{  
# Y= teneur en protéines du blé  
# mu=paramètres d'espérance  
# prec= paramètre de précision = 1/variance  
  
#Modèle d'observations  
for (i in 1:N) {  
    Y[i] ~ dnorm(mu,prec)  
    loglik[i] <- logdensity.norm(Y[i], mu, prec)  
}  
  
#Lois a priori  
mu~dnorm(0.0,1.0E-6)  
sd ~ dunif(0,100)  
prec<- 1/pow(sd,2)  
}
```

Questions

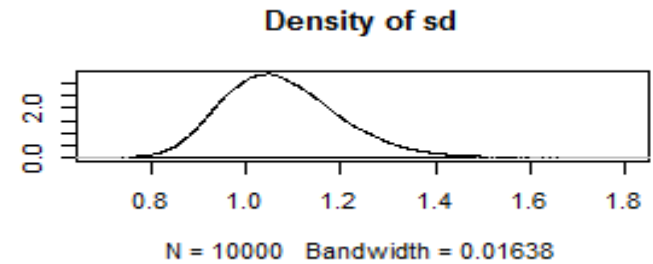
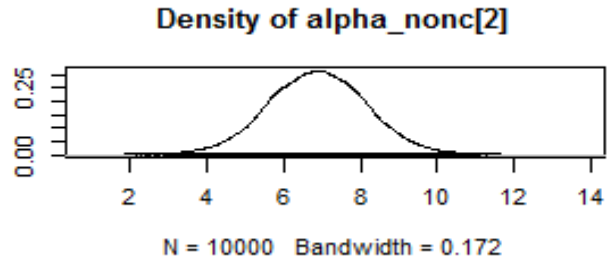
8- Représenter les densités a posteriori des paramètres et résumer ces densités par leurs moyennes, médianes, et quelques percentiles.

9- Calculer le DIC des différents modèles proposés et identifier le modèle qui a le DIC le plus faible

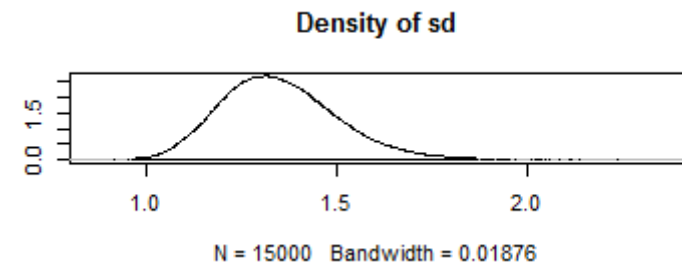
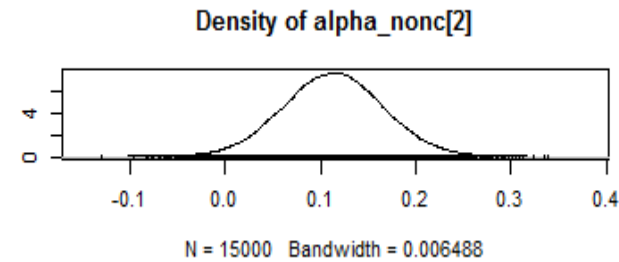
10- Calculer le WAIC des différents modèles proposés et identifier le modèle qui a le WAIC le plus faible

Densités a posteriori

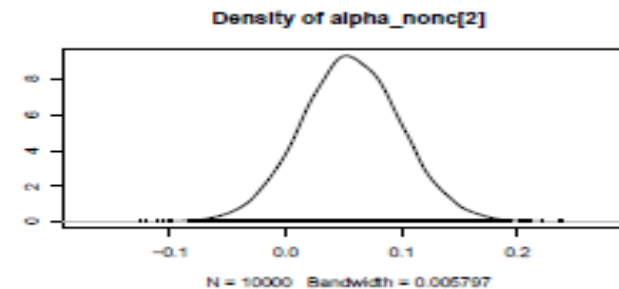
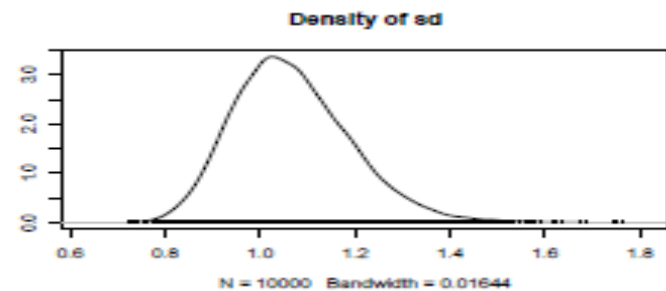
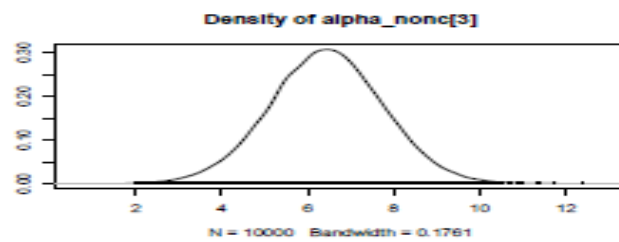
M₁
INN



M₁
SPAD



M₂
INN+SPAD



Statistiques a posteriori

node	mean	sd	MC error	2.5%	median	97.5%
M₁ - INN						
alpha_nonc[1]	6.09	1.08	0.006	3.96	6.09	8.20
alpha_nonc[2]-INN	6.89	1.28	0.007	4.38	6.89	9.42
sd	1.08	0.12	0.0007	0.87	1.07	1.35
M₁ - SPAD						
alpha_nonc[1]	6.68	2.41	0.011	1.95	6.68	11.4
alpha_nonc[2]-SPAD	0.11	0.05	1.7E-4	0.01	0.11	0.22
sd	1.35	0.15	5.2E-4	1.09	1.34	1.69
M₂ - INN + SPAD						
alpha_nonc[1]	3.90	1.98	0.011	-0.05	3.90	7.77
alpha_nonc[3] - INN	6.44	1.32	0.008	3.83	6.44	9.04
alpha_nonc[2]-SPAD	0.06	0.04	0.0002	-0.03	0.06	0.14
sd	1.07	0.12	0.0007	0.86	1.06	1.34
M₃ - sanscov						
sd	1.41	0.16	0.0009	1.14	1.40	1.78

DIC et WAIC

	M_1 INN	M_1 SPAD	M_2 INN+SPA D	M_0
DIC	130.2	149.6	130.3	152.4
WAIC	130.3	151.0	131.5	152.6

Exemple sortie DIC sous rjags

Mean deviance: 146.4
penalty 3.178
Penalized deviance: 149.6

Questions

11 - Utiliser le modèle sélectionné pour déterminer la densité prédictive a posteriori de la teneur en protéines d'une nouvelle parcelle caractérisée par SPAD=50 et/ou INN=0.95

Prédiction a posteriori selon le modèle M_1 INN

Le script du modèle univarié n'est pas modifié! Ce qui change :

➤ **Les données – covariable INN**

```
INNpred_cr<- (0.95-mean(INN))/sd(INN)
```

```
data =list(Y=c(Y,NA),covar_cr=c(INN_cr,INNpred_cr),N=44,m=mean(INN), s=sd(INN))
```

➤ **Les valeurs initiales (exemple pour une chaîne)**

```
list(Y=c(rep(NA,times=43),10),alpha=c(-5,-5), sd=10)
```

Densité prédictive a posteriori de la teneur en protéines des grains de blé pour INN=0.95 (44^{ème} donnée)

1. Empirical mean and standard deviation for each variable plus standard error of the mean:

Mean	SD	Naive SE	Time-series SE
12.63531	1.10496	0.00638	0.00634

2. Quantiles for each variable:

2.5%	25%	50%	75%	97.5%
10.5	11.9	12.6	13.4	14.8

