TD « Teneur en protéines »

Régression linéaire pour prédire la teneur en protéines du blé

Corrections package rstan

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.)

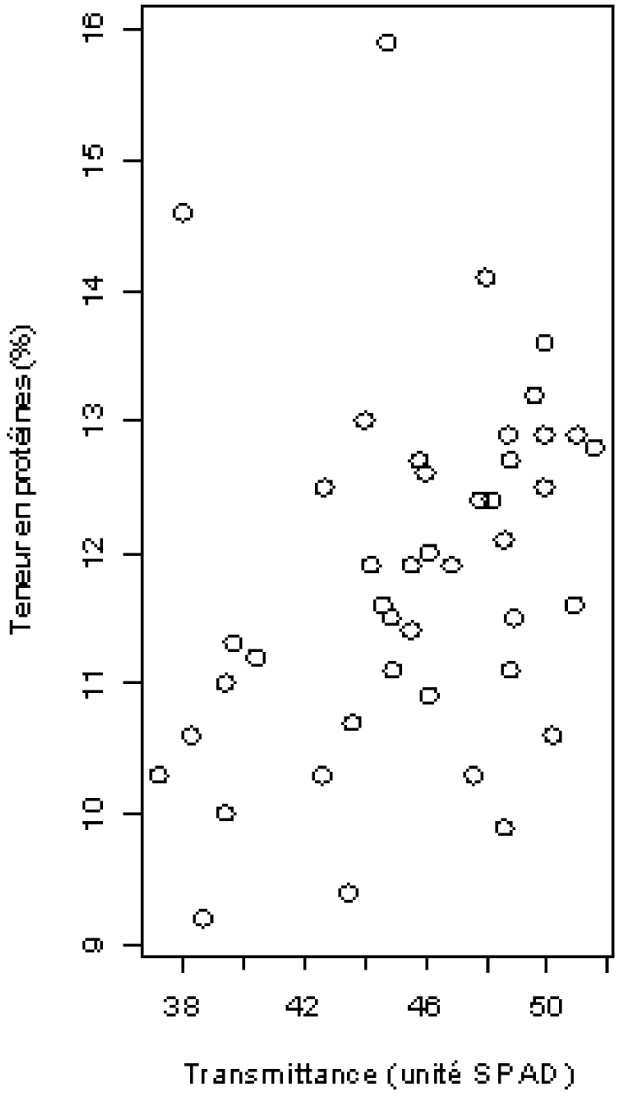
- Important de pouvoir prédire, avant la récolte, la qualité du blé afin d'organiser le stockage des grains en silos et de passer des contrats.

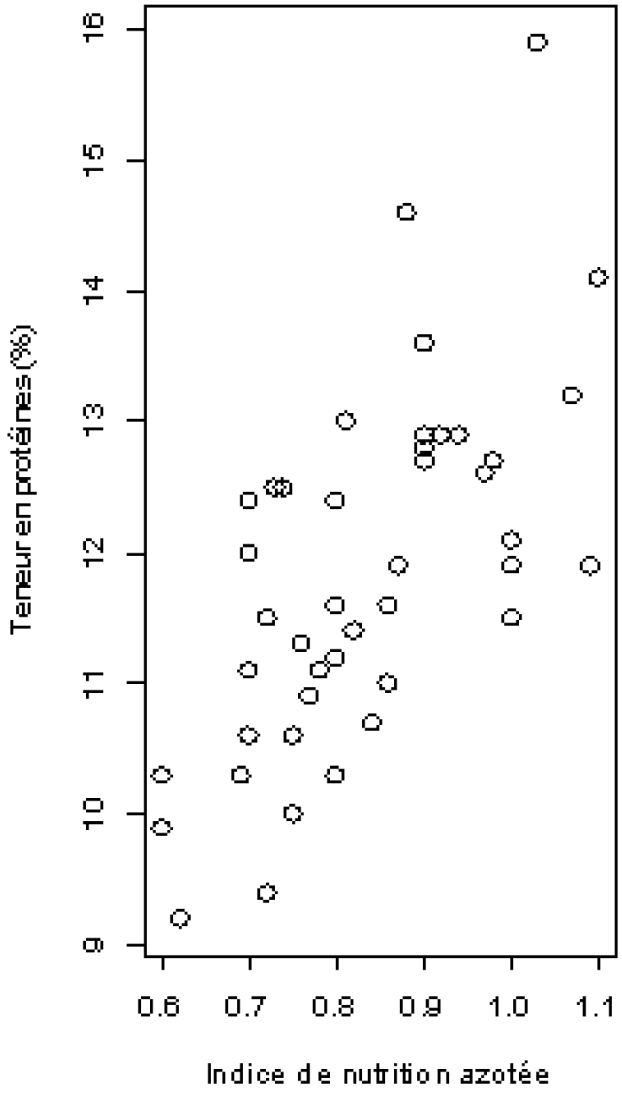
Objectif

- 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.


Teneur en protéines (%)
9
10
11
12
13
14
15
16
38
42
46
50
Transmittance (unité SPAD)


Teneur en protéines (%)
9
10
11
12
13
14
15
16
0.6
0.7
0.8
0.9
1.0
1.1
Indice de nutrition azotée

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

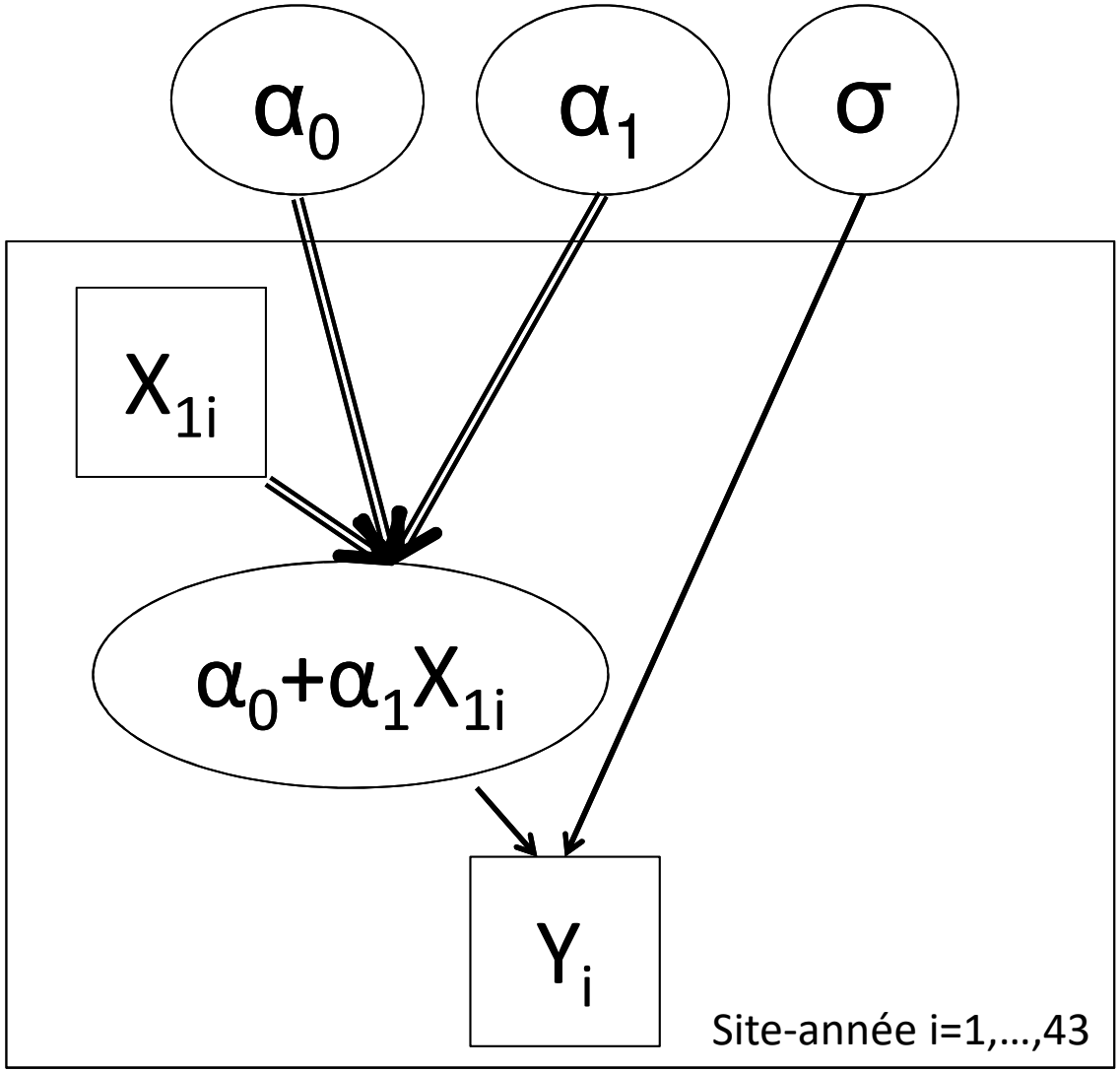
- Modèle d'observations

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

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

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



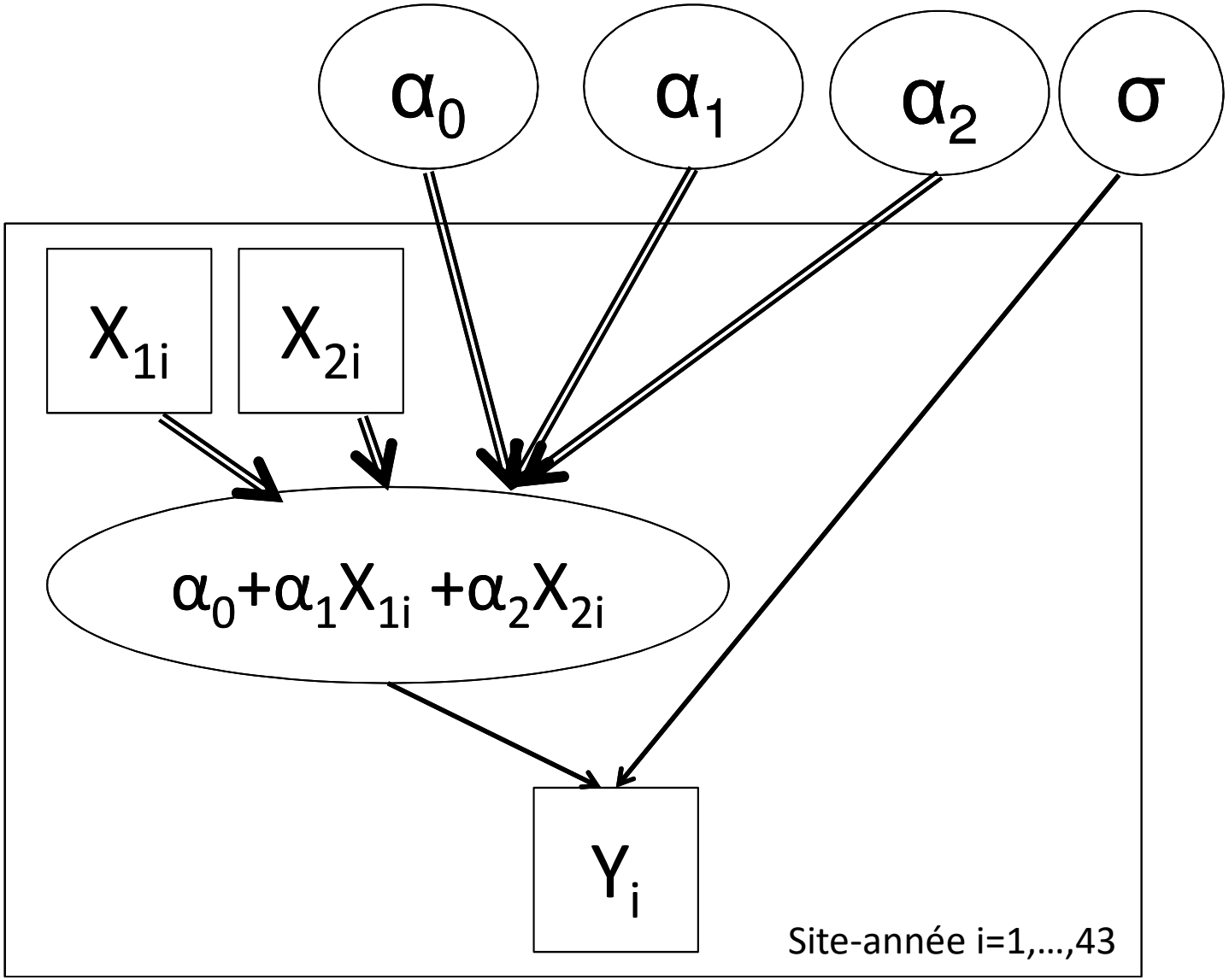
Directed Acyclic Graph (DAG) du modèle M_1 de teneur en protéines du blé.
Les variables observables sont représentées dans des rectangles et les paramètres inconnus dans des ellipses. Les double flèches correspondent à des relations déterministes. Les flèches en trait plein indiquent une relation stochastique.

Modèle à deux variables explicatives

- Modèle d'observations

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

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

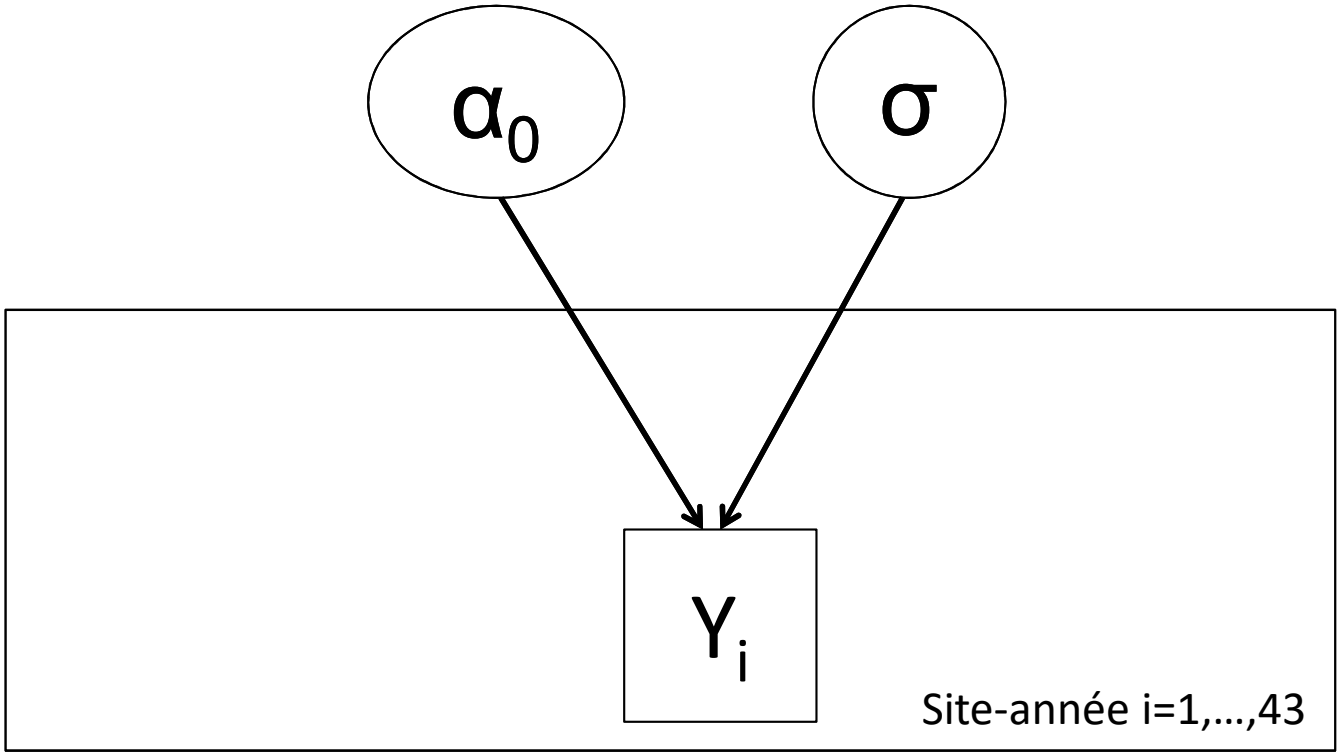
(M_0)

$$Y_i = \alpha_0 + \varepsilon_i$$

$$\varepsilon_i \sim^{i.i.d} N(0, \sigma^2)$$

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

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

```
teneur_univarie <- "
data {
        int<lower=0> N;
        vector[N] covar_cr;  //centré réduit
        vector[N] Y;
        real m;
        real<lower=0> s;
        vector[2] mu;
        matrix[2,2] W;
}

parameters {
        vector[2] alpha;
         real<lower=0, upper=100> sigma;
//Loi a priori Uniforme sur [0,100]
}

model{
  Y ~normal(alpha[1]+alpha[2]*covar_cr,sigma);
  alpha ~ multi_normal(mu,W);
}

generated quantities {
        vector[2] alpha_nonc;
        vector[N] loglike;

        alpha_nonc[1] = alpha[1]-alpha[2]*(m/s);
        alpha_nonc[2] = alpha[2]/s;

        //Log-vraisemblance pour l'observation i
        for(i in 1:N){
                loglike[i] = normal_lpdf(Y[i]
        alpha[1]+alpha[2]*covar_cr[i], sigma);
        }
}
"
```

Les données

#covar=Inn

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(1.26708817,-0.26123697, -0.26123697,… , 1.80200197),

N=as.integer(43),m=mean(INN),s=sd(INN),mu=c(0,0),W=matrix(c(100,0,0,100),nr=2,ncol=2))

#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=as.integer(43),m=mean(SPAD),s=sd(SPAD),mu=c(0,0),W=matrix(c(100,0,0,100),nr=2,ncol=2))

Les valeurs initiales (3 chaînes)

```
 init = list(

#Chaîne 1
list(alpha_c=c(0,0), sigma=1),

#Chaîne 2
list(alpha_c= c(5,5), sigma = 0.1),

#Chaîne 3
list(alpha_c= c(-5,-5), sigma = 10))
```

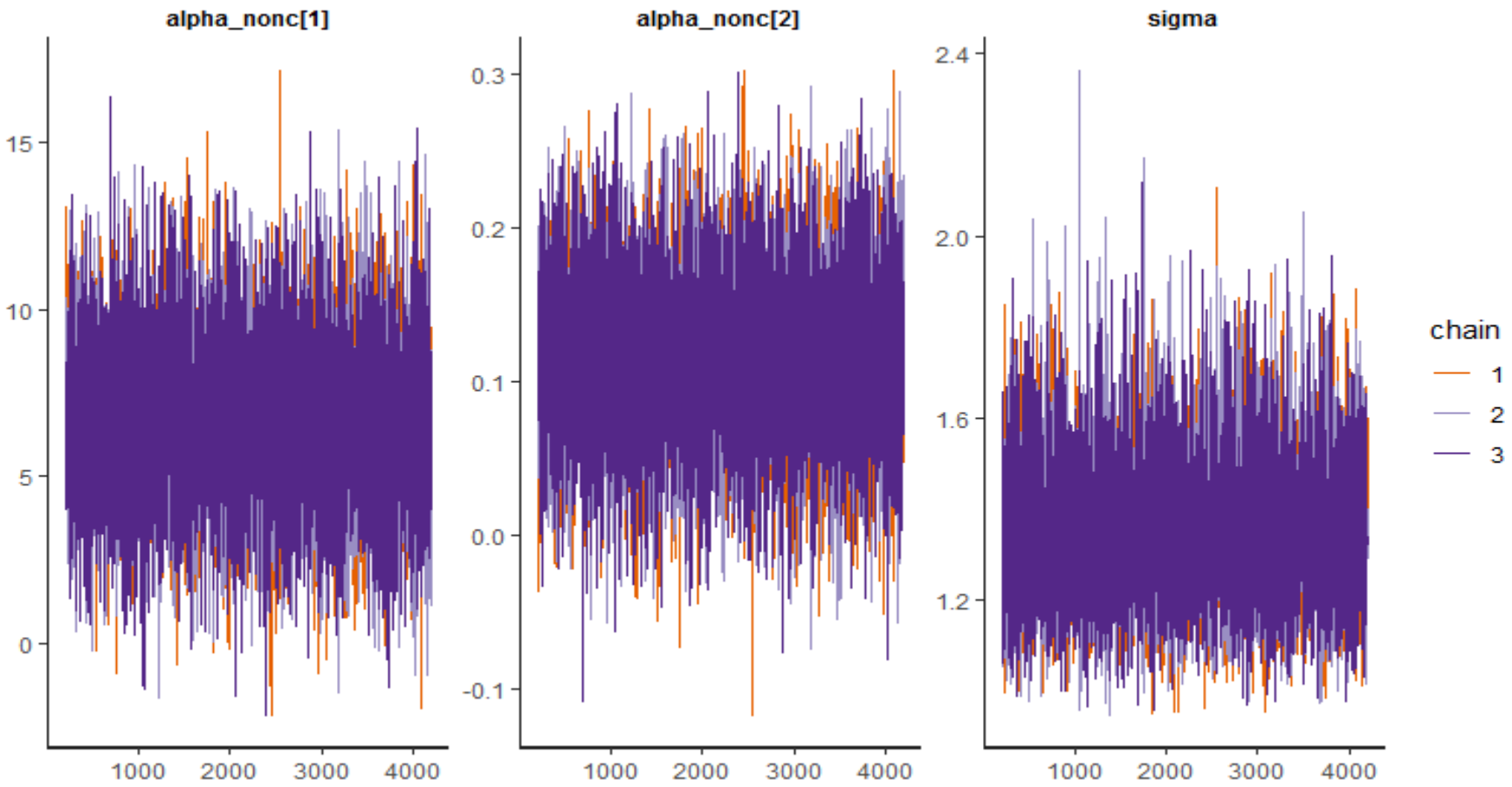
Questions

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.

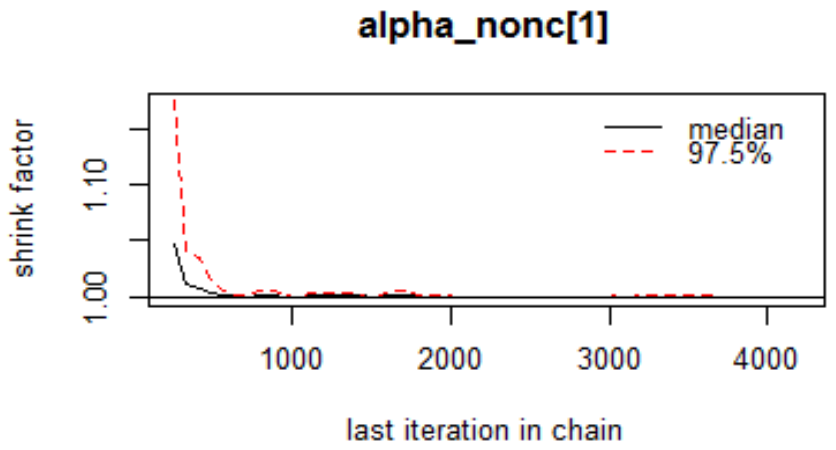
Résultats (SPAD)

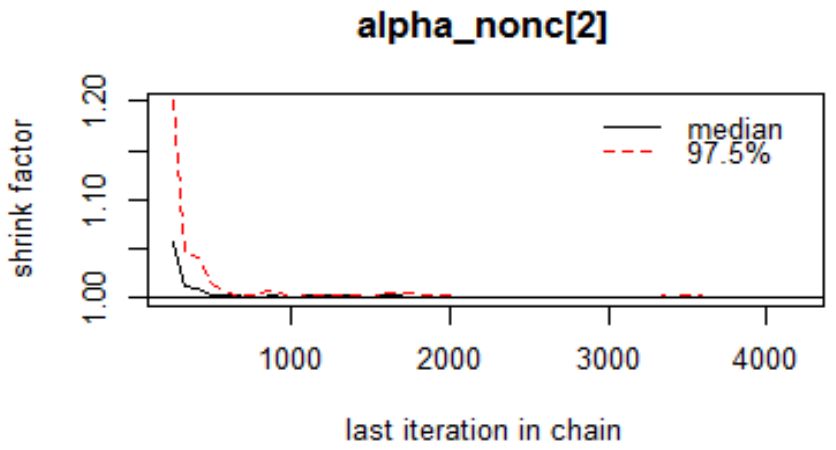
Visualisation des 3 chaînes de Markov

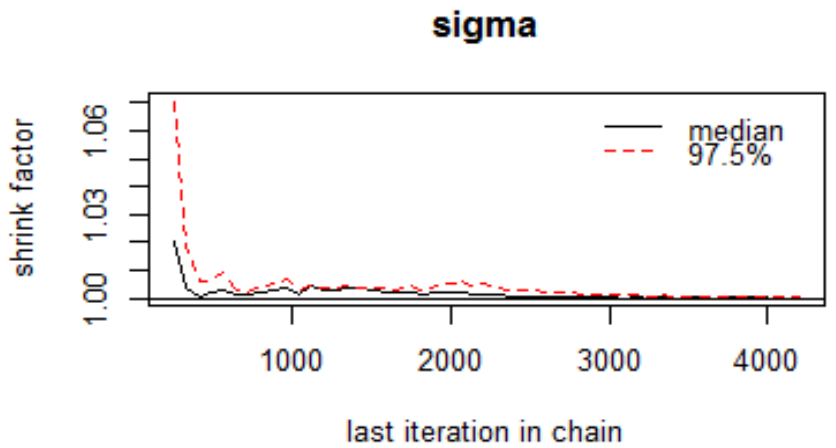


Résultats (SPAD)

Statistique de Gelman-Rubin

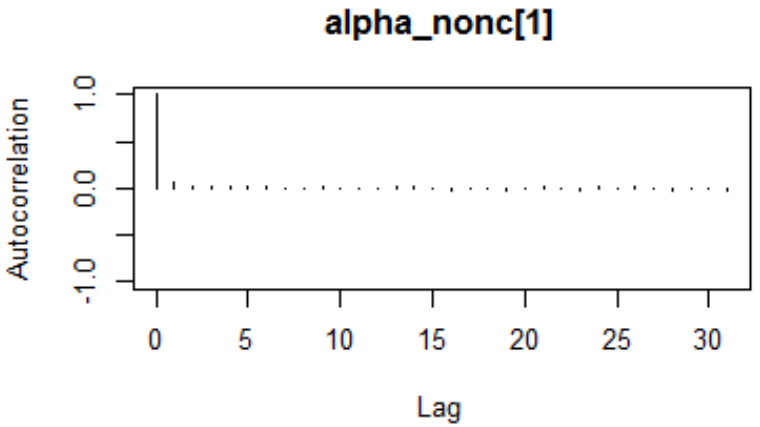


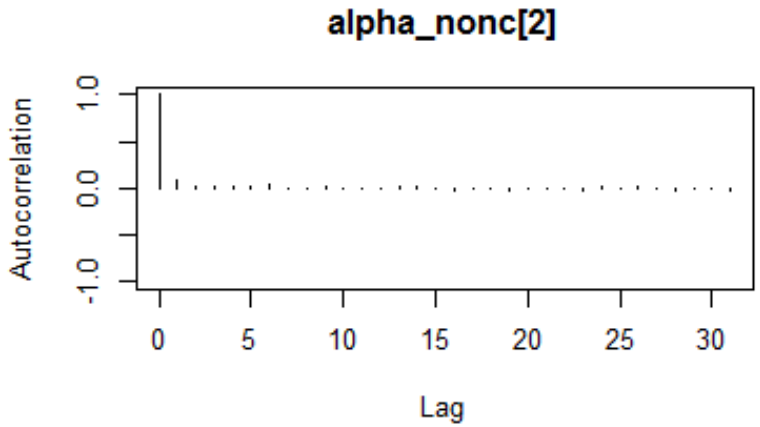


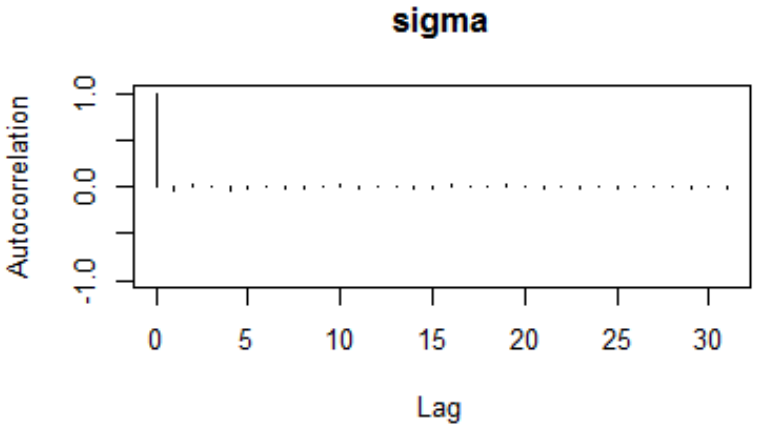


Résultats (SPAD)

Autocorrélations intra-chaîne (chaîne 1)







Modèle à 2 variables explicatives

```
teneur_bivarie <- "

data {
        int<lower=0> N;
        vector[N] INN_cr;
        vector[N] SPAD_cr;
        vector[N] Y;
        real m_INN;
        real<lower=0> s_INN;
        real m_SPAD;
        real<lower=0> s_SPAD;
        vector[3] mu;
        matrix[3,3] W;
}


parameters {
        vector[3] alpha;
        real<lower=0, upper=100> sigma;
//Loi a priori Uniforme sur [0,100]
}
model{
        Y ~normal(alpha[1]+alpha[2]*SPAD_cr+
                alpha[3]*INN_cr,sigma);
        alpha ~ multi_normal(mu,W);
 //Loi a priori normales
}
generated quantities {
        vector[3] alpha_nonc;
        vector[N] loglike;

        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(i in 1:N){
        loglike[i] = normal_lpdf(Y[i] |
        alpha[1]+alpha[2]*SPAD_cr[i]+
                alpha[3]*INN_cr[i], sigma);
}}
"
```

Modèle sans variable explicative

```
### code du modele

teneur_sanscov <- "
data {
        int<lower=0> N;
        vector[N] Y;
        real a;
        real b;
}

parameters {
        real mu;
        real<lower=0, upper=100> sigma;
}

model{
        Y ~ normal(mu,sigma);
        mu ~ normal(a,b);
}

generated quantities {
        vector[N] loglike;

        for(i in 1:N){
                loglike[i] = normal_lpdf(Y[i] |
                        mu, sigma);
        }
}

"
```

Questions

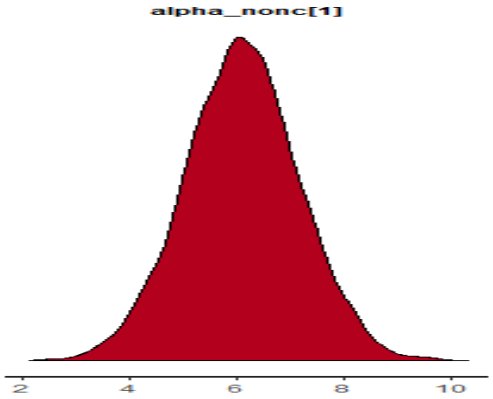
8- Représenter les densités a posteriori des paramètres et résumer ces densités par leurs moyennes, médianes, et quelques percentiles.

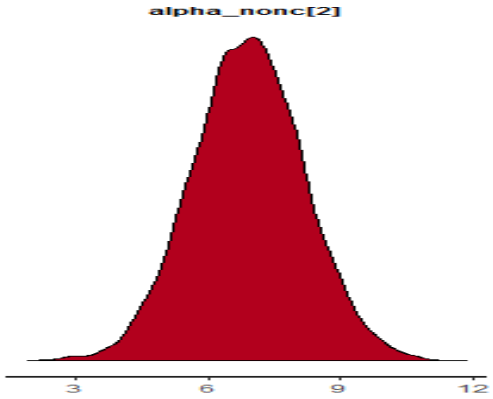
9- Calculer le DIC des différents modèles proposés et identifier le modèle qui a le DIC le plus faible

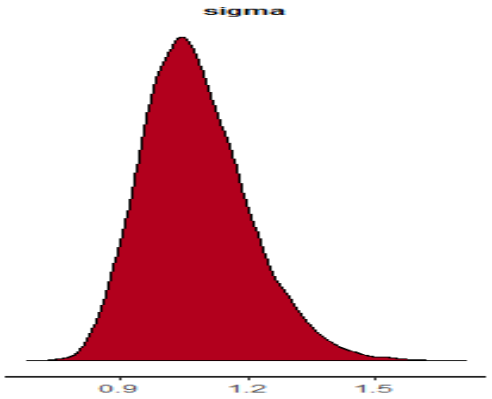
10- Calculer le WAIC des différents modèles proposés et identifier le modèle qui a le WAIC le plus faible

Densités a posteriori

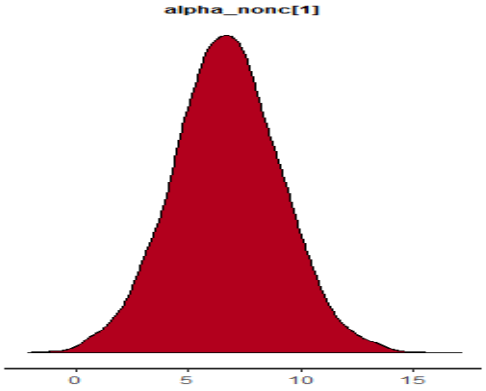
M_1
INN

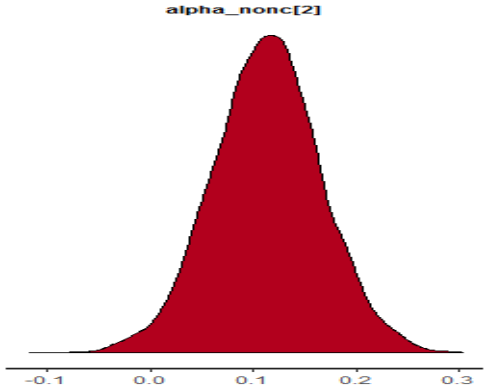


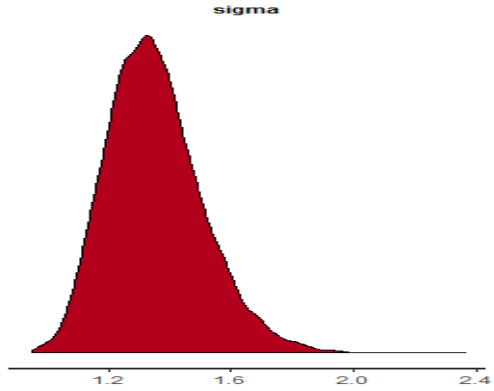




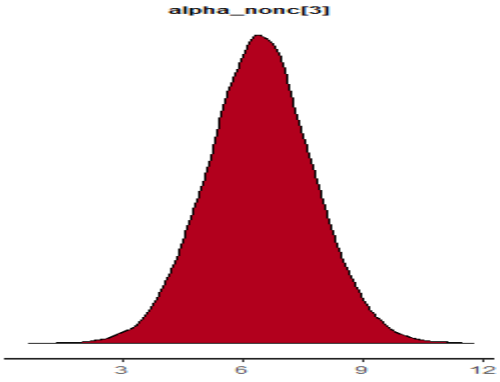
M_1
SPAD

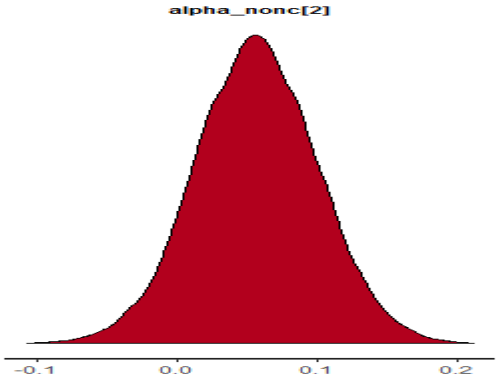


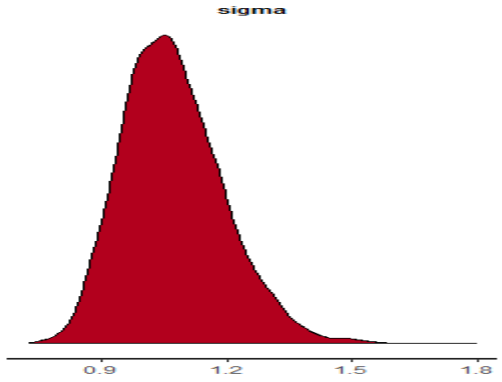




M_2
INN+SPAD







Statistiques a posteriori

node	mean	sd	2.5%	median	97.5%
M_1 - INN					
alpha_nonc[1]	6.09	1.07	3.98	6.08	8.20
alpha_nonc[2]-INN	6.90	1.26	4.42	6.90	9.38
sd	1.08	0.12	0.87	1.07	1.35
M_1 - SPAD					
alpha_nonc[1]	6.69	2.38	2.04	6.67	11.3
alpha_nonc[2]-SPAD	0.11	0.05	0.01	0.11	0.21
sd	1.35	0.15	1.09	1.34	1.69
M_2 - INN + SPAD					
alpha_nonc[1]	3.90	1.99	-0.05	3.91	7.73
alpha_nonc[3] - INN	6.43	1.33	3.84	6.42	9.04
alpha_nonc[2]-SPAD	0.06	0.04	-0.03	0.06	0.14
sd	1.07	0.12	0.86	1.06	1.33
M_3 - sanscov					
sd	1.41	0.16	1.14	1.40	1.77

DIC et WAIC

	M_1 INN	M_1 SPAD	M_2 INN+SPAD	M_0
DIC	129.8	149.2	130.0	152.3
WAIC	130.2	150.8	131.4	152.7

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

```
teneur_pred <- «

data {
        int<lower=0> N;
        vector[N] covar_cr;
        vector[N] Y;
        real covar_pred;
        real m;
        real<lower=0> s;
        vector[2] mu;
        matrix[2,2] W;
}

parameters {
        vector[2] alpha;
        real<lower=0, upper=100> sigma;
}

model{
        Y ~ normal(alpha[1]+alpha[2]*covar_cr,sigma);
        alpha ~ multi_normal(mu,W);
}

generated quantities {
        real Y_pred;
        Y_pred = normal_rng(alpha[1]+alpha[2]*((covar_pred
-m)/s),sigma);
        }
"
```

```
data =list(Y=Y,covar_cr=INN_cr,N=as.integer(43),m=mean(INN),s=sd(INN),covar_pred=0.95,
mu=c(0,0),W=matrix(c(1000,0,0,1000),nr=2,ncol=2)),
```

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

	mean	sd	2.5%	50%	97.5%
Y_pred	12.63	1.13	10.42	12.62	14.90