

Statistique bayésienne

Executive Master Statistics and Big Data

Yannick Le Pen

06/06/2022

Contents

1	Introduction	2
2	Données	2
3	Régression linéaire initiale	4
4	Recherche d'une meilleure spécification	7
4.1	Préselection des variables	7
4.2	Recherche d'une spécification	12
4.2.1	Régression 1	12
4.2.2	Régression 2	15
4.2.3	Régression 3	17
4.3	Regression finale	19
4.3.1	Comparaison des modèles par les facteurs de Bayes	20
4.4	Effet des matières et des Etablissements	21
4.5	Comparaison avec l'estimation par les MCO	22
5	Difference des conditions entre Anglais et Maths	23
5.1	Barre des mutations pour l'anglais	23
5.2	Barre des mutations pour les maths	24
6	Partie II Loi de Pareto	27
6.1	Loi de Pareto	27
6.2	Loi a posteriori du paramètre α	28
6.3	Echantillon tire de la loi a posteriori	28
6.4	Anglais et Mathématiques	30
6.4.1	Simulation de la distribution a posteriori pour $\alpha_{anglais}$	30
6.5	Simulation de la distribution a posteriori pour α_{math} et conclusion	32

1 Introduction

Nous voulons expliquer le nombre de points (variable Barre) nécessaires pour obtenir une mutation dans un lycée de l'Académie de Versailles. Nous disposons de variables de plusieurs types :

- ville
- établissement
- effectifs présents en séries l, es et s
- taux bruts de réussite en séries l, es et s
- effectifs en seconde et en première
- taux d'accès brut et attendu au bac en seconde et première
- taux brut et attendu de réussite pour toutes les séries

2 Données

Nous représentons les statistiques descriptives des variables numériques. Les valeurs de la variable Barre sont d'un ordre de grandeur plus élevé que les autres variables, dont beaucoup sont des taux. Nous décidons de transformer la variable Barre en prenant son logarithme. Cela change l'interprétation des coefficients.

Barre	effectif_presents_serie_l	effectif_presents_serie_es	effectif_presents_serie_s
Min. : 21.0	Min. : 6.00	Min. : 10.00	Min. : 13.0
1st Qu.: 111.0	1st Qu.: 18.00	1st Qu.: 53.00	1st Qu.: 64.0
Median : 196.0	Median : 30.00	Median : 69.00	Median :100.0
Mean : 321.9	Mean : 34.24	Mean : 74.42	Mean :106.1
3rd Qu.: 292.0	3rd Qu.: 47.00	3rd Qu.: 99.00	3rd Qu.:140.0
Max. :2056.0	Max. :133.00	Max. :192.00	Max. :328.0

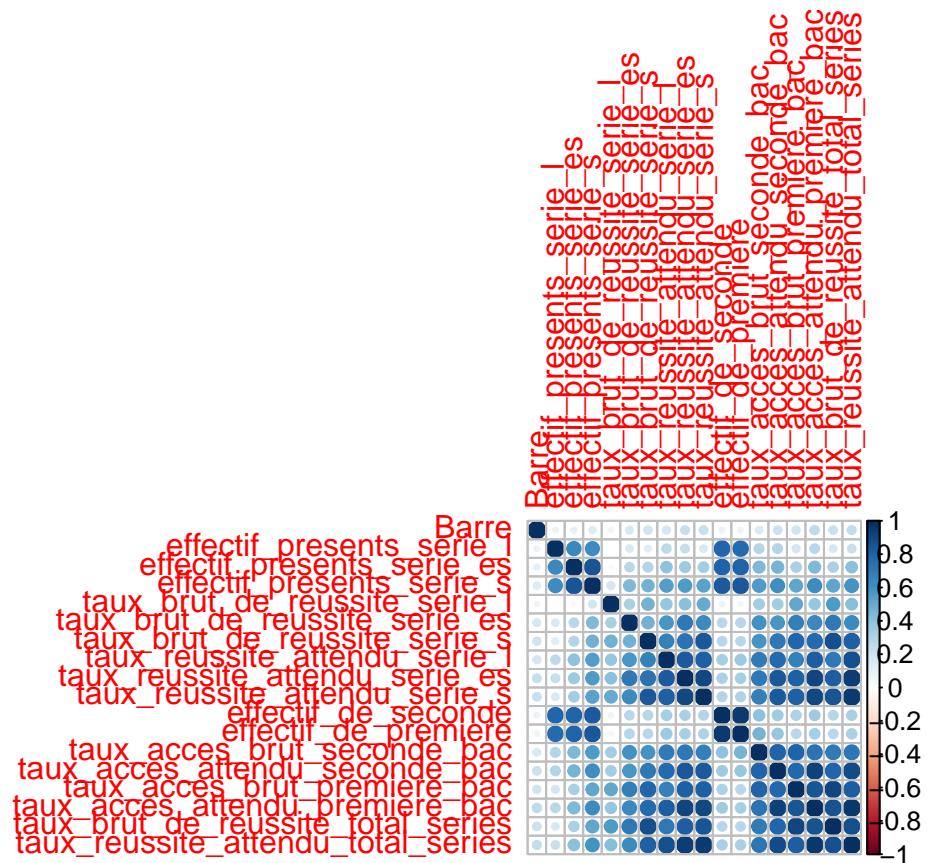
taux_brut_de_reussite_serie_l	taux_brut_de_reussite_serie_es	taux_brut_de_reussite_serie_s
Min. : 36.00	Min. : 51.0	Min. :50.00
1st Qu.: 82.00	1st Qu.: 81.0	1st Qu.:81.00
Median : 89.00	Median : 88.0	Median :88.00
Mean : 86.35	Mean : 86.4	Mean :86.23
3rd Qu.: 94.00	3rd Qu.: 94.0	3rd Qu.:93.00
Max. :100.00	Max. :100.0	Max. :99.00

taux_reussite_attendu_serie_l	taux_reussite_attendu_serie_es	taux_reussite_attendu_serie_s
Min. :65.00	Min. :61.00	Min. :61.00
1st Qu.:84.00	1st Qu.:86.00	1st Qu.:86.00
Median :89.00	Median :90.00	Median :89.00
Mean :86.91	Mean :87.97	Mean :87.39
3rd Qu.:92.00	3rd Qu.:94.00	3rd Qu.:94.00
Max. :98.00	Max. :98.00	Max. :98.00

effectif_de_seconde	effectif_de_premiere	taux_acces_brut_seconde_bac	taux_acces_attendu_seconde_bac
Min. : 36.0	Min. : 36.0	Min. :49.00	Min. :50.00
1st Qu.:268.0	1st Qu.:226.5	1st Qu.:64.00	1st Qu.:64.00
Median :336.0	Median :289.0	Median :71.00	Median :69.00
Mean :351.6	Mean :307.7	Mean :69.61	Mean :68.47
3rd Qu.:415.0	3rd Qu.:364.0	3rd Qu.:76.00	3rd Qu.:73.00
Max. :764.0	Max. :691.0	Max. :87.00	Max. :83.00

taux_acces_brut_premiere_bac	taux_acces_attendu_premiere_bac	taux_brut_de_reussite_total_series
Min. :65.00	Min. :70.00	Min. :64.00
1st Qu.:82.00	1st Qu.:81.00	1st Qu.:82.00
Median :85.00	Median :85.00	Median :86.00
Mean :84.53	Mean :84.19	Mean :85.46
3rd Qu.:89.25	3rd Qu.:89.00	3rd Qu.:91.00
Max. :97.00	Max. :94.00	Max. :98.00

x
Min. :67.0
1st Qu.:84.0
Median :88.0
Mean :86.8
3rd Qu.:92.0
Max. :98.0



Nous analysons les corrélations entre les variables numériques. La variable Barre (en logarithme) est positivement corrélée avec les autres variables mais, d'après le graphique, le niveau des corrélations se situe autour de 0.2. Il n'y a donc pas de variable qui serait fortement corrélé à Barre. Nous pouvons remarquer différents blocs dans l'ensemble des corrélations :

- Les effectifs présents en séries 1, es et l sont corrélés logiquement avec les effectifs en seconde,
- Les taux de réussite bruts et attendus sont corrélés avec les taux d'accès bruts et attendus
- Les taux de réussite bruts sont corrélés entre eux
- Les taux d'accès et de réussite sont corrélés entre eux

Dans chaque cas, les corrélations sont positives et assez élevées (entre 0.6 et 0.8). Nous pouvons en déduire que des variables sont redondantes dans l'explication de Barre.

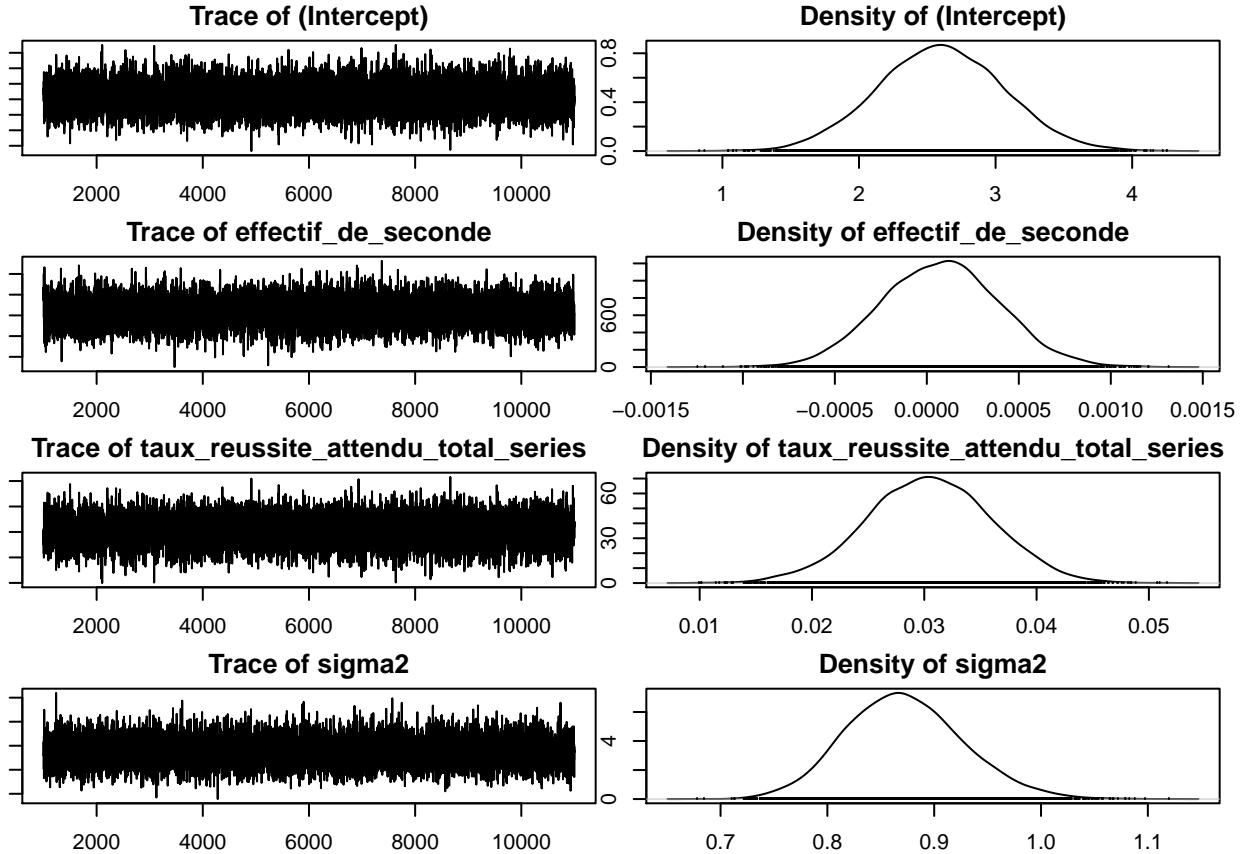
3 Régression linéaire initiale

Nous effectuons une première régression linéaire. Nous régressons la variable Barre (en logarithme) sur deux variables représentatives de nos données :

- effectif_de_seconde,
- taux_reussite_attendu_total_series.

Nous nous plaçons dans le cas où les priors des coefficients des regresseurs suivent une loi normale et celui de la variance du résidu une loi Gamma, ces deux lois étant indépendantes. La fonction mcmcregress du package MCMCpack utilise l'algorithme Gibbs sampler pour simuler la distribution a posteriori. Nous conservons les valeurs par défaut du nombre d'itération soit 1000 pour le burnin et 10000 pour le nombre d'itérations après le burnin. Afin de pouvoir par la suite calculer les facteurs de Bayes, nous spécifions des priors. Dans ces priors nous supposons que la moyenne des coefficients de la régression est égale à 0 et

```
##  
## Iterations = 1001:11000  
## Thinning interval = 1  
## Number of chains = 1  
## Sample size per chain = 10000  
  
##  
## 1. Empirical mean and standard deviation for each variable,  
## plus standard error of the mean:  
##  
##  
##  
## (Intercept) 2.5979872 0.4591701 4.592e-03 4.602e-03  
## effectif_de_seconde 0.0000749 0.0003215 3.215e-06 3.215e-06  
## taux_reussite_attendu_total_series 0.0304247 0.0055432 5.543e-05 5.575e-05  
## sigma2 0.8719626 0.0548676 5.487e-04 5.487e-04  
  
##  
## 2. Quantiles for each variable:  
##  
##  
## (Intercept) 2.5% 25% 50% 75%  
## (Intercept) 1.7098830 2.2867311 2.596e+00 2.9087285  
## effectif_de_seconde -0.0005447 -0.0001476 7.824e-05 0.0002881  
## taux_reussite_attendu_total_series 0.0193549 0.0266697 3.042e-02 0.0341522  
## sigma2 0.7703096 0.8336968 8.696e-01 0.9071835  
  
## 97.5%  
## (Intercept) 3.5039633  
## effectif_de_seconde 0.0007151  
## taux_reussite_attendu_total_series 0.0410998  
## sigma2 0.9870214
```



Les distributions des coefficients a posteriori montrent que la moyenne a posteriori des deux coefficients est positive, avec un ordre de grandeur bien plus élevé pour `taux_reussite_attendu_total_series`. Si nous comparons les moyennes a posteriori avec les écarts types correspondants, nous voyons l'écart type pour `effectif_de_seconde` est nettement plus élevé que la moyenne a posteriori, ce qui nous conduit à considérer cette variable comme peu significative. Dans le cas du `taux_reussite_attendu_total_series`, l'écart type estimé nettement 18 fois plus petit, ce qui est favorable à la significativité de cette variable. Ces constats sont confirmés par l'étude des quartiles qui sont tous positif pour `taux_reussite_attendu_total_series` et de signes négatifs, au moins jusqu'au quantile d'ordre 25%, pour `effectif_de_seconde` et positif ensuite. L'étude des densités a posteriori montrent que la densité de `effectif_de_seconde` est presque centrée en 0. La densité de `taux_de_reussite_attendu` est centrée autour de 0.3 et nettement au-dessus de zéro.

Les résultats précédents sont conditionnés à la précision de l'estimation de la distribution a posteriori. Concernant la précision des l'estimation de la moyenne a posteriori des coefficients, nous observons que les écarts types mesurant cette précision, reportés dans Naive SE et Time-Series SE sont suffisamment petits (environ 100 plus petits que les écarts types des distributions a posteriori) pour que la qualité de l'estimation soit jugée comme suffisante.

Les graphiques “trace” des chaines MCMC montrent que les trajectoires de ces chaines balayent bien l'espace des valeurs possibles des coefficients.

Les diagnostics relatifs au nombre d'itérations minimum montrent que les paramètres retenus pour la taille du burnin et l'estimation de la distribution a posteriori sont suffisants.

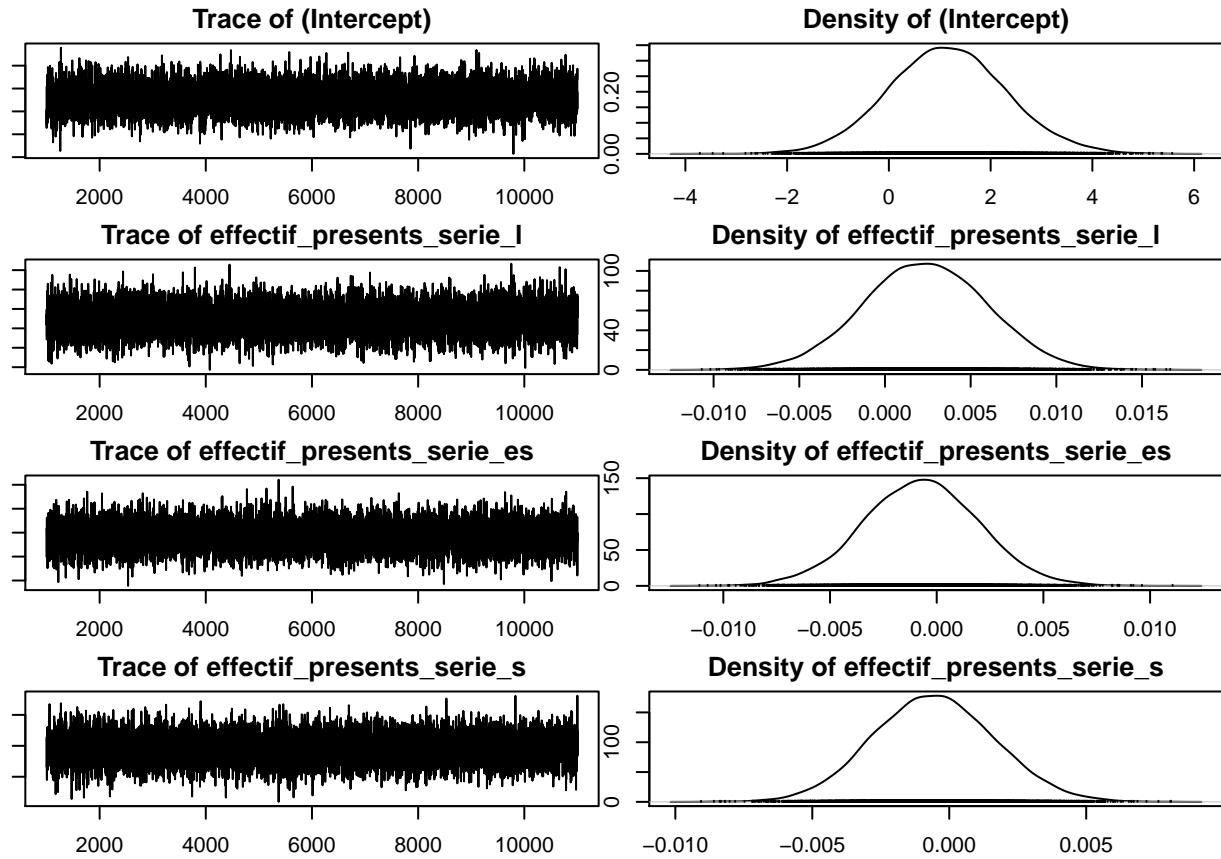
```
##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
##                                     Burn-in   Total Lower bound Dependence
##
```

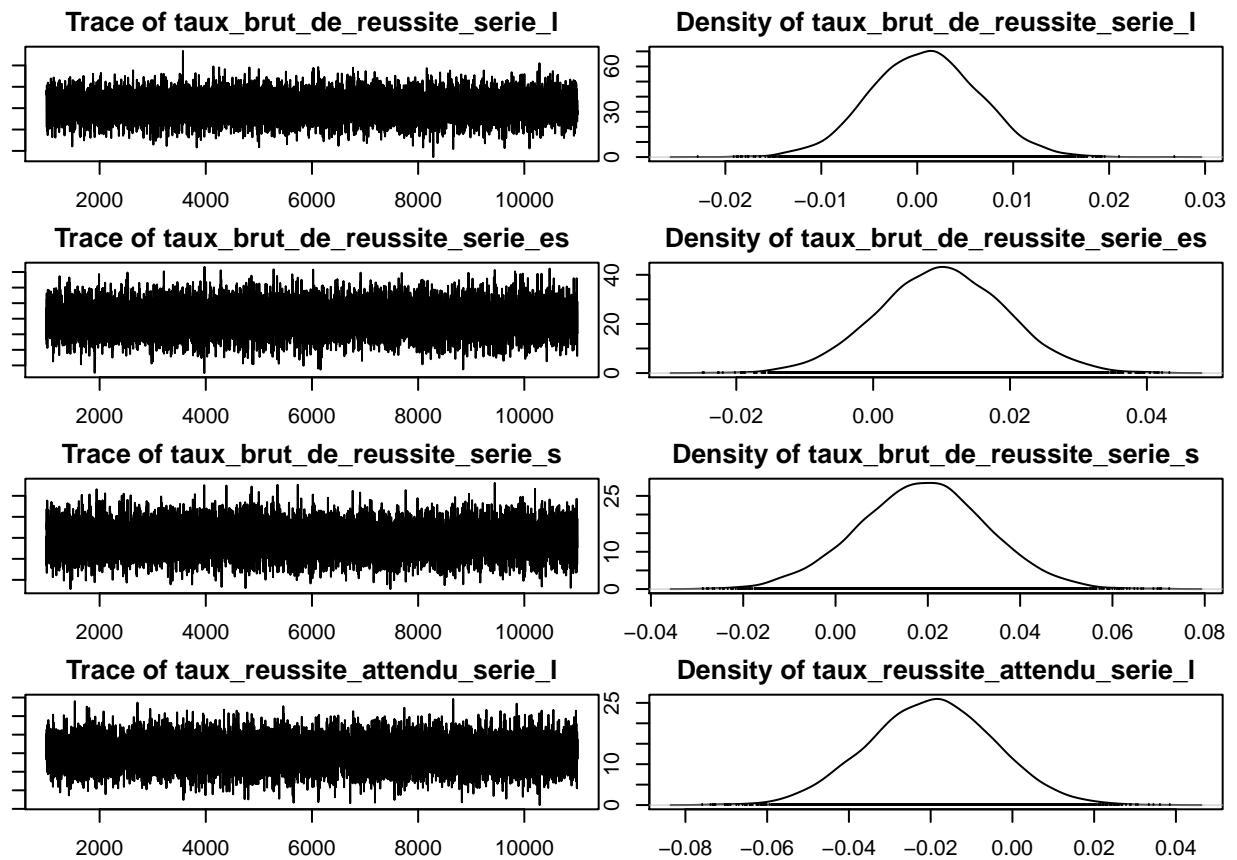
```
##                               (M)      (N)    (Nmin)      factor (I)
## (Intercept)                  2       3741   3746       0.999
## effectif_de_seconde         2       3741   3746       0.999
## taux_reussite_attendu_total_series 2       3650   3746       0.974
## sigma2                      2       3834   3746       1.020
##                               (Intercept)      effectif_de_seconde
##                               9953.473        10000.000
## taux_reussite_attendu_total_series
##                               9885.599        10000.000
```

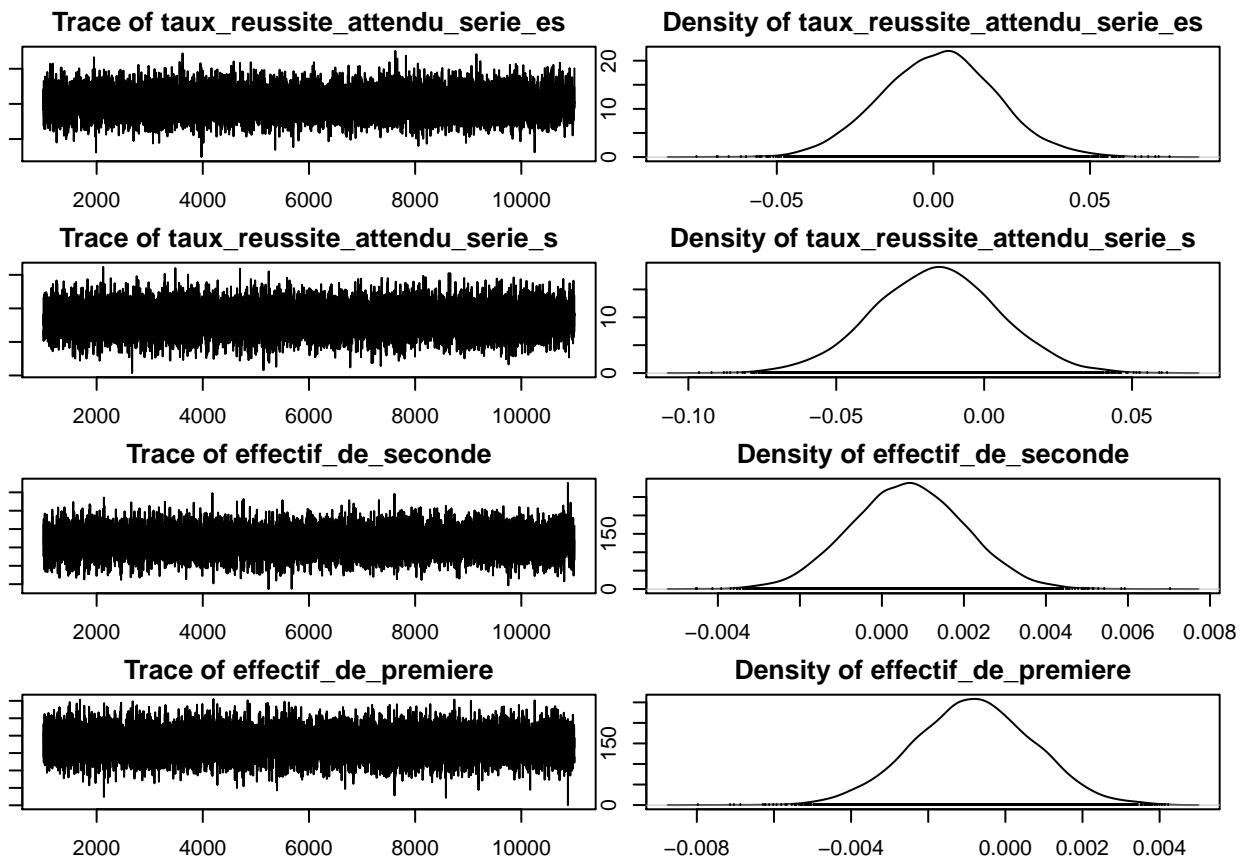
4 Recherche d'une meilleure spécification

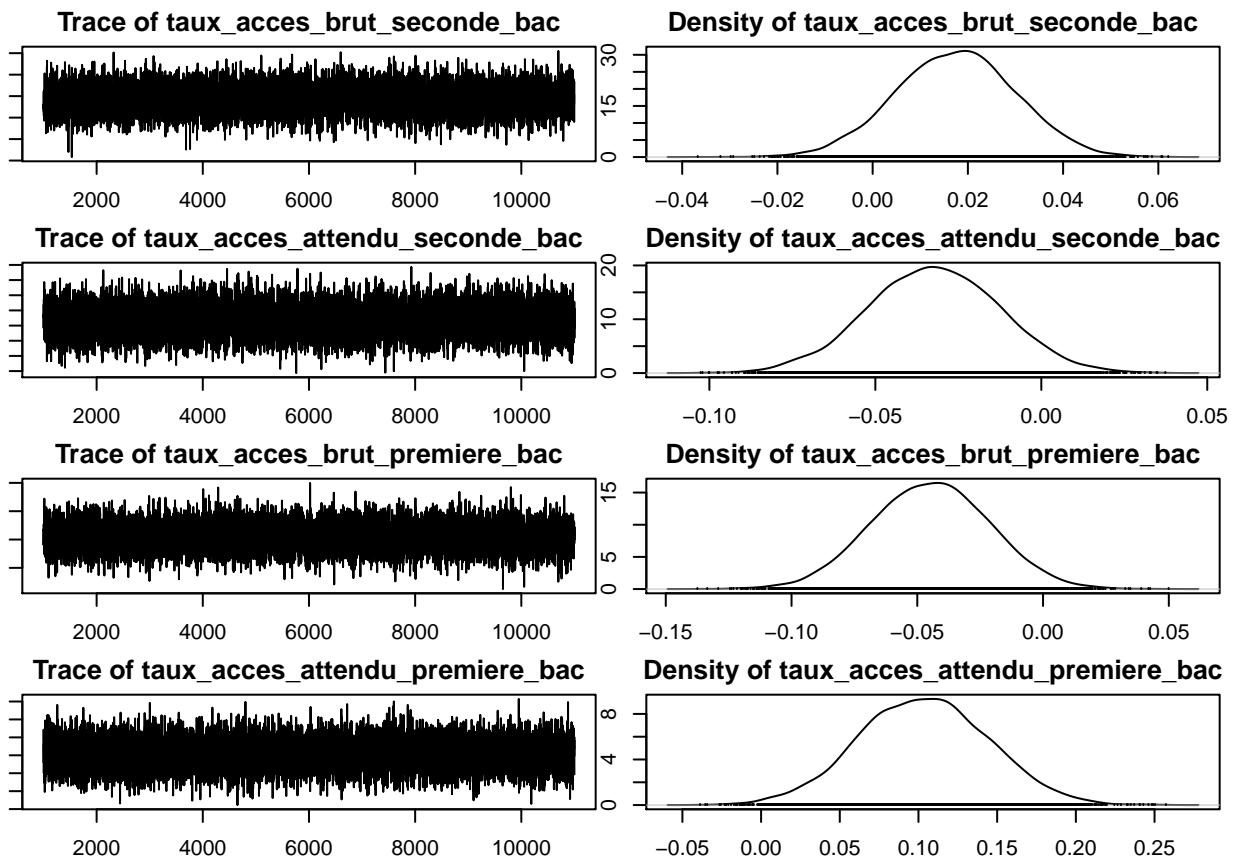
4.1 Préselection des variables

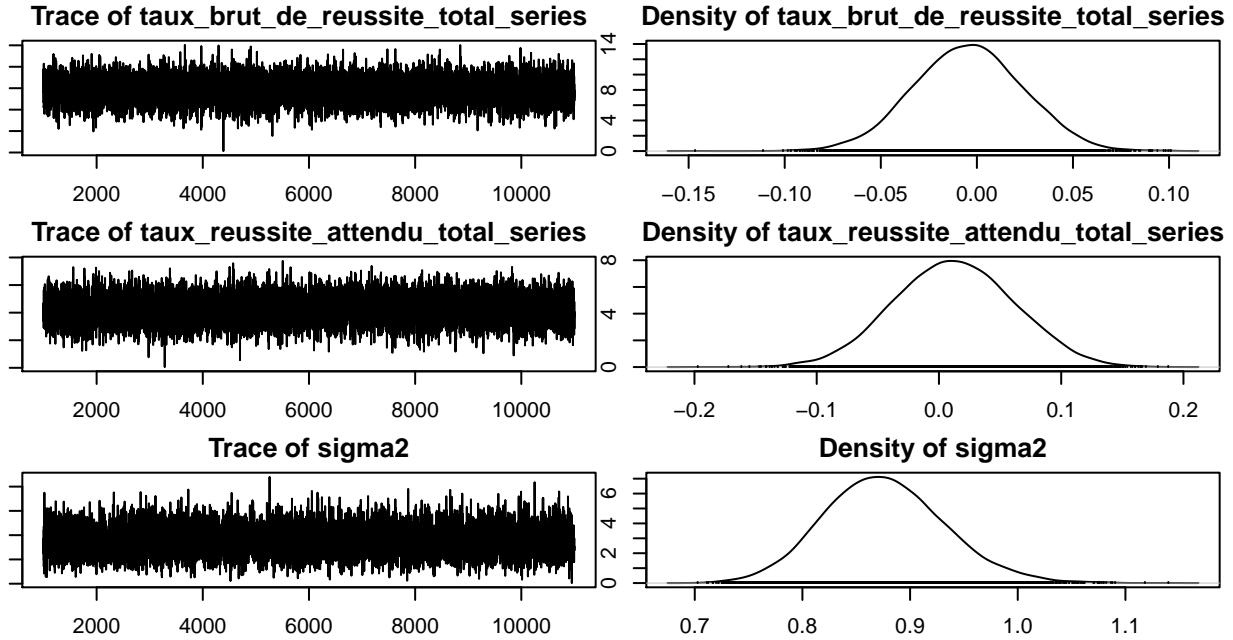
Nous commençons par une estimation de la régression de la variable Barre (en logarithme) sur toutes les autres variables numériques simultanément. Nous représentons juste les densités a posteriori des coefficients.











Les distributions a posteriori montrent que :

- Les effectifs présents par série et les effectifs en seconde et en première ont tous des distributions a posteriori qui sont centrées autour de 0.
- La variable taux_acces_brut_seconde_bac a un effet positif.
- Les variables taux_acces_attendu_seconde_bac et taux_acces_brut_premiere_bac ont un effet négatif.
- La variable taux_acces_attendu_premiere_bac a un effet positif et significatif.
- Les deux variables taux_reussite brut et attendu totaux ou par série n'ont apparemment pas d'effet lorsqu'elles sont considérées isolément.
- Ce constat est confirmé par l'étude des quantiles et du graphique de la distribution a posteriori. Nous observons que pour "taux_acces_attendu_premiere_bac" seulement, tous les quantiles sont positifs.

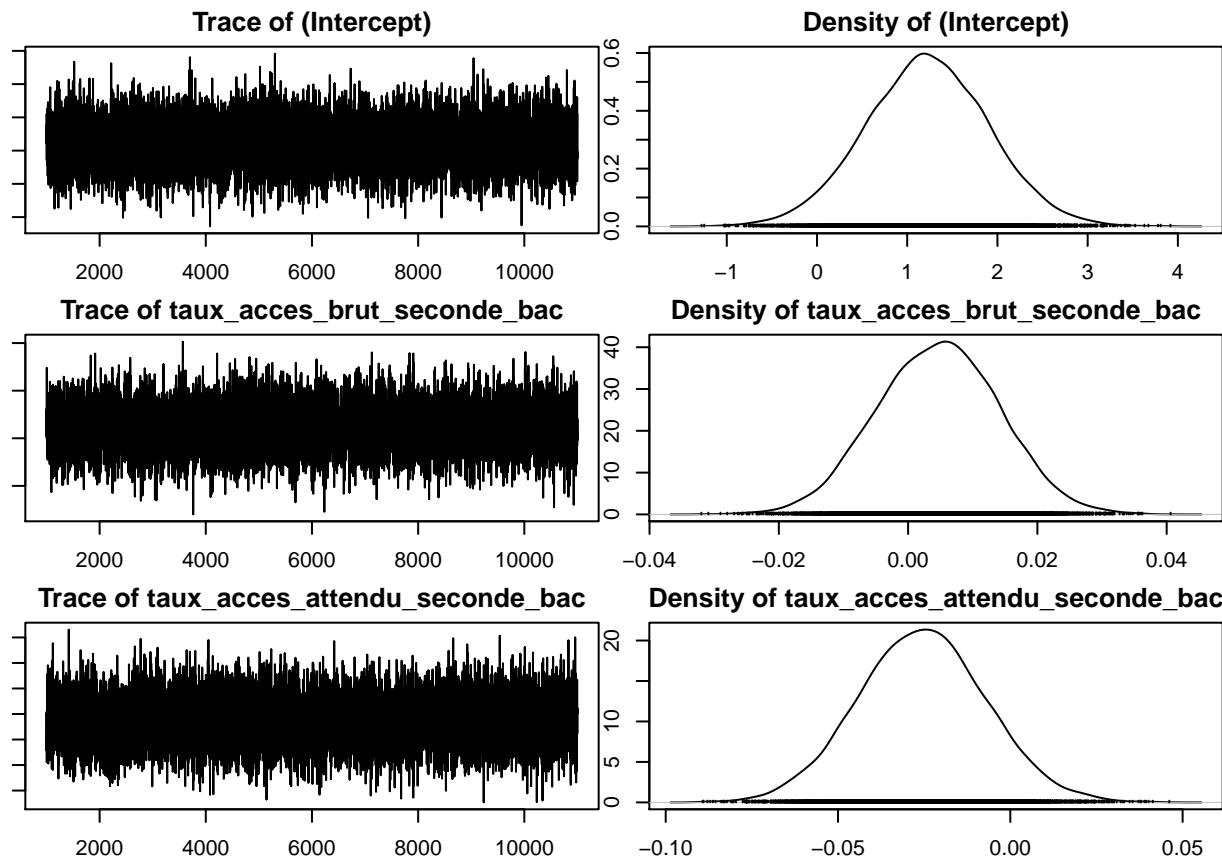
Si nous régressions la variable Barre sur chaque variable numérique. Les distributions a posteriori (non représentées par souci de place) montrent que les effectifs présents comme les effectifs en seconde et en première sont assez peu significatif. La distribution a posteriori des coefficients de ces variables est très concentrée autour de 0. Nous décidons donc des les exclure. Nous remarquons par contre que la distribution des variables des taux d'accès et des taux de réussite attendu montrent que les coefficients sont très probablement supérieurs à 0.

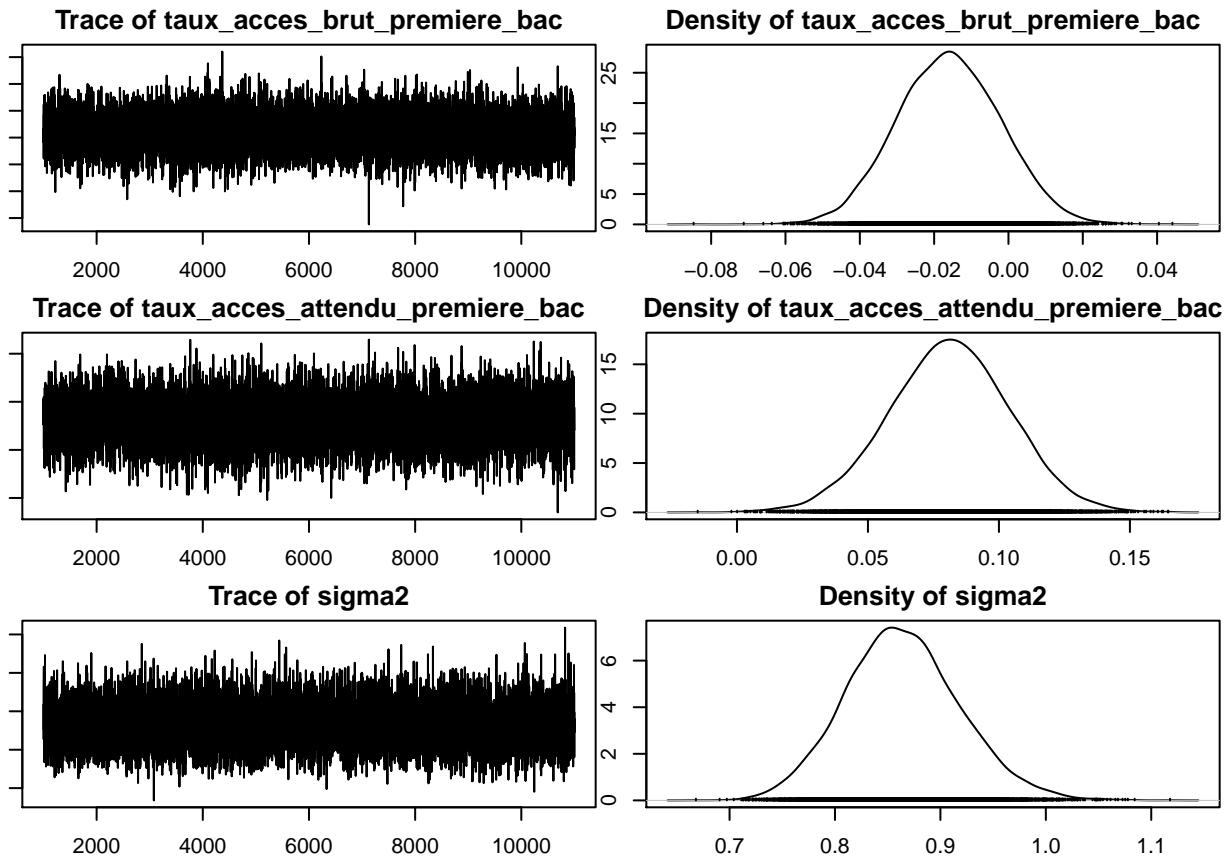
4.2 Recherche d'une spécification

4.2.1 Regression 1

On considère une première régression sur les taux d'accès au bac, en première ou en seconde, brut ou attendu, soit les variables qui sont apparues comme les plus significatives.

```
##  
## Iterations = 1001:11000  
## Thinning interval = 1  
## Number of chains = 1  
## Sample size per chain = 10000  
##  
## 1. Empirical mean and standard deviation for each variable,  
##     plus standard error of the mean:  
##  
##  
## (Intercept)      Mean        SD  Naive SE Time-series SE  
## taux_acces_brut_seconde_bac 0.004635 0.009457 9.457e-05 9.457e-05  
## taux_acces_attendu_seconde_bac -0.025571 0.018485 1.849e-04 1.849e-04  
## taux_acces_brut_premiere_bac -0.015959 0.013897 1.390e-04 1.390e-04  
## taux_acces_attendu_premiere_bac 0.081042 0.023055 2.305e-04 2.305e-04  
## sigma2           0.864241 0.054367 5.437e-04 5.437e-04  
##  
## 2. Quantiles for each variable:  
##  
##  
## (Intercept)      2.5%       25%       50%       75%    97.5%  
## taux_acces_brut_seconde_bac -0.12180  0.762609  1.218586  1.680688 2.55884  
## taux_acces_attendu_seconde_bac -0.01377 -0.001789  0.004778  0.011074 0.02295  
## taux_acces_brut_premiere_bac -0.06156 -0.038147 -0.025618 -0.013391 0.01101  
## taux_acces_attendu_premiere_bac -0.04230 -0.025515 -0.016051 -0.006519 0.01120  
## sigma2           0.03517  0.065836  0.081289  0.096388 0.12590  
##  
## sigma2           0.76259  0.827292  0.862156  0.898794 0.97752
```





```
##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
##                                     Burn-in   Total Lower bound Dependence
##                                     (M)      (N)    (Nmin) factor (I)
## (Intercept)                      2        3710  3746     0.990
## taux_acces_brut_seconde_bac     2        3834  3746     1.020
## taux_acces_attendu_seconde_bac  2        3802  3746     1.010
## taux_acces_brut_premiere_bac   2        3929  3746     1.050
## taux_acces_attendu_premiere_bac 2        3710  3746     0.990
## sigma2                           2        3680  3746     0.982
##
## (Intercept)      taux_acces_brut_seconde_bac
##                   10000                            10000
## taux_acces_attendu_seconde_bac  taux_acces_brut_premiere_bac
##                   10000                            10000
## taux_acces_attendu_premiere_bac          sigma2
##                   10000                            10000
```

Les résultats des estimations permettent de tirer les conclusions suivantes:

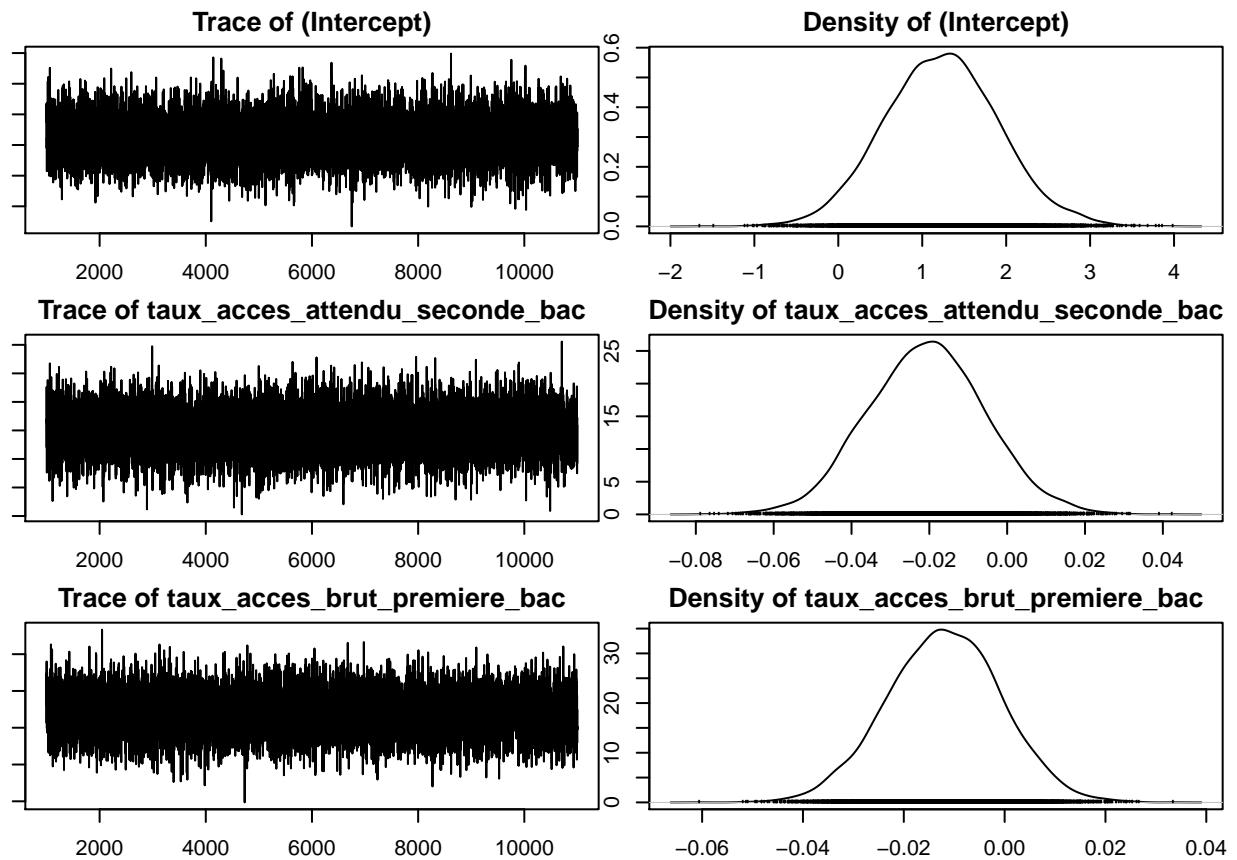
- la variable taux_acces_brut_premiere_bac possède une distribution a posteriori nettement au dessus de zero. On peut donc conclure que l'effet de cette variable est significativement positif.
- la variable taux_acces_brut_seconde_bac a une distribution a posteriori centrée autour de 0
- les deux autres variables ont plutôt un effet négatif même si une petite partie de leur distribution se situe au-dessus de 0.

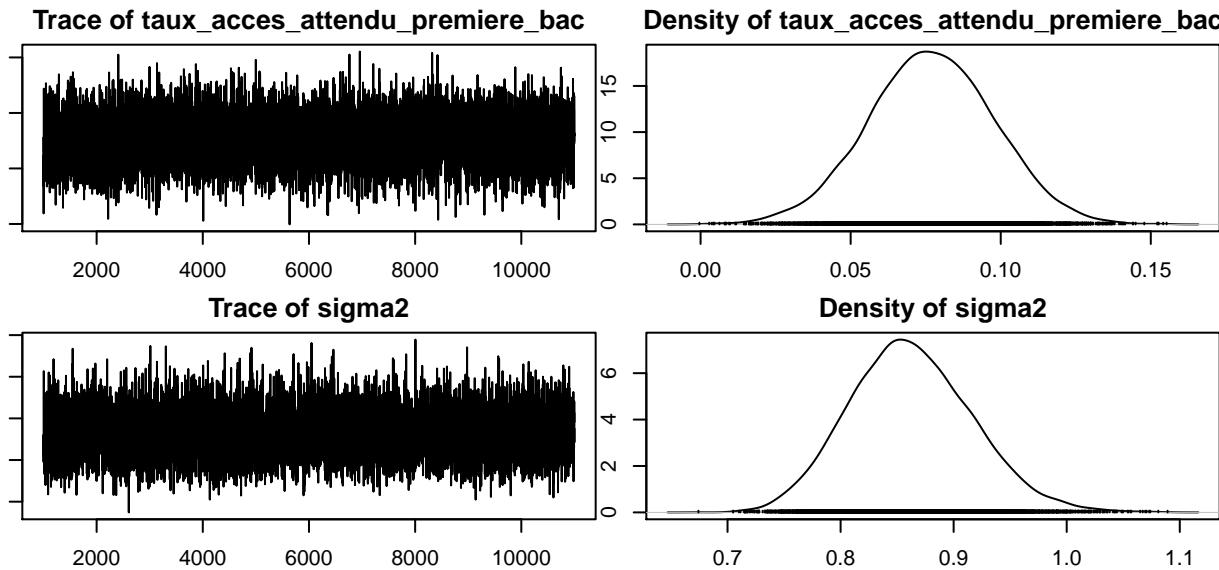
Les différents diagnostics sur la précision de l'estimation et l'algorithme de simulation ne révèlent pas de problème particulier.

4.2.2 Régression 2

Nous estimons le modèle en excluant taux_acces_brut_seconde. Les résultats pour les autres variables ne sont pas très différents de ceux obtenus précédemment.

```
##  
## Iterations = 1001:11000  
## Thinning interval = 1  
## Number of chains = 1  
## Sample size per chain = 10000  
##  
## 1. Empirical mean and standard deviation for each variable,  
## plus standard error of the mean:  
##  
##  
##  
## (Intercept) 1.23002 0.68084 0.0068084 0.0068084  
## taux_acces_attendu_seconde_bac -0.02073 0.01513 0.0001513 0.0001513  
## taux_acces_brut_premiere_bac -0.01202 0.01124 0.0001124 0.0001108  
## taux_acces_attendu_premiere_bac 0.07688 0.02104 0.0002104 0.0002104  
## sigma2 0.86231 0.05414 0.0005414 0.0005607  
##  
## 2. Quantiles for each variable:  
##  
## 2.5% 25% 50% 75% 97.5%  
## (Intercept) -0.07741 0.76964 1.23413 1.681684 2.603087  
## taux_acces_attendu_seconde_bac -0.04996 -0.03092 -0.02059 -0.010614 0.009350  
## taux_acces_brut_premiere_bac -0.03410 -0.01963 -0.01193 -0.004358 0.009696  
## taux_acces_attendu_premiere_bac 0.03525 0.06284 0.07684 0.091048 0.118060  
## sigma2 0.76220 0.82454 0.86004 0.897426 0.974716
```





```
##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
##                                     Burn-in   Total Lower bound Dependence
##                                     (M)      (N)    (Nmin)   factor (I)
## (Intercept)                      2       3802   3746     1.010
## taux_acces_attendu_seconde_bac  2       3834   3746     1.020
## taux_acces_brut_premiere_bac   2       3741   3746     0.999
## taux_acces_attendu_premiere_bac 2       3929   3746     1.050
## sigma2                           2       3710   3746     0.990
##
## (Intercept)  taux_acces_attendu_seconde_bac
## 10000.000          10000.000
## taux_acces_brut_premiere_bac  taux_acces_attendu_premiere_bac
## 10293.597           10000.000
## sigma2
## 9326.557
```

4.2.3 Régression 3

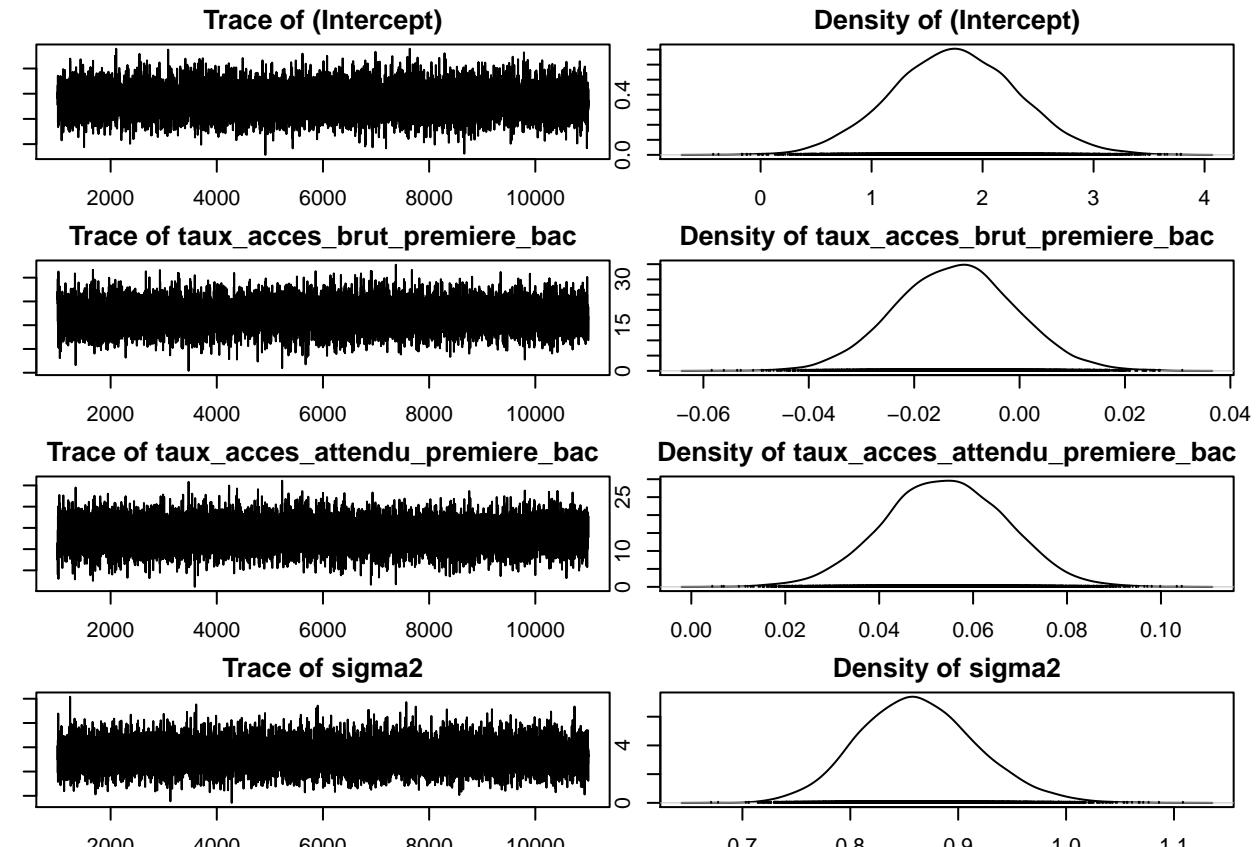
Nous excluons maintenant taux_acces_attendu_seconde_bac, ce qui n'a pas d'effet sur la distribution a posteriori du coefficient des deux autres variables.

```
##
## Iterations = 1001:11000
```

```

## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##     plus standard error of the mean:
##
##                                     Mean      SD  Naive SE Time-series SE
## (Intercept)                 1.74857 0.56501 0.0056501      0.0056635
## taux_acces_brut_premiere_bac -0.01217 0.01135 0.0001135      0.0001135
## taux_acces_attendu_premiere_bac 0.05399 0.01296 0.0001296      0.0001296
## sigma2                      0.86298 0.05430 0.0005430      0.0005430
##
## 2. Quantiles for each variable:
##
##                                     2.5%    25%    50%    75%   97.5%
## (Intercept)                 0.65580 1.36604 1.74593 2.130965 2.86329
## taux_acces_brut_premiere_bac -0.03425 -0.01988 -0.01205 -0.004637 0.01018
## taux_acces_attendu_premiere_bac 0.02866 0.04524 0.05398 0.062940 0.07881
## sigma2                      0.76240 0.82510 0.86063 0.897860 0.97690

```



```

##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
```

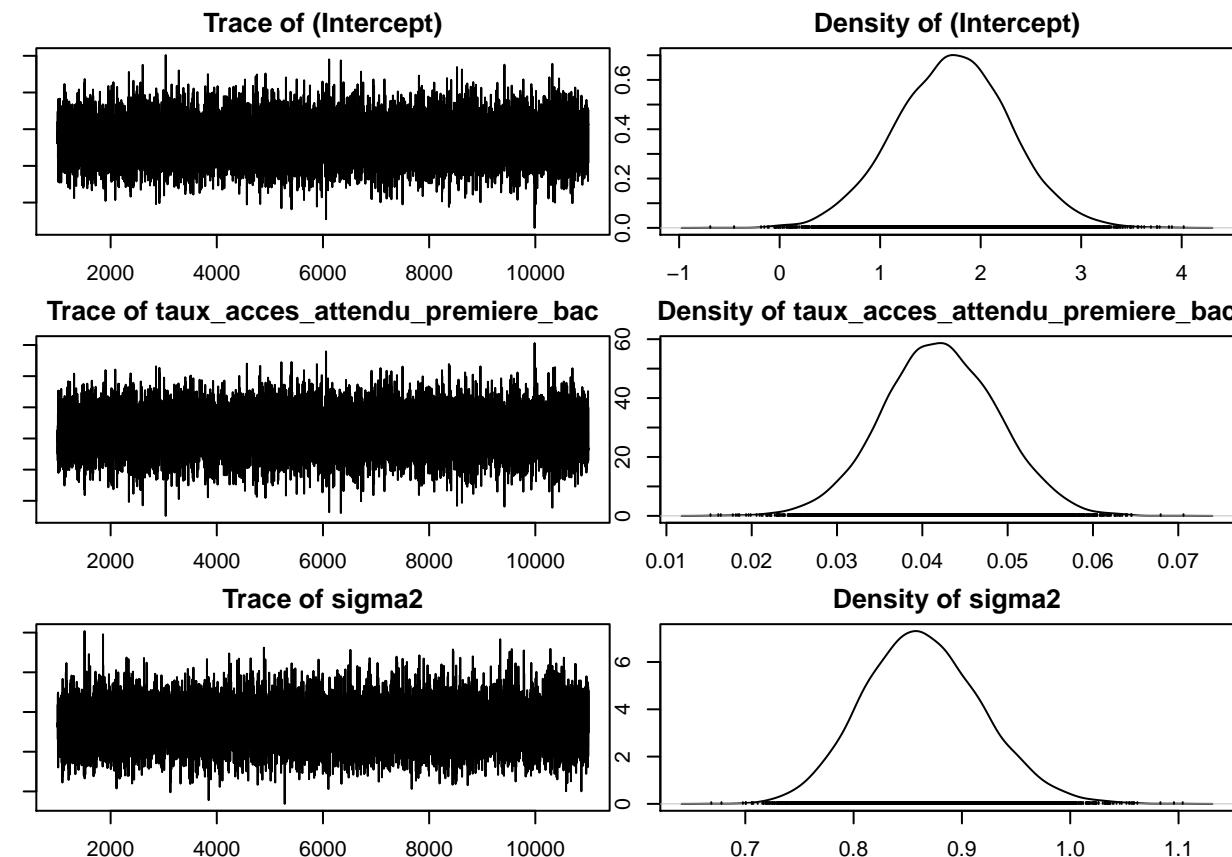
```

##                                     Burn-in   Total Lower bound Dependence
##                                     (M)      (N)    (Nmin) factor (I)
## (Intercept)                      2       3741   3746     0.999
## taux_acces_brut_premiere_bac   2       3834   3746     1.020
## taux_acces_attendu_premiere_bac 2       3771   3746     1.010
## sigma2                           2       3834   3746     1.020
##                               (Intercept)  taux_acces_brut_premiere_bac
##                               9952.575          10000.000
## taux_acces_attendu_premiere_bac           sigma2
##                               10000.000          10000.000

```

4.3 Regression finale

Nous considérons une dernière régression sur `taux_acces_attendu_premiere_bac` uniquement. La loi a posteriori du coefficient est très nettement au-dessus de zéro avec une valeur moyenne a posteriori égale à 0.04



```

##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                     Mean      SD  Naive SE Time-series SE

```

```

## (Intercept) 1.72270 0.566668 5.667e-03 5.667e-03
## taux_acces_attendu_premiere_bac 0.04208 0.006723 6.723e-05 6.723e-05
## sigma2 0.86410 0.054450 5.445e-04 5.445e-04
##
## 2. Quantiles for each variable:
##
## 2.5% 25% 50% 75% 97.5%
## (Intercept) 0.60950 1.34152 1.72687 2.10310 2.83509
## taux_acces_attendu_premiere_bac 0.02893 0.03757 0.04201 0.04663 0.05525
## sigma2 0.76305 0.82627 0.86202 0.90001 0.97622

##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
## Burn-in Total Lower bound Dependence
## (M) (N) (Nmin) factor (I)
## (Intercept) 2 3741 3746 0.999
## taux_acces_attendu_premiere_bac 2 3741 3746 0.999
## sigma2 2 3802 3746 1.010

## (Intercept) taux_acces_attendu_premiere_bac
## 10000 10000
## sigma2
## 10000

```

4.3.1 Comparaison des modèles par les facteurs de Bayes

Afin de déterminer le modèle préféré nous calculons les facteurs de Bayes permettant de comparer les 4 dernières régressions. Les valeurs obtenues sont très nettement favorables à la dernière spécification correspondant à la régression de Barre sur taux_acces_attendu_premiere_bac.

```

## Loading required package: coda
## Loading required package: MASS
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
## ## Copyright (C) 2003-2022 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
## ##
## ## Support provided by the U.S. National Science Foundation
## ## (Grants SES-0350646 and SES-0350613)
## ##

## The matrix of Bayes Factors is:
## reg1 reg2 reg3 reg4
## reg1 1 3.32e-03 3.96e-05 2.49e-07
## reg2 301 1.00e+00 1.19e-02 7.49e-05
## reg3 25239 8.39e+01 1.00e+00 6.29e-03
## reg4 4015148 1.34e+04 1.59e+02 1.00e+00
##

## The matrix of the natural log Bayes Factors is:
## reg1 reg2 reg3 reg4
## reg1 0.00 -5.71 -10.14 -15.21
## reg2 5.71 0.00 -4.43 -9.50

```

```

## reg3 10.14  4.43   0.00  -5.07
## reg4 15.21  9.50   5.07   0.00
##
##  reg1 :
##  call =
## MCMCregress(formula = Barre ~ taux_acces_brut_seconde_bac + taux_acces_attendu_seconde_bac +
##              taux_acces_brut_premiere_bac + taux_acces_attendu_premiere_bac,
##              data = d, mcmc = 10000, b0 = 0, B0 = 0.1, c0 = 1, d0 = 0.2,
##              marginal.likelihood = "Laplace")
##
##  log marginal likelihood = -722.3017
##
##
##  reg2 :
##  call =
## MCMCregress(formula = Barre ~ taux_acces_attendu_seconde_bac +
##              taux_acces_brut_premiere_bac + taux_acces_attendu_premiere_bac,
##              data = d, mcmc = 10000, b0 = 0, B0 = 0.1, c0 = 1, d0 = 0.2,
##              marginal.likelihood = "Laplace")
##
##  log marginal likelihood = -716.5953
##
##
##  reg3 :
##  call =
## MCMCregress(formula = Barre ~ taux_acces_brut_premiere_bac +
##              taux_acces_attendu_premiere_bac, data = d, mcmc = 10000,
##              b0 = 0, B0 = 0.1, c0 = 1, d0 = 0.2, marginal.likelihood = "Laplace")
##
##  log marginal likelihood = -712.1655
##
##
##  reg4 :
##  call =
## MCMCregress(formula = Barre ~ taux_acces_attendu_premiere_bac,
##              data = d, b0 = 0, B0 = 0.1, c0 = 1, d0 = 0.2, marginal.likelihood = "Laplace")
##
##  log marginal likelihood = -707.0961

```

4.4 Effet des matières et des Etablissements

Nous ajoutons à la regression précédente les différentes matières (reg5). Nous observons que certaines matières exercent une effet négatif indubitable sur la barre de mutation : anglais, biologie biochimie, maths...

En ce qui concerne les établissements, nous n'obtenons aucun effet significatif à l'exception d'un effet positif pour le lycée Condorcet.

L'étude des facteurs de Bayes conduit cependant à privilégier le modèle avec taux_acces_attendu_primaire seulement.

```

## The matrix of Bayes Factors is:
##      reg4    reg5    reg6
## reg4 1.00e+00 5.55e+07 1.75e+60
## reg5 1.80e-08 1.00e+00 3.15e+52
## reg6 5.72e-61 3.18e-53 1.00e+00
##
```

```

## The matrix of the natural log Bayes Factors is:
##      reg4   reg5 reg6
## reg4    0.0   17.8 139
## reg5   -17.8   0.0 121
## reg6  -138.7 -120.9    0
##
## reg4 :
##   call =
## MCMCregress(formula = Barre ~ taux_acces_attendu_premiere_bac,
##   data = d, b0 = 0, B0 = 0.1, c0 = 1, d0 = 0.2, marginal.likelihood = "Laplace")
##
##   log marginal likelihood = -707.0961
##
##
## reg5 :
##   call =
## MCMCregress(formula = Barre ~ taux_acces_attendu_premiere_bac +
##   as.factor(Matiere), data = d, b0 = 0, B0 = 0.1, c0 = 1, d0 = 0.2,
##   marginal.likelihood = "Laplace")
##
##   log marginal likelihood = -724.9287
##
##
## reg6 :
##   call =
## MCMCregress(formula = Barre ~ taux_acces_attendu_premiere_bac +
##   as.factor(etablissement), data = d, b0 = 0, B0 = 0.1, c0 = 1,
##   d0 = 0.2, marginal.likelihood = "Laplace")
##
##   log marginal likelihood = -845.8094

```

4.5 Comparaison avec l'estimation par les MCO

Nous estimons la régression par les MCO de la variable Barre sur taux_acces_attendu_première_bac. Nous remarquons que le coefficient estimé est significativement différent de zéro pour un risque de première espèce de 1%. Nous remarquons aussi que sa valeur est très proche de la moyenne a posteriori du coefficient obtenu par la méthode bayésienne.

```

##
## Call:
## lm(formula = Barre ~ taux_acces_attendu_premiere_bac, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50342 -0.52743 -0.06094  0.42639  2.66655
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.778911  0.576646  3.085  0.00215 ***
## taux_acces_attendu_premiere_bac 0.041418  0.006832  6.062  2.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9284 on 514 degrees of freedom
## Multiple R-squared:  0.06672,    Adjusted R-squared:  0.06491

```

```
## F-statistic: 36.75 on 1 and 514 DF, p-value: 2.605e-09
```

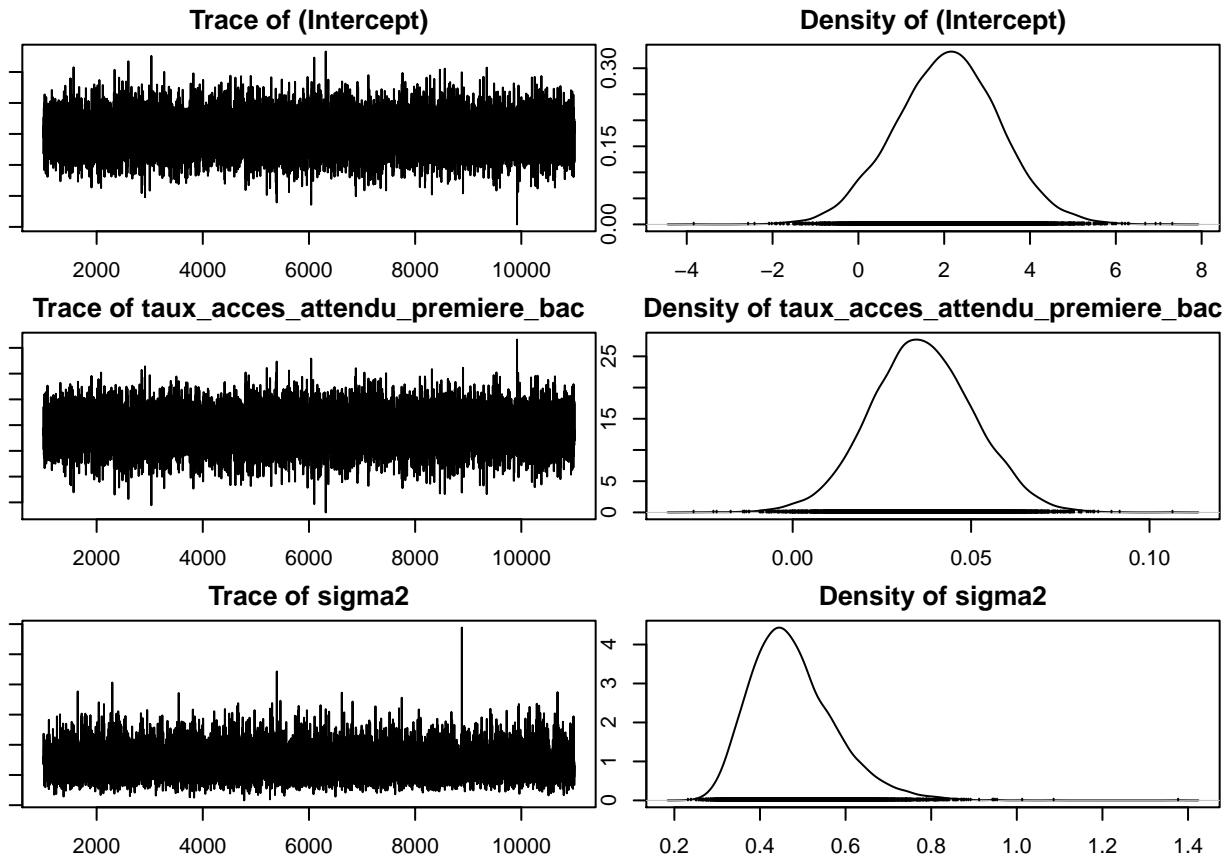
5 Difference des conditions entre Anglais et Maths

Nous séparons l'échantillon initial en deux sous-échantillons correspondant aux deux matières.

5.1 Barre des mutations pour l'anglais

Nous effectuons la régression de la variable Barre sur le taux d'accès attendu première bac. La distribution a posteriori montre que l'effet de cette variables positif et "significatif", la moyenne du coefficient a posteriori étant égale à 0.03. Nous pouvons formuler cette conclusion d'une part parce l'écart type estimé de ce coefficient est égale à 0.01 et surtout la distribution a posteriori est bien au-dessus de zéro comme le montre les quartiles et la représentation graphique de la densité.

```
##  
## Iterations = 1001:11000  
## Thinning interval = 1  
## Number of chains = 1  
## Sample size per chain = 10000  
##  
## 1. Empirical mean and standard deviation for each variable,  
## plus standard error of the mean:  
##  
##  
##  
## (Intercept) 2.0707 1.21860 0.0121860 0.0121860  
## taux_acces_attendu_premiere_bac 0.0362 0.01453 0.0001453 0.0001453  
## sigma2 0.4765 0.09883 0.0009883 0.0010227  
##  
## 2. Quantiles for each variable:  
##  
##  
## (Intercept) -0.322072 1.27268 2.09261 2.89038 4.4472  
## taux_acces_attendu_premiere_bac 0.007913 0.02649 0.03595 0.04583 0.0648  
## sigma2 0.321972 0.40644 0.46350 0.53258 0.7059
```



```
##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
##                                     Burn-in   Total Lower bound Dependence
##                                     (M)      (N)    (Nmin)   factor (I)
## (Intercept)                      2       3771   3746     1.010
## taux_acces_attendu_premiere_bac 2       3771   3746     1.010
## sigma2                           2       3741   3746     0.999
##
## (Intercept)  taux_acces_attendu_premiere_bac
##           10000.000                  10000.000
##           sigma2
##           9338.753
```

5.2 Barre des mutations pour les maths

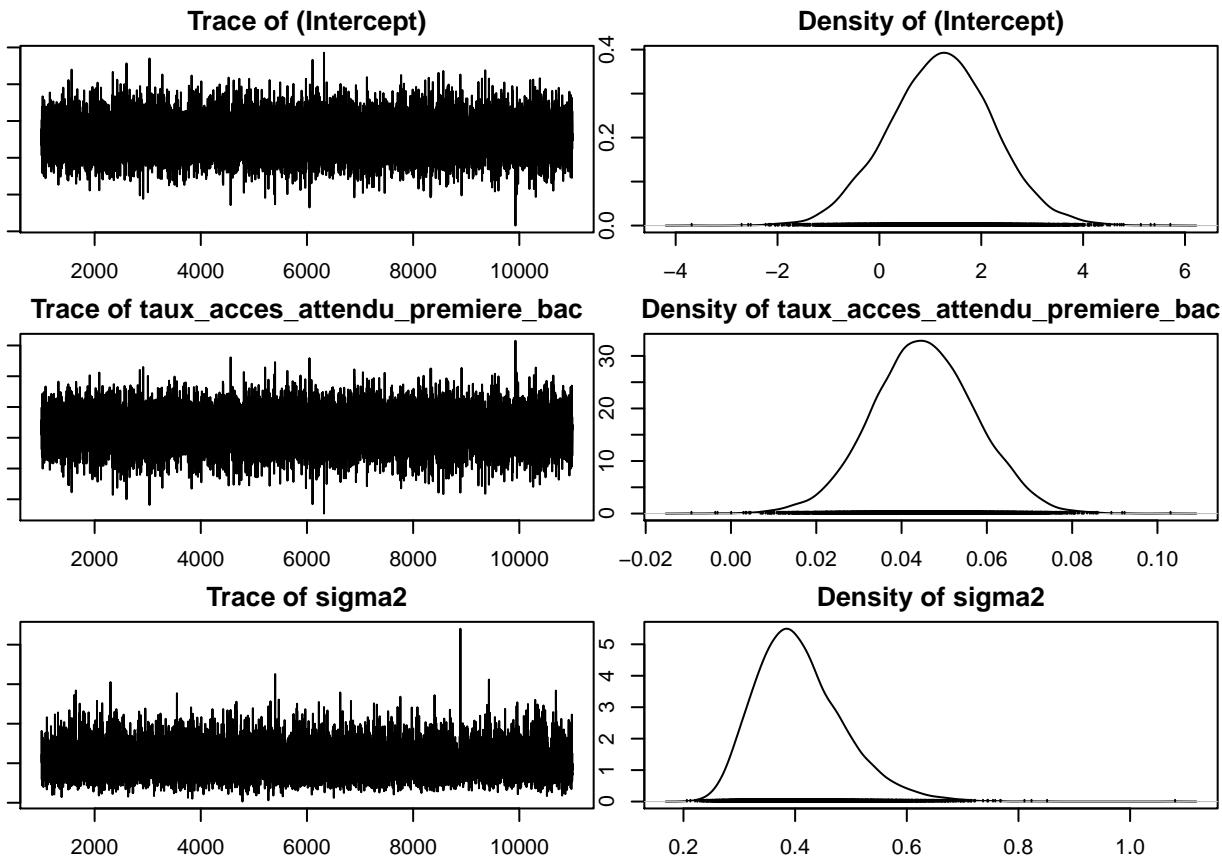
Nous estimons la même régression pour le sous-échantillon correspondant à la matière Math. Nous obtenons un moyenne a posteriori de 0.04 pour le coefficient du taux d'accès attendu. Les quantiles comme la représentation graphique de la densité de la distribution a posteriori du coefficient montre sont situé à droite de ceux obtenu pour la matière Anglais. Nous pouvons donc en déduire raisonnablement que l'effet du taux d'accès est plus important en math qu'en anglais.

```
##
## Iterations = 1001:11000
## Thinning interval = 1
```

```

## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##     plus standard error of the mean:
##
##                                     Mean      SD  Naive SE Time-series SE
## (Intercept)                 1.21256 1.02856 0.0102856      0.0102856
## taux_acces_attendu_premiere_bac 0.04524 0.01219 0.0001219      0.0001219
## sigma2                      0.40842 0.07910 0.0007910      0.0008029
##
## 2. Quantiles for each variable:
##
##                                     2.5%    25%    50%    75%   97.5%
## (Intercept)                -0.8027 0.53858 1.22349 1.90036 3.21275
## taux_acces_attendu_premiere_bac 0.0215 0.03715 0.04509 0.05328 0.06912
## sigma2                     0.2828 0.35255 0.39852 0.45405 0.59029

```



```

##
## Quantile (q) = 0.025
## Accuracy (r) = +/- 0.005
## Probability (s) = 0.95
##
##                                     Burn-in Total Lower bound Dependence
##                                     (M)     (N)   (Nmin)        factor (I)
## (Intercept)                   2       3802  3746          1.01

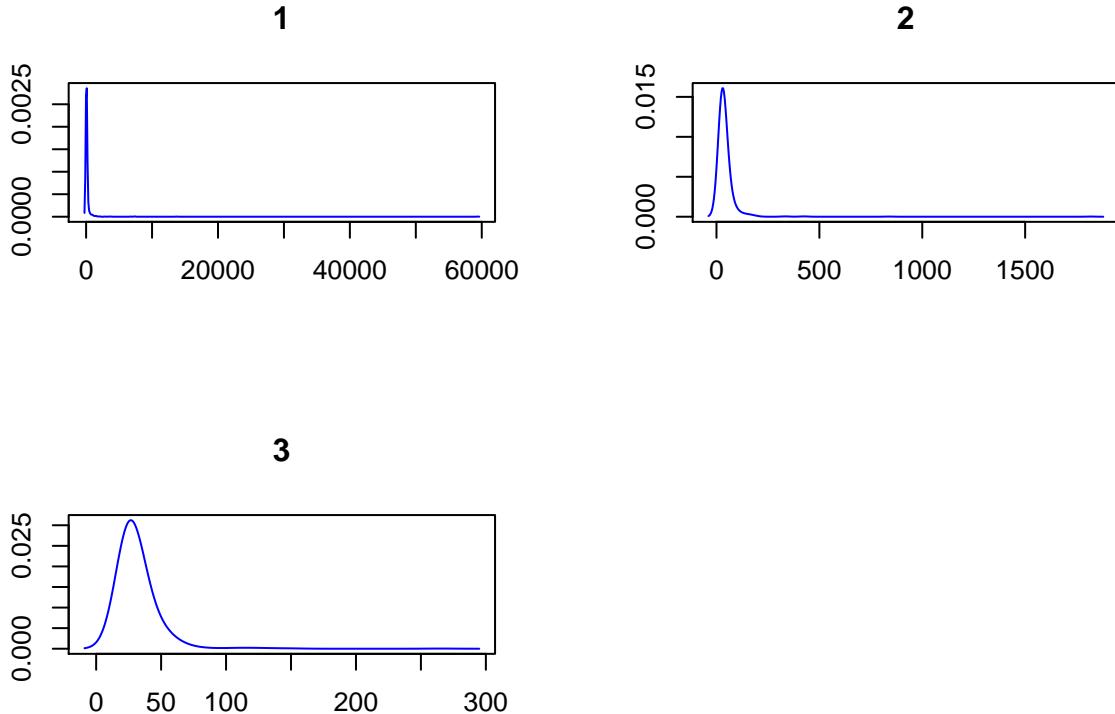
```

```
##   taux_acces_attendu_premiere_bac 2      3802 3746      1.01
##   sigma2                          2      3710 3746      0.99
##                               (Intercept) taux_acces_attendu_premiere_bac
##                               10000.000                  10000.000
##                               sigma2
##                               9704.564
```

6 Partie II Loi de Pareto

6.1 Loi de Pareto

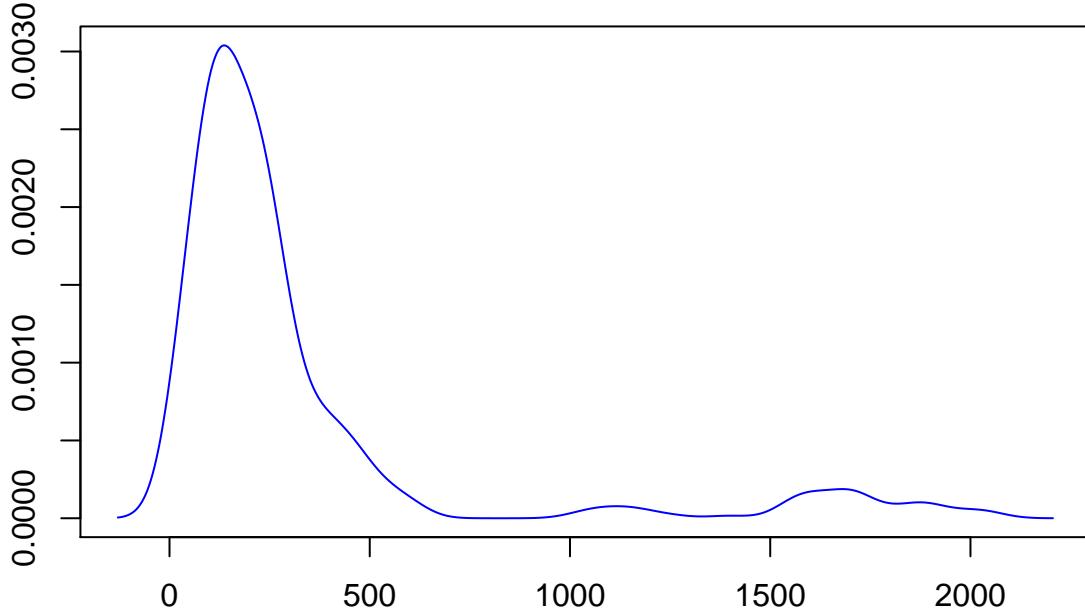
Nous représentons des distributions simulées de la loi de Pareto $Z \sim \text{Pareto}(m, \alpha)$ pour $m = 21$ et $\alpha = 1, 2, 3$.



A mesure que le paramètre *alpha* augmente, l'ensemble des valeurs de z pour lesquelles la densité est différente de zéro se réduit.

Nous pouvons comparer ces différentes distribution à la densité estimée de la variable Barre. La ressemblance justifie de choisir une loi de Pareto. Nous pouvons aussi anticiper que la valeur du paramètre α doit être inférieure à 1.

density Barre



Loi a priori pour α . Dans la mesure où le paramètre α doit être strictement supérieur à 0, un choix possible serait de prendre une loi exponentielle $\alpha \sim \mathcal{E}(\lambda)$. Le prior du paramètre α s'écrirait $f(\alpha) = \lambda e^{-\lambda\alpha}$. Nous en déduisons que l'espérance de α a priori est $E(\alpha) = \frac{1}{\lambda}$ et sa variance $V(\alpha) = \frac{1}{\lambda^2}$

6.2 Loi a posteriori du paramètre α

La loi a posteriori est construite à partir du produit de la log-vraisemblance de l'échantillon et du prior :

$$\prod_{i=1}^N \left(\alpha \frac{m^\alpha}{z_i^{\alpha+1}} \mathbb{1}_{z_i > m} \right) \times \lambda e^{-\lambda\alpha} = \alpha^N \frac{m^N}{\left(\prod_{i=1}^N z_i \right)^{\alpha+1}} \lambda e^{-\lambda\alpha} \propto \frac{e^{N \ln(\alpha) - \lambda\alpha}}{\left(\prod_{i=1}^N z_i \right)^{\alpha+1}}$$

On en déduit l'expression de la densité a posteriori en logarithme

$$N \ln(\alpha) - \alpha\lambda - (\alpha + 1) \left(\sum_{i=1}^N \ln(z_i) \right) + N \ln(m) + \ln(\lambda)$$

6.3 Echantillon tire de la loi a posteriori

Dans un premier temps, on construit la fonction logfit de la densité a posteriori du paramètre α en logarithme.

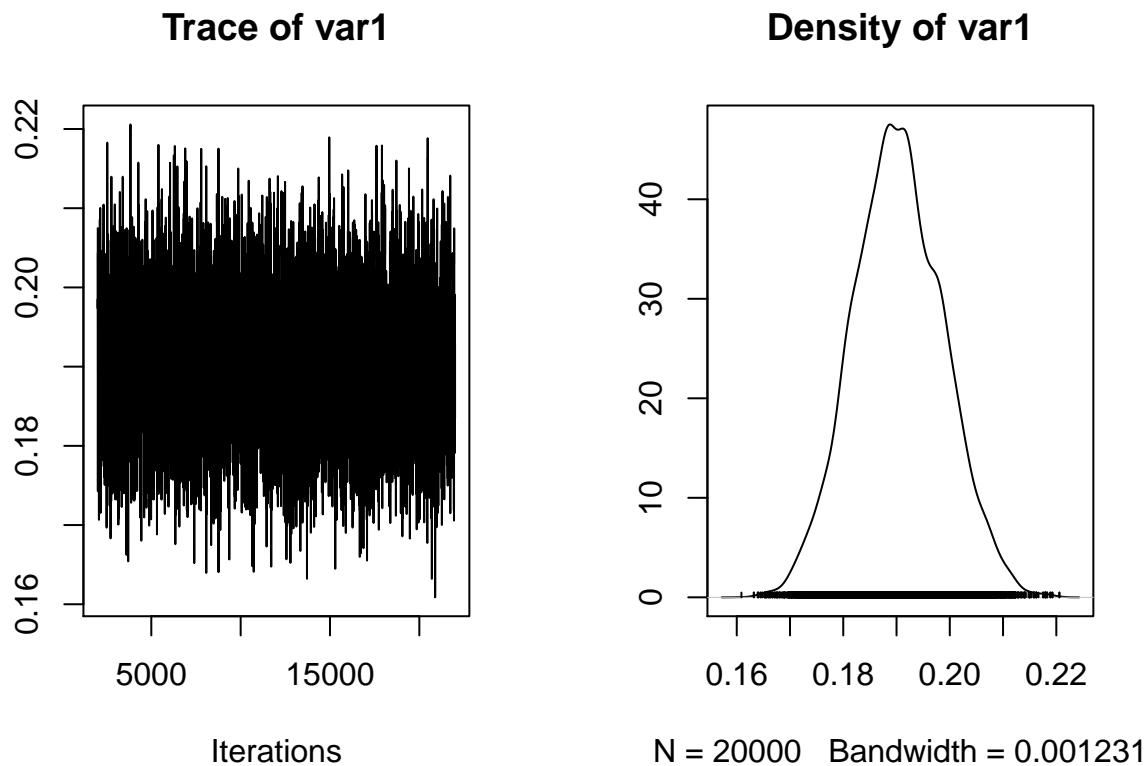
```
logfit <- function(alpha, z, X){
  m=X[1]
  lambda = X[2]
  N<-length(z)
  ll<-N*log(abs(alpha))-abs(alpha)*lambda-(abs(alpha)+1)*sum(log(z))+N*log(m)+log(lambda)
```

```

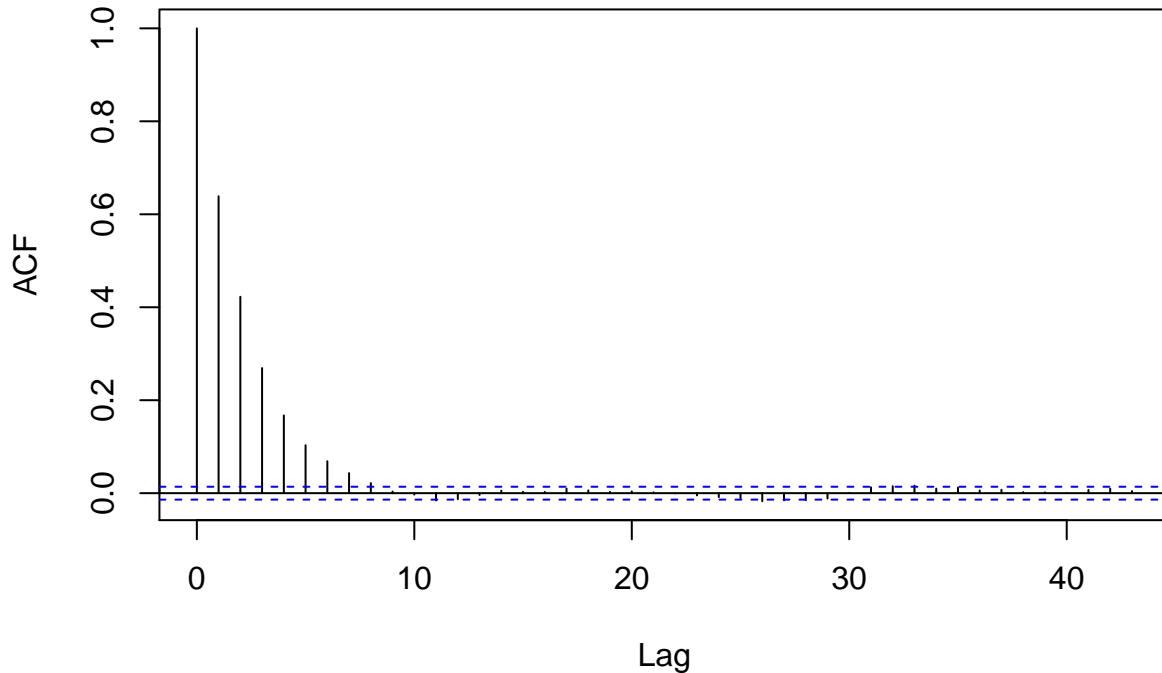
    return(l1)
}

##
## 
## 000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## The Metropolis acceptance rate was 0.37155
## 000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##
## Iterations = 2001:22000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 20000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean           SD      Naive SE Time-series SE
## 1.903e-01   8.418e-03   5.953e-05   1.278e-04
##
## 2. Quantiles for each variable:
##
##    2.5%     25%     50%     75%   97.5%
## 0.1740  0.1845  0.1901  0.1961  0.2071

```



Series 1



Les critères d'évaluation de la chaîne MCMC ne semblent pas montrer de problème particulier :

- le taux d'acceptation est égal à 0.37 ce qui correspond aux valeurs préconisées,
- le graphique de la chaîne de Markov montrent que l'on balaye assez bien l'espace des valeurs possibles,
- l'autocorrelogramme montre que l'autocorrélation s'annule assez rapidement (elle est pratiquement nulle pour 10 retards). Nous pouvons en déduire que nous estimons assez bien les paramètres de la distribution a posteriori du paramètre α .

L'estimation donne une espérance a posteriori de 0.1903. L'écart type estimé est égal à 8.418.e-03. Un intervalle de crédibilité à 95% est [0.1740, 0.2071]

6.4 Anglais et Mathématiques

On reproduit l'analyse précédente pour l'anglais et les mathématiques.

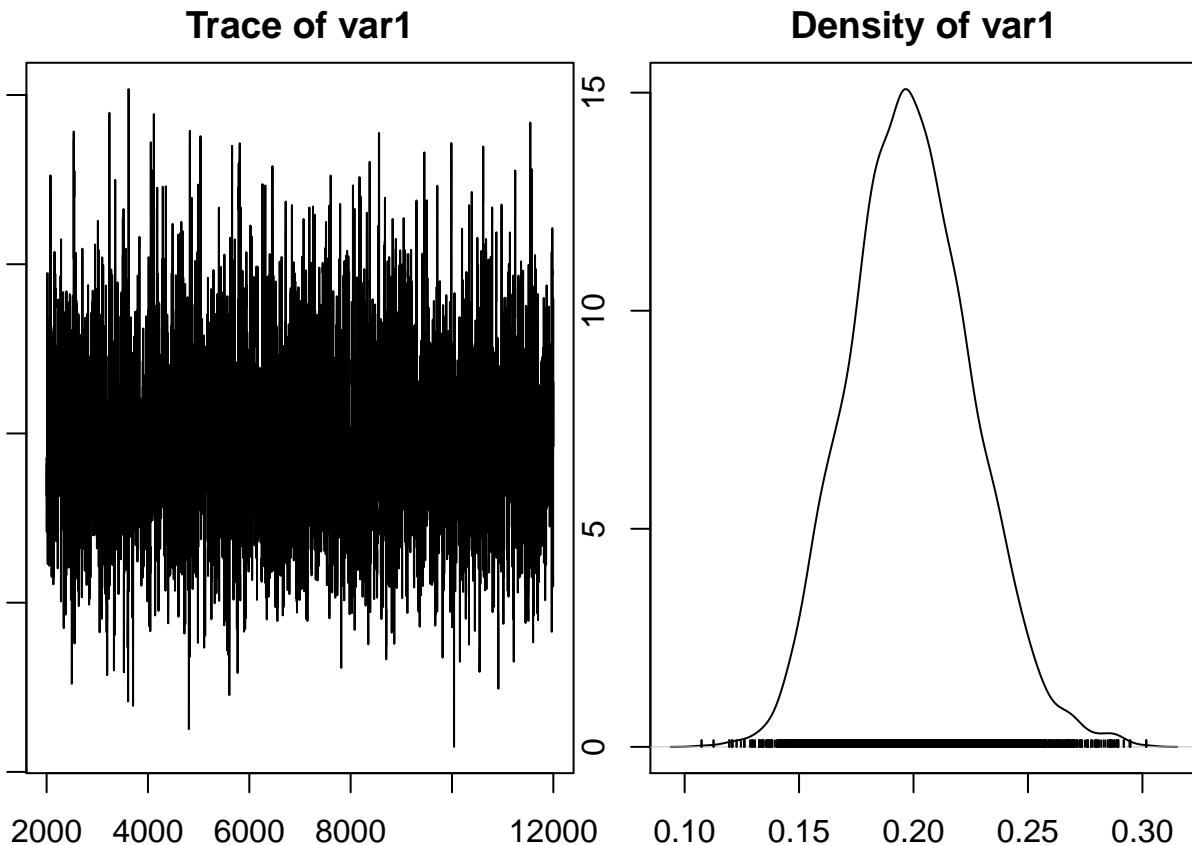
6.4.1 Simulation de la distibution a posteriori pour $\alpha_{anglais}$

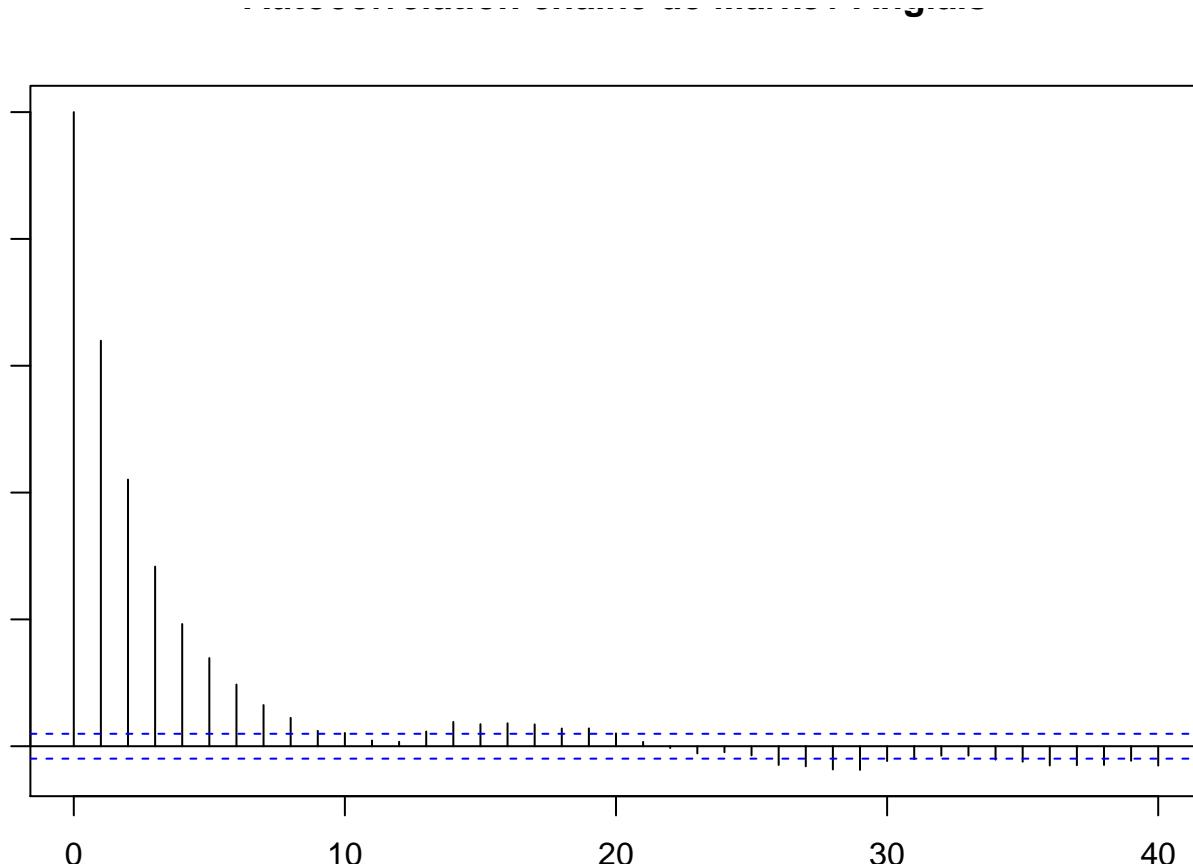
```
##  
##  
## 000000000000000000000000000000000000000000000000000000000000000  
## The Metropolis acceptance rate was 0.49642  
## 000000000000000000000000000000000000000000000000000000000000000  
  
##  
## Iterations = 2001:12000  
## Thinning interval = 1  
## Number of chains = 1  
## Sample size per chain = 10000
```

```

## 
## 1. Empirical mean and standard deviation for each variable,
## plus standard error of the mean:
## 
##      Mean          SD      Naive SE Time-series SE
## 0.1993023   0.0268853   0.0002689   0.0005848
## 
## 2. Quantiles for each variable:
## 
##    2.5%    25%    50%    75%  97.5%
## 0.1505 0.1809 0.1982 0.2168 0.2540

```

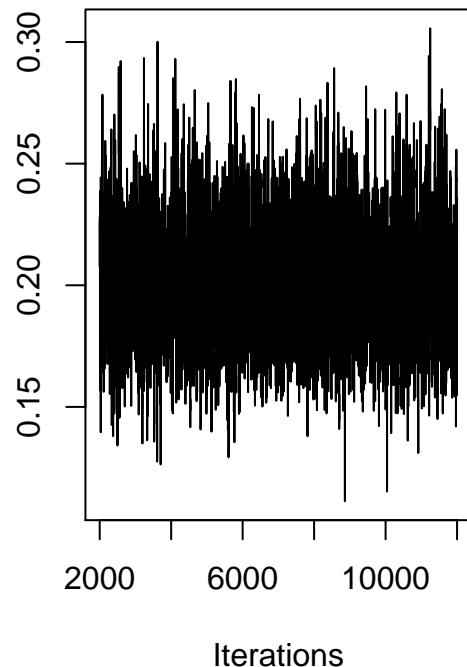




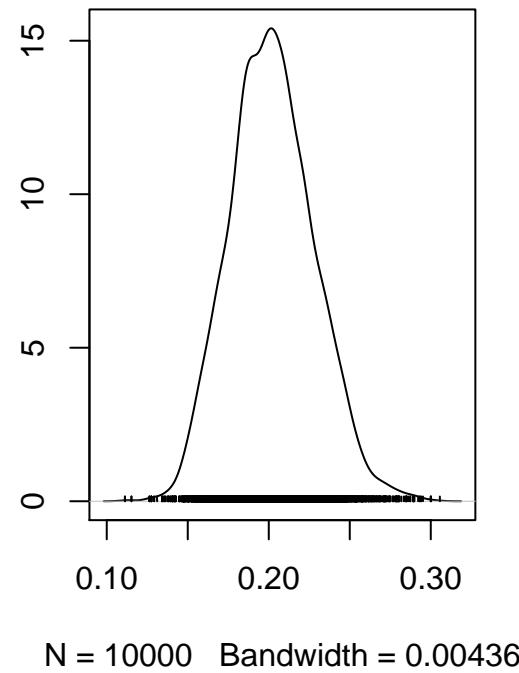
6.5 Simulation de la distribution a posteriori pour α_{math} et conclusion

```
##
##   #####
## The Metropolis acceptance rate was 0.49592
##   #####
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##     plus standard error of the mean:
##
##                Mean           SD      Naive SE Time-series SE
##    0.2021350    0.0259497    0.0002595    0.0005740
##
## 2. Quantiles for each variable:
##
##    2.5%    25%    50%    75%  97.5%
## 0.1547  0.1844  0.2011  0.2194  0.2543
```

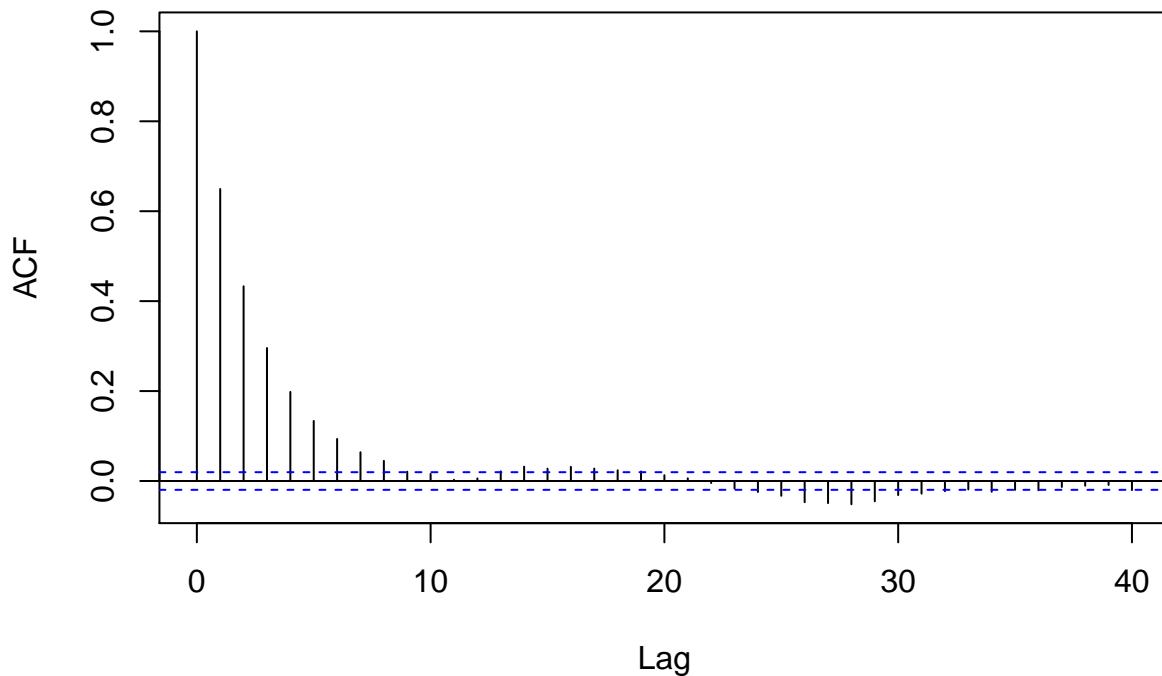
Trace of var1



Density of var1



Autocorrélation chaine de Markov Maths



Dans les deux cas, les diagnostics concernant la chaîne de Markov n'indiquent pas de problème particulier : les taux d'acceptation sont respectivement égaux à 0.4964 pour Anglais et 0.4959 pour maths. Les chaînes de Markov balayent les valeurs possibles du paramètre α et l'autocorrélation décroît assez rapidement.

L'espérance estimée de la distribution a posteriori est égale à 0.1993 pour anglais et 0.2021 pour maths. Elles sont donc très proches. De plus les intervalles de crédibilité sont respectivement égaux à $[0.1505, 0.2560]$ pour l'anglais et $[0.1547, 0.2543]$ pour maths. Nous pouvons raisonnablement accepter l'hypothèse $\alpha_{anglais} = \alpha_{maths}$.