Statistique Bayésienne

Haeji Yun

2023-07-27

Dans cette étude, nous nous intéressons au nombre de points nécessaires pour obtenir une mutation professionnelle dans les lycées de l'académie de Versailles en 2012. Nous cherchons à expliquer ces points en nous basant sur les différents caratéristiques de lycée tels que l'effectif dans différentes séries, le taux de réussite, le taux d'accès etc.

Notre jeu de données contient 516 observations et 23 variables. Les observations correspondent aux couples établissement - discipline et les variables correspondent à différentes caractéristiques de lycées. Nous considérons uniquement les filières du lycée général.

0. Chargement de données

Notre variable d'intérêt est la variable quantitative *Barre* qui correspond au nombre de points.

Nous avons 5 covariables qualitatives qui sont :
- Le code et le nom d'établissement
- Le code et le nom de ville
- La matière

Les autres 17 covariables quantitatives correspondent à différentes caractéristiques :
- les effectifs dans les différentes series
- les taux de résussite brut et attendu de chaque série
- les taux d'accès brut et attendu en seconde
- les taux d'accès brut et attendu en première
- les taux de réussite totaux d'accès brut et attendu

```
##  [1] "code_etablissement"                "ville"
##  [3] "etablissement"                     "commune"
##  [5] "Matiere"                           "Barre"
##  [7] "effectif_presents_serie_l"         "effectif_presents_serie_es"
##  [9] "effectif_presents_serie_s"         "taux_brut_de_reussite_serie_l"
## [11] "taux_brut_de_reussite_serie_es"    "taux_brut_de_reussite_serie_s"
## [13] "taux_reussite_attendu_serie_l"     "taux_reussite_attendu_serie_es"
## [15] "taux_reussite_attendu_serie_s"     "effectif_de_seconde"
## [17] "effectif_de_premiere"              "taux_acces_brut_seconde_bac"
## [19] "taux_acces_attendu_seconde_bac"    "taux_acces_brut_premiere_bac"
## [21] "taux_acces_attendu_premiere_bac"   "taux_brut_de_reussite_total_series"
## [23] "taux_reussite_attendu_total_series"
```

1. Etude Exploratoire

Aperçu

À part les variables qualitatives qui donnent l'information sur l'identification du lycée et la discipline, le jeu de données comporte principalement des données quantitatives. Elles sont exprimées en nombre d'élèves pour les effectifs et en pourcentage pour les taux.

Nous observons une grande variabilité des effectifs globaux dans toutes les series avec leurs valeurs maximales 20 fois plus grandes de leurs valeurs minimales. Nous pouvons supposer qu'il y a une différence de taille entre les lycées.

Nous remarquons une légère variabilité de taux de réussite entre les différentes séries. Cette variabilité est presque inexistante pour les taux attendus.

Il y a également une grande variabilité dans la variable *Barre*. Avec un grand écart entre les 3ème et 4ème quartiles, nous pouvons supposer qu'il y a quelques valeurs particulièrement élevées par rapport au reste.

```
## # A tibble: 18 x 8
##    var                                 min   q25 median   q75   max  mean     sd
##    <chr>                             <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
##  1 Barre                                21  111     196 292    2056 322.  424.
##  2 effectif_presents_serie_l             6   18      30  47     133  34.2  21.0
##  3 effectif_presents_serie_es           10   53      69  99     192  74.4  34.4
##  4 effectif_presents_serie_s            13   64     100 140     328 106.   58.0
##  5 taux_brut_de_reussite_serie_l        36   82      89  94     100  86.3  11.6
##  6 taux_brut_de_reussite_serie_es       51   81      88  94     100  86.4   9.86
##  7 taux_brut_de_reussite_serie_s        50   81      88  93      99  86.2   9.10
##  8 taux_reussite_attendu_serie_l        65   84      89  92      98  86.9   7.42
##  9 taux_reussite_attendu_serie_es       61   86      90  94      98  88.0   8.48
## 10 taux_reussite_attendu_serie_s        61   86      89  94      98  87.4   9.39
## 11 effectif_de_seconde                  36  268     336 415     764 352.  136.
## 12 effectif_de_premiere                 36  226.    289 364     691 308.  126.
## 13 taux_acces_brut_seconde_bac          49   64      71  76      87  69.6   9.09
## 14 taux_acces_attendu_seconde_bac       50   64      69  73      83  68.5   7.22
## 15 taux_acces_brut_premiere_bac         65   82      85  89.2    97  84.5   6.88
## 16 taux_acces_attendu_premiere_bac      70   81      85  89      94  84.2   5.99
## 17 taux_brut_de_reussite_total_seri~    64   82      86  91      98  85.5   7.39
## 18 taux_reussite_attendu_total_seri~    67   84      88  92      98  86.8   7.72
```

Dans notre jeu de données, nous n'avons pas de données manquantes mais il existe 6 doublons. En supprimant les doublons, nous nous retrouvons avec 510 observations et toujours 23 variables.

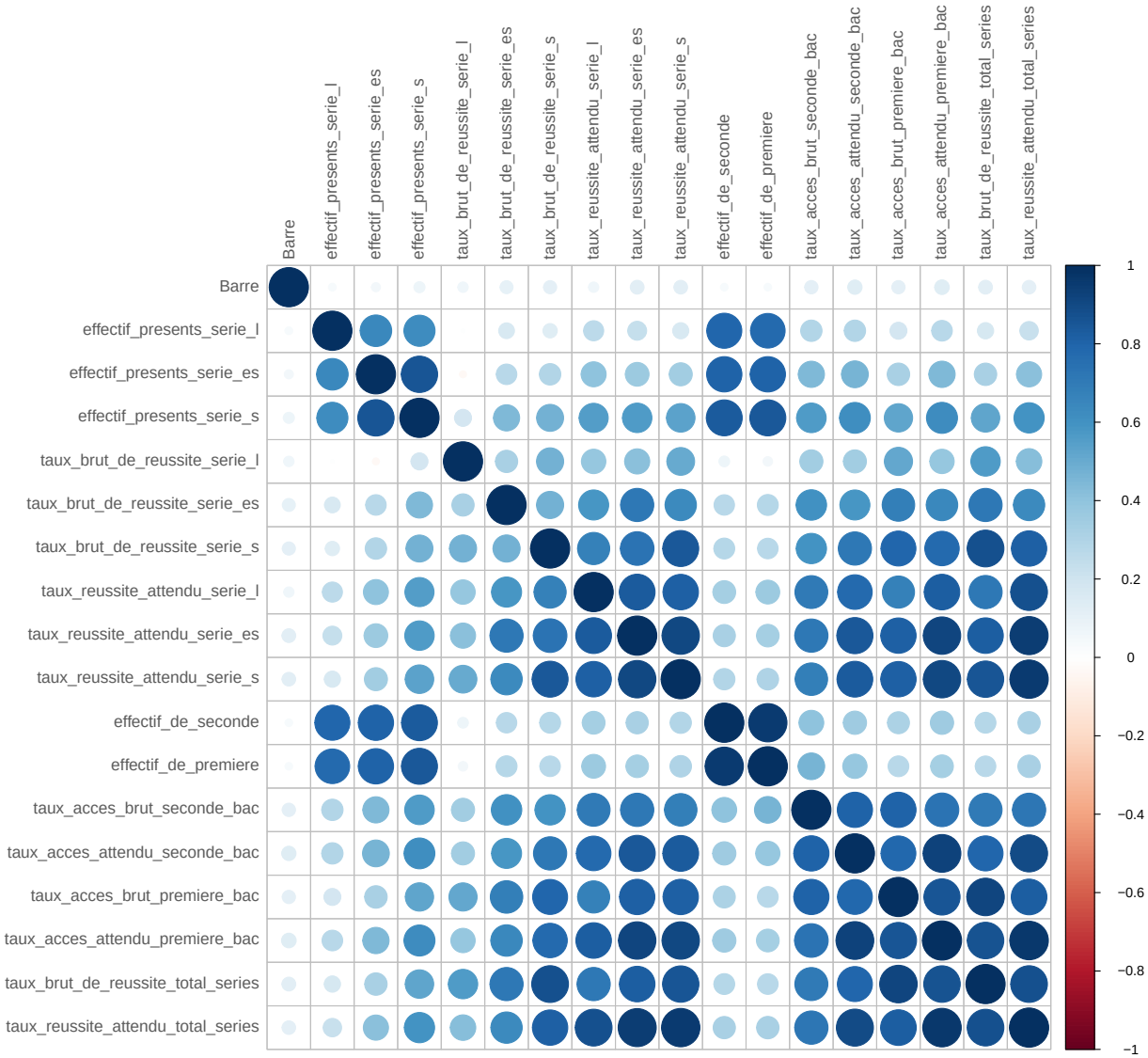
```
##     code_etablissement                  ville commune    Matiere
## 14            0781512V MONTIGNY LE BRETONNEUX   78423   ALLEMAND
## 16            0781512V MONTIGNY LE BRETONNEUX   78423   ALLEMAND
## 60            0781951X            MAGNANVILLE   78354   ALLEMAND
## 61            0781951X            MAGNANVILLE   78354   ALLEMAND
## 393           0950646L                GONESSE   95277 LETT CLASS
## 395           0950646L                GONESSE   95277 LETT CLASS
## 409           0950650R              SARCELLES   95585 LETT CLASS
## 413           0950650R              SARCELLES   95585 LETT CLASS
## 446           0951147F           L ISLE ADAM    95313 LETT CLASS
## 450           0951147F           L ISLE ADAM    95313 LETT CLASS
## 459           0951710T                VAUREAL   95637 LETT CLASS
## 461           0951710T                VAUREAL   95637 LETT CLASS
```

Variables explicatives

Variables quantitatives

Nous pouvons étudier la corrélation des variables quantitatives avec une matrice de corrélation. Nous observons des corrélations sur des blocs de variables :

- Le bloc des effectifs : l'effectif de seconde, l'effectif de primière, l'effectif de série L, l'effectif de série ES, et l'esffectif de serie S sont corréelés.
- Le bloc des taux : le taux brut de réussite de série S, le taux de réussite attendu de série L, le taux de réusssite attendu de série ES, le taux de réussite attendu de séries S, le taux d'accès brut de seconde, le taux d'accès attendu de seconde, le taux d'accès brut de première, le taux d'accès attendu de première, le taux brut de résussite total, et le taux réussite total sont corrélés.

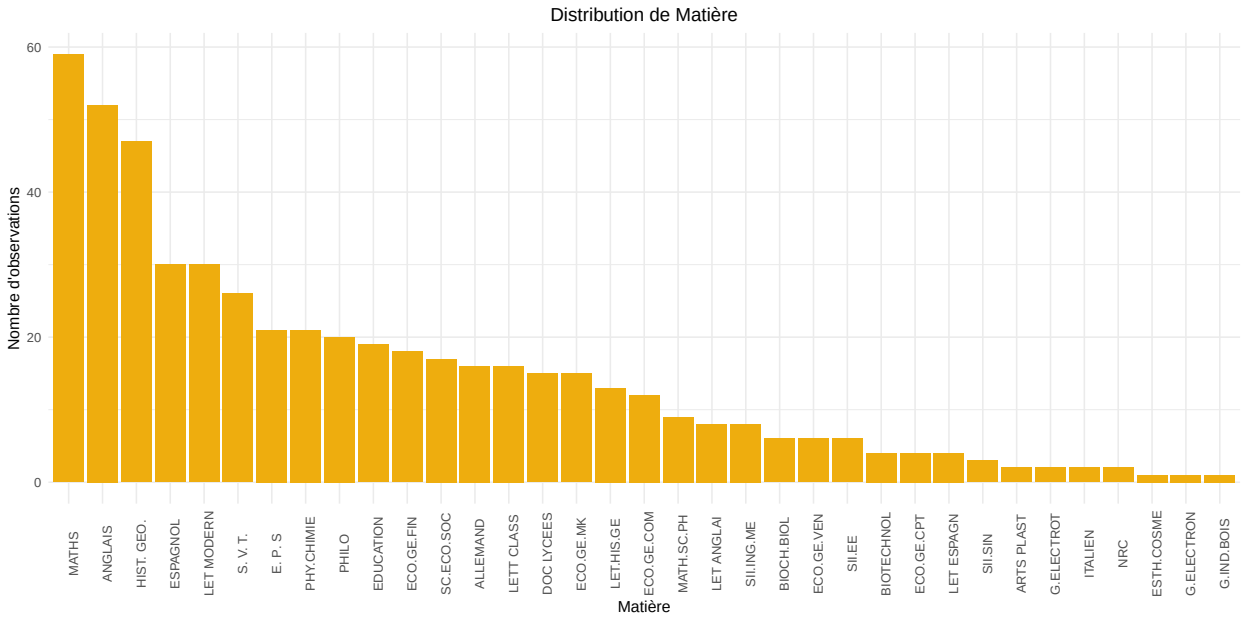


Variables qualitatives

Parmi les variables qualitatives, le code d'établissement, la ville, l'établissement, et la commune donnent l'information sur l'identité de l'établissement. Nous ne les utiliseront pas comme des covariables.

Pour chaque établissemnt, nous avons mêmes valeurs pour les variables explicatives quantitatives quelques soit la matière. Donc la matière semble clé parmi les variables qualitatives. Néanmoins, nous avons 36 matières différentes avec des matières qui sont très peu observées.

Par exemple, nous avons les matières *G.IND.BOIS*, *G.ELECTRON*, et *ESTH.COSME* qui sont observées une fois. Les matières *NRC*, *ITALIEN*, *G.ELECTROT*, et *ARTSPLAST* sont observées deux fois dans tout le jeu de données.



Puisque les variables très peu observées n'apportent pas d'information de qualité, nous allons supprimer les variables qui ont moins de 3 observations.

De plus, nous avons des matières qui n'apartiennent pas aux filières du lycée générale. Nous allons également les supprimer car nous nous intéressons qu'aux lycées générales.

Nous nous retrouvons donc avec 427 observations et 20 matières différentes.

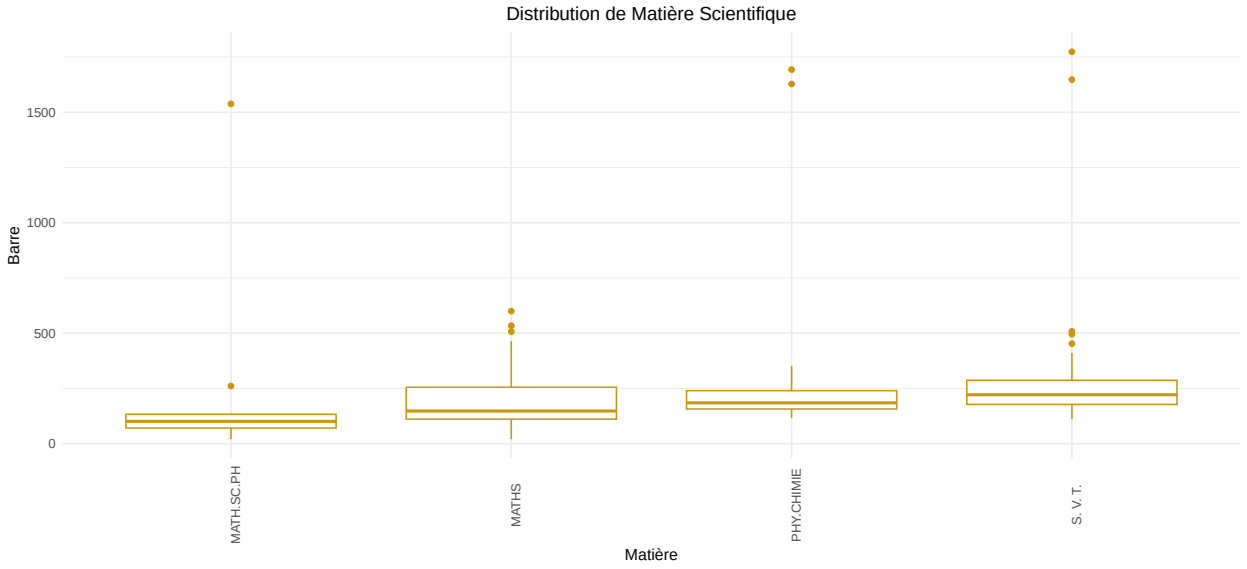
Néanmoins, nous remarquons que le regroupement des matières est différent selon les lycées. Bien qu'il y a des lycées qui consière les matières uniques, il y a des lycées qui regroupe plusieurs matières dans un groupe pour leur attribuer un point de mutation unique au groupe.

- Il y a 9 lycées qui regroupent les *MATHS*, *S.V.T*, et *PHY.CHIMIE* en un seul groupe de *MATH.SC.PH*.
- Il y a 13 lycées qui regroupent les *HIST.GEO*, *LET MODERN*, *LETT CLASS* en un groupe de *LET.HIS.GE*.
- Il y a 8 lycées qui regroupent les *ANGLAIS*, *LET MODERN*, *LETT CLASS* en un groupe de *LET ANGLAI*.
- Il y a 4 lycées qui regroupent les *ESPAGNOL*, *LET MODERN*, *LETT CLASS* en un groupe de *LET ESPAGN*.

```
##
##      MATHS    ANGLAIS HIST. GEO.   ESPAGNOL LET MODERN   S. V. T.    E. P. S
##         59         52         47         30         30         26         21
## PHY.CHIMIE      PHILO  EDUCATION SC.ECO.SOC   ALLEMAND LETT CLASS DOC LYCEES
##         21         20         19         17         16         16         15
## LET.HIS.GE MATH.SC.PH LET ANGLAI LET ESPAGN
##         13          9          8          4
```

Pour éviter que les mêmes matières soient considérées différentes, nous allons harmoniser le regroupement des maitères. Pour la simplicité, nous allons nous baser seulement sur la distribution de *Barre* de chaque regroupement pour l'harmonisation.

Tout d'abord pour les matières scientifiques, il est difficile d'affecter le regroupement *MATH.SC.PH* à une des maitères *MATHS*, *S.V.T*, ou *PHY.CHIMIE* car nous n'observons pas de similitude particulière avec une des matières. Nous allons garder le regroupement tel qu'il est.

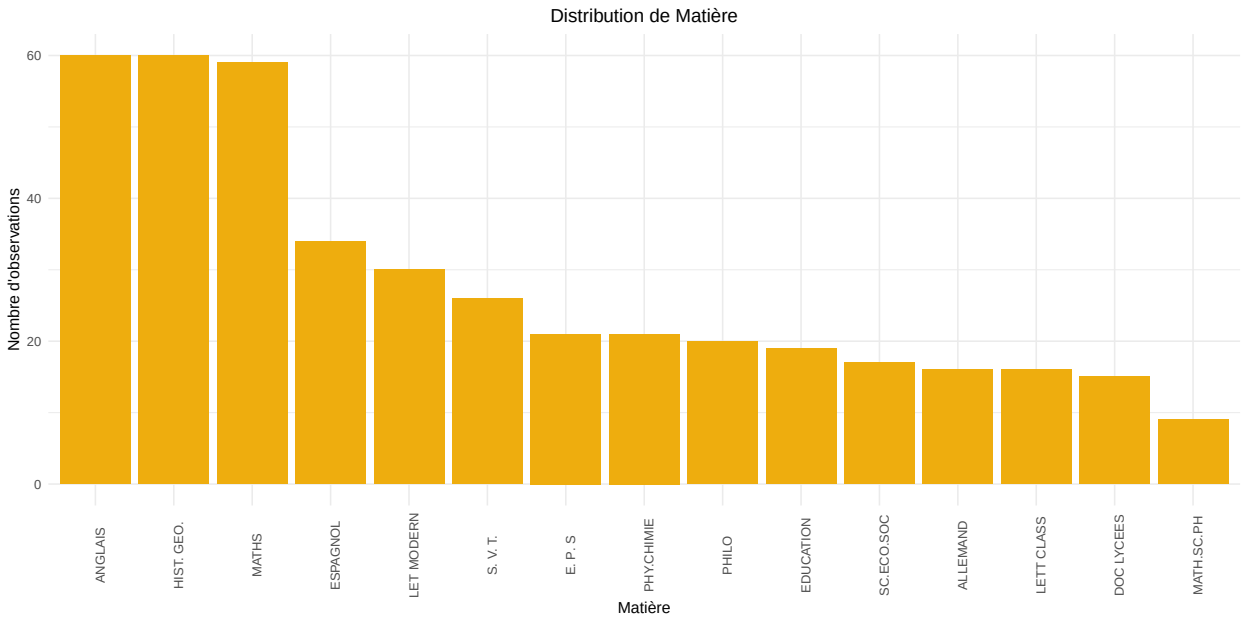


Pour les matières littéraires, nous allons affecter les matières regroupées à une seule matière qui leur ressemble le plus en terme de distribution de *Barre* :

- Nous allons remplacer le groupe *LET.HIS.GE* par la matière *HIST.GEO.*
- La maitère *ANGLAIS* remplace le groupe *LET ANGLAI.*
- La maitère *ESPAGNOL* remplace le groupe *LET ESPAGN.*

```
## # A tibble: 7 x 6
##   Matiere      min `q-25` `q-50` `q-75`   max
##   <chr>      <dbl>  <dbl>  <dbl>  <dbl> <dbl>
## 1 ANGLAIS       21  111    158     229. 1718
## 2 HIST. GEO.    21  124    210     261. 1807.
## 3 LET ANGLAI    21   26.2   36.5   278. 1709
## 4 LET.HIS.GE    21   58     81.2   143. 1396
## 5 LETT CLASS    38  174.   232.    300   443
## 6 LET MODERN    58  200.   261     451. 1935.
## 7 LET ESPAGN    71   76.2  135     207.  258
```

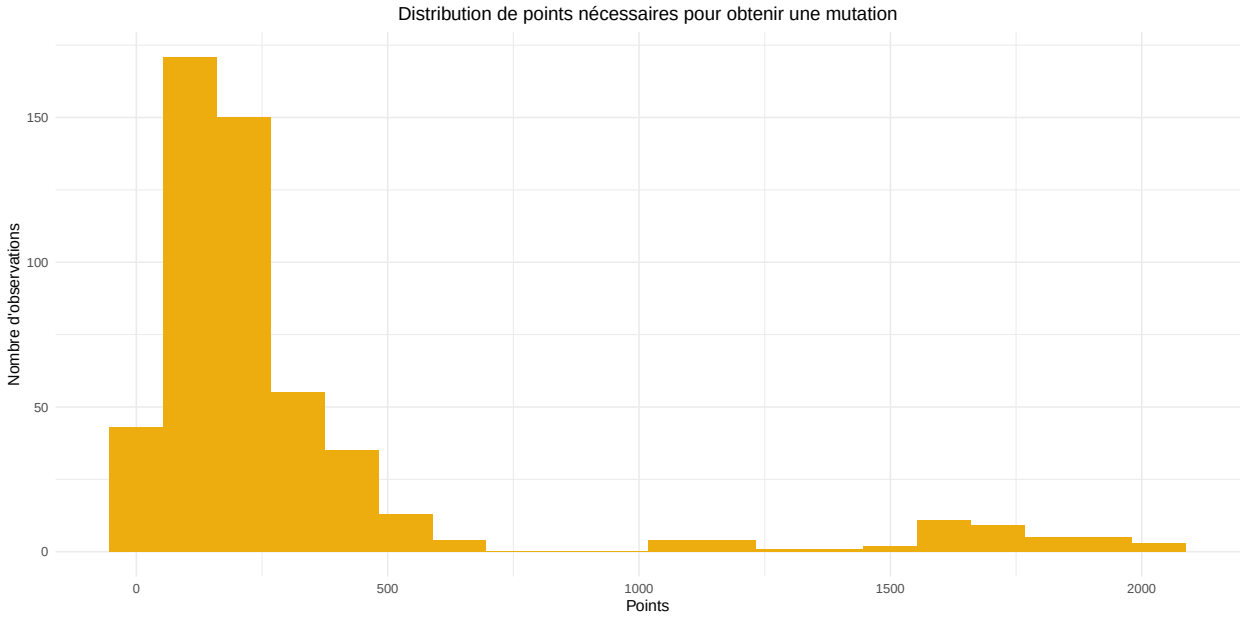
Au final, nous avons 15 matières avec 423 observations.



Variable cible

Notre variable d'intérêt est *Barre*.

Nous avons une distribution avec longue queue à droit. Il y a beaucoup d'observations avec les valeurs entre 50 et 250. La plupart des observations se trouvent en dessus de 500 et nous avons quelques observations jusqu'à l'entour de 2100.



2. Régression Linéaire Bayésienne

Dans un premier temps, nous pouvons appliquer la régression bayésienne sur tout notre jeu de données avec 10.000 itérations. En utilisant la fonction MCMCRegress, nous pouvons obtenir un échantillon simulé à partir de la distribution posterior du modèle régression linéaire guassien en utilisant Gibbs sampling, une méthode d'une chaîne de Markov.

Nous avons pas mal de variables explicatives avec des quantiles qui contiennent le 0. En effet, la présence de 0 entre les quantiles 2,5% et 97,5% signifie qu'il y a une grande probabilité que le coefficient de ces variables peuvent être nul. Donc nous préférons d'exclure telles covariables que de l'inclure à tort dans le modèle.

Les variables qui ne contiennent pas de 0 dans les quantiles 2,5% - 97,5% sont les matières *anglais*, *doc lycées*, *éducation*, *espangol*, *histo & géo*, *lettres classiques*, *math & science & physique*, *maths*, *physique & chimie*, et *économie sociale*.

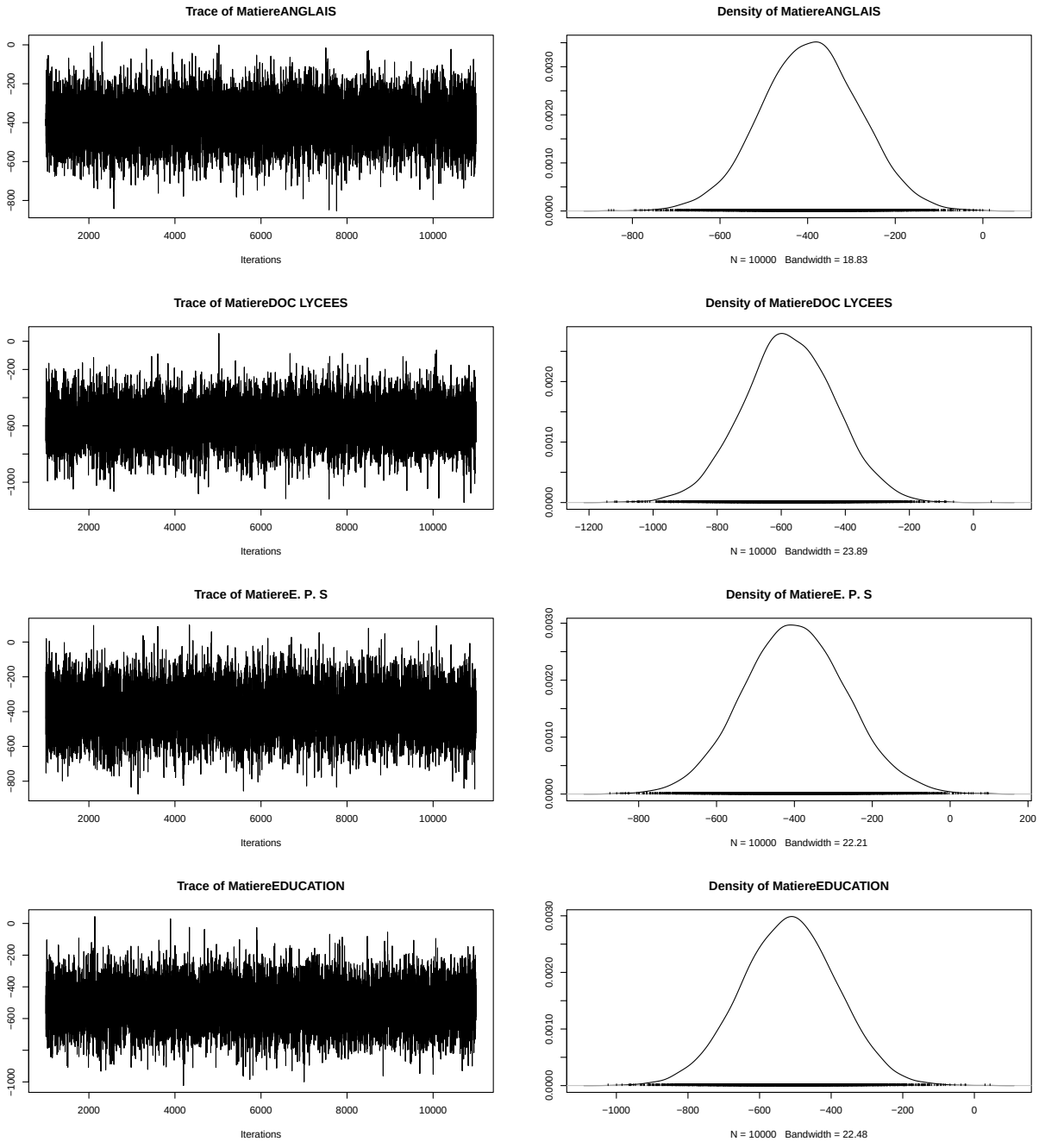
Le modèle considère que les variables matières comme significatives.

```
## $quantiles
##                                      2.5%      25%      50%      75%    97.5%
## (Intercept)                       -1.1e+03 -2.9e+02  1.0e+02  4.9e+02  1.3e+03
## MatiereANGLAIS                    -6.2e+02 -4.7e+02 -3.9e+02 -3.2e+02 -1.7e+02
## MatiereDOC LYCEES                 -8.6e+02 -6.7e+02 -5.8e+02 -4.8e+02 -2.9e+02
## MatiereE. P. S                    -6.5e+02 -4.9e+02 -4.0e+02 -3.1e+02 -1.3e+02
## MatiereEDUCATION                  -7.8e+02 -6.1e+02 -5.2e+02 -4.3e+02 -2.6e+02
## MatiereESPAGNOL                   -7.4e+02 -5.8e+02 -5.1e+02 -4.3e+02 -2.7e+02
## MatiereHIST. GEO.                 -6.2e+02 -4.7e+02 -4.0e+02 -3.2e+02 -1.8e+02
## MatiereLET MODERN                 -4.3e+02 -2.6e+02 -1.8e+02 -9.2e+01  6.6e+01
## MatiereLETT CLASS                 -7.5e+02 -5.6e+02 -4.7e+02 -3.7e+02 -1.9e+02
## MatiereMATH.SC.PH                 -6.5e+02 -4.2e+02 -3.0e+02 -1.8e+02  4.2e+01
## MatiereMATHS                      -7.1e+02 -5.7e+02 -4.9e+02 -4.2e+02 -2.8e+02
## MatierePHILO                      -4.4e+02 -2.7e+02 -1.8e+02 -9.5e+01  7.1e+01
## MatierePHY.CHIMIE                 -6.2e+02 -4.5e+02 -3.7e+02 -2.8e+02 -1.1e+02
## MatiereS. V. T.                   -5.9e+02 -4.2e+02 -3.4e+02 -2.6e+02 -1.0e+02
## MatiereSC.ECO.SOC                 -5.7e+02 -3.9e+02 -3.0e+02 -2.1e+02 -3.1e+01
## effectif_presents_serie_l         -2.4e+00 -2.4e-01  8.2e-01  1.9e+00  3.9e+00
## effectif_presents_serie_es        -2.3e+00 -7.2e-01  1.2e-01  9.4e-01  2.5e+00
## effectif_presents_serie_s         -8.9e-01  4.5e-01  1.1e+00  1.8e+00  3.1e+00
## taux_brut_de_reussite_serie_l     -4.4e+00 -1.1e+00  5.9e-01  2.3e+00  5.7e+00
## taux_brut_de_reussite_serie_es    -7.2e+00 -1.2e+00  2.0e+00  5.0e+00  1.1e+01
## taux_brut_de_reussite_serie_s     -7.6e+00  1.3e+00  5.7e+00  1.0e+01  1.9e+01
## taux_reussite_attendu_serie_l     -2.3e+01 -1.4e+01 -8.8e+00 -3.9e+00  5.5e+00
## taux_reussite_attendu_serie_es    -1.6e+01 -5.3e+00  3.6e-01  6.2e+00  1.7e+01
## taux_reussite_attendu_serie_s     -2.6e+01 -1.3e+01 -6.8e+00 -3.5e-02  1.2e+01
## effectif_de_seconde               -1.2e+00 -3.5e-01  9.3e-02  5.3e-01  1.3e+00
## effectif_de_premiere              -2.3e+00 -1.3e+00 -8.0e-01 -2.8e-01  6.8e-01
## taux_acces_brut_seconde_bac       -7.5e+00  8.7e-02  4.0e+00  7.9e+00  1.6e+01
## taux_acces_attendu_seconde_bac    -2.3e+01 -1.1e+01 -4.5e+00  1.7e+00  1.3e+01
## taux_acces_brut_premiere_bac      -3.7e+01 -2.2e+01 -1.5e+01 -7.9e+00  6.2e+00
## taux_acces_attendu_premiere_bac   -9.5e+00  1.6e+01  3.0e+01  4.4e+01  7.1e+01
## taux_brut_de_reussite_total_series -2.5e+01 -7.5e+00  1.5e+00  1.0e+01  2.7e+01
## taux_reussite_attendu_total_series -4.6e+01 -1.7e+01 -1.4e+00  1.4e+01  4.4e+01
## sigma2                             1.1e+05  1.2e+05  1.3e+05  1.4e+05  1.5e+05
```

Pour chaque variable, nous pouvons visualiser l'estimation de sa densité et la trace de toutes les sorties issues de l'échantillonnage.

Voici l'estimation de quelques variables. Les graphiques de gauche montre la trace des valeurs prises par la chaîne à chaque itération. Nous observons que la chaîne mélange bien et se déplace bien dans la loi a posteriori sans être coincé à une partie de la chaîne. Cela explique que le modèle a bien convergé.

Les graphiques de droite montre la densité. Nous pouvons comprendre quelles valeurs chaque paramètre peut prendre.



Avec le diagnostique de raftery, nous remarquons il nous faut à peu près 3.900 itérations. Nous avons fait 10.000 itérations qui est un nombre largement suffisant.

```
##
## 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        3994  3746         1.070
##  MatiereANGLAIS                    2        3865  3746         1.030
##  MatiereDOC LYCEES                 2        3741  3746         0.999
##  MatiereE. P. S                    2        3802  3746         1.010
##  MatiereEDUCATION                  2        3865  3746         1.030
##  MatiereESPAGNOL                   2        3710  3746         0.990
##  MatiereHIST. GEO.                 2        3802  3746         1.010
##  MatiereLET MODERN                 2        3771  3746         1.010
##  MatiereLETT CLASS                 2        3771  3746         1.010
##  MatiereMATH.SC.PH                 2        3771  3746         1.010
##  MatiereMATHS                      2        3680  3746         0.982
##  MatierePHILO                      2        3865  3746         1.030
##  MatierePHY.CHIMIE                 2        3710  3746         0.990
##  MatiereS. V. T.                   2        3650  3746         0.974
##  MatiereSC.ECO.SOC                 2        3834  3746         1.020
##  effectif_presents_serie_l         2        3865  3746         1.030
##  effectif_presents_serie_es        2        3636  3746         0.971
##  effectif_presents_serie_s         2        3897  3746         1.040
##  taux_brut_de_reussite_serie_l     2        3802  3746         1.010
##  taux_brut_de_reussite_serie_es    2        3788  3746         1.010
##  taux_brut_de_reussite_serie_s     2        3680  3746         0.982
##  taux_reussite_attendu_serie_l     2        3802  3746         1.010
##  taux_reussite_attendu_serie_es    2        3710  3746         0.990
##  taux_reussite_attendu_serie_s     2        3710  3746         0.990
##  effectif_de_seconde               2        3680  3746         0.982
##  effectif_de_premiere              2        3802  3746         1.010
##  taux_acces_brut_seconde_bac       2        3620  3746         0.966
##  taux_acces_attendu_seconde_bac    2        3834  3746         1.020
##  taux_acces_brut_premiere_bac      2        3771  3746         1.010
##  taux_acces_attendu_premiere_bac   2        3771  3746         1.010
##  taux_brut_de_reussite_total_series 2       3741  3746         0.999
##  taux_reussite_attendu_total_series 2       3834  3746         1.020
##  sigma2                            2        3929  3746         1.050
```

Choix de covariables

Meilleur Modèle Bayésien

Avec la fonction BMS, qui simule toutes les combinaisons possibles de modèle par MCMC, nous pouvons obtenir les meilleurs modèles bayesiens. Ici, nous allons garder l'information de 500 meilleurs modèles.

Voici les 5 meilleures modèles obtenus. Les variables ayant le coefficiet 1 sont les variables prises par chaque modèle. Le meilleur modèle prend que la matière *allemand* comme la variable explicative. Le deuxième meilleur modèle ne considère aucune variable comme significative. Dans le reste, les modèles prennent deux covariables incluant l'*allemand*. La matière *allemend* semble avoir un impact.

Nous remarquons également que les meilleurs modèles prennent moins de covariables, 1 ou 2, voire 0

```
##                                    000004000 000000000  000004080  000024000  000004008
## effectif_presents_serie_l          0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## effectif_presents_serie_es         0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## effectif_presents_serie_s          0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_l      0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_es     0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_s      0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_l      0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_es     0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_s      0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## effectif_de_seconde                0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## effectif_de_premiere               0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_acces_brut_seconde_bac        0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_acces_attendu_seconde_bac     0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_acces_brut_premiere_bac       0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_acces_attendu_premiere_bac    0.0000000 0.0000000 0.00000000 1.00000000 0.00000000
## taux_brut_de_reussite_total_series 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## taux_reussite_attendu_total_series 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_ALLEMAND                   1.0000000 0.0000000 1.00000000 1.00000000 1.00000000
## Matiere_ANGLAIS                    0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_DOC LYCEES                 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_E. P. S                    0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_EDUCATION                  0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_ESPAGNOL                   0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_HIST. GEO.                 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_LET MODERN                 0.0000000 0.0000000 1.00000000 0.00000000 0.00000000
## Matiere_LETT CLASS                 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_MATH.SC.PH                 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_MATHS                      0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_PHILO                      0.0000000 0.0000000 0.00000000 0.00000000 1.00000000
## Matiere_PHY.CHIMIE                 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_S. V. T.                   0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## Matiere_SC.ECO.SOC                 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## PMP (Exact)                        0.3177472 0.1036293 0.10206987 0.08147405 0.03874128
## PMP (MCMC)                         0.2673333 0.1340000 0.04866667 0.06966667 0.01800000
```

3. Analyse Fréquentiste

Nous pouvons maintenant effectuer l'analyse fréquentiste pour la comparaison. Le modèle linéaire fréquentiste reprend les mêmes covariables significatives que le premier modèle bayésien, c'est à dire *anglais*, *doc lycées*, *éducation*, *histo & géo*, *lettres classiques*, *math & science & physique*, *maths*, *physique & chimie*, et *économie sociale*. Le modèle considère en plus *eps* et *svt* comme significatives.

```
##
## Call:
## lm(formula = Barre ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -494.11 -170.29  -73.45   33.24 1453.42
##
## Coefficients:
```

```
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         89.79177  583.22802   0.154 0.877734
## MatiereANGLAIS                    -394.40680  111.01959  -3.553 0.000435 ***
## MatiereDOC LYCEES                 -574.57582  141.61687  -4.057 6.14e-05 ***
## MatiereE. P. S                    -399.68080  132.24504  -3.022 0.002696 **
## MatiereEDUCATION                  -519.00550  133.14417  -3.898 0.000117 ***
## MatiereESPAGNOL                   -505.80981  117.58736  -4.302 2.21e-05 ***
## MatiereHIST. GEO.                 -397.09462  111.19124  -3.571 0.000406 ***
## MatiereLET MODERN                 -177.22386  123.00414  -1.441 0.150549
## MatiereLETT CLASS                 -469.64430  142.80274  -3.289 0.001110 **
## MatiereMATH.SC.PH                 -301.83755  174.45527  -1.730 0.084494 .
## MatiereMATHS                      -493.76273  109.85874  -4.495 9.53e-06 ***
## MatierePHILO                      -182.53625  129.02227  -1.415 0.158039
## MatierePHY.CHIMIE                 -365.35499  128.72133  -2.838 0.004803 **
## MatiereS. V. T.                   -339.57334  123.56257  -2.748 0.006308 **
## MatiereSC.ECO.SOC                 -299.09780  137.14752  -2.181 0.029870 *
## effectif_presents_serie_l            0.82613    1.60181   0.516 0.606363
## effectif_presents_serie_es           0.12472    1.23154   0.101 0.919396
## effectif_presents_serie_s            1.10927    1.01837   1.089 0.276798
## taux_brut_de_reussite_serie_l        0.63144    2.58147   0.245 0.806910
## taux_brut_de_reussite_serie_es       1.88917    4.60726   0.410 0.682029
## taux_brut_de_reussite_serie_s        5.62571    6.71039   0.838 0.402410
## taux_reussite_attendu_serie_l       -8.80520    7.22499  -1.219 0.223785
## taux_reussite_attendu_serie_es       0.40585    8.57454   0.047 0.962276
## taux_reussite_attendu_serie_s       -6.76482    9.79658  -0.691 0.490326
## effectif_de_seconde                  0.08219    0.64355   0.128 0.898445
## effectif_de_premiere                -0.78748    0.76280  -1.032 0.302635
## taux_acces_brut_seconde_bac          4.06426    5.82439   0.698 0.485772
## taux_acces_attendu_seconde_bac      -4.76196    9.39474  -0.507 0.612566
## taux_acces_brut_premiere_bac       -15.14591   11.08541  -1.366 0.172737
## taux_acces_attendu_premiere_bac     30.54235   20.43223   1.495 0.135878
## taux_brut_de_reussite_total_series   1.58838   13.17477   0.121 0.904108
## taux_reussite_attendu_total_series  -1.44471   22.98294  -0.063 0.949914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 361.9 on 345 degrees of freedom
## Multiple R-squared:  0.1593, Adjusted R-squared:  0.0838
## F-statistic: 2.109 on 31 and 345 DF,  p-value: 0.0007087
```

En terme de AIC, le meilleur modèle fréquentiste propose encore plus de covariables, avec le *taux d'accès attendu du premier* en plus.

```
##
## Call:
## lm(formula = Barre ~ Matiere + effectif_presents_serie_s + taux_reussite_attendu_serie_l +
##     effectif_de_premiere + taux_acces_attendu_premiere_bac, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -491.99 -173.70  -70.47   32.76 1466.01
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                    258.8121   391.1493   0.662 0.508608
## MatiereANGLAIS                -402.4733   108.2724  -3.717 0.000234 ***
## MatiereDOC LYCEES             -556.7035   137.8119  -4.040 6.56e-05 ***
## MatiereE. P. S                -394.0332   128.5070  -3.066 0.002332 **
## MatiereEDUCATION              -523.3222   129.9572  -4.027 6.90e-05 ***
## MatiereESPAGNOL               -496.9857   114.7427  -4.331 1.93e-05 ***
## MatiereHIST. GEO.             -394.3406   108.2436  -3.643 0.000309 ***
## MatiereLET MODERN             -185.7340   119.4630  -1.555 0.120891
## MatiereLETT CLASS             -448.6883   138.2483  -3.246 0.001283 **
## MatiereMATH.SC.PH             -327.0728   168.0238  -1.947 0.052366 .
## MatiereMATHS                  -493.3207   107.3807  -4.594 6.03e-06 ***
## MatierePHILO                  -167.4880   125.9240  -1.330 0.184341
## MatierePHY.CHIMIE             -362.7356   125.1618  -2.898 0.003985 **
## MatiereS. V. T.               -333.9796   119.6655  -2.791 0.005537 **
## MatiereSC.ECO.SOC             -307.6045   133.4036  -2.306 0.021692 *
## effectif_presents_serie_s        1.1938     0.7957   1.500 0.134405
## taux_reussite_attendu_serie_l   -7.7528     4.6361  -1.672 0.095349 .
## effectif_de_premiere            -0.5428     0.3141  -1.728 0.084792 .
## taux_acces_attendu_premiere_bac 13.3777     6.6602   2.009 0.045330 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357.2 on 358 degrees of freedom
## Multiple R-squared:  0.1502, Adjusted R-squared:  0.1075
## F-statistic: 3.515 on 18 and 358 DF,  p-value: 2.43e-06
```

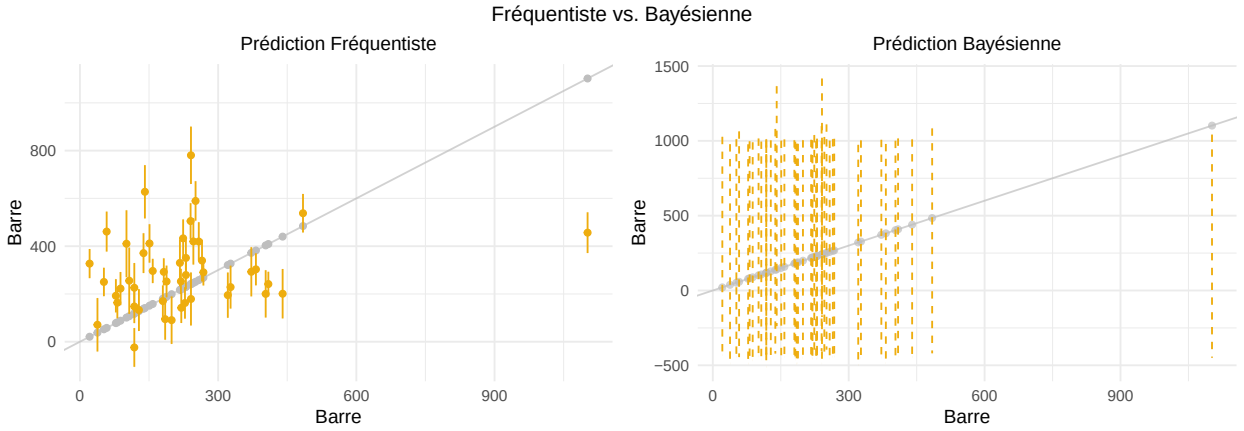
4. Prédiction

Nous pouvos comparer la prédiction du modèle bayésien et du modèle fréquentiste sur les observations de test qui correspondent à 10% de notre jeu de données.

Sur les deux graphiques, les points gris correspondent aux vraies observations.

Dans la prédiction fréquentiste, nous remarquons qu'il y a une grande partie d'observations qui n'est pas dans l'intervalle de confiance prédit. L'incertitude n'est pas assez forte dans le cadre fréquentiste.

Dans la prédiction bayésienne, presque toutes les observations sont dans l'intervalle de crédibilité prédit par le modèle. L'incertitude est beaucoup plus large.



5. Mathématiques & Anglais

Maintenant, nous allons nous concentrer uniquement sur la mutation en mathématiques, puis sur la mutation en anglais.

Mutation en mathématiques

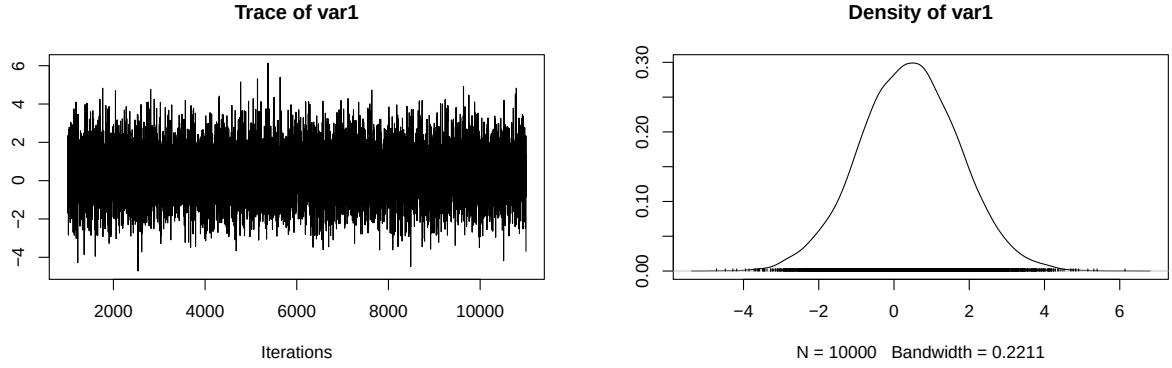
Approche Bayésienne

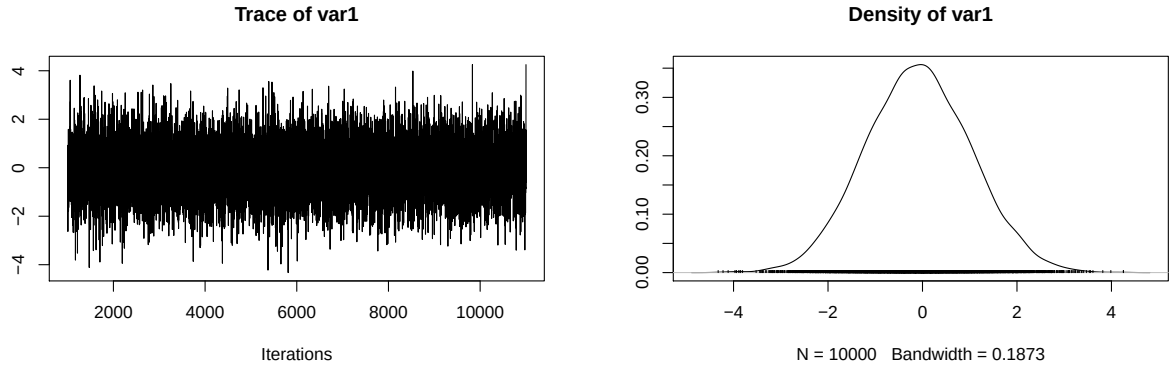
Nous allons effectuer la régression bayésienne avec 10.000 itérations que sur les observations concernant les mathématiques. Donc ici, la variable matière n'est plus incluse dans les covariables.

Le modèle nous donne deux variables significatives : Le *taux de réussite attendu de série L* et le *taux d'accès brut de première*.

```
## $quantiles
##                                       2.5%      25%      50%      75%    97.5%
## (Intercept)                       -1.6e+03 -8.4e+02 -4.2e+02 -9.4e+00  8.2e+02
## effectif_presents_serie_l         -2.7e+00 -4.7e-01  7.0e-01  1.9e+00  4.1e+00
## effectif_presents_serie_es        -2.2e+00 -4.7e-01  4.1e-01  1.3e+00  3.0e+00
## effectif_presents_serie_s         -2.2e+00 -8.5e-01 -9.0e-02  6.7e-01  2.1e+00
## taux_brut_de_reussite_serie_l     -1.5e+00  2.1e+00  4.0e+00  5.8e+00  9.4e+00
## taux_brut_de_reussite_serie_es    -4.4e+00  1.6e+00  4.7e+00  7.9e+00  1.4e+01
## taux_brut_de_reussite_serie_s     -4.2e+00  5.0e+00  9.7e+00  1.4e+01  2.3e+01
## taux_reussite_attendu_serie_l     -3.1e+01 -2.1e+01 -1.6e+01 -1.1e+01 -1.7e+00
## taux_reussite_attendu_serie_es    -1.4e+01 -2.7e+00  3.6e+00  9.5e+00  2.2e+01
## taux_reussite_attendu_serie_s     -2.5e+01 -1.1e+01 -3.8e+00  3.2e+00  1.7e+01
## effectif_de_seconde               -1.1e+00 -3.0e-01  1.6e-01  6.1e-01  1.5e+00
## effectif_de_premiere              -2.0e+00 -9.9e-01 -4.7e-01  5.6e-02  1.0e+00
## taux_acces_brut_seconde_bac        3.1e-02  8.1e+00  1.2e+01  1.7e+01  2.5e+01
## taux_acces_attendu_seconde_bac    -2.6e+01 -1.3e+01 -6.5e+00  3.2e-01  1.3e+01
## taux_acces_brut_premiere_bac      -4.7e+01 -3.2e+01 -2.4e+01 -1.6e+01 -9.3e-01
## taux_acces_attendu_premiere_bac   -5.1e+00  2.1e+01  3.6e+01  5.0e+01  7.8e+01
## taux_brut_de_reussite_total_series -3.3e+01 -1.6e+01 -6.1e+00  3.3e+00  2.1e+01
## taux_reussite_attendu_total_series -4.9e+01 -1.9e+01 -2.7e+00  1.3e+01  4.3e+01
## sigma2                             1.7e+05  1.8e+05  1.9e+05  2.0e+05  2.2e+05
```

Avec les graphqiues, nous pouvons remarquer que le modèle a bien convergé.





Les 5 meilleures modèles obtenus avec la fonction BMS, nous avons deux modèles qui prennent chacun l'une des deux variables données par le modèle précédant.

```
##                                      00000      00004     00010      00200      00100
## effectif_presents_serie_l          0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## effectif_presents_serie_es         0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## effectif_presents_serie_s          0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_l      0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_es     0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_s      0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_l      0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_es     0.0000000 0.00000000 0.0000000 1.00000000 0.00000000
## taux_reussite_attendu_serie_s      0.0000000 0.00000000 0.0000000 0.00000000 1.00000000
## effectif_de_seconde                0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## effectif_de_premiere               0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_acces_brut_seconde_bac        0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_acces_attendu_seconde_bac     0.0000000 0.00000000 1.0000000 0.00000000 0.00000000
## taux_acces_brut_premiere_bac       0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_acces_attendu_premiere_bac    0.0000000 1.00000000 0.0000000 0.00000000 0.00000000
## taux_brut_de_reussite_total_series 0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_reussite_attendu_total_series 0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## PMP (Exact)                        0.4839580 0.11787945 0.1041284 0.04914207 0.04684361
## PMP (MCMC)                         0.5283333 0.08166667 0.0990000 0.04966667 0.03233333
```

Approche fréquentiste

Avec l'analyse fréquentiste, les variables significatives sont le *taux de réussite attendu de série L*, le *taux d'accès brut de seconde*, et le *taux d'accès brut de première*.

Le modèle fréquentiste garde toujours plus de covariables que les modèles bayésiens, incluant celles données par les modèles bayésiens.

```
##
## Call:
## lm(formula = Barre ~ ., data = train_math)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -421.10 -212.58 -131.63  -13.48 1648.02
##
## Coefficients:
```

```
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -417.8391   616.4411  -0.678   0.4982
## effectif_presents_serie_l           0.7285     1.7423   0.418   0.6760
## effectif_presents_serie_es          0.4063     1.3245   0.307   0.7592
## effectif_presents_serie_s          -0.0764     1.1079  -0.069   0.9451
## taux_brut_de_reussite_serie_l       4.0197     2.8018   1.435   0.1521
## taux_brut_de_reussite_serie_es      4.8355     4.6658   1.036   0.3006
## taux_brut_de_reussite_serie_s       9.7882     7.0373   1.391   0.1649
## taux_reussite_attendu_serie_l     -16.3719     7.4195  -2.207   0.0278 *
## taux_reussite_attendu_serie_es      3.2919     9.1750   0.359   0.7199
## taux_reussite_attendu_serie_s      -3.9356    10.5125  -0.374   0.7083
## effectif_de_seconde                 0.1473     0.6692   0.220   0.8258
## effectif_de_premiere               -0.4682     0.7805  -0.600   0.5489
## taux_acces_brut_seconde_bac        12.4518     6.2688   1.986   0.0476 *
## taux_acces_attendu_seconde_bac     -6.5992     9.9418  -0.664   0.5072
## taux_acces_brut_premiere_bac      -23.9057    11.7713  -2.031   0.0428 *
## taux_acces_attendu_premiere_bac    35.8359    21.0279   1.704   0.0890 .
## taux_brut_de_reussite_total_series -6.5211    14.0837  -0.463   0.6436
## taux_reussite_attendu_total_series -2.2162    23.6589  -0.094   0.9254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 436 on 461 degrees of freedom
## Multiple R-squared:  0.0425, Adjusted R-squared:  0.007191
## F-statistic: 1.204 on 17 and 461 DF,  p-value: 0.257
```

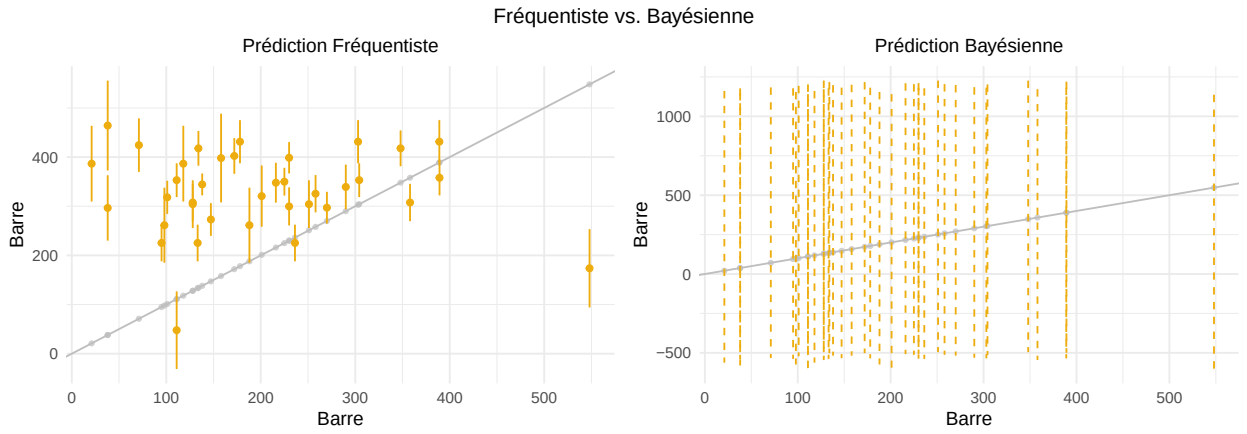
Le meilleur modèle en terme de AIC propose exactement les même covariables que le modèle bayésien, proposant le *taux de réussite attendu de série L* et le *taux d'accès brut de première*.

```
##
## Call:
## lm(formula = Barre ~ taux_brut_de_reussite_serie_l + taux_reussite_attendu_serie_l +
##     taux_acces_brut_seconde_bac + taux_acces_brut_premiere_bac +
##     taux_acces_attendu_premiere_bac, data = train_math)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -429.51 -220.51 -136.96    0.77 1662.89
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -357.540    302.855  -1.181   0.2384
## taux_brut_de_reussite_serie_l     3.122      2.107   1.482   0.1390
## taux_reussite_attendu_serie_l   -13.168      5.203  -2.531   0.0117 *
## taux_acces_brut_seconde_bac       6.923      4.090   1.693   0.0912 .
## taux_acces_brut_premiere_bac    -12.888      7.537  -1.710   0.0879 .
## taux_acces_attendu_premiere_bac  25.792      8.585   3.004   0.0028 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432.3 on 473 degrees of freedom
## Multiple R-squared:  0.0343, Adjusted R-squared:  0.02409
## F-statistic:  3.36 on 5 and 473 DF,  p-value: 0.005411
```

Prédiction

Nous pouvons comparer la prédiction qui combine les meilleurs modèles bayésien et celle du modèle fréquentiste.

Le modèle fréquentiste surestime dans la plupart du temps et il y a peu d'observations qui sont incluses dans l'intervalle de confiance. Quant au modèle bayésien, l'incertitude est très forte. Nous avons tous les observations qui se trouvent dans l'intervalle de crédibilité du modèle.



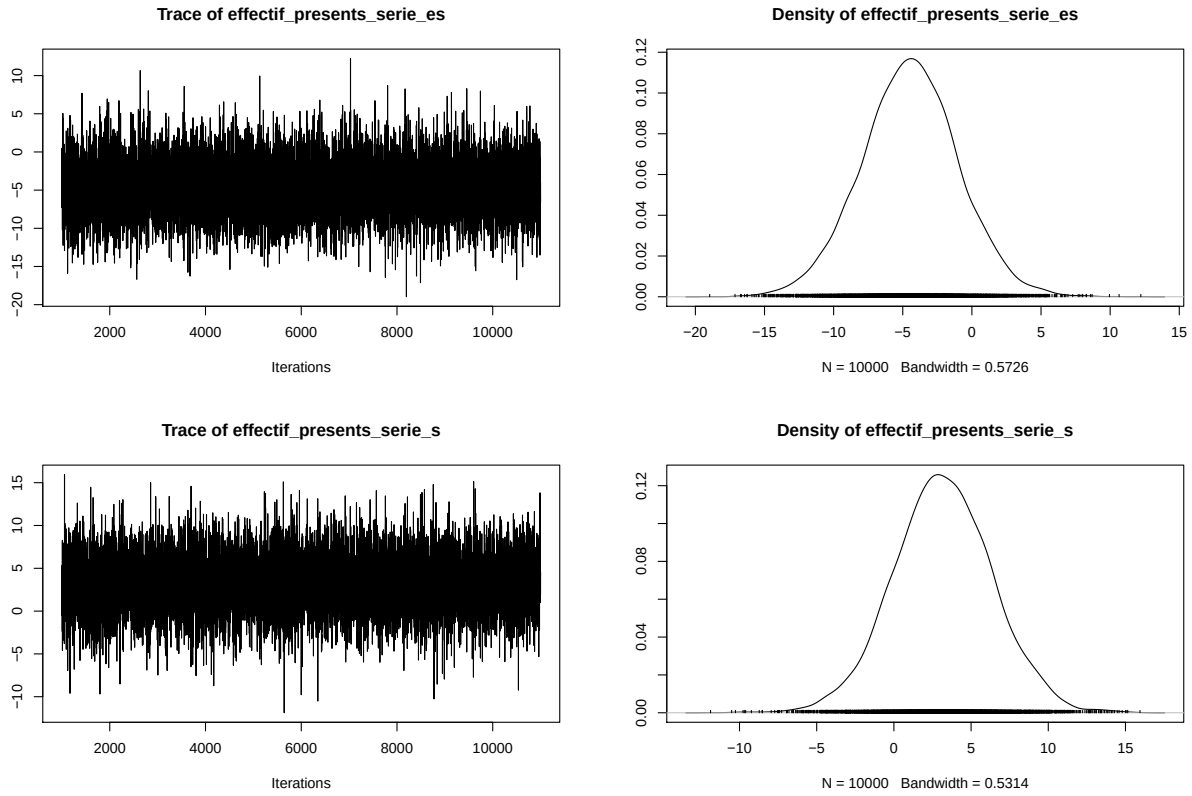
Mutation en anglais

Approche Bayésienne

Nous allons également effectuer la régression bayésienne avec 10.000 itérations uniquement sur les observations concernant l'anglais. Nous obtenons qu'une seule variable significative : le *taux d'accès brut de première*.

```
## $quantiles
##                                        2.5%     25%      50%      75%    97.5%
## (Intercept)                        -4932.3 -1284.3  5.5e+02   2330.7   6087.3
## effectif_presents_serie_l            -18.4    -8.1 -2.9e+00      2.3     12.7
## effectif_presents_serie_es           -11.4    -6.7 -4.5e+00     -2.2      2.4
## effectif_presents_serie_s             -3.1     1.1  3.1e+00      5.3      9.5
## taux_brut_de_reussite_serie_l        -19.4    -6.6  2.8e-01      6.9     19.9
## taux_brut_de_reussite_serie_es       -13.5     2.3  9.7e+00     17.4     32.4
## taux_brut_de_reussite_serie_s        -25.4     2.5  1.6e+01     30.6     59.2
## taux_reussite_attendu_serie_l        -90.7   -51.9 -3.3e+01    -13.0     27.1
## taux_reussite_attendu_serie_es        -6.2    23.1  3.8e+01     52.2     82.9
## taux_reussite_attendu_serie_s        -86.9   -48.4 -2.9e+01     -9.8     30.2
## effectif_de_seconde                   -2.5     0.3  1.7e+00      3.1      6.0
## effectif_de_premiere                  -8.8    -4.4 -2.3e+00     -0.2      3.9
## taux_acces_brut_seconde_bac          -12.0    17.5  3.2e+01     47.2     76.1
## taux_acces_attendu_seconde_bac      -111.7   -70.4 -4.9e+01    -28.7     12.4
## taux_acces_brut_premiere_bac        -148.3   -98.5 -7.5e+01    -50.4     -2.5
## taux_acces_attendu_premiere_bac      -57.0    38.7  8.7e+01    135.2    227.2
## taux_brut_de_reussite_total_series   -71.1   -11.5  1.7e+01     47.6    107.1
## taux_reussite_attendu_total_series  -173.1   -71.9 -1.9e+01     31.8    137.4
## sigma2                             61042.6 85258.7  1.0e+05 128658.0 198479.8
```

Avec les graphique, nous pouvons remarquer que le modèle a bien convergé.



Les meilleurs modèles BMS incluent différentes variables telles que l'*effectuf de première*, le *taux de réussite attendu de série L*, et le *taux brut de réussite de série ES*.

```
##                                    00000      00040     00400      01000      10000
## effectif_presents_serie_l         0.0000000 0.00000000 0.0000000 0.00000000 1.00000000
## effectif_presents_serie_es        0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## effectif_presents_serie_s         0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_l     0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_brut_de_reussite_serie_es    0.0000000 0.00000000 0.0000000 1.00000000 0.00000000
## taux_brut_de_reussite_serie_s     0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_l     0.0000000 0.00000000 1.0000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_es    0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_reussite_attendu_serie_s     0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## effectif_de_seconde               0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## effectif_de_premiere              0.0000000 1.00000000 0.0000000 0.00000000 0.00000000
## taux_acces_brut_seconde_bac       0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_acces_attendu_seconde_bac    0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_acces_brut_premiere_bac      0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_acces_attendu_premiere_bac   0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_brut_de_reussite_total_series 0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## taux_reussite_attendu_total_series 0.0000000 0.00000000 0.0000000 0.00000000 0.00000000
## PMP (Exact)                       0.8197812 0.01608057 0.0144797 0.01362732 0.01258426
## PMP (MCMC)                        0.7980000 0.03466667 0.0110000 0.01033333 0.02166667
```

Avec l'approche fréquentiste, le résultat est en accord avec le première modèle bayésien, en gardant le *taux d'accès brut de première* comme la variable significative unique.

```
##
## Call:
## lm(formula = Barre ~ ., data = train_en)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -350.09 -134.96    2.22  106.30  957.29
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       561.5705  2630.5410   0.213   0.8328
## effectif_presents_serie_l          -2.8690     7.4589  -0.385   0.7040
## effectif_presents_serie_es         -4.5042     3.3913  -1.328   0.1972
## effectif_presents_serie_s           3.1986     3.0921   1.034   0.3117
## taux_brut_de_reussite_serie_l       0.3276     9.6515   0.034   0.9732
## taux_brut_de_reussite_serie_es      9.7989    11.0125   0.890   0.3828
## taux_brut_de_reussite_serie_s      16.6961    20.5027   0.814   0.4238
## taux_reussite_attendu_serie_l     -32.3508    28.6420  -1.129   0.2703
## taux_reussite_attendu_serie_es     37.7404    21.5270   1.753   0.0929 .
## taux_reussite_attendu_serie_s     -28.9546    28.1220  -1.030   0.3139
## effectif_de_seconde                 1.6739     2.0244   0.827   0.4168
## effectif_de_premiere               -2.3214     3.0281  -0.767   0.4511
## taux_acces_brut_seconde_bac        32.1002    21.5659   1.488   0.1502
## taux_acces_attendu_seconde_bac     -49.5218    30.2916  -1.635   0.1157
## taux_acces_brut_premiere_bac       -74.7307    35.7530  -2.090   0.0478 *
## taux_acces_attendu_premiere_bac     86.1518    69.5049   1.240   0.2277
## taux_brut_de_reussite_total_series  18.5959    43.0499   0.432   0.6698
## taux_reussite_attendu_total_series -19.8925    75.0881  -0.265   0.7934
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 317.4 on 23 degrees of freedom
## Multiple R-squared:  0.4581, Adjusted R-squared:  0.05751
## F-statistic: 1.144 on 17 and 23 DF,  p-value: 0.3758
```

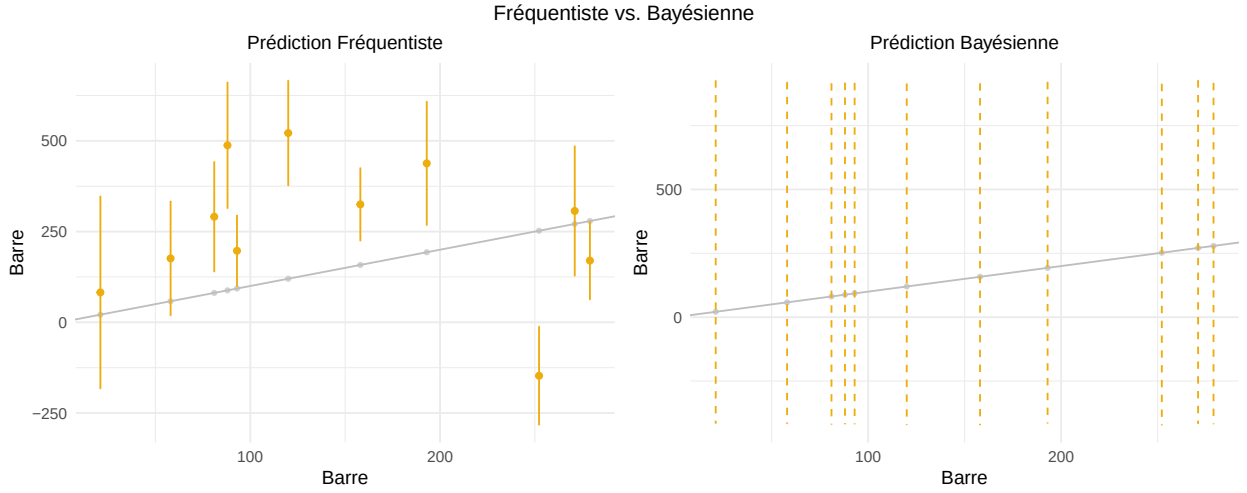
Néanmins, 5 variables sont considérés commes significatives dans le meilleur modèle AIC : le *taux brut de réussite de série ES*, le *taux de réussite attendu de série L*, le *taux de réussite attendu de série S*, et le *taux d'accès brut de première*

```
##
## Call:
## lm(formula = Barre ~ effectif_presents_serie_es + taux_brut_de_reussite_serie_es +
##     taux_brut_de_reussite_serie_s + taux_reussite_attendu_serie_l +
##     taux_reussite_attendu_serie_es + taux_reussite_attendu_serie_s +
##     taux_acces_brut_seconde_bac + taux_acces_attendu_seconde_bac +
##     taux_acces_brut_premiere_bac + taux_acces_attendu_premiere_bac,
##     data = train_en)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -386.43 -153.57  -15.04  107.62 1002.21
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1048.914   1008.155  -1.040  0.30645
## effectif_presents_serie_es        -2.781      1.742  -1.596  0.12096
## taux_brut_de_reussite_serie_es    14.652      6.679   2.194  0.03614 *
## taux_brut_de_reussite_serie_s     21.474     11.904   1.804  0.08129 .
## taux_reussite_attendu_serie_l    -33.320     13.768  -2.420  0.02178 *
## taux_reussite_attendu_serie_es    33.790     18.699   1.807  0.08079 .
## taux_reussite_attendu_serie_s    -42.527     17.944  -2.370  0.02442 *
## taux_acces_brut_seconde_bac       21.957     14.201   1.546  0.13255
## taux_acces_attendu_seconde_bac   -47.740     24.826  -1.923  0.06402 .
## taux_acces_brut_premiere_bac     -57.223     24.293  -2.356  0.02523 *
## taux_acces_attendu_premiere_bac  102.553     36.266   2.828  0.00827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 295.4 on 30 degrees of freedom
## Multiple R-squared:  0.3876, Adjusted R-squared:  0.1835
## F-statistic: 1.899 on 10 and 30 DF,  p-value: 0.08526
```

Prédiction

Dans la prédiction, le modèle fréquentiste surestime toujours dans la plupart du temps et il y a peu d'observations qui se trouvent dans l'intervalle de confiance. Dans le cas bayésien, l'incertitude est toujours très forte. Nous avons tous les observations qui se trouvent dans l'intervalle de crédibilité du modèle.



Pour les mutations en mathématiques en anglais, les covariables significatives sont différentes. Les covariables n'agissent pas de la même manière dans les deux disciplines.

6. Conclusion

Dans notre étude, nous retrouvons les covariables similaires dans les approches bayésiens et fréquentistes. Néanmoins lorsque nous avons beaucoup de variables, le modèle fréquentiste a tendance à garder plus de variables que le modèle bayésien.

L'incertitude est plus large dans le cas bayésien qui inclut la grande partie des vraies valeurs observées dans son intervalle de crédibilité. L'écart entre les vraies valeurs et la prédiction du modèle fréquentiste reste grande par rapport au modèle bayésien.