Projet Bayesien - Mutations -

Miora Tsiry R.

6/07/2022

Introduction

Le jeu de données “Mutations” contient les informations relatives aux mutations des enseignants de collèges et lycées français. Ce jeu de données retrace le nombre de points requis dans différentes disciplines de différents lieux d'établissements permettant à l'enseignant d'obtenir sa mutation professionnelle dans le lycée/collège désiré.
Notre base de données possède 516 lignes et 23 colonnes. Dans les 5 premières colonnes, nous retrouverons les caractéristiques propres au lieu d'établissement (code de l'établissement, ville, commune, nom de l'établissement) qui sont repértoriés dans des variables qualitatives. Dans les autres variables (quantitatives), nous recensons le nombre de points (colonne Barre) qui est l'objet de notre analyse prédictive, et également plusieurs variables relatant les effectifs, les taux de réussite au bac, les taux d'accès aux niveaux/filières respectifs.

Nous allons commencer par charger puis observer le résumé des données.

```
##  code_etablissement    ville           etablissement         commune
##  Length:516         Length:516         Length:516         Min.   :78005
##  Class :character   Class :character   Class :character   1st Qu.:91027
##  Mode  :character   Mode  :character   Mode  :character   Median :92012
##                                                           Mean   :89739
##                                                           3rd Qu.:95018
##                                                           Max.   :95637
##    Matiere              Barre        effectif_presents_serie_l
##  Length:516         Min.   :  21.0   Min.   :  6.00
##  Class :character   1st Qu.: 111.0   1st Qu.: 18.00
##  Mode  :character   Median : 196.0   Median : 30.00
##                     Mean   : 321.9   Mean   : 34.24
##                     3rd Qu.: 292.0   3rd Qu.: 47.00
##                     Max.   :2056.0   Max.   :133.00
##  effectif_presents_serie_es effectif_presents_serie_s
##  Min.   : 10.00             Min.   : 13.0
##  1st Qu.: 53.00             1st Qu.: 64.0
##  Median : 69.00             Median :100.0
##  Mean   : 74.42             Mean   :106.1
##  3rd Qu.: 99.00             3rd Qu.:140.0
##  Max.   :192.00             Max.   :328.0
##  taux_brut_de_reussite_serie_l taux_brut_de_reussite_serie_es
##  Min.   : 36.00                Min.   : 51.0
##  1st Qu.: 82.00                1st Qu.: 81.0
##  Median : 89.00                Median : 88.0
##  Mean   : 86.35                Mean   : 86.4
```

```
##  3rd Qu.: 94.00                3rd Qu.: 94.0
##  Max.   :100.00                Max.   :100.0
##  taux_brut_de_reussite_serie_s taux_reussite_attendu_serie_l
##  Min.   :50.00                 Min.   :65.00
##  1st Qu.:81.00                 1st Qu.:84.00
##  Median :88.00                 Median :89.00
##  Mean   :86.23                 Mean   :86.91
##  3rd Qu.:93.00                 3rd Qu.:92.00
##  Max.   :99.00                 Max.   :98.00
##  taux_reussite_attendu_serie_es taux_reussite_attendu_serie_s
##  Min.   :61.00                  Min.   :61.00
##  1st Qu.:86.00                  1st Qu.:86.00
##  Median :90.00                  Median :89.00
##  Mean   :87.97                  Mean   :87.39
##  3rd Qu.:94.00                  3rd Qu.:94.00
##  Max.   :98.00                  Max.   :98.00
##  effectif_de_seconde effectif_de_premiere taux_acces_brut_seconde_bac
##  Min.   : 36.0       Min.   : 36.0        Min.   :49.00
##  1st Qu.:268.0       1st Qu.:226.5        1st Qu.:64.00
##  Median :336.0       Median :289.0        Median :71.00
##  Mean   :351.6       Mean   :307.7        Mean   :69.61
##  3rd Qu.:415.0       3rd Qu.:364.0        3rd Qu.:76.00
##  Max.   :764.0       Max.   :691.0        Max.   :87.00
##  taux_acces_attendu_seconde_bac taux_acces_brut_premiere_bac
##  Min.   :50.00                  Min.   :65.00
##  1st Qu.:64.00                  1st Qu.:82.00
##  Median :69.00                  Median :85.00
##  Mean   :68.47                  Mean   :84.53
##  3rd Qu.:73.00                  3rd Qu.:89.25
##  Max.   :83.00                  Max.   :97.00
##  taux_acces_attendu_premiere_bac taux_brut_de_reussite_total_series
##  Min.   :70.00                   Min.   :64.00
##  1st Qu.:81.00                   1st Qu.:82.00
##  Median :85.00                   Median :86.00
##  Mean   :84.19                   Mean   :85.46
##  3rd Qu.:89.00                   3rd Qu.:91.00
##  Max.   :94.00                   Max.   :98.00
##  taux_reussite_attendu_total_series
##  Min.   :67.0
##  1st Qu.:84.0
##  Median :88.0
##  Mean   :86.8
##  3rd Qu.:92.0
##  Max.   :98.0
```

Sur notre variable réponse Barre, nous nous apercevons par lecture que son min est à 21 et son max à 2056. Nous verrons qu'il y a un couple Etablissement/Matiere qui nécessite 2056 points, le maximum de points. L'étendue des valeurs est assez importante. Une distribution de cette variable sera tracée plus loin dans l'analyse descriptive. Les taux d'accès et/ou de réussite ne sont pas incohérents car ont leurs valeurs comprises entre 0 et 100 (il s'agit d'un pourcentage).

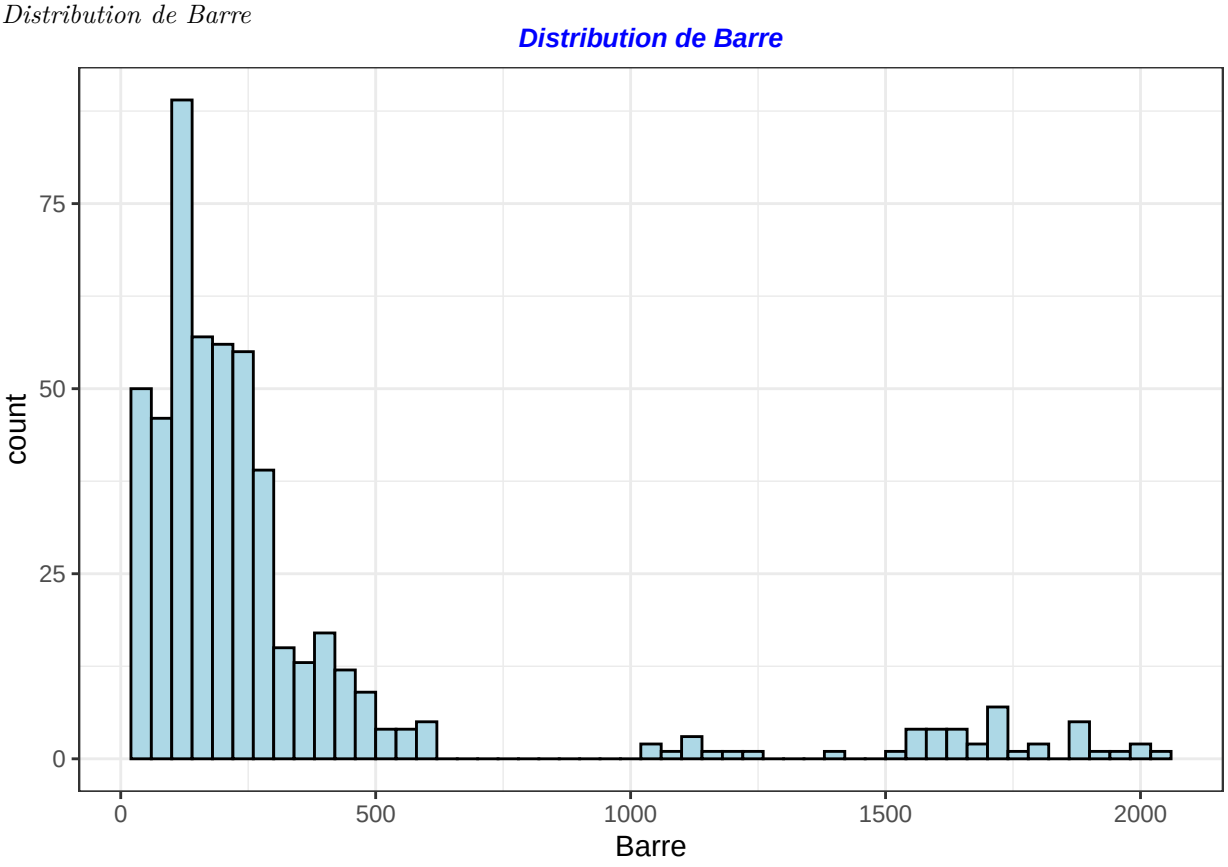
I. Statistiques descriptives

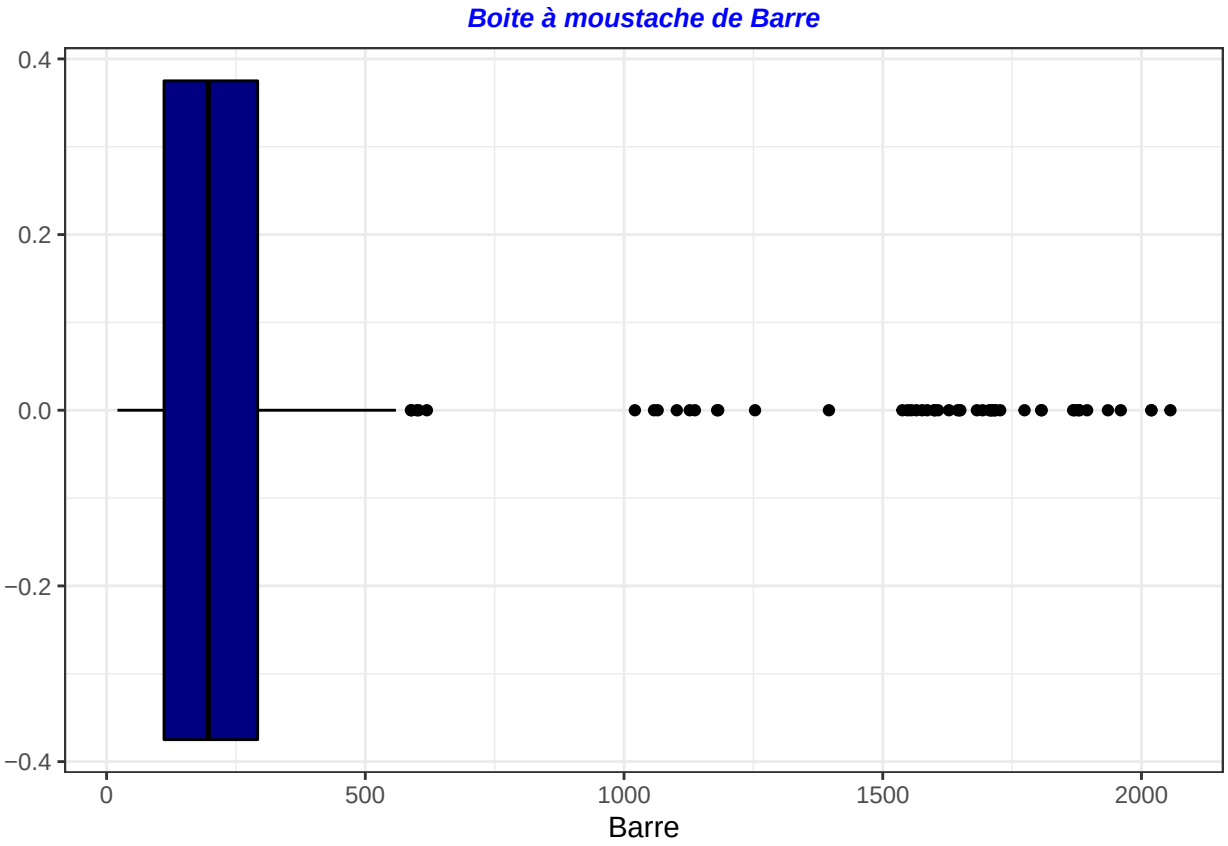
```
sum(is.na(datamutations))
```

```
## [1] 0
```

Il n'y a pas de données manquantes ce qui est rassurant.
Regardons la distribution de la variable Barre.

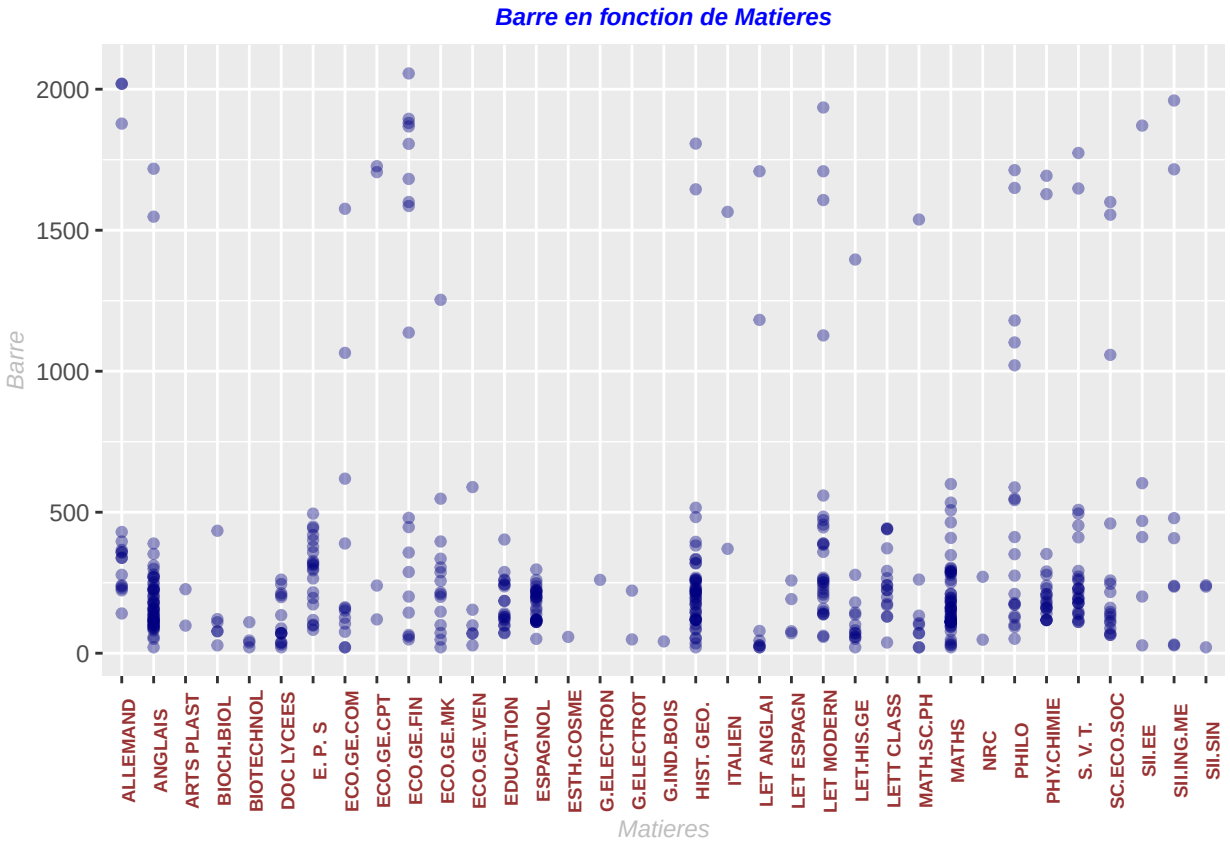
Distribution de Barre





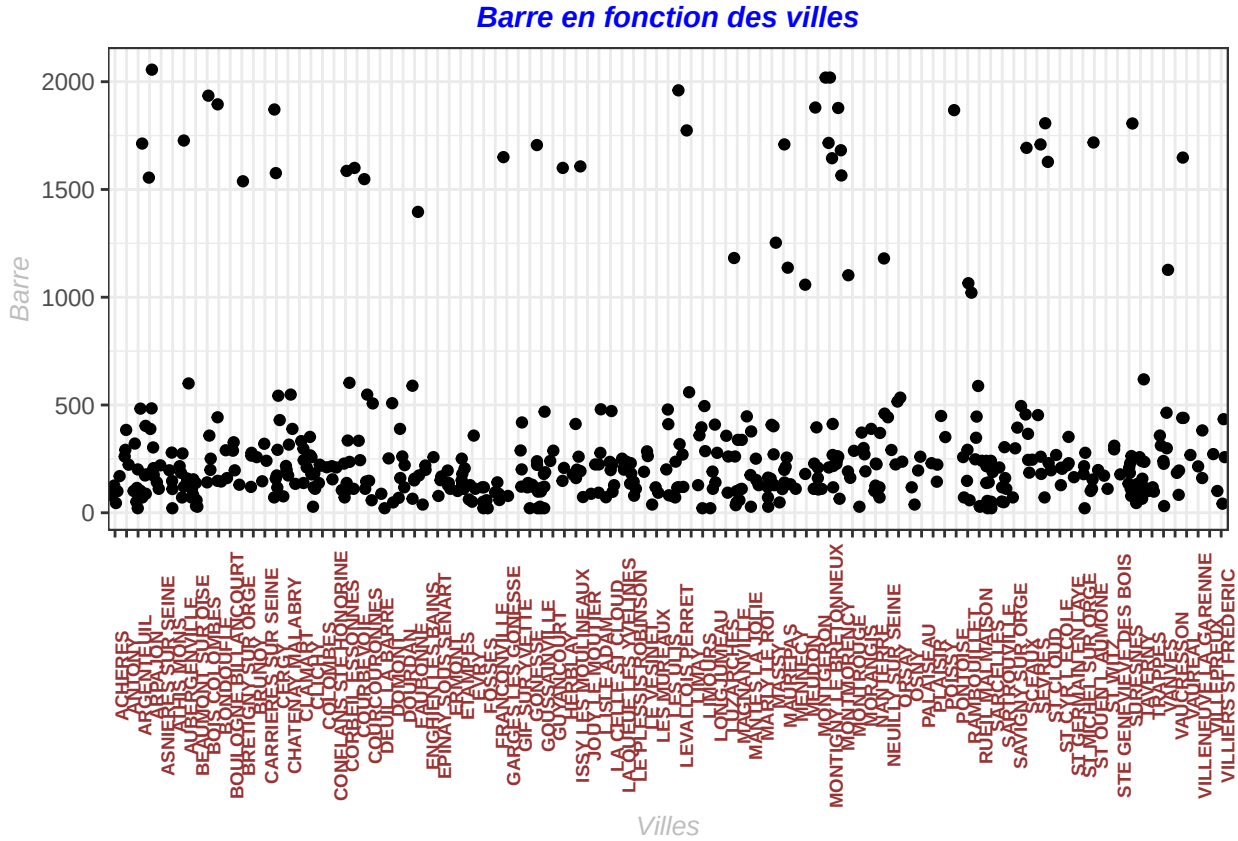
La distribution générale de Barre montre une forte asymétrie à droite (vue sur l'histogramme et sur la boite à moustache). La plupart des établissements exigent des points en dessous de 500, seuls certains sont hautement stricts sur des disciplines et vont jusqu'à 2000 points requis. Nous remarquons également que les moustaches sont courtes, les plages des 25% inférieurs et 25% supérieurs des valeurs s'étendent entre 21 le minimum et un peu plus de 500. Aussi, nous voyons un certain nombre de valeurs aberrantes en dehors de la boite et des moustaches, le maximum étant atteint en 2056 points. L'allure de cette distribution illustre également une queue épaisse visiblement, et peut évoquer une loi de Pareto.

Relation de Barre en fonction des Matieres



De façon globale, les nombres de points (Barre) dans les disciplines exigés se situent en grande majorité entre 0 et un peu plus de 500 points, il y a certains points au delà de 1000 et jusqu'à 2000. Ces établissements sont ceux qui exigent le plus grand nombre de points dans des disciplines précises. Par exemple, en Allemand, un lycée en particulier exige bien plus de 2000 points ; de même, en éco-ge-fin, il y a un établissement qui exige un nombre de points supérieur à 2000. Dans les matières phares comme l'anglais, l'eps, l'histoire-géo ou les maths en particulier, la majorité des établissements exigent un nombre de points en dessous ou égal à 500 environ.

Relation de Barre en fonction des Villes

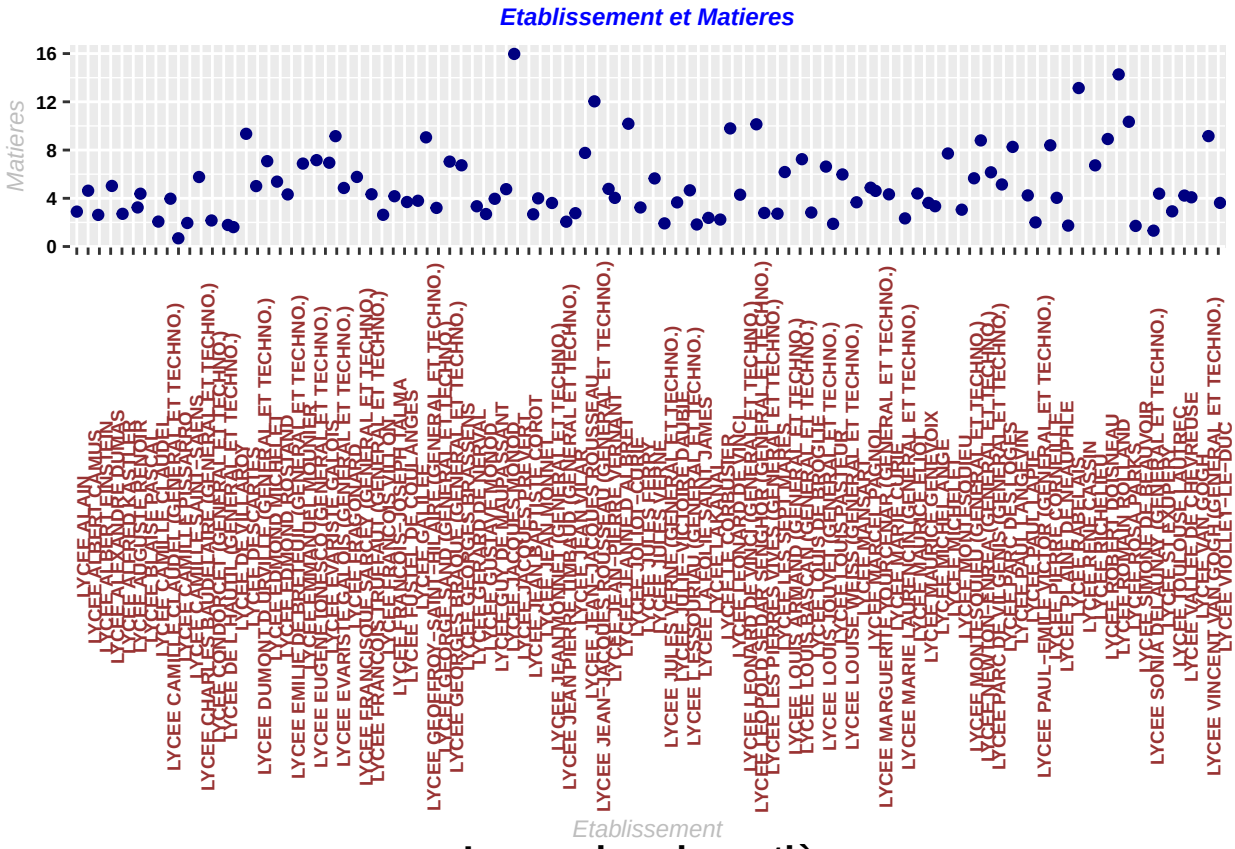



```
##                     ville  Barre
## 195                ARPAJON 2056.0
## 14  MONTIGNY LE BRETONNEUX 2019.2
## 16  MONTIGNY LE BRETONNEUX 2019.2
## 347       LEVALLOIS PERRET 1960.0
## 254          BOIS COLOMBES 1935.2
## 218              BONDOUFLE 1895.0
```

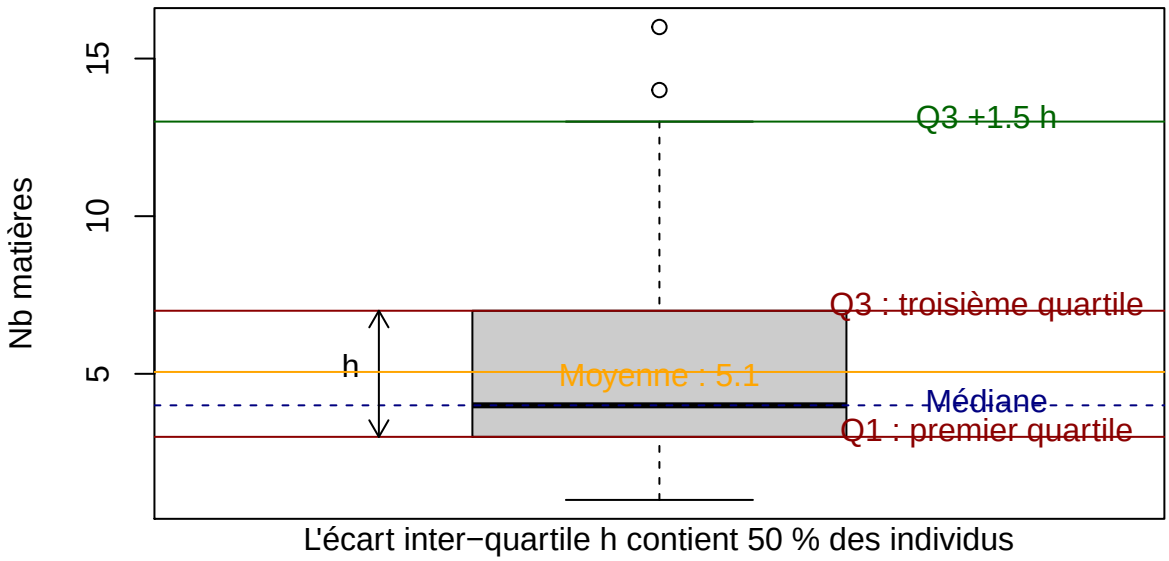
Le top 5 des villes où se situent les établissements les plus exigeants sont Arpajon, Montigny-le-Bretonneux, Levallois Perret, Bois Colombes et Bondoufle.

Comme nous allons nous focaliser sur les couples Etablissements/Matiere, illustrons les données de ces couples.

Relation de Etablissement en fonction des Matieres



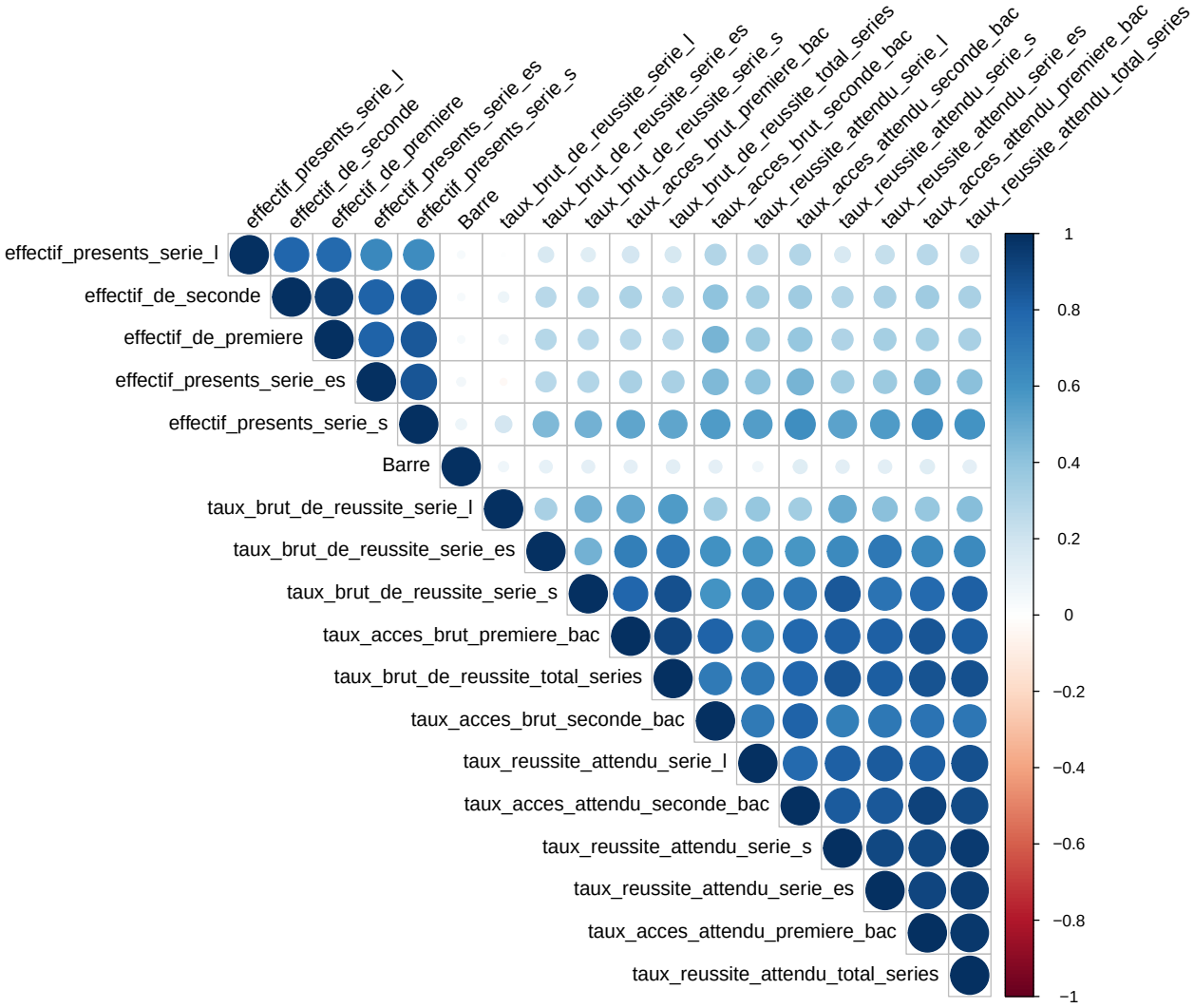
Le nombre de matières



Sur ces graphes sont illustrés le nombre de matières enseignées par établissement. Ainsi, en moyenne, il y a 5,1 matières par établissement renseignées dans notre échantillon. Pour quelques établissements, nous pouvons apercevoir qu'il y a 16 matières renseignées auxquelles peuvent prétendre les enseignants. Avec la boite à moustaches, nous lisons que 75% des établissements (Q3) ont un nombre de matières entre 3 et 6.

Investiguons la corrélation sur les variables quantitatives uniquement afin de mettre en évidence la/les variables qui peuvent potentiellement avoir un lien avec notre variable cible Barre.

Matrice de corrélation sur variables quantitatives



La variable Barre a assez peu de corrélation avec les variables explicatives évoquées. Elle présente de très légères corrélations toutes négatives avec ces dernières, dont la plus importante est effectif_presents_serie_s (-0.6).
En revanche, les autres variables ont de très fortes corrélations entre elles.
Sur le groupe des effectifs notamment, nous voyons une forte corrélation positive entre les effectifs d'un niveau inférieur et supérieur, et également entre les séries.
De la même façon pour les taux de réussite sur les séries qui ont une influence entre eux, puisque cela concerne le même établissement.
On pourrait ainsi se contenter de retenir par exemple une seule variable du groupe des variables effectifs et une seule variable du groupe taux de réussite, mais on va quand même les conserver ainsi.
A noter au passage que la variable taux_brut_de_reussite_serie_l n'a que très peu de corrélations avec les autres variables.
Si l'on souhaitait réduire le nombre de variables, certaines variables qui n'apportent que peu d'informations pourraient être écartées.

Après le constat de toutes ces informations, nous pouvons passer à l'étape de régression.

Procédons à présent à la régression linéaire bayesienne.

II. Régression linéaire

Ce que l'on cherche à expliquer ici c'est donc la variable réponse Barre en fonction par les variables caractéristiques du lycée.

Rappel de la méthode de régression linéaire ordinaire

Comme vu en cours, le modèle linéaire gaussien souhaite expliquer les observations (y_i) par des covariables $(x^1, ...x^p)$ avec le modèle :

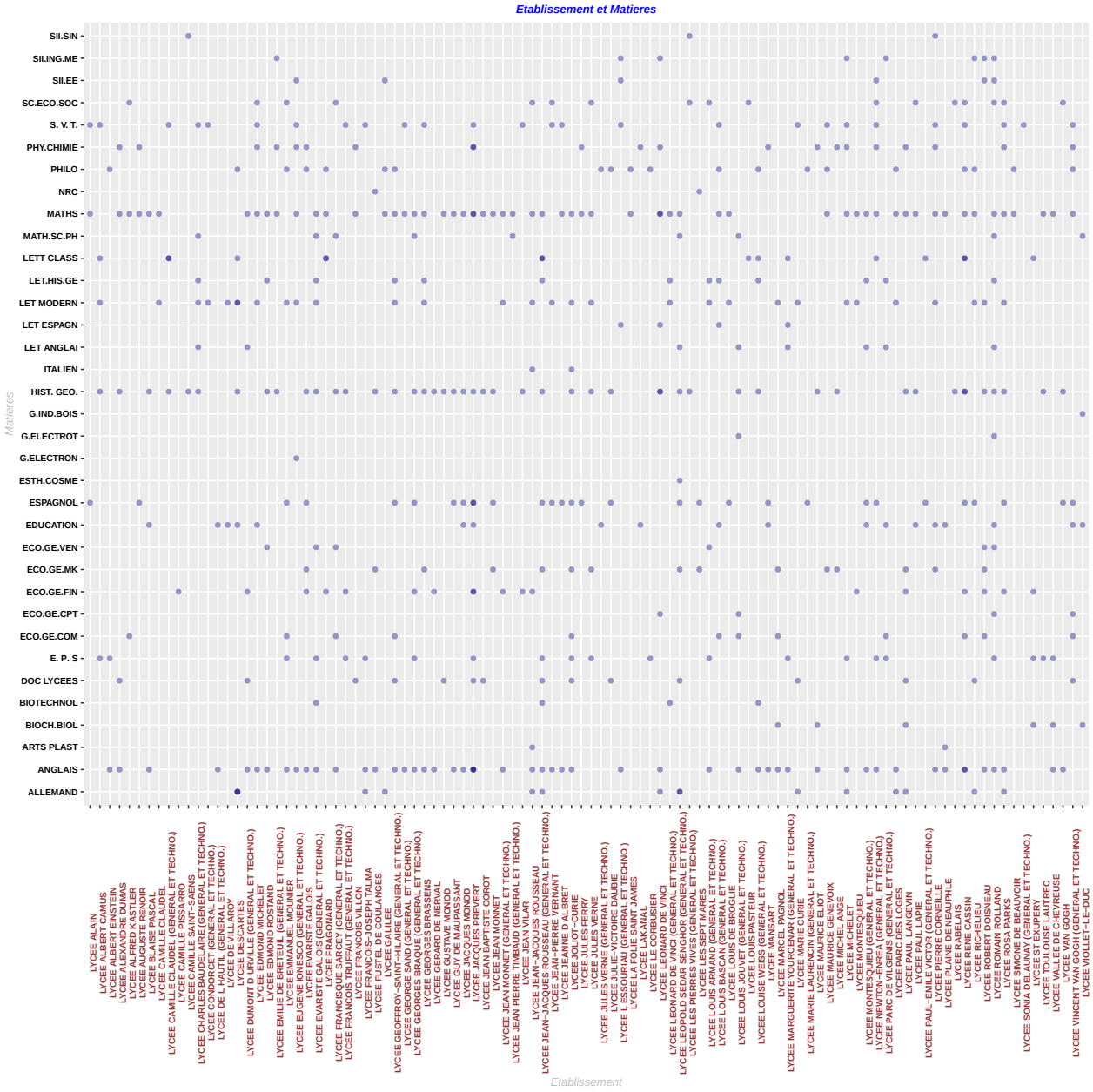
$y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + \epsilon_i$ $\epsilon_i \sim N(0, \sigma^2)$ iid.

On note y le vecteur des observations $(y_1, ..., y_n)$ et X la matrice des covariables. Dans le cadre fréquentiste, nous maximisons la vraisemblance
$L(\beta, \sigma^2 \mid y, X) = (2\pi\sigma^2)^{-\frac{n}{2}} exp[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)])$
et on a :
$\hat{\beta} = (X^T X)^{-1} X^T y$
$\hat{\sigma^2} = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$
En notation matricielle, cela se traduit par la formule suivante :
$y \mid \alpha, \beta, \sigma^2 \sim N_n(\alpha 1_n + X\beta, \sigma^2 I_n)$

Les y_i suivent des lois normales indépendantes où :
$E(y_i \mid \alpha, \beta, \sigma^2) = \alpha + \sum_{j=1}^{p} \beta_j x_{ij}$ $V(y_i \mid \alpha, \beta, \sigma^2) = \sigma^2$

II.1 Régression linéaire bayesienne

Au préalable, nous allons retraiter notre base de données pour nous assurer qu'il n'y ait pas de redondance d'informations (càd de doublons de lignes).



En regardant très attentivement, nous voyons que certains points apparaissent plus foncés, car il y a plus d'une observation concernée. En effet, certaines lignes (censée apparaître une seule fois) sont redondantes, entre autres cela concerne ANGLAIS, ALLEMAND, LETT CLASS...
Pour pallier à ces doublons, nous allons supprimer les lignes en double.

```
datamutations_nodup = datamutations %>% distinct()
```

Ainsi, nous avons supprimé 6 lignes, il y a donc à présent 510 lignes.

Rappel sur le contexte bayesien

Le choix de la loi a priori est une étape fondamentale dans la régression bayesienne.
Dans notre cas, nous allons choisir une loi a priori de Zellner à partir du moment où on considère qu'aucune information n'est disponible sur la loi a priori. L'avantage de cette loi a priori est qu'elle permet d'introduire

des informations (très faibles) sur le paramètre de localisation de régression g et surtout d'éviter l'écueil principal de la prior à savoir la structure de corrélation.

Ainsi, nous prenons comme loi a priori

$\beta \mid \sigma^2, X \sim N_{k+1}(\tilde{\beta}, \sigma^2 M^{-1})$
$\sigma^2 \mid X \sim IG(a, b)$
où M est une matrice symétrique définie positive de taille (k+1)x(k+1)

Il faut ainsi fixer M de sorte que :
$\beta \mid \sigma^2, X \sim N_{k+1}(\tilde{\beta}, g\sigma^2({}^tXX)^{-1})$
$\sigma^2 \sim \pi(\sigma^2 \mid X) \propto \sigma^{-2}$

Il faut choisir le paramètre g, g=1 ou g=n selon le poids accordé à la prior.
Pour l'espérance à priori $\tilde{\beta}$ ou pourra la prendre = 0 comme nous n'avons pas d'information a priori.
Ainsi, la loi a posteriori se définit alors comme suit :

$\beta \mid \sigma^2, y, X \sim N_{k+1}(\frac{g}{g+1}\hat{\beta}, \frac{\sigma^2 g}{g+1}({}^tXX)^{-1})$ $\sigma^2 \mid y, X \sim IG(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(g+1)}(-\hat{\beta}^T){}^tXX(-\hat{\beta}))$

donc : $\beta \mid y, X \sim Student_{k+1}(n, \frac{g}{g+1}\hat{\beta}, \frac{g(s^2+((\hat{\beta})^T{}^tXX\hat{\beta})/(g+1))}{n(g+1)}({}^tXX)^{-1})$

Nous allons opérer la transformation log(Barre) dans notre régression linéaire bayésienne et standardiser la matrice de design X, car nous utiliserons la fonction BayesReg plus tard.

On cherche à calculer la moyenne à priori, à partir de la formule suivante: $E^{\pi}(\beta \mid y) = \frac{g}{g+1}\left(\hat{\beta} + \tilde{\beta}/g\right)$ Où $\hat{\beta}$ est le vecteur des coefficients du modèle linéaire ordinaire obtenu par maximum de vraisemblance.

Nous n'allons travailler que sur les variables quantitatives.

```
##
## Call:
## lm(formula = Y ~ X - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -464.20 -200.61 -120.79    0.27 1649.40
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## Xeffectif_presents_serie_l         1.0273     1.5824   0.649   0.5165
## Xeffectif_presents_serie_es        0.2531     1.2198   0.208   0.8357
## Xeffectif_presents_serie_s         0.4378     0.8361   0.524   0.6008
## Xtaux_brut_de_reussite_serie_l     2.6351     2.4441   1.078   0.2815
## Xtaux_brut_de_reussite_serie_es    4.5524     4.1671   1.092   0.2752
## Xtaux_brut_de_reussite_serie_s     8.0597     6.2337   1.293   0.1966
## Xtaux_reussite_attendu_serie_l   -13.8616     6.7846  -2.043   0.0416 *
## Xtaux_reussite_attendu_serie_es    4.5967     8.1920   0.561   0.5750
## Xtaux_reussite_attendu_serie_s    -0.8590     9.0050  -0.095   0.9240
## Xeffectif_de_seconde               0.2301     0.6021   0.382   0.7024
```

```
## Xeffectif_de_premiere                -0.6820     0.6193  -1.101   0.2713
## Xtaux_acces_brut_seconde_bac         10.1694     5.5315   1.838   0.0666 .
## Xtaux_acces_attendu_seconde_bac      -3.3842     8.4377  -0.401   0.6885
## Xtaux_acces_brut_premiere_bac       -20.9481    10.4912  -1.997   0.0464 *
## Xtaux_acces_attendu_premiere_bac     23.2834    13.8433   1.682   0.0932 .
## Xtaux_brut_de_reussite_total_series  -3.8579    12.5706  -0.307   0.7591
## Xtaux_reussite_attendu_total_series  -4.9329    21.8389  -0.226   0.8214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 418.5 on 493 degrees of freedom
## Multiple R-squared:   0.39,  Adjusted R-squared:  0.369
## F-statistic: 18.54 on 17 and 493 DF,  p-value: < 2.2e-16
##
##            Xeffectif_presents_serie_l Xeffectif_presents_serie_es
## Intercept                    5.258490                   5.2584901
## mbetabayes                   1.025252                   0.2526441
##            Xeffectif_presents_serie_s Xtaux_brut_de_reussite_serie_l
## Intercept                     5.25849                       5.258490
## mbetabayes                    0.43694                       2.629913
##            Xtaux_brut_de_reussite_serie_es Xtaux_brut_de_reussite_serie_s
## Intercept                         5.258490                       5.258490
## mbetabayes                        4.543503                       8.043969
##            Xtaux_reussite_attendu_serie_l Xtaux_reussite_attendu_serie_es
## Intercept                         5.25849                       5.258490
## mbetabayes                      -13.83448                       4.587743
##            Xtaux_reussite_attendu_serie_s Xeffectif_de_seconde
## Intercept                       5.2584901           5.2584901
## mbetabayes                     -0.8573405           0.2296972
##            Xeffectif_de_premiere Xtaux_acces_brut_seconde_bac
## Intercept              5.2584901                     5.25849
## mbetabayes            -0.6806685                    10.14950
##            Xtaux_acces_attendu_seconde_bac Xtaux_acces_brut_premiere_bac
## Intercept                         5.258490                       5.25849
## mbetabayes                       -3.377625                     -20.90711
##            Xtaux_acces_attendu_premiere_bac Xtaux_brut_de_reussite_total_series
## Intercept                           5.25849                             5.258490
## mbetabayes                         23.23788                            -3.850342
##            Xtaux_reussite_attendu_total_series
## Intercept                             5.258490
## mbetabayes                           -4.923234
##          Xeffectif_presents_serie_l        Xeffectif_presents_serie_es
##                           1.0272628                          0.2531395
##          Xeffectif_presents_serie_s      Xtaux_brut_de_reussite_serie_l
##                           0.4377968                          2.6350694
##     Xtaux_brut_de_reussite_serie_es      Xtaux_brut_de_reussite_serie_s
##                           4.5524115                          8.0597418
##      Xtaux_reussite_attendu_serie_l     Xtaux_reussite_attendu_serie_es
##                          -13.8616056                          4.5967385
##      Xtaux_reussite_attendu_serie_s                Xeffectif_de_seconde
##                          -0.8590216                          0.2301476
##               Xeffectif_de_premiere        Xtaux_acces_brut_seconde_bac
##                          -0.6820032                         10.1694025
##     Xtaux_acces_attendu_seconde_bac       Xtaux_acces_brut_premiere_bac
```

```
##                           -3.3842478                           -20.9481073
##    Xtaux_acces_attendu_premiere_bac Xtaux_brut_de_reussite_total_series
##                           23.2834432                            -3.8578916
## Xtaux_reussite_attendu_total_series
##                           -4.9328877
```

Nous obtenons des coefficients assez proches sur les 2 méthodes.

Pour choisir les covariables significatives, nous allons nous servir des facteurs de Bayes de la fonction BayesReg évoquée précédemment mais en l'adaptant.

```
##                            colnames(X) bfactor
## 1            effectif_presents_serie_l -1.3515
## 2           effectif_presents_serie_es -1.3535
## 3            effectif_presents_serie_s -1.3542
## 4         taux_brut_de_reussite_serie_l -1.3541
## 5        taux_brut_de_reussite_serie_es -1.3455
## 6         taux_brut_de_reussite_serie_s -1.3434
## 7         taux_reussite_attendu_serie_l -1.3436
## 8        taux_reussite_attendu_serie_es -1.3542
## 9         taux_reussite_attendu_serie_s -1.3515
## 10                 effectif_de_seconde -1.3519
## 11                effectif_de_premiere -1.3511
## 12          taux_acces_brut_seconde_bac -1.3435
## 13       taux_acces_attendu_seconde_bac -1.3398
## 14         taux_acces_brut_premiere_bac -1.3313
## 15      taux_acces_attendu_premiere_bac -1.3216
## 16 taux_brut_de_reussite_total_series -1.3540
## 17 taux_reussite_attendu_total_series -1.3537
##                            colnames(X) bfactor
## 1            effectif_presents_serie_l -0.1492
## 2           effectif_presents_serie_es -0.1502
## 3            effectif_presents_serie_s -0.1505
## 4         taux_brut_de_reussite_serie_l -0.1505
## 5        taux_brut_de_reussite_serie_es -0.1461
## 6         taux_brut_de_reussite_serie_s -0.1451
## 7         taux_reussite_attendu_serie_l -0.1452
## 8        taux_reussite_attendu_serie_es -0.1505
## 9         taux_reussite_attendu_serie_s -0.1491
## 10                 effectif_de_seconde -0.1494
## 11                effectif_de_premiere -0.1490
## 12          taux_acces_brut_seconde_bac -0.1452
## 13       taux_acces_attendu_seconde_bac -0.1433
## 14         taux_acces_brut_premiere_bac -0.1390
## 15      taux_acces_attendu_premiere_bac -0.1342
## 16 taux_brut_de_reussite_total_series -0.1504
## 17 taux_reussite_attendu_total_series -0.1502
```

En accordant plus de poids à la loi a priori (avec g=1), certaines variables deviennent significatives au sens de Jeffreys, notamment les variables 6,7,12,13,14 et 15.(les plus bas qui se rapprochent de 0)

Nous pouvons sélectionner ces 6 variables ressortant significatives qui sont :
- la 6 : taux_brut_de_reussite_serie_s

- la 7 : taux_reussite_attendu_serie_l
- la 12 : taux_acces-brut_seconde_bac
- la 13 : taux_acces_attendu_seconde_bac
- la 14 : taux_acces_brut_premiere_bac
- la 15 : taux_acces_attendu_premiere_bac

Investiguons à présent le choix de modèles par un échantillonneur de Gibbs basé sur la fonction ModChoBayesReg réadaptée, de la librairie BayesReg.

```
##
## bCalc + false
## Model posterior probabilities are calculated by Gibbs
##
##    Top10Models PostProb
## 1           15   0.3691
## 2           17   0.0479
## 3        13 15   0.0399
## 4        14 15   0.0289
## 5        12 15   0.0255
## 6         4 15   0.0230
## 7        15 17   0.0212
## 8         7 15   0.0207
## 9         6 15   0.0207
## 10        5 15   0.0199
##
##
## $top10models
##  [1] "15"    "17"    "13 15" "14 15" "12 15" "4 15"  "15 17" "7 15"  "6 15"
## [10] "5 15"
##
## $postprobtop10
##  [1] 0.3690875 0.0479000 0.0399375 0.0288500 0.0254500 0.0230000 0.0212375
##  [8] 0.0207000 0.0207000 0.0199500
```

Le meilleur modèle est le modèle avec la variable 15 uniquement dont la probabilité est la plus importante. Cette variable est **taux_acces_attendu_premiere_bac**.
Cette même variable semble être sélectionnée également dans tous les autres hormis le 2ème meilleur modèle. En deuxième position, il y a justement le modèle constitué de la seule variable n°17 qui se positionne, c'est la variable taux_reussite_attendu_total_series. Puis, apparaissent également d'autres modèles avec les variables 14, 7, 12. Nous retrouvons presque les variables qui ont été sélectionnés par les facteurs de Bayes.

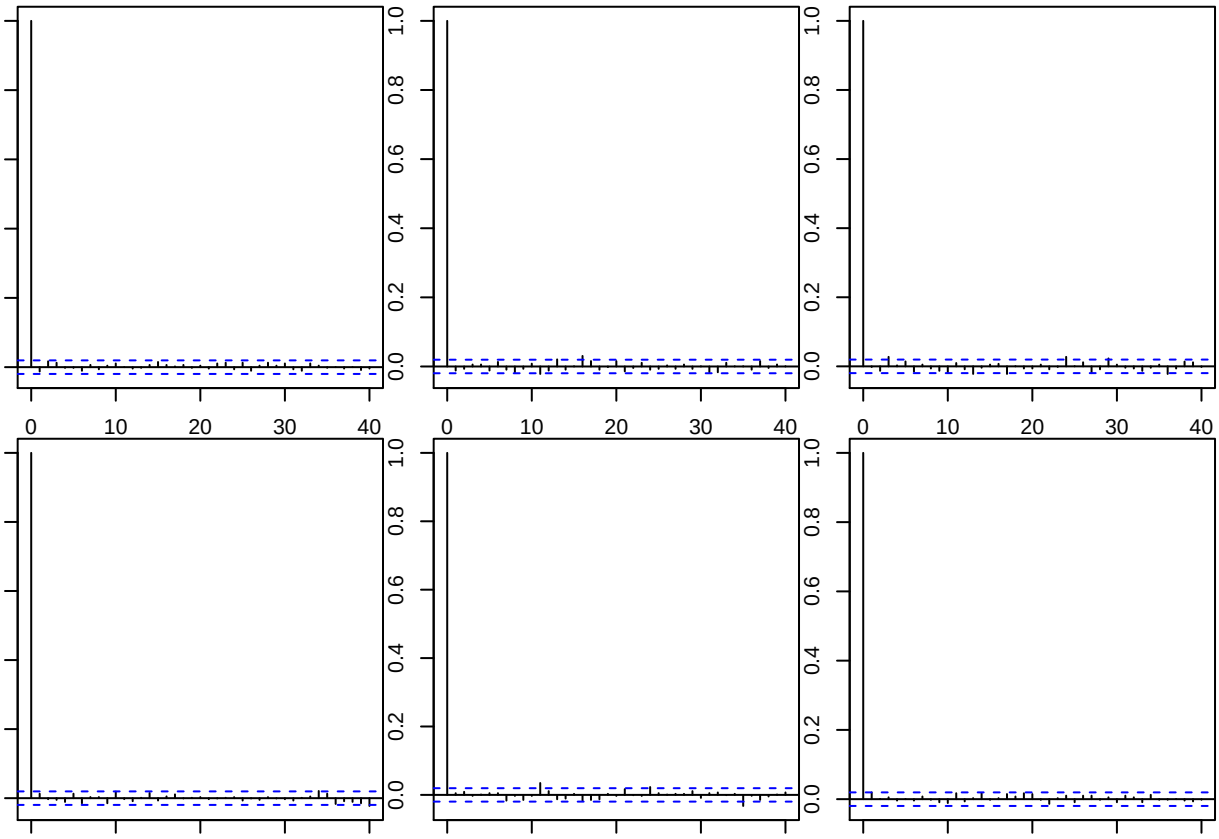
A présent, essayons l'algorithme de l'échantillonneur de Gibbs défini par la méthode vue en cours pour voir la sélection proposée.

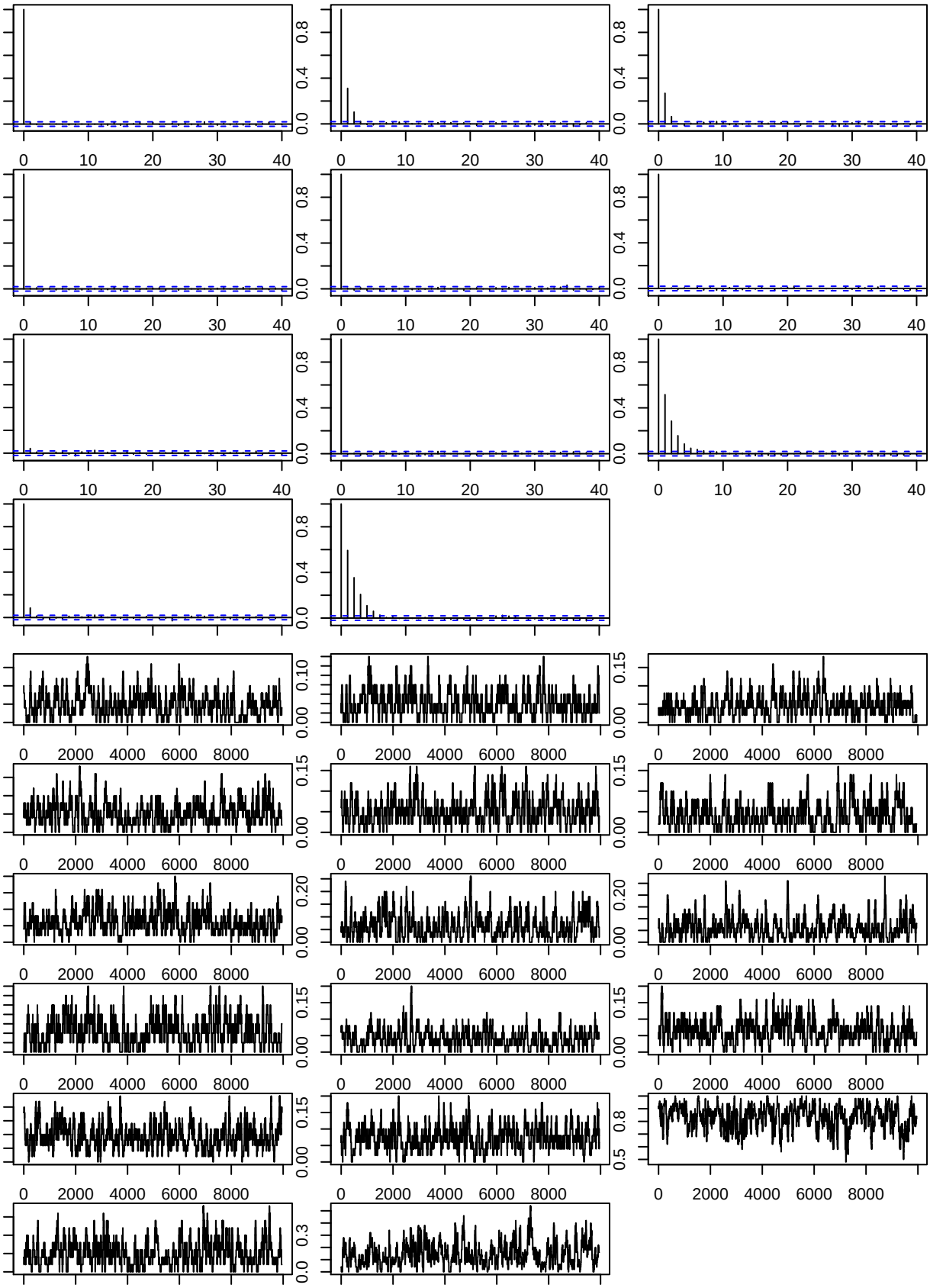
```
##                                    x gamma.mean
## 15    taux_acces_attendu_premiere_bac     0.8276
## 17 taux_reussite_attendu_total_series     0.1492
## 13     taux_acces_attendu_seconde_bac     0.0948
## 14       taux_acces_brut_premiere_bac     0.0753
## 8      taux_reussite_attendu_serie_es     0.0672
## 12        taux_acces_brut_seconde_bac     0.0649
## 9       taux_reussite_attendu_serie_s     0.0639
```

```
## 7       taux_reussite_attendu_serie_l     0.0638
## 4       taux_brut_de_reussite_serie_l     0.0534
## 5      taux_brut_de_reussite_serie_es     0.0534
## 16 taux_brut_de_reussite_total_series     0.0533
## 1           effectif_presents_serie_l     0.0500
## 3           effectif_presents_serie_s     0.0467
## 10                effectif_de_seconde     0.0457
## 6       taux_brut_de_reussite_serie_s     0.0444
## 11               effectif_de_premiere     0.0425
## 2          effectif_presents_serie_es     0.0421
```

Les résultats nous donnent une prédominance de la variable n°15 (taux_acces_attendu_premiere_bac), puis la n°17 en 2ème position, et la 13 et 14, et les prochaines variables à peu près dans le même ordre.

Nous devons maintenant vérifier la convergence de notre chaîne de Markov, car les premières itérations ne suivent généralement pas la loi cible. Nous allons donc observer la trace de la chaîne c'est-à-dire sa valeur prise à chaque itération pour détecter à quel moment la chaîne atteint sa loi limite.





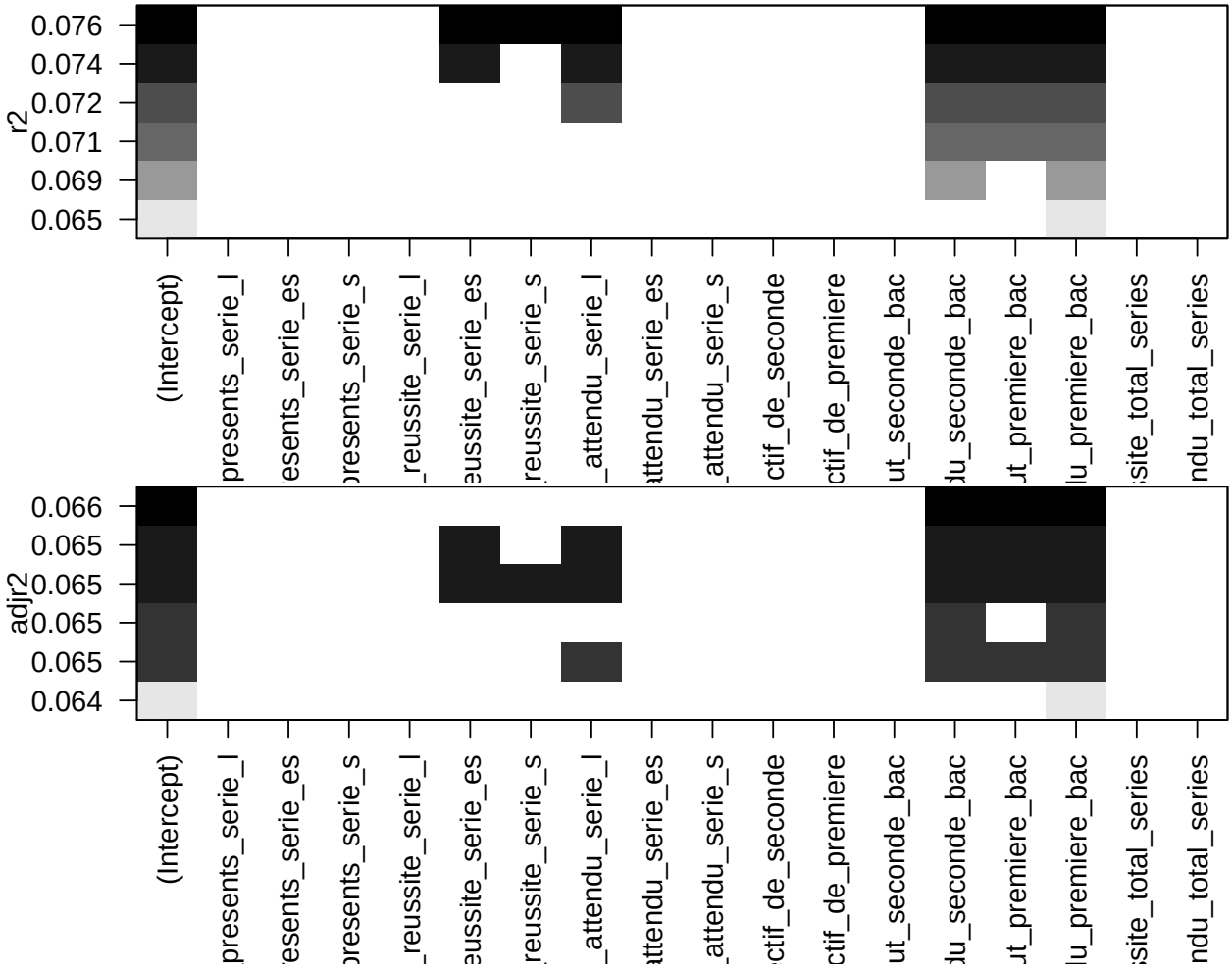
Dans chaque graphe d'autocorrélation, nous voyons que la courbe décroit très rapidement. De ce fait, notre

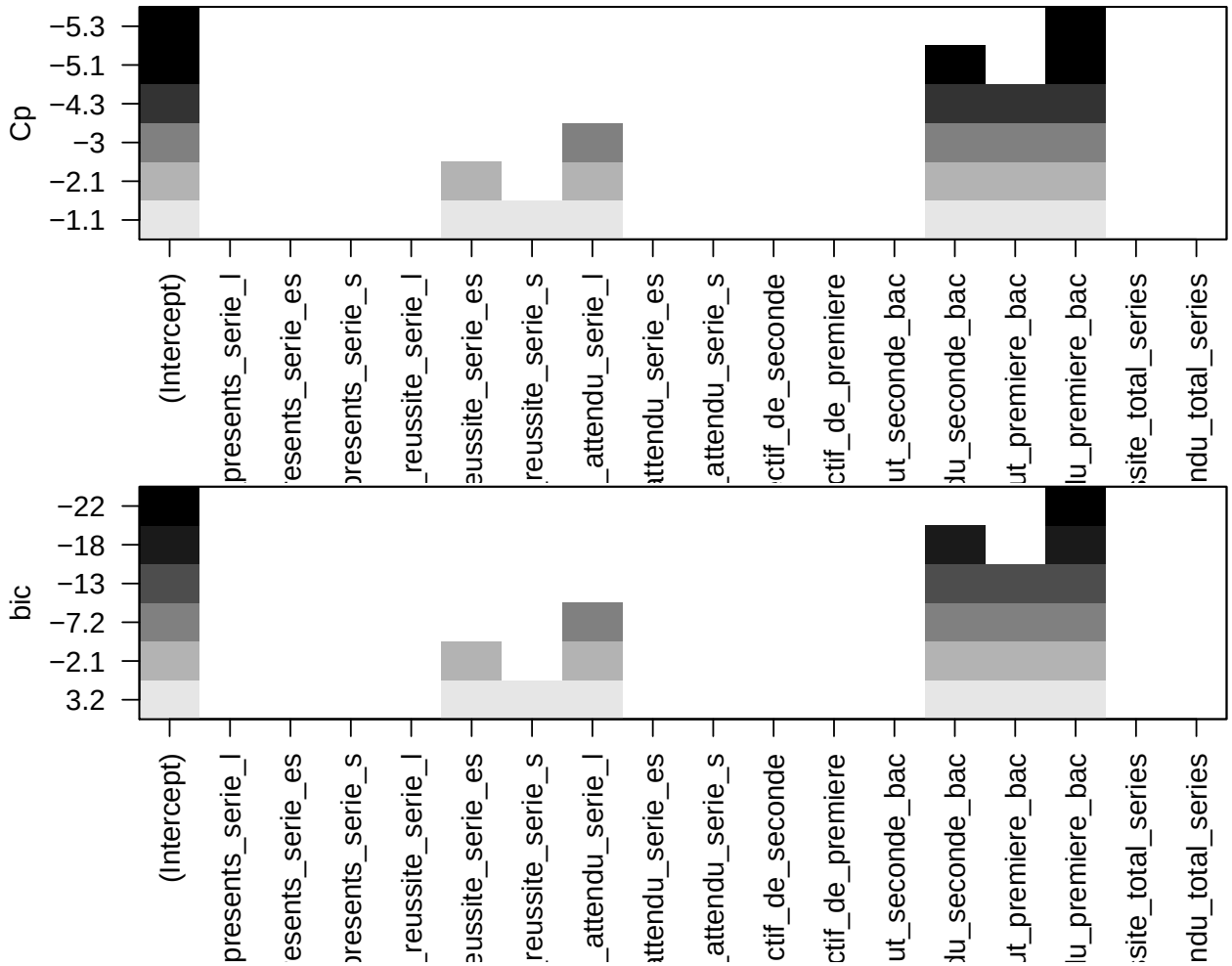
chaîne converge donc très rapidement, et ce dès les premières itérations.

Sur les graphes de trace, nous constatons que les courbes sont relativement stables autour d'une valeur moyenne pour chaque variable, cela montre que l'algorithme fonctionne correctement.

II.2 Régression linéaire classique

Comparons à présent ce modèle avec une approche fréquentiste en appliquant une recherche exhaustive sur la régression linéaire et une recherche pas à pas stepwise.





En réalisant la régression linéaire ordinaire, la seule variable qui est significative au sens de Wald est taux_acces_attendu_premiere_bac (la n°15 comme déjà vu avec le modèle bayesien).

Suivant les différents critères (Cp de Mallow's, Bic,R2, R2adj) la variable qui apparaît toujours est taux_acces_attendu_premiere_bac. Puis, selon Cp et Bic, les meilleurs modèles font apparaître également en plus de la n°15 la 5, 6, 7, 13 et 14.

Si nous envisageaons une recherche pas à pas (stepwise) basé sur le critère Aic par exemple, voici ce que cela donnerait :

Avec cette méthode stepwise, le meilleur modèle qui apparaît est celui avec la seule variable taux_acces_attendu_premiere_bac, tout comme parmi les meilleurs modèles de la recherche exhaustive.

Ainsi, la sélection de covariables a été réalisée par ces différentes méthodes (facteur de Bayes et échantillonnage de Gibbs).

Nous allons maintenant nous intéresser particulièrement aux 2 matières: Maths et Anglais.

III. Régression linéaire bayesienne sur Maths et Anglais

III.1 Régression bayesienne sur Maths et Anglais

Dans la régression avec g=n pour les maths, la variable 11 (effectif_de_premiere) est la seule significative.
Dans la régression avec g=1 pour les maths, il y a la n°4,5,10,11,12,13 et 14 qui sont significatives.

Dans la régression avec g=n pour l'anglais, la seule variable significative est la n°15 (taux_acces_attendu_premiere_bac).
Nous retrouvons encore cette variable prépondérante ici.
Dans la régression avec g=1 pour l'anglais, les variables n° 1,4,8,13 et 15 qui sont significatives.

Procédons à la sélection de modèles par l'échantillonneur de Gibbs proposé dans la fonction ModChoBayesReg (avec g=n).

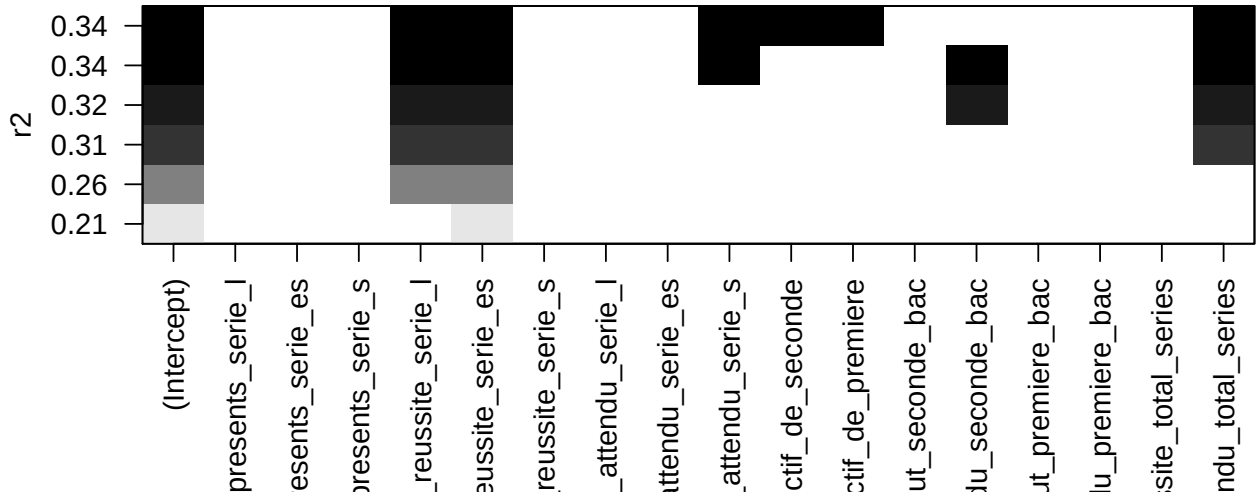
```
##
## Number of variables greather than 15
## Model posterior probabilities are estimated by using an MCMC algorithm
##
##    Top10Models PostProb
## 1            5   0.0236
## 2          4 5   0.0192
## 3       4 5 17   0.0132
## 4       4 5 15   0.0119
## 5       4 5 16   0.0117
## 6        4 5 8   0.0109
## 7          4 8   0.0106
## 8            8   0.0094
## 9        4 5 7   0.0076
## 10       4 5 6   0.0073
##
##
## $top10models
##  [1] "5"      "4 5"    "4 5 17" "4 5 15" "4 5 16" "4 5 8"  "4 8"    "8"
##  [9] "4 5 7"  "4 5 6"
##
## $postprobtop10
##  [1] 0.0236125 0.0191875 0.0131750 0.0118625 0.0116625 0.0109000 0.0105625
##  [8] 0.0094250 0.0076000 0.0072500
##
##
## Number of variables greather than 15
## Model posterior probabilities are estimated by using an MCMC algorithm
##
##    Top10Models PostProb
## 1            5   0.0162
## 2          4 5   0.0073
## 3    4 5 12 15   0.0072
## 4           15   0.0062
## 5    4 5 13 15   0.0061
## 6                0.0053
```

```
## 7      4 13 15   0.0048
## 8         4 15   0.0046
## 9          4 8   0.0038
## 10           8   0.0034
##
##
## $top10models
##  [1] "5"         "4 5"       "4 5 12 15" "15"        "4 5 13 15" ""
##  [7] "4 13 15"   "4 15"      "4 8"       "8"
##
## $postprobtop10
##  [1] 0.0162125 0.0073375 0.0071500 0.0061875 0.0061000 0.0053375 0.0047750
##  [8] 0.0046375 0.0037750 0.0033875
```

Dans les 2 cas, nous retrouvons le top 1 des modèles avec la seule variable n°5 (taux_brut_de_reussite_serie_es). En maths, la variable n°4 est également très souvent présente dans tous les autres modèles. (taux_brut_de_reussite_serie_l).
En anglais, en plus de la n°5, ce sont les variables n°4 et 15 qui sont souvent rencontrées dans les autres modèles. (taux_brut_de_reussite_serie_l et taux_acces_attendu_premiere_bac).

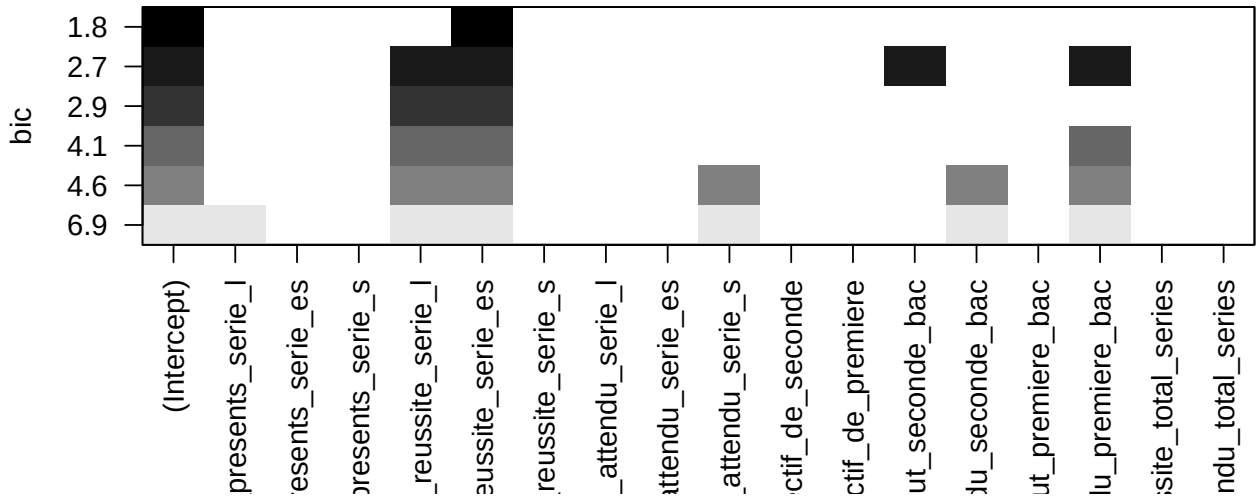
III.2 Régression linéaire sur Maths et Anglais

Revenons à un modèle linéaire gaussien ordinaire pour procéder à la comparaison.



adjr2
0.27
0.27
0.27
0.27
0.24
0.2
Cp
−3.3
−2.2
−2.1
−1.1
−0.41
0.32
bic
−5.8
−5.8
−5.6
−2.6
0.33
3.6
(Intercept)
presents_serie_l
esents_serie_es
presents_serie_s
reussite_serie_l
eussite_serie_es
reussite_serie_s
_attendu_serie_l
ttendu_serie_es
_attendu_serie_s
ctif_de_seconde
ctif_de_premiere
ut_seconde_bac
lu_seconde_bac
ut_premiere_bac
lu_premiere_bac
site_total_series
ndu_total_series

r2
0.33
0.31
0.28
0.2
0.16
0.11
adjr2
0.24
0.23
0.22
0.15
0.12
0.094
Cp
−1.5
−1
−0.27
0.91
1.4
2
(Intercept)
presents_serie_l
esents_serie_es
presents_serie_s
reussite_serie_l
eussite_serie_es
reussite_serie_s
_attendu_serie_l
ttendu_serie_es
_attendu_serie_s
ctif_de_seconde
ctif_de_premiere
ut_seconde_bac
lu_seconde_bac
ut_premiere_bac
lu_premiere_bac
site_total_series
ndu_total_series



La méthode stepwise pas à pas sélectionne un modèle, pour les maths, avec 3 variables dont 2 sont significatives et sont taux_brut_de_reussite_serie_es et taux_brut_de_reussite_serie_l.

Pour l'anglais, le modèle sélectionné par la méthode possède 4 variables qui sont toutes significatives à savoir, les 2 similaires à celui des maths et taux_acces_attendu_premiere_bac et taux_acces_brut_seconde_bac.

Conclusion

Dans un premier contexte général, toutes matières confondues, l'approche fréquentiste ou bayesienne sélectionne comme meilleur modèle celui ayant comme unique variable **taux_acces_attendu_premiere_bac**.

Dans les bases ne contenant que les matières maths ou anglais, les 2 approches sont moins en accord. L'approche bayesienne va privilégier le modèle avec l'unique variable **taux_brut_de_reussite_serie_es** tandis que les modèles de régression linéaire ordinaires seront moins parcimonieux avec au moins 3 variables dont taux_brut_de_reussite_serie_es, taux_brut_de_reussite_serie_l et taux_acces_attendu_premiere_bac.