Rapport - Statistique bayésienne

Philippe Real

26 février, 2020

Contents

1 Introduction

1.1 Lecture des données - description statistique

- Rennomage des colonnes

Nouveau Nom	Ancien Nom
Prs_l	effectif_presents_serie_l
prs_es	effectif_presents_serie_es
Prs_s	effectif_presents_serie_s
Eff_2nd	effectif_de_seconde
Eff_1er	effectif_de_premiere
Suc.brt_l	taux_brut_de_reussite_serie_l
Suc.brt_es	taux_brut_de_reussite_serie_es
Suc.brt_s	taux_brut_de_reussite_serie_s
Suc.att_l	taux_reussite_attendu_serie_l
Suc.att_es	taux_reussite_attendu_serie_es
Suc.att_s	taux_reussite_attendu_serie_s
Acc.brt_bac.2	taux_acces_brut_seconde_bac
Acc.brt_bac.1	taux_acces_brut_premiere_bac
Acc.att_bac.1	taux_acces_attendu_premiere_bac)
Acc.att_bac.2	taux_acces_attendu_seconde_bac)
Suc.brt_Tot	taux_brut_de_reussite_total_series)
Suc.att_Tot	taux_reussite_attendu_total_series)

```
##  code_etablissement           ville
##  0950667J: 14       GOUSSAINVILLE: 14
##  0950650R: 12       ARPAJON      : 13
##  0781951X: 10       SARCELLES    : 12
##  0910625K: 10       TAVERNY      : 12
##  0920141D: 10       ARGENTEUIL   : 11
##  0781859X:  9       MAGNANVILLE  : 10
##  (Other) :451       (Other)      :444
##                                          etablissement    commune
##  LYCEE JACQUES PREVERT                          : 16   Min.   :78005
##  LYCEE ROMAIN ROLLAND                           : 14   1st Qu.:91027
##  LYCEE RENE CASSIN                              : 13   Median :92012
##  LYCEE JEAN-JACQUES ROUSSEAU (GENERAL ET TECHNO.): 12   Mean   :89739
##  LYCEE JOLIOT-CURIE                             : 10   3rd Qu.:95018
##  LYCEE LEONARD DE VINCI                         : 10   Max.   :95637
##  (Other)                                        :441
##        Matiere        Barre            Prs_l            Prs_es
##  MATHS     : 59   Min.   :  21.0   Min.   :  6.00   Min.   : 10.00
##  ANGLAIS   : 52   1st Qu.: 111.0   1st Qu.: 18.00   1st Qu.: 53.00
##  HIST. GEO.: 47   Median : 196.0   Median : 30.00   Median : 69.00
##  ESPAGNOL  : 30   Mean   : 321.9   Mean   : 34.24   Mean   : 74.42
##  LET MODERN: 30   3rd Qu.: 292.0   3rd Qu.: 47.00   3rd Qu.: 99.00
##  S. V. T.  : 26   Max.   :2056.0   Max.   :133.00   Max.   :192.00
##  (Other)   :272
##      Prs_s         Suc.brt_l        Suc.brt_es      Suc.brt_s
##  Min.   : 13.0   Min.   : 36.00   Min.   : 51.0   Min.   :50.00
##  1st Qu.: 64.0   1st Qu.: 82.00   1st Qu.: 81.0   1st Qu.:81.00
```

```
##  Median :100.0   Median : 89.00   Median : 88.0   Median :88.00
##  Mean   :106.1   Mean   : 86.35   Mean   : 86.4   Mean   :86.23
##  3rd Qu.:140.0   3rd Qu.: 94.00   3rd Qu.: 94.0   3rd Qu.:93.00
##  Max.   :328.0   Max.   :100.00   Max.   :100.0   Max.   :99.00
##
##    Suc.att_l       Suc.att_es      Suc.att_s        Eff_2nd
##  Min.   :65.00   Min.   :61.00   Min.   :61.00   Min.   : 36.0
##  1st Qu.:84.00   1st Qu.:86.00   1st Qu.:86.00   1st Qu.:268.0
##  Median :89.00   Median :90.00   Median :89.00   Median :336.0
##  Mean   :86.91   Mean   :87.97   Mean   :87.39   Mean   :351.6
##  3rd Qu.:92.00   3rd Qu.:94.00   3rd Qu.:94.00   3rd Qu.:415.0
##  Max.   :98.00   Max.   :98.00   Max.   :98.00   Max.   :764.0
##
##     Eff_1er      Acc.brt_bac.2   Acc.att_bac.2   Acc.brt_bac.1
##  Min.   : 36.0   Min.   :49.00   Min.   :50.00   Min.   :65.00
##  1st Qu.:226.5   1st Qu.:64.00   1st Qu.:64.00   1st Qu.:82.00
##  Median :289.0   Median :71.00   Median :69.00   Median :85.00
##  Mean   :307.7   Mean   :69.61   Mean   :68.47   Mean   :84.53
##  3rd Qu.:364.0   3rd Qu.:76.00   3rd Qu.:73.00   3rd Qu.:89.25
##  Max.   :691.0   Max.   :87.00   Max.   :83.00   Max.   :97.00
##
##  Acc.att_bac.1    Suc.brt_Tot     Suc.att_Tot
##  Min.   :70.00   Min.   :64.00   Min.   :67.0
##  1st Qu.:81.00   1st Qu.:82.00   1st Qu.:84.0
##  Median :85.00   Median :86.00   Median :88.0
##  Mean   :84.19   Mean   :85.46   Mean   :86.8
##  3rd Qu.:89.00   3rd Qu.:91.00   3rd Qu.:92.0
##  Max.   :94.00   Max.   :98.00   Max.   :98.0
##
```

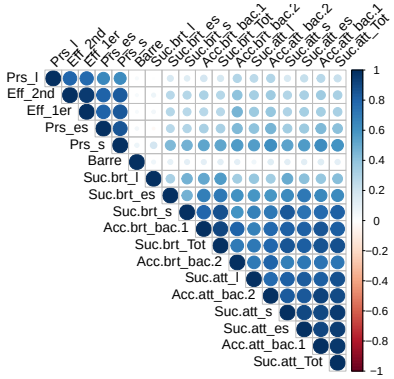
```
## Warning in as.data.frame.integer(length(colnames(data.mutations)),
## colnames(data.mutations)): 'row.names' is not a character vector of length
## 1 -- omitting it. Will be an error!
```

```
## [1] "Barre"
```

```
##   Barre Prs_l Prs_es Prs_s Suc.brt_l Suc.brt_es Suc.brt_s Suc.att_l
## 1 118.0    25     54    97        56         85        80        72
## 2  93.0    25     54    97        56         85        80        72
## 3  38.0    25     54    97        56         85        80        72
## 4 199.0    34     47    47        79         98        85        87
## 5  48.0    34     47    47        79         98        85        87
## 6 256.2    34     47    47        79         98        85        87
##   Suc.att_es Suc.att_s Eff_2nd Eff_1er Acc.brt_bac.2 Acc.att_bac.2
## 1         86        75     304     222            61            64
## 2         86        75     304     222            61            64
## 3         86        75     304     222            61            64
## 4         93        91     194     168            80            69
## 5         93        91     194     168            80            69
## 6         93        91     194     168            80            69
##   Acc.brt_bac.1 Acc.att_bac.1 Suc.brt_Tot Suc.att_Tot
## 1            84            81          81          79
## 2            84            81          81          79
```

```
## 3            84            81            81            79
## 4            92            87            88            89
## 5            92            87            88            89
## 6            92            87            88            89
```

- Corrélations 2 à 2 entre les variables



On a de fortes corrélations entre les groupes de variables. Effectifs (Eff_2nd/Eff_1e) et Effectifs présents (Prs_l/Prs_es/Prs_s) Succés brute (Suc.brt_l/Suc.brt_es/Suc.brt_s) et Succés Attentus (Suc.att_l/Suc.att_es/Suc.att_s) On remarque que le taux de réussite brute série L $Suc.brt_l$ est moins corrélés aux autres varuiables, et semble avoir une certaine indépendance.

La variable $Acc.brt_bac.2$ est très corrélé avec la variable $Acc.att_bac.2$ et de même pour $Acc.brt_bac.1$ et $Acc.att_bac.1$ On pourrait ne considérer que les variables Accès brute.

les covariables $Suc.brt_Tot$ et $Suc.att_Tot$ sont évidemment fortement corrélés avec les groupes Réussites et Réussites attendus.

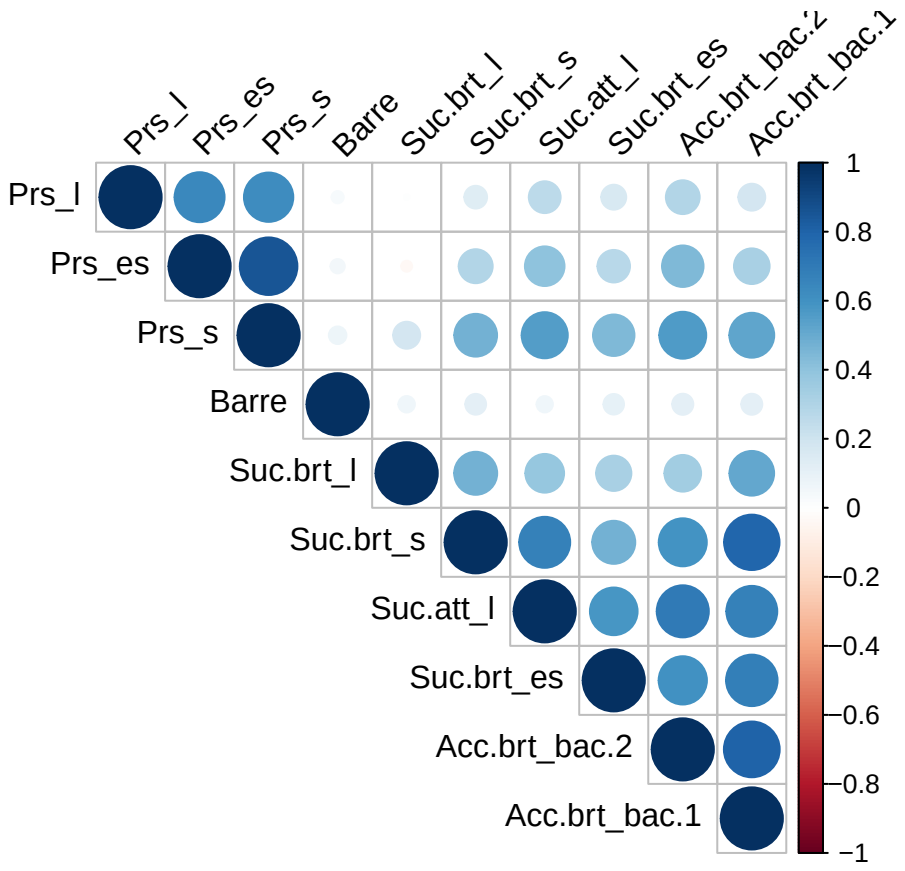
La variable a expliquer *Barre* n'est pas corrélés avec les caractéristiques de l'établissement.

On pourrait imaginer, de ne considérer que les variables covaraiables : Effectifs présents: Prs_l/Prs_es/Prs_s Succés brute: Suc.brt_l/Suc.brt_es/Suc.brt_s on garderait aussi Suc.att_l Accès brute: Acc.brt_bac.2/Acc.brt_bac.1

15 taux_acces_attendu_premiere_bac 0.3369
13 taux_acces_attendu_seconde_bac 0.1957
7 taux_reussite_attendu_serie_l 0.1224
17 taux_reussite_attendu_total_series 0.1200
8 taux_reussite_attendu_serie_es 0.1183
16 taux_brut_de_reussite_total_series 0.1161
9 taux_reussite_attendu_serie_s 0.1025
12 taux_acces_brut_seconde_bac 0.0898
5 taux_brut_de_reussite_serie_es 0.0821
6 taux_brut_de_reussite_serie_s 0.0776

```
##   Barre Prs_l Prs_es Prs_s Suc.brt_l Suc.brt_es Suc.brt_s Suc.att_l
## 1 118.0    25     54    97        56         85        80        72
## 2  93.0    25     54    97        56         85        80        72
## 3  38.0    25     54    97        56         85        80        72
## 4 199.0    34     47    47        79         98        85        87
## 5  48.0    34     47    47        79         98        85        87
## 6 256.2    34     47    47        79         98        85        87
##   Acc.brt_bac.2 Acc.brt_bac.1
## 1            61            84
## 2            61            84
```

```
## 3          61          84
## 4          80          92
## 5          80          92
## 6          80          92
```



Le résultat n'est pas très convaincant, il semble difficile de supprimer des variables.

2 Régression linéaire

On cherche à expliquer le nombre de points nécessaire à une mutation (colonne Barre) par les caractéristiques du lycée. On considère un modéle de régression linéaire gaussien, que l'on rappelle ici.

2.1 Rappels définitions et notations

2.1.1 Modèle linéaire Gaussien

Le modèle linéaire, tente d'expliquer les observations (input) (y_i) par des covariables $(x_1, ..., x_p)$ à partir du modèle suivant :

$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$ où $\epsilon_i \sim N(0, \sigma^2)$ et iid.

On note $y = (y_1, .., y_n)$ le vecteur des observations et $X = (x_{ik})_{1 \leq i \leq n, 1 \leq k \leq p}$ la matrice des covaraiables ou de design (predictor).

La réponse pour l'individus y_i est donnée par (variable Barre dans notre exemple).

En notation matricielles le modèle se réécrit de la manière suivante:

$$y \mid \alpha, \beta, \sigma^2 \sim N_n \alpha 1_n + X\beta, \sigma^2 I_n)$$

où N_n est la distribution de la loi normale en dimension n.

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

2.1.2 Contexte bayésien

On rappelle ici la formulation de la régression linaire dans le contexte bayésien.

On se place dans le cadre d'une expérience statistique paramétrique, où le vecteur des observations $Y = (y_1, ..., y_n)$ est iid et les $y_i \sim P_\theta$ une loi de paramètre θ.

Dans le contexte bayésien, on suppose que le paramètre inconnu θ est une v.a dont la loi de probabilité représente notre incertitude sur les valeurs possibles.

- Loi à priori $\pi(\theta)$

Cette loi du paramètre θ est la loi à priori, notée: $\pi(\theta)$. Elle représente "l'appriori" ou la croyance du statisticien avant le début de l'expérience. Sont choix est important, et on doit la choisir demanière à obtenir : une loi conjuguée pour faciliter les calculs, ou bien non informative (à priori de Jeffreys), fournit par un expert...

- Loi à postériori $\pi(\theta, y)$

On appelle la loi à postériori de θ sachant $y_1, y_2, ..., y_n$ la loi de distribution $\pi(\theta \mid Y) \propto \pi(\theta) L(\theta \mid Y)$

Cette définition découle de la formule de Bayes: $\pi(\theta \mid y) = \frac{\pi(\theta) f_{Y|\theta}(y|\theta)}{f_Y(y)}$

On retrouve l'équivalence des écritures avec $f_{Y|\theta}(y \mid \theta) = L(\theta \mid Y)$ Et $f_Y(y)$ ne dépend pas du paramètre θ, c'est une constante de normalisation qui est unique et que l'on peut retrouver une fois la loi à postériori déterminer analytiquement, qui doit s'intégrer à 1.

2.1.3 Régression linaire Bayésienne - Inférence bayésienne à l'aide de la loi a priori g de Zellner

On reprend les hypothèses et le contexte de définition du modèle linéaire gaussien, que l'on réinterprète avec l'approche Bayésienne. On considère la loi à priori $\pi(\theta)$ définit à partir des deux lois suivantes :

$$\beta \mid \sigma^2, X \sim N_{k+1}(\tilde{\beta}, \sigma^2 M^{-1})$$

$\sigma^2 \mid X \sim IG(a, b)$

En fixant la matrice M de la manière suivante, on obtient la g-prior ou loi informative de Zellner :

$$\beta \mid \sigma^2, X \sim N_{k+1}(\tilde{\beta}, g\sigma^2 ({}^tXX)^{-1})$$

$$\sigma^2 \sim \pi(\sigma^2 \mid X) \propto \sigma^{-2}$$

Il reste à choisir le paramètre g, souvent g=1 ou g=n en fonction du poids que l'on veut accorder à la prior. Si g=2 celà revient à donner à la prior le même poids que 50% de l'échantilon. Avec g=n on donne à la loi à priori le même poids que 1-observation.

Pour l'espérance à priori $\tilde{\beta}$ ou pourra la prendre $= 0$ si l'on n'a pas d'information à priori.

La loi à priori $\pi(\theta)$ se déduit simplement à partir des deux lois précédentes:

$$\pi(\theta) = \pi(\beta, \sigma^2 \mid X) = \pi(\beta \mid \sigma^2, X)\pi(\sigma^2 \mid X)$$

Cette loi à la propriété remarquable d'être une loi conjugué et sa loi à postériori associée a l'expression analytique suivnate:

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

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

donc :

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

2.2 Résultats et interprétation des coéfficients

2.2.1 Calcul explicite des coefficients

- Hypothèses Zellner G-prior

```
g=length(y)
betatilde=rep(0,dim(X)[2])
```

- calcul de $\hat{\beta}$ coefficient du modèle linéaire

$\hat{\beta} = (X^TX)^{-1}X^Ty$

```
beta0.lm=mean(y)
beta.lm=(solve(t(X)%*%X)%*%t(X))%*%(y)
betahat=rbind(Intercept=beta0.lm,beta.lm)
betahat
```

```
##                                       [,1]
## Intercept                       321.9155039
## effectif_presents_serie_l         1.1024250
## effectif_presents_serie_es        0.1440384
## effectif_presents_serie_s         0.4972839
## taux_brut_de_reussite_serie_l     2.5390471
## taux_brut_de_reussite_serie_es    4.4383584
## taux_brut_de_reussite_serie_s     8.4469888
## taux_reussite_attendu_serie_l   -15.0920251
## taux_reussite_attendu_serie_es    4.2333500
## taux_reussite_attendu_serie_s    -1.6603243
## effectif_de_seconde               0.1615037
## effectif_de_premiere             -0.6530472
## taux_acces_brut_seconde_bac      11.5971588
## taux_acces_attendu_seconde_bac   -4.4885363
## taux_acces_brut_premiere_bac    -21.8154757
## taux_acces_attendu_premiere_bac  23.3327884
## taux_brut_de_reussite_total_series  -3.4297701
## taux_reussite_attendu_total_series  -2.4005239
```

On pourrait aussi retrouver les coéfficients $\hat{\beta}$ à partir de la fonction lm.
On remarque cependant une différence assez significative entre les deux appproches, bien que l'odre de grandeur des coefficients est comparable.

```
reg.lm=lm(y~X)
summary(reg.lm)
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -429.72 -205.90 -122.25   -8.55 1645.96
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -4.725e+02  5.586e+02  -0.846   0.3980
## Xeffectif_presents_serie_l       7.781e-01  1.638e+00   0.475   0.6351
## Xeffectif_presents_serie_es      2.924e-01  1.234e+00   0.237   0.8128
## Xeffectif_presents_serie_s       9.694e-03  1.019e+00   0.010   0.9924
## Xtaux_brut_de_reussite_serie_l   3.122e+00  2.559e+00   1.220   0.2232
## Xtaux_brut_de_reussite_serie_es  4.811e+00  4.205e+00   1.144   0.2531
## Xtaux_brut_de_reussite_serie_s   9.385e+00  6.383e+00   1.470   0.1421
## Xtaux_reussite_attendu_serie_l  -1.428e+01  6.879e+00  -2.077   0.0383
## Xtaux_reussite_attendu_serie_es  3.814e+00  8.261e+00   0.462   0.6445
## Xtaux_reussite_attendu_serie_s  -4.299e+00  9.586e+00  -0.448   0.6540
## Xeffectif_de_seconde             4.306e-02  6.229e-01   0.069   0.9449
## Xeffectif_de_premiere           -3.521e-01  7.182e-01  -0.490   0.6242
## Xtaux_acces_brut_seconde_bac     1.074e+01  5.655e+00   1.900   0.0580
## Xtaux_acces_attendu_seconde_bac -7.077e+00  9.038e+00  -0.783   0.4340
## Xtaux_acces_brut_premiere_bac   -2.039e+01  1.071e+01  -1.904   0.0575
## Xtaux_acces_attendu_premiere_bac 3.444e+01  1.916e+01   1.797   0.0729
## Xtaux_brut_de_reussite_total_series -5.392e+00  1.288e+01  -0.419   0.6757
## Xtaux_reussite_attendu_total_series -4.072e+00  2.202e+01  -0.185   0.8534
##
## (Intercept)
## Xeffectif_presents_serie_l
## Xeffectif_presents_serie_es
## Xeffectif_presents_serie_s
## Xtaux_brut_de_reussite_serie_l
## Xtaux_brut_de_reussite_serie_es
## Xtaux_brut_de_reussite_serie_s
## Xtaux_reussite_attendu_serie_l      *
## Xtaux_reussite_attendu_serie_es
## Xtaux_reussite_attendu_serie_s
## Xeffectif_de_seconde
## Xeffectif_de_premiere
## Xtaux_acces_brut_seconde_bac        .
## Xtaux_acces_attendu_seconde_bac
## Xtaux_acces_brut_premiere_bac       .
## Xtaux_acces_attendu_premiere_bac    .
## Xtaux_brut_de_reussite_total_series
## Xtaux_reussite_attendu_total_series
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 422.4 on 498 degrees of freedom
## Multiple R-squared:  0.04068,    Adjusted R-squared:  0.007931
## F-statistic: 1.242 on 17 and 498 DF,  p-value: 0.2267
```

- Calcul de $E^{\pi}(\beta \mid y, X) = \frac{g}{g+1}(\hat{\beta} + \frac{\tilde{\beta}}{g})$ G-prior informative de Zellner

```
mbetabayes=g/(g+1)*(beta.lm+betatilde/g)
postmean=rbind(Intercept=beta0.lm,mbetabayes)
postmean
```

```
##                                       [,1]
## Intercept                        321.9155039
## effectif_presents_serie_l          1.1002926
## effectif_presents_serie_es         0.1437598
## effectif_presents_serie_s          0.4963221
## taux_brut_de_reussite_serie_l      2.5341360
## taux_brut_de_reussite_serie_es     4.4297736
## taux_brut_de_reussite_serie_s      8.4306503
## taux_reussite_attendu_serie_l    -15.0628335
## taux_reussite_attendu_serie_es     4.2251617
## taux_reussite_attendu_serie_s     -1.6571128
## effectif_de_seconde                0.1611913
## effectif_de_premiere              -0.6517840
## taux_acces_brut_seconde_bac       11.5747272
## taux_acces_attendu_seconde_bac    -4.4798544
## taux_acces_brut_premiere_bac     -21.7732794
## taux_acces_attendu_premiere_bac   23.2876573
## taux_brut_de_reussite_total_series -3.4231361
## taux_reussite_attendu_total_series -2.3958807
```

2.2.2 Calcul de $\hat{\beta}$

Pour estimer les β à postériori, on va utiliser la fonction (modifiée) BayesReg du package Bayess issue du livre de Marin et Robert : Bayesian Essentials with R. On comparera le résultat obtenu avec le résultat renvoyé par la fonction du livre de P. Hoff: A First Course in Bayesian Statistical Methods.

- Bayes Regression 1 : $FonctionBayesReg$

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept 321.9155     23.2203
## x1          1.1003      1.9872 -1.2631
## x2          0.1438      1.5245 -1.3575
## x3          0.4963      1.0485 -1.2426
## x4          2.5341      3.0754 -1.2552
## x5          4.4298      5.2174 -1.2177
## x6          8.4307      7.8419 -1.1432
## x7        -15.0628      8.4990 -0.6244
## x8          4.2252     10.2884  -1.315
```

```
## x9         -1.6571     11.3088 -1.3622
## x10         0.1612      0.7573  -1.337
## x11        -0.6518      0.7784  -1.108
## x12        11.5747      6.9422 -0.6919
## x13        -4.4799     10.6099 -1.3485
## x14       -21.7733     13.1896 -0.7114
## x15        23.2877     17.4102 -1.2203
## x16        -3.4231     15.8117 -1.3545
## x17        -2.3959     27.3661 -1.3567
##
##
## Posterior Mean of Sigma2: 278217.8422
## Posterior StError of Sigma2: 393843.496


## $postmeancoeff
##  [1] 321.9155039   1.1002926   0.1437598   0.4963221   2.5341360
##  [6]   4.4297736   8.4306503 -15.0628335   4.2251617  -1.6571128
## [11]   0.1611913  -0.6517840  11.5747272  -4.4798544 -21.7732794
## [16]  23.2876573  -3.4231361  -2.3958807
##
## $postsqrtcoeff
##                                           effectif_presents_serie_l
##                           23.2202899                           1.9871983
##         effectif_presents_serie_es          effectif_presents_serie_s
##                           1.5244679                           1.0485202
##      taux_brut_de_reussite_serie_l     taux_brut_de_reussite_serie_es
##                           3.0754024                           5.2173996
##      taux_brut_de_reussite_serie_s      taux_reussite_attendu_serie_l
##                           7.8418759                           8.4989746
##     taux_reussite_attendu_serie_es      taux_reussite_attendu_serie_s
##                          10.2883711                          11.3087683
##                effectif_de_seconde               effectif_de_premiere
##                           0.7572988                           0.7784006
##        taux_acces_brut_seconde_bac     taux_acces_attendu_seconde_bac
##                           6.9421508                          10.6098775
##       taux_acces_brut_premiere_bac    taux_acces_attendu_premiere_bac
##                          13.1896055                          17.4102349
## taux_brut_de_reussite_total_series taux_reussite_attendu_total_series
##                          15.8116686                          27.3661012
##
## $log10bf
##  [1] -1.2631090 -1.3575334 -1.2425776 -1.2551568 -1.2177032 -1.1431858
##  [7] -0.6244135 -1.3150201 -1.3622209 -1.3369829 -1.1080139 -0.6918887
## [13] -1.3484836 -0.7114405 -1.2203473 -1.3545095 -1.3566625
##
## $postmeansigma2
## [1] 278217.8
##
## $postvarsigma2
## [1] 155112699325
```

Les Log10 bayes factors sont tous négatifs, aucunes des variables ne se dégage.

2.2.3 Autres méthodes et packages

- Bayes Regression 2

```
##                          (Intercept)          Xeffectif_presents_serie_l
##                        -4.725281e+02                        7.780548e-01
##         Xeffectif_presents_serie_es          Xeffectif_presents_serie_s
##                         2.924407e-01                        9.694315e-03
##      Xtaux_brut_de_reussite_serie_l     Xtaux_brut_de_reussite_serie_es
##                         3.121824e+00                        4.811087e+00
##      Xtaux_brut_de_reussite_serie_s      Xtaux_reussite_attendu_serie_l
##                         9.384937e+00                       -1.428483e+01
##     Xtaux_reussite_attendu_serie_es      Xtaux_reussite_attendu_serie_s
##                         3.814350e+00                       -4.299223e+00
##                Xeffectif_de_seconde               Xeffectif_de_premiere
##                         4.306141e-02                       -3.520708e-01
##        Xtaux_acces_brut_seconde_bac     Xtaux_acces_attendu_seconde_bac
##                         1.074444e+01                       -7.077486e+00
##       Xtaux_acces_brut_premiere_bac    Xtaux_acces_attendu_premiere_bac
##                        -2.038674e+01                        3.444343e+01
## Xtaux_brut_de_reussite_total_series Xtaux_reussite_attendu_total_series
##                        -5.392357e+00                       -4.071861e+00
```

```
##                                     betaMean
## effectif_presents_serie_l           1.1275322
## effectif_presents_serie_es          0.1471268
## effectif_presents_serie_s           0.5060717
## taux_brut_de_reussite_serie_l       2.5601219
## taux_brut_de_reussite_serie_es      4.4900605
## taux_brut_de_reussite_serie_s       8.4237253
## taux_reussite_attendu_serie_l     -14.9749586
## taux_reussite_attendu_serie_es      4.1552169
## taux_reussite_attendu_serie_s      -1.5903452
## effectif_de_seconde                0.1581856
## effectif_de_premiere              -0.6549692
## taux_acces_brut_seconde_bac       11.6072687
## taux_acces_attendu_seconde_bac    -4.6317879
## taux_acces_brut_premiere_bac     -21.8544023
## taux_acces_attendu_premiere_bac   23.4973083
## taux_brut_de_reussite_total_series -3.5783871
## taux_reussite_attendu_total_series -2.4378076
```

On trouve des résultats relativement proches avec la fonction zlm

```
## Coefficients
##                                       Exp.Val.    St.Dev.
## (Intercept)                       -4.709915e+02         NA
## effectif_presents_serie_l          7.765499e-01  1.6126850
## effectif_presents_serie_es         2.918751e-01  1.2151292
## effectif_presents_serie_s          9.675564e-03  1.0031695
## taux_brut_de_reussite_serie_l      3.115785e+00  2.5194645
## taux_brut_de_reussite_serie_es     4.801782e+00  4.1392797
## taux_brut_de_reussite_serie_s      9.366785e+00  6.2827255
```

```
## taux_reussite_attendu_serie_l      -1.425720e+01  6.7711046
## taux_reussite_attendu_serie_es      3.806972e+00  8.1320817
## taux_reussite_attendu_serie_s      -4.290907e+00  9.4362277
## effectif_de_seconde                 4.297812e-02  0.6131953
## effectif_de_premiere               -3.513898e-01  0.7070083
## taux_acces_brut_seconde_bac         1.072366e+01  5.5664718
## taux_acces_attendu_seconde_bac     -7.063796e+00  8.8967805
## taux_acces_brut_premiere_bac       -2.034731e+01 10.5384879
## taux_acces_attendu_premiere_bac     3.437680e+01 18.8645303
## taux_brut_de_reussite_total_series -5.381927e+00 12.6826322
## taux_reussite_attendu_total_series -4.063985e+00 21.6790804
##
##  Log Marginal Likelihood:
## -4765.986
##  g-Prior: UIP
## Shrinkage Factor: 0.998
```

2.3 Choix des covariables et comparaison au résultat obtenu par une analyse fréquentiste.

Choisir les covariables significatives. Comparer au résultat obtenu par une analyse fréquentiste. Afin de réduire le coût computationnel, il peut être intéressant d'éffectuer une présélection des covariables considérées.

2.3.1 Choix des covariables

Bayes Factors et comparaison de modèles Pour comparer les modèles on peut utiliser les facteurs de Bayes

- Test d'hypothèse $H_0 : \beta_i = 0i$

On test l'hypothèse H_0, $\forall i = 1, ..., 17$ et on calcul le Bayes Factor à partir de la formule du cours (TP4)

- A partir de la fonction $CalcBayesFactor$ pour $g = n$

```
##                            colnames(X) bfactor
## 7       taux_reussite_attendu_serie_l  2.4506
## 12        taux_acces_brut_seconde_bac  2.3254
## 14       taux_acces_brut_premiere_bac  2.3064
## 15    taux_acces_attendu_premiere_bac  1.9811
## 6       taux_brut_de_reussite_serie_s  1.7605
## 5      taux_brut_de_reussite_serie_es  1.6087
## 11               effectif_de_premiere  1.6018
## 4       taux_brut_de_reussite_serie_l  1.5941
## 1           effectif_presents_serie_l  1.4640
## 3           effectif_presents_serie_s  1.4351
## 13     taux_acces_attendu_seconde_bac  1.4191
## 8      taux_reussite_attendu_serie_es  1.4158
## 16 taux_brut_de_reussite_total_series  1.3731
## 10                effectif_de_seconde  1.3726
## 9       taux_reussite_attendu_serie_s  1.3643
## 2          effectif_presents_serie_es  1.3599
## 17 taux_reussite_attendu_total_series  1.3594
```

- A partir de la fonction $BayesReg2$ pour $g = n$

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept 321.9155     23.2203
## x1          1.1003      1.9872 -1.2631
## x2          0.1438      1.5245 -1.3575
## x3          0.4963      1.0485 -1.2426
## x4          2.5341      3.0754 -1.2552
## x5          4.4298      5.2174 -1.2177
## x6          8.4307      7.8419 -1.1432
## x7        -15.0628      8.4990 -0.6244
## x8          4.2252     10.2884  -1.315
## x9         -1.6571     11.3088 -1.3622
## x10         0.1612      0.7573  -1.337
## x11        -0.6518      0.7784  -1.108
## x12        11.5747      6.9422 -0.6919
## x13        -4.4799     10.6099 -1.3485
## x14       -21.7733     13.1896 -0.7114
## x15        23.2877     17.4102 -1.2203
## x16        -3.4231     15.8117 -1.3545
## x17        -2.3959     27.3661 -1.3567
##
##
## Posterior Mean of Sigma2: 278217.8422
## Posterior StError of Sigma2: 393843.496

## $postmeancoeff
##  [1] 321.9155039   1.1002926   0.1437598   0.4963221   2.5341360
##  [6]   4.4297736   8.4306503 -15.0628335   4.2251617  -1.6571128
## [11]   0.1611913  -0.6517840  11.5747272  -4.4798544 -21.7732794
## [16]  23.2876573  -3.4231361  -2.3958807
##
## $postsqrtcoeff
##                                             effectif_presents_serie_l
##                          23.2202899                          1.9871983
##         effectif_presents_serie_es          effectif_presents_serie_s
##                           1.5244679                          1.0485202
##      taux_brut_de_reussite_serie_l     taux_brut_de_reussite_serie_es
##                           3.0754024                          5.2173996
##      taux_brut_de_reussite_serie_s      taux_reussite_attendu_serie_l
##                           7.8418759                          8.4989746
##     taux_reussite_attendu_serie_es      taux_reussite_attendu_serie_s
##                          10.2883711                         11.3087683
##                effectif_de_seconde               effectif_de_premiere
##                           0.7572988                          0.7784006
##        taux_acces_brut_seconde_bac     taux_acces_attendu_seconde_bac
##                           6.9421508                         10.6098775
##       taux_acces_brut_premiere_bac    taux_acces_attendu_premiere_bac
##                          13.1896055                         17.4102349
## taux_brut_de_reussite_total_series taux_reussite_attendu_total_series
##                          15.8116686                         27.3661012
##
## $log10bf
```

```
##  [1] -1.2631090 -1.3575334 -1.2425776 -1.2551568 -1.2177032 -1.1431858
##  [7] -0.6244135 -1.3150201 -1.3622209 -1.3369829 -1.1080139 -0.6918887
## [13] -1.3484836 -0.7114405 -1.2203473 -1.3545095 -1.3566625
##
## $postmeansigma2
## [1] 278217.8
##
## $postvarsigma2
## [1] 155112699325
```

Même fonction mais avec des données centrées et réduites

```
##
##            PostMean PostStError Log10bf EvidAgaH0
## Intercept  321.9155     18.3206
## x1          16.3295     33.9119 -1.3062
## x2          10.0287     41.7513 -1.3442
## x3           0.5605     58.1159 -1.3567
## x4          36.0144     29.1217 -1.0238
## x5          47.3120     40.7843 -1.0638
## x6          85.1944     57.1437 -0.8733
## x7        -105.7570     50.2267 -0.3944
## x8          32.2582     68.9069  -1.309
## x9         -40.2692     88.5569 -1.3117
## x10          5.8227     83.0766 -1.3557
## x11        -44.4039     89.3421 -1.3029
## x12         97.3432     50.5293 -0.5506
## x13        -50.9800     64.2088 -1.2194
## x14       -139.8800     72.4481  -0.547
## x15        205.6277    112.8398 -0.6352
## x16        -39.7571     93.6884 -1.3175
## x17        -31.3305    167.1307 -1.3491
##
##
## Posterior Mean of Sigma2: 173193.2688
## Posterior StError of Sigma2: 245171.3446
```

```
## $postmeancoeff
##  [1]  321.915504   16.329487   10.028689    0.560527   36.014371
##  [6]   47.311975   85.194353 -105.757008   32.258228  -40.269226
## [11]    5.822742  -44.403881   97.343237  -50.979986 -139.880026
## [16]  205.627719  -39.757068  -31.330506
##
## $postsqrtcoeff
##                                          effectif_presents_serie_l
##                            18.32064                          33.91195
##         effectif_presents_serie_es          effectif_presents_serie_s
##                            41.75126                          58.11585
##      taux_brut_de_reussite_serie_l     taux_brut_de_reussite_serie_es
##                            29.12169                          40.78434
##      taux_brut_de_reussite_serie_s      taux_reussite_attendu_serie_l
##                            57.14370                          50.22669
##     taux_reussite_attendu_serie_es      taux_reussite_attendu_serie_s
```

```
##                            68.90688                            88.55694
##                effectif_de_seconde               effectif_de_premiere
##                            83.07664                            89.34214
##        taux_acces_brut_seconde_bac     taux_acces_attendu_seconde_bac
##                            50.52925                            64.20878
##       taux_acces_brut_premiere_bac    taux_acces_attendu_premiere_bac
##                            72.44811                           112.83976
## taux_brut_de_reussite_total_series taux_reussite_attendu_total_series
##                            93.68842                           167.13066
##
## $log10bf
##  [1] -1.3062110 -1.3441685 -1.3567250 -1.0238434 -1.0637704 -0.8732526
##  [7] -0.3944166 -1.3089805 -1.3116783 -1.3556744 -1.3029098 -0.5506177
## [13] -1.2194081 -0.5470375 -0.6351706 -1.3174966 -1.3490849
##
## $postmeansigma2
## [1] 173193.3
##
## $postvarsigma2
## [1] 60108988208
```

- A partir de la fonction $CalcBayesFactor$ pour $g = 1$

```
##                            colnames(X) bfactor
## 7       taux_reussite_attendu_serie_l  0.5671
## 12        taux_acces_brut_seconde_bac  0.5193
## 14       taux_acces_brut_premiere_bac  0.5121
## 15    taux_acces_attendu_premiere_bac  0.3880
## 6       taux_brut_de_reussite_serie_s  0.3040
## 5      taux_brut_de_reussite_serie_es  0.2463
## 11               effectif_de_premiere  0.2436
## 4       taux_brut_de_reussite_serie_l  0.2407
## 1           effectif_presents_serie_l  0.1912
## 3           effectif_presents_serie_s  0.1803
## 13     taux_acces_attendu_seconde_bac  0.1742
## 8      taux_reussite_attendu_serie_es  0.1729
## 16 taux_brut_de_reussite_total_series  0.1567
## 10                effectif_de_seconde  0.1565
## 9       taux_reussite_attendu_serie_s  0.1534
## 2          effectif_presents_serie_es  0.1517
## 17 taux_reussite_attendu_total_series  0.1515
```

- A partir de la fonction $BayesReg2$ pour $g = 1$

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept 321.9155     25.4292
## x1          0.5512      1.5403 -0.1002
## x2          0.0720      1.1816  -0.152
## x3          0.2486      0.8127 -0.0756
## x4          1.2695      2.3838 -0.1274
## x5          2.2192      4.0441 -0.0999
## x6          4.2235      6.0784 -0.0771
```

```
## x7         -7.5460      6.5877  0.1767       (*)
## x8          2.1167      7.9747  -0.131
## x9         -0.8302      8.7657  -0.157
## x10         0.0808      0.5870 -0.1381
## x11        -0.3265      0.6034 -0.0066
## x12         5.7986      5.3810  0.1529       (*)
## x13        -2.2443      8.2239 -0.1598
## x14       -10.9077     10.2236  0.1415       (*)
## x15        11.6664     13.4951  -0.199
## x16        -1.7149     12.2560 -0.1529
## x17        -1.2003     21.2121 -0.1511
##
##
## Posterior Mean of Sigma2: 333667.6519
## Posterior StError of Sigma2: 472337.9115

## $postmeancoeff
##  [1] 321.91550388   0.55121249   0.07201918   0.24864197   1.26952357
##  [6]   2.21917919   4.22349439  -7.54601254   2.11667501  -0.83016213
## [11]   0.08075185  -0.32652358   5.79857942  -2.24426814 -10.90773784
## [16]  11.66639422  -1.71488505  -1.20026194
##
## $postsqrtcoeff
##                                          effectif_presents_serie_l
##                           25.4291710                           1.5403208
##         effectif_presents_serie_es          effectif_presents_serie_s
##                           1.1816484                           0.8127309
##      taux_brut_de_reussite_serie_l     taux_brut_de_reussite_serie_es
##                           2.3838117                           4.0441205
##      taux_brut_de_reussite_serie_s      taux_reussite_attendu_serie_l
##                           6.0784094                           6.5877410
##     taux_reussite_attendu_serie_es      taux_reussite_attendu_serie_s
##                           7.9747413                           8.7656735
##                effectif_de_seconde               effectif_de_premiere
##                           0.5869989                           0.6033553
##        taux_acces_brut_seconde_bac     taux_acces_attendu_seconde_bac
##                           5.3810129                           8.2239480
##       taux_acces_brut_premiere_bac    taux_acces_attendu_premiere_bac
##                          10.2235516                          13.4950537
## taux_brut_de_reussite_total_series taux_reussite_attendu_total_series
##                          12.2559700                          21.2120634
##
## $log10bf
##  [1] -0.100226477 -0.151978763 -0.075620854 -0.127361053 -0.099906540
##  [6] -0.077138798  0.176720441 -0.131017110 -0.157028390 -0.138145531
## [11] -0.006578394  0.152933126 -0.159797897  0.141537344 -0.199012597
## [16] -0.152902215 -0.151141268
##
## $postmeansigma2
## [1] 333667.7
##
## $postvarsigma2
## [1] 223103102624
```

- Conclusion les 7ème (Suc.att_l), 12ème (Acc.brt_bac.2) et 14ème (Acc.brt_bac.1) variables sont les plus significatives.
- Choix de modèle : calcul exact

A partir de la méthoide vue en TP, on va considérer les 4 variables les plus significatives

```
## [1] 0.000 0.230 0.433 0.021 0.193 0.018 0.098 0.007
```

c'est le modèle (T,F, F) qui est de loin le plus probable a posteriori le modèle avec la covariable: taux_reussite_attendu_serie_l (Suc.att_l)

A partir de la fonction (modifée) - ModChoBayesReg du package Bayess

```
##
## Number of variables less than 18
## Model posterior probabilities are calculated exactly
##
##    Top10Models  PostProb
## 1              -2051.412
## 2            1 -2091.963
## 3            3 -2093.878
## 4          1 3 -2095.955
## 5            2 -2096.343
## 6          2 3 -2097.355
## 7           11 -2097.730
## 8          1 2 -2097.819
## 9           10 -2098.380
## 10        3 11 -2098.557
```

```
## $top10models
##  [1] ""     "1"    "3"    "1 3"  "2"    "2 3"  "11"   "1 2"  "10"   "3 11"
##
## $postprobtop10
##  [1] -2051.412 -2091.963 -2093.878 -2095.955 -2096.342 -2097.355 -2097.730
##  [8] -2097.819 -2098.380 -2098.557
```

- Choix de modèle par échantillonnage de Gibbs

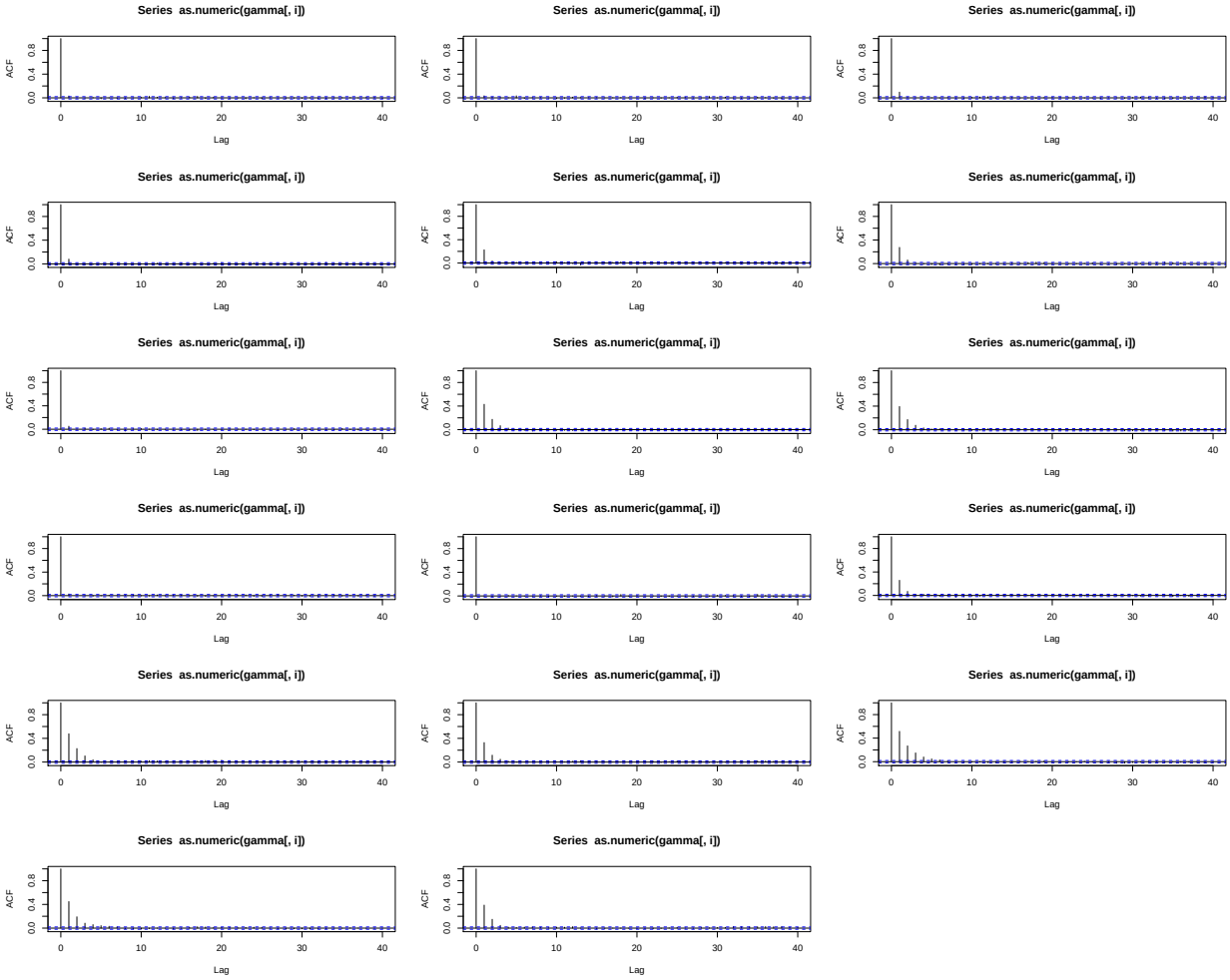
Avec la fonction utilsée en TP

[1] 0.1927 0.2517 0.5165 0.1815 0.2063 0.2628 0.4438 0.4137

```
##                                     X gamma.mean
## 15    taux_acces_attendu_premiere_bac     0.3245
## 13     taux_acces_attendu_seconde_bac     0.1848
## 17 taux_reussite_attendu_total_series     0.1324
## 16 taux_brut_de_reussite_total_series     0.1243
## 7       taux_reussite_attendu_serie_l     0.1234
## 8      taux_reussite_attendu_serie_es     0.1202
## 9       taux_reussite_attendu_serie_s     0.1053
## 5      taux_brut_de_reussite_serie_es     0.0872
## 12        taux_acces_brut_seconde_bac     0.0840
```

```
## 14       taux_acces_brut_premiere_bac     0.0807
## 6       taux_brut_de_reussite_serie_s     0.0764
## 3          effectif_presents_serie_s     0.0528
## 4       taux_brut_de_reussite_serie_l     0.0506
## 2          effectif_presents_serie_es     0.0476
## 11               effectif_de_premiere     0.0473
## 10                effectif_de_seconde     0.0464
## 1           effectif_presents_serie_l     0.0461
```

On regarde la convergence de la méthode :



- Vérifions la convergence + le mélange à l'aide de la trace (on utilise une moyenne glissante puisque les valeurs sont binaires).
- Prédiction

2.3.2 Comparaison au résultat obtenu par une analyse fréquentiste

- Analyse fréquentiste

On considère un modéle de régression linéaire gaussiennne i.e

$$y \mid \alpha, \beta, \sigma^2 \sim N_n(\alpha 1_n + X\beta, \sigma^2 I_n)$$

où N_n est la distribution de la loi normale en dimension n.

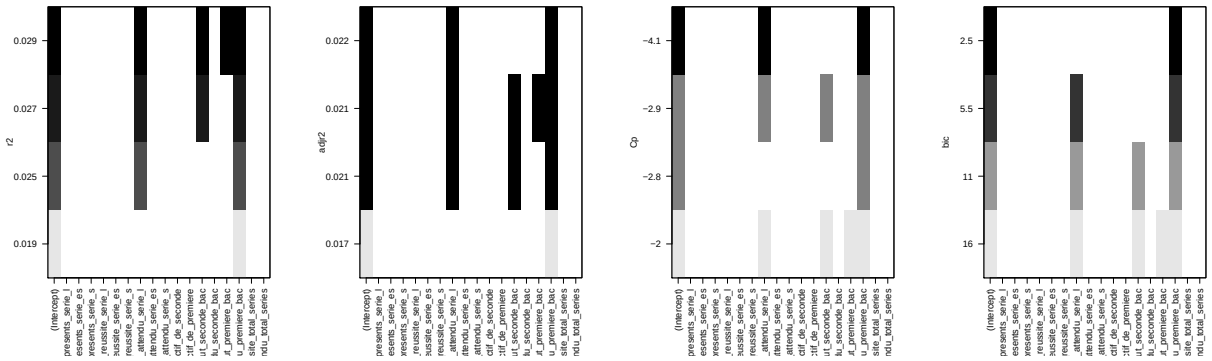
Ainsi les y_i suivent des lois normales indépendantes avec :

$$E(y_i \mid \alpha, \beta, \sigma^2) = \alpha + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$V(y_i \mid \alpha, \beta, \sigma^2) = \sigma^2$$

```
##
## Call:
## lm(formula = Barre ~ ., data = data.mutations)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -429.72 -205.90 -122.25   -8.55 1645.96
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -4.725e+02  5.586e+02  -0.846   0.3980
## effectif_presents_serie_l        7.781e-01  1.638e+00   0.475   0.6351
## effectif_presents_serie_es       2.924e-01  1.234e+00   0.237   0.8128
## effectif_presents_serie_s        9.694e-03  1.019e+00   0.010   0.9924
## taux_brut_de_reussite_serie_l    3.122e+00  2.559e+00   1.220   0.2232
## taux_brut_de_reussite_serie_es   4.811e+00  4.205e+00   1.144   0.2531
## taux_brut_de_reussite_serie_s    9.385e+00  6.383e+00   1.470   0.1421
## taux_reussite_attendu_serie_l   -1.428e+01  6.879e+00  -2.077   0.0383
## taux_reussite_attendu_serie_es   3.814e+00  8.261e+00   0.462   0.6445
## taux_reussite_attendu_serie_s   -4.299e+00  9.586e+00  -0.448   0.6540
## effectif_de_seconde              4.306e-02  6.229e-01   0.069   0.9449
## effectif_de_premiere            -3.521e-01  7.182e-01  -0.490   0.6242
## taux_acces_brut_seconde_bac      1.074e+01  5.655e+00   1.900   0.0580
## taux_acces_attendu_seconde_bac  -7.077e+00  9.038e+00  -0.783   0.4340
## taux_acces_brut_premiere_bac    -2.039e+01  1.071e+01  -1.904   0.0575
## taux_acces_attendu_premiere_bac  3.444e+01  1.916e+01   1.797   0.0729
## taux_brut_de_reussite_total_series -5.392e+00  1.288e+01  -0.419   0.6757
## taux_reussite_attendu_total_series -4.072e+00  2.202e+01  -0.185   0.8534
##
## (Intercept)
## effectif_presents_serie_l
## effectif_presents_serie_es
## effectif_presents_serie_s
## taux_brut_de_reussite_serie_l
## 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
## effectif_de_seconde
## effectif_de_premiere
## taux_acces_brut_seconde_bac        .
## taux_acces_attendu_seconde_bac
## taux_acces_brut_premiere_bac       .
## taux_acces_attendu_premiere_bac    .
```

```
## taux_brut_de_reussite_total_series
## taux_reussite_attendu_total_series
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 422.4 on 498 degrees of freedom
## Multiple R-squared:  0.04068,    Adjusted R-squared:  0.007931
## F-statistic: 1.242 on 17 and 498 DF,  p-value: 0.2267
```




```
summary(step_mod)
```

```
##
## Call:
## lm(formula = Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac,
##     data = data.mutations)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -387.32 -196.56 -130.83  -14.95 1696.20
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -494.324    260.593  -1.897  0.05840 .
## taux_reussite_attendu_serie_l     -7.882      4.360  -1.808  0.07124 .
## taux_acces_attendu_premiere_bac   17.833      5.407   3.298  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 419.5 on 513 degrees of freedom
## Multiple R-squared:  0.02539,    Adjusted R-squared:  0.02159
## F-statistic: 6.681 on 2 and 513 DF,  p-value: 0.001366
```

Les 3 covariables qui se dégagent :

- taux_reussite_attendu_serie_l
- taux_acces_attendu_premiere_bac
- taux_acces_brut_seconde_bac

nettement - “taux_acces_brut_brute_bac”

- On considère les 2 modèles suivants :

taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac + taux_acces_brut_seconde_bac + taux_acces_brut_premiere_bac

```
#reg.mod2 = lm(Barre ~ Suc.att_l + Acc.att_bac.1  + Acc.brt_bac.1 +  Acc.brt_bac.2, data=dataMutations_
reg.mod2 = lm(Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac + taux_acces_brut_
summary(reg.mod2)
```

```
##
## Call:
## lm(formula = Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac +
##     taux_acces_brut_seconde_bac + taux_acces_brut_premiere_bac,
##     data = dataMutations_d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -410.82 -203.23 -128.06   -4.57 1670.03
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -356.286    279.244  -1.276  0.20257
## taux_reussite_attendu_serie_l   -10.207      4.671  -2.185  0.02934 *
## taux_acces_attendu_premiere_bac  20.776      7.694   2.700  0.00716 **
## taux_acces_brut_seconde_bac       4.986      3.708   1.345  0.17930
## taux_acces_brut_premiere_bac     -6.280      6.129  -1.025  0.30600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 419.5 on 511 degrees of freedom
## Multiple R-squared:  0.02905,    Adjusted R-squared:  0.02145
## F-statistic: 3.822 on 4 and 511 DF,  p-value: 0.004514
```

taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac + taux_acces_brut_seconde_bac

```
reg.mod1 = lm(Barre ~ taux_reussite_attendu_serie_l
              + taux_acces_attendu_premiere_bac
              +  taux_acces_brut_seconde_bac, data=dataMutations_d)
summary(reg.mod1)
```

```
##
## Call:
## lm(formula = Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac +
##     taux_acces_brut_seconde_bac, data = dataMutations_d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -379.54 -206.00 -132.06   -2.57 1674.19
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -422.564    271.663  -1.555  0.12045
## taux_reussite_attendu_serie_l    -8.952      4.508  -1.986  0.04759 *
## taux_acces_attendu_premiere_bac  15.685      5.875   2.670  0.00783 **
## taux_acces_brut_seconde_bac       2.903      3.101   0.936  0.34963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 419.5 on 512 degrees of freedom
## Multiple R-squared:  0.02705,    Adjusted R-squared:  0.02135
## F-statistic: 4.745 on 3 and 512 DF,  p-value: 0.002831
```

- On réalise maintenant des tests entre modèles emboîtés :

```
anova(reg.mod2,reg.mod1)
```

```
## Analysis of Variance Table
##
## Model 1: Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac +
##     taux_acces_brut_seconde_bac + taux_acces_brut_premiere_bac
## Model 2: Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac +
##     taux_acces_brut_seconde_bac
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1    511 89918104
## 2    512 90102863 -1   -184758 1.05  0.306
```

Au vu des p-valeurs des tests de Fisher, on peut envisager de se passer de la variable : taux_acces_brut_premiere_bac
On conserve le plus petit modèle : reg.mod1

On réalise à nouveaux un test anova, maintenant entre reg.mod1 et step_mod.

```
anova(step_mod,reg.mod1)
```

```
## Analysis of Variance Table
##
## Model 1: Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac
## Model 2: Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac +
##     taux_acces_brut_seconde_bac
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    513 90257096
## 2    512 90102863  1    154234 0.8764 0.3496
```

Au vu des p-valeurs des tests de Fisher, on peut envisager de se passer de la variable : taux_acces_brut_seconde_bac
On conserve le plus petit modèle : step_mod

Un estimateur sans biais de σ^2 est donnée par la formule suivante:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}(y - \hat{\alpha}\mathbb{1}_n - X\hat{\beta})^T(y - \hat{\alpha}\mathbb{1}_n - X\hat{\beta}) = \frac{s^2}{n-p-1}$$

on obtient σ^2

```
##           [,1]
## [1,] 181239.1
```

et les estimations par les moindres carrés des coéfficients de régression :

```
##
## Call:
## lm(formula = Barre ~ taux_reussite_attendu_serie_l + taux_acces_attendu_premiere_bac,
##     data = data.mutations)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -387.32 -196.56 -130.83  -14.95 1696.20
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -494.324    260.593  -1.897  0.05840 .
## taux_reussite_attendu_serie_l     -7.882      4.360  -1.808  0.07124 .
## taux_acces_attendu_premiere_bac   17.833      5.407   3.298  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 419.5 on 513 degrees of freedom
## Multiple R-squared:  0.02539,    Adjusted R-squared:  0.02159
## F-statistic: 6.681 on 2 and 513 DF,  p-value: 0.001366
```

effectif_presents_serie_l
effectif_presents_serie_es taux_reussite_attendu_serie_l
taux_brut_de_reussite_total_series

2.3.3 Préselection des covariables

2.3.4 Conclusion

2.4 Mutations en mathématiques et anglais

2.4.1 Calcul explicite des coefficients

- G-prior informative de Zellner Hypothèses Zellner G-prior calcul de $\hat{\beta}$ coefficient du modèle linéaire: $\hat{\beta} = (X^T X)^{-1} X^T y$ Calcul de $E^{\pi}(\beta \mid y, X) = \frac{g}{g+1}(\hat{\beta} + \frac{\tilde{\beta}}{g})$
- Mutations - Mathématiques

```
y<-y.math
X<-X.math

#X=scale(X)
g=length(y)
betatilde=rep(0,dim(X)[2])
beta0.lm=mean(y)
beta.lm=(solve(t(X)%*%X)%*%t(X))%*%(y)
betahat=rbind(Intercept=beta0.lm,beta.lm)
```

```r
#betahat
mbetabayes=g/(g+1)*(beta.lm+betatilde/g)
postmean=rbind(Intercept=beta0.lm,mbetabayes)
postmean
```

```
##                         [,1]
## Intercept      8.610169e+01
## Prs_l         -6.904238e-12
## Prs_es         7.453031e-14
## Prs_s          7.482883e-14
## Suc.brt_l      1.497580e-12
## Suc.brt_es    -6.605994e-13
## Suc.brt_s      9.833333e-01
## Suc.att_l      2.614668e-13
## Suc.att_es     1.959800e-12
## Suc.att_s     -2.478203e-12
## Eff_2nd        3.916202e-14
## Eff_1er       -4.285681e-14
## Acc.brt_bac.2 -9.653528e-14
## Acc.att_bac.2  5.790479e-13
## Acc.brt_bac.1 -4.294824e-13
## Acc.att_bac.1 -7.925882e-14
## Suc.brt_Tot    3.286075e-13
## Suc.att_Tot   -7.655136e-13
```

On pourrait aussi retrouver les coéfficients $\hat{\beta}$ à partir de la fonction lm.
On remarque cependant une différence assez significative entre les deux appproches, bien que l'odre de grandeur des coefficients est comparable.

```r
reg.lm=lm(y~X)
summary(reg.lm)
```

```
## Warning in summary.lm(reg.lm): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.508e-15 -2.737e-16 -4.241e-17  2.947e-16  1.015e-15
##
## Coefficients:
##                  Estimate Std. Error    t value Pr(>|t|)
## (Intercept)     6.505e-15  2.217e-15  2.935e+00 0.005449 **
## XPrs_l          2.797e-18  5.646e-18  4.950e-01 0.623040
## XPrs_es         6.693e-18  4.356e-18  1.536e+00 0.132117
## XPrs_s          4.341e-18  3.830e-18  1.133e+00 0.263637
## XSuc.brt_l      3.290e-17  9.622e-18  3.420e+00 0.001431 **
## XSuc.brt_es     4.985e-17  1.886e-17  2.643e+00 0.011596 *
## XSuc.brt_s      1.000e+00  2.534e-17  3.947e+16  < 2e-16 ***
```

```
## XSuc.att_l     -4.132e-17  2.559e-17 -1.615e+00 0.114076
## XSuc.att_es     3.778e-17  3.248e-17  1.163e+00 0.251469
## XSuc.att_s      9.380e-18  3.788e-17  2.480e-01 0.805651
## XEff_2nd        2.611e-18  2.372e-18  1.101e+00 0.277444
## XEff_1er       -1.900e-18  2.863e-18 -6.640e-01 0.510607
## XAcc.brt_bac.2  7.361e-17  2.811e-17  2.619e+00 0.012307 *
## XAcc.att_bac.2 -1.234e-16  3.307e-17 -3.732e+00 0.000576 ***
## XAcc.brt_bac.1 -7.267e-17  4.750e-17 -1.530e+00 0.133679
## XAcc.att_bac.1  1.164e-16  6.787e-17  1.716e+00 0.093776 .
## XSuc.brt_Tot   -6.019e-18  4.990e-17 -1.210e-01 0.904589
## XSuc.att_Tot   -5.238e-17  8.027e-17 -6.520e-01 0.517727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.275e-16 on 41 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.143e+33 on 17 and 41 DF,  p-value: < 2.2e-16
```

- Mutations - Anglais

```
y<-y.en
X<-X.en

#X=scale(X)
g=length(y)
betatilde=rep(0,dim(X)[2])
beta0.lm=mean(y)
beta.lm=(solve(t(X)%*%X)%*%t(X))%*%(y)
betahat=rbind(Intercept=beta0.lm,beta.lm)
mbetabayes=g/(g+1)*(beta.lm+betatilde/g)
postmean=rbind(Intercept=beta0.lm,mbetabayes)
postmean
```

```
##                        [,1]
## Intercept      8.513462e+01
## Prs_l          3.045396e-12
## Prs_es         1.726689e-13
## Prs_s         -2.983934e-14
## Suc.brt_l      1.449118e-12
## Suc.brt_es     1.282731e-12
## Suc.brt_s      9.811321e-01
## Suc.att_l     -2.266455e-12
## Suc.att_es     4.594564e-13
## Suc.att_s     -4.516136e-13
## Eff_2nd       -4.166223e-14
## Eff_1er        4.336678e-14
## Acc.brt_bac.2  2.810331e-13
## Acc.att_bac.2 -6.993148e-13
## Acc.brt_bac.1 -1.014769e-12
## Acc.att_bac.1  1.474443e-12
## Suc.brt_Tot   -2.053502e-12
## Suc.att_Tot    1.897518e-12
```

On pourrait aussi retrouver les coéfficients $\hat{\beta}$ à partir de la fonction lm.
On remarque cependant une différence assez significative entre les deux appproches, bien que l'odre de grandeur des coefficients est comparable.

```
reg.lm=lm(y~X)
summary(reg.lm)
```

```
## Warning in summary.lm(reg.lm): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.221e-15 -1.049e-16  6.021e-17  1.795e-16  5.505e-16
##
## Coefficients:
##                   Estimate Std. Error    t value Pr(>|t|)
## (Intercept)     1.523e-16  2.404e-15  6.300e-02   0.9499
## XPrs_l          3.857e-17  7.276e-18  5.301e+00 6.99e-06 ***
## XPrs_es        -1.923e-17  3.553e-18 -5.412e+00 5.02e-06 ***
## XPrs_s          0.000e+00  3.021e-18  0.000e+00   1.0000
## XSuc.brt_l      0.000e+00  7.509e-18  0.000e+00   1.0000
## XSuc.brt_es     0.000e+00  1.151e-17  0.000e+00   1.0000
## XSuc.brt_s      1.000e+00  1.907e-17  5.243e+16  < 2e-16 ***
## XSuc.att_l     -3.319e-18  2.519e-17 -1.320e-01   0.8960
## XSuc.att_es    -5.924e-18  2.262e-17 -2.620e-01   0.7950
## XSuc.att_s     -1.204e-17  2.598e-17 -4.630e-01   0.6461
## XEff_2nd       -4.930e-18  1.974e-18 -2.498e+00   0.0175 *
## XEff_1er        6.087e-18  2.537e-18  2.399e+00   0.0221 *
## XAcc.brt_bac.2  4.600e-18  1.769e-17  2.600e-01   0.7964
## XAcc.att_bac.2 -7.892e-17  2.973e-17 -2.655e+00   0.0120 *
## XAcc.brt_bac.1  4.957e-17  3.265e-17  1.518e+00   0.1382
## XAcc.att_bac.1  7.629e-17  6.763e-17  1.128e+00   0.2672
## XSuc.brt_Tot   -6.055e-17  4.062e-17 -1.490e+00   0.1453
## XSuc.att_Tot    9.347e-18  7.012e-17  1.330e-01   0.8947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.72e-16 on 34 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.978e+33 on 17 and 34 DF,  p-value: < 2.2e-16
```

2.4.2 Choix des covariables à l'aide des Bayes factor

Bayes Factors et comparaison de modèles Pour comparer les modèles on peut utiliser les facteurs de Bayes On test l'hypothèse H_0, $\forall i = 1, ..., 17$ et on calcul le Bayes Factor à partir de la fonction $BayesReg2$ pour $g = n$

- Mutations en mathématiques - A partir de la fonction $BayesReg2$ pour $g = n$

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept  86.1017      0.1652
## x1          0.0000      0.3021 -0.8891
## x2          0.0000      0.3863 -0.8891
## x3          0.0000      0.5842 -0.8891
## x4          0.0000      0.2491 -0.8891
## x5          0.0000      0.4013 -0.8891
## x6          9.4156      0.5787 21.0919    (****)
## x7          0.0000      0.4491 -0.8891
## x8          0.0000      0.6531 -0.8891
## x9          0.0000      0.8822 -0.8891
## x10         0.0000      0.7836 -0.8891
## x11         0.0000      0.8959 -0.8891
## x12         0.0000      0.5886 -0.8891
## x13         0.0000      0.6433 -0.8891
## x14         0.0000      0.7929 -0.8891
## x15         0.0000      1.0439 -0.8891
## x16         0.0000      0.9570 -0.8891
## x17         0.0000      1.5394 -0.8891
##
##
## Posterior Mean of Sigma2: 1.6099
## Posterior StError of Sigma2: 2.2974

## $postmeancoeff
##  [1]  8.610169e+01 -1.192157e-13 -4.148533e-15  2.467286e-13 -4.174735e-13
##  [6] -6.899666e-13  9.415619e+00  1.402641e-12  1.866731e-12  1.727537e-12
## [11] -3.034980e-13  2.744582e-13  4.464040e-13 -1.411811e-12 -2.179072e-12
## [16]  3.036501e-12  3.657260e-12 -6.452498e-12
##
## $postsqrtcoeff
##                      Prs_l        Prs_es         Prs_s     Suc.brt_l
##     0.1651880     0.3020953     0.3862564     0.5841860     0.2491424
##    Suc.brt_es     Suc.brt_s     Suc.att_l    Suc.att_es     Suc.att_s
##     0.4013360     0.5786809     0.4490823     0.6530826     0.8822358
##       Eff_2nd       Eff_1er Acc.brt_bac.2 Acc.att_bac.2 Acc.brt_bac.1
##     0.7836040     0.8959226     0.5886290     0.6432968     0.7928656
## Acc.att_bac.1   Suc.brt_Tot   Suc.att_Tot
##     1.0439175     0.9570449     1.5393626
##
## $log10bf
##  [1] -0.8890756 -0.8890756 -0.8890756 -0.8890756 -0.8890756 21.0919120
##  [7] -0.8890756 -0.8890756 -0.8890756 -0.8890756 -0.8890756 -0.8890756
## [13] -0.8890756 -0.8890756 -0.8890756 -0.8890756 -0.8890756
##
## $postmeansigma2
## [1] 1.609937
##
## $postvarsigma2
## [1] 5.278048
```

- Mutations en mathématiques - A partir de la fonction $BayesReg2$ pour $g = 1$

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept  86.1017      0.9048
## x1          0.0000      1.1799 -0.1505
## x2          0.0000      1.5086 -0.1505
## x3          0.0000      2.2816 -0.1505
## x4          0.0000      0.9731 -0.1505
## x5          0.0000      1.5675 -0.1505
## x6          4.7876      2.2601  0.8203      (**)
## x7          0.0000      1.7540 -0.1505
## x8          0.0000      2.5507 -0.1505
## x9          0.0000      3.4457 -0.1505
## x10         0.0000      3.0605 -0.1505
## x11         0.0000      3.4992 -0.1505
## x12         0.0000      2.2990 -0.1505
## x13         0.0000      2.5125 -0.1505
## x14         0.0000      3.0967 -0.1505
## x15         0.0000      4.0772 -0.1505
## x16         0.0000      3.7379 -0.1505
## x17         0.0000      6.0122 -0.1505
##
##
## Posterior Mean of Sigma2: 48.2981
## Posterior StError of Sigma2: 68.922

## $postmeancoeff
##  [1]  8.610169e+01 -6.061818e-14 -2.109424e-15  1.254552e-13 -2.122746e-13
##  [6] -3.508305e-13  4.787603e+00  7.132073e-13  9.491852e-13  8.784085e-13
## [11] -1.543210e-13  1.395550e-13  2.269851e-13 -7.178702e-13 -1.108003e-12
## [16]  1.543984e-12  1.859624e-12 -3.280931e-12
##
## $postsqrtcoeff
##                      Prs_l        Prs_es         Prs_s     Suc.brt_l
##     0.9047719     1.1798837     1.5085890     2.2816360     0.9730674
##    Suc.brt_es     Suc.brt_s     Suc.att_l    Suc.att_es     Suc.att_s
##     1.5674849     2.2601349     1.7539660     2.5507234     3.4457197
##       Eff_2nd       Eff_1er Acc.brt_bac.2 Acc.att_bac.2 Acc.brt_bac.1
##     3.0604967     3.4991755     2.2989892     2.5125034     3.0966693
## Acc.att_bac.1   Suc.brt_Tot   Suc.att_Tot
##     4.0771944     3.7378988     6.0122382
##
## $log10bf
##  [1] -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150  0.8202562
##  [7] -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150
## [13] -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150
##
## $postmeansigma2
## [1] 48.29812
##
## $postvarsigma2
## [1] 4750.243
```

- Mutations en anglais - A partir de la fonction $BayesReg2$ pour $g = n$

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept  85.1346      0.1856
## x1          0.0000      0.4680 -0.8621
## x2          0.0000      0.4393 -0.8621
## x3          0.0000      0.6283 -0.8621
## x4          0.0000      0.3186 -0.8621
## x5          0.0000      0.4189 -0.8621
## x6          9.2800      0.6429 17.5057    (****)
## x7          0.0000      0.6551 -0.8621
## x8          0.0000      0.6518 -0.8621
## x9          0.0000      0.8641 -0.8621
## x10         0.0000      0.9345 -0.8621
## x11         0.0000      1.0954 -0.8621
## x12         0.0000      0.5832 -0.8621
## x13         0.0000      0.7794 -0.8621
## x14         0.0000      0.7603 -0.8621
## x15         0.0000      1.4612 -0.8621
## x16         0.0000      1.0542 -0.8621
## x17         0.0000      1.9166 -0.8621
##
##
## Posterior Mean of Sigma2: 1.7913
## Posterior StError of Sigma2: 2.5596

## $postmeancoeff
##  [1]  8.513462e+01  4.217674e-13 -2.740072e-13  8.348207e-13 -3.180684e-14
##  [6] -2.614261e-15  9.280008e+00 -4.814597e-13 -5.803659e-13  9.794765e-13
## [11]  9.034450e-13 -1.847411e-12  5.751374e-13  7.529072e-13 -3.241684e-13
## [16] -3.585895e-12 -3.415968e-13  2.966315e-12
##
## $postsqrtcoeff
##                      Prs_l        Prs_es         Prs_s     Suc.brt_l
##     0.1856030     0.4679936     0.4393209     0.6282758     0.3185938
##    Suc.brt_es     Suc.brt_s     Suc.att_l    Suc.att_es     Suc.att_s
##     0.4188504     0.6429324     0.6550707     0.6517884     0.8641411
##       Eff_2nd       Eff_1er Acc.brt_bac.2 Acc.att_bac.2 Acc.brt_bac.1
##     0.9344800     1.0953708     0.5832310     0.7793598     0.7602870
## Acc.att_bac.1   Suc.brt_Tot   Suc.att_Tot
##     1.4611691     1.0541526     1.9165925
##
## $log10bf
##  [1] -0.8621379 -0.8621379 -0.8621379 -0.8621379 -0.8621379 17.5056625
##  [7] -0.8621379 -0.8621379 -0.8621379 -0.8621379 -0.8621379 -0.8621379
## [13] -0.8621379 -0.8621379 -0.8621379 -0.8621379 -0.8621379
##
## $postmeansigma2
## [1] 1.79132
##
## $postvarsigma2
## [1] 6.551355
```

- Mutations en anglais - A partir de la fonction $BayesReg2$ pour $g = 1$

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept  85.1346      0.9554
## x1          0.0000      1.7198 -0.1505
## x2          0.0000      1.6145 -0.1505
## x3          0.0000      2.3088 -0.1505
## x4          0.0000      1.1708 -0.1505
## x5          0.0000      1.5392 -0.1505
## x6          4.7292      2.3627  0.7199      (**)
## x7          0.0000      2.4073 -0.1505
## x8          0.0000      2.3952 -0.1505
## x9          0.0000      3.1756 -0.1505
## x10         0.0000      3.4341 -0.1505
## x11         0.0000      4.0254 -0.1505
## x12         0.0000      2.1433 -0.1505
## x13         0.0000      2.8641 -0.1505
## x14         0.0000      2.7940 -0.1505
## x15         0.0000      5.3696 -0.1505
## x16         0.0000      3.8739 -0.1505
## x17         0.0000      7.0433 -0.1505
##
##
## Posterior Mean of Sigma2: 47.47
## Posterior StError of Sigma2: 67.8284

## $postmeancoeff
##  [1]  8.513462e+01  2.149392e-13 -1.396383e-13  4.254375e-13 -1.620926e-14
##  [6] -1.332268e-15  4.729235e+00 -2.453593e-13 -2.957634e-13  4.991563e-13
## [11]  4.604095e-13 -9.414691e-13  2.930989e-13  3.836931e-13 -1.652012e-13
## [16] -1.827427e-12 -1.740830e-13  1.511680e-12
##
## $postsqrtcoeff
##                      Prs_l        Prs_es         Prs_s     Suc.brt_l
##     0.9554497     1.7198243     1.6144555     2.3088435     1.1707966
##    Suc.brt_es     Suc.brt_s     Suc.att_l    Suc.att_es     Suc.att_s
##     1.5392285     2.3627050     2.4073121     2.3952498     3.1756225
##       Eff_2nd       Eff_1er Acc.brt_bac.2 Acc.att_bac.2 Acc.brt_bac.1
##     3.4341100     4.0253659     2.1433090     2.8640606     2.7939703
## Acc.att_bac.1   Suc.brt_Tot   Suc.att_Tot
##     5.3696341     3.8738935     7.0432642
##
## $log10bf
##  [1] -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150  0.7198731
##  [7] -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150
## [13] -0.1505150 -0.1505150 -0.1505150 -0.1505150 -0.1505150
##
## $postmeansigma2
## [1] 47.46998
##
## $postvarsigma2
## [1] 4600.689
```

- Conclusion

Critère de choix : Succés brute S

2.4.3 Choix de modèles par test de tous les modèles ou Gibbs-sampler

On utilse la fonctionModChoBayesReg du package Bayess

- Mutations en Math

```
##
## Number of variables greather than 15
## Model posterior probabilities are estimated by using an MCMC algorithm
##
##    Top10Models PostProb
## 1            6   0.1499
## 2         6 12   0.0202
## 3         6 15   0.0201
## 4          3 6   0.0200
## 5         6 11   0.0194
## 6          2 6   0.0192
## 7          4 6   0.0186
## 8          6 9   0.0185
## 9         6 17   0.0183
## 10         5 6   0.0182
```

```
## $top10models
##  [1] "6"    "6 12" "6 15" "3 6"  "6 11" "2 6"  "4 6"  "6 9"  "6 17" "5 6"
##
## $postprobtop10
##  [1] 0.1498625 0.0202250 0.0200875 0.0200000 0.0193750 0.0192000 0.0185625
##  [8] 0.0185125 0.0182750 0.0182000
```

La 6ème covariable est omniprésente dans tous les modèles. La probabilité à piriori du modèle constitué de cette seule variable est écrasante

- Mutations en Anglais

```
##
## Number of variables greather than 15
## Model posterior probabilities are estimated by using an MCMC algorithm
##
##    Top10Models PostProb
## 1            6   0.1273
## 2         6 17   0.0200
## 3         6 12   0.0195
## 4          6 9   0.0185
## 5          1 6   0.0178
## 6          2 6   0.0177
## 7         6 14   0.0177
## 8          6 8   0.0176
## 9         6 15   0.0176
## 10         5 6   0.0168
```

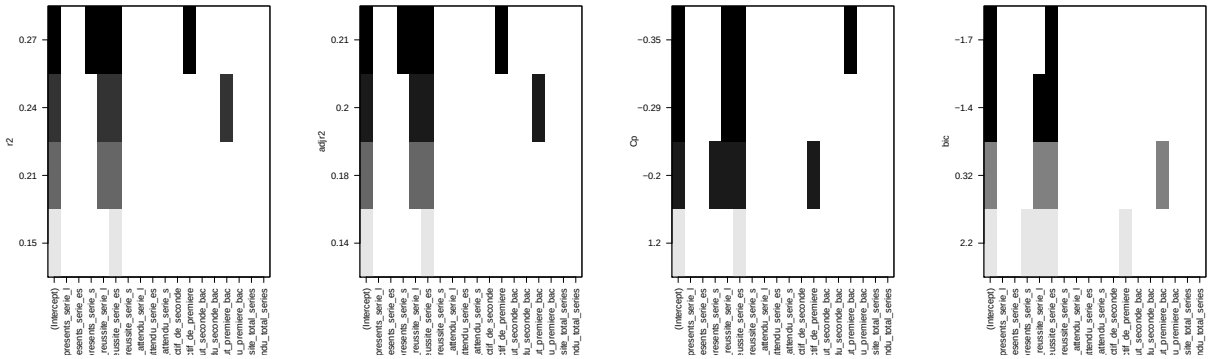
```
## $top10models
##  [1] "6"    "6 17" "6 12" "6 9"  "1 6"  "2 6"  "6 14" "6 8"  "6 15" "5 6"
##
## $postprobtop10
##  [1] 0.1273125 0.0200250 0.0194625 0.0184750 0.0177625 0.0177500 0.0176875
##  [8] 0.0176125 0.0175625 0.0167625
```

On retrouve la encore la prédominance de la 6ème variable : $Suc.brt_s$ = Réussite brute terminale s.

2.4.4 Comparaison au résultat obtenu par une analyse fréquentiste

- Analyse fréquentiste - Mutations en mathématiques

```
##
## Call:
## lm(formula = Barre ~ ., data = d.math.reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -173.08  -69.00  -21.11   75.56  273.03
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       65.7874   489.3234   0.134   0.8937
## effectif_presents_serie_l          0.4061     1.2463   0.326   0.7462
## effectif_presents_serie_es        -0.3858     0.9616  -0.401   0.6903
## effectif_presents_serie_s          1.1721     0.8454   1.386   0.1731
## taux_brut_de_reussite_serie_l     -3.6683     2.1240  -1.727   0.0917 .
## taux_brut_de_reussite_serie_es     7.2467     4.1641   1.740   0.0893 .
## taux_brut_de_reussite_serie_s      1.4955     5.5932   0.267   0.7905
## taux_reussite_attendu_serie_l     -4.1272     5.6497  -0.731   0.4692
## taux_reussite_attendu_serie_es    -6.2008     7.1700  -0.865   0.3922
## taux_reussite_attendu_serie_s     -2.0550     8.3611  -0.246   0.8071
## effectif_de_seconde                0.8035     0.5236   1.535   0.1325
## effectif_de_premiere              -1.4042     0.6320  -2.222   0.0319 *
## taux_acces_brut_seconde_bac       13.0976     6.2043   2.111   0.0409 *
## taux_acces_attendu_seconde_bac    -9.2312     7.2998  -1.265   0.2132
## taux_acces_brut_premiere_bac     -12.5976    10.4840  -1.202   0.2364
## taux_acces_attendu_premiere_bac    5.6413    14.9814   0.377   0.7084
## taux_brut_de_reussite_total_series  1.4551   11.0152   0.132   0.8956
## taux_reussite_attendu_total_series 11.3206   17.7198   0.639   0.5265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 116.4 on 41 degrees of freedom
## Multiple R-squared:  0.3826, Adjusted R-squared:  0.1267
## F-statistic: 1.495 on 17 and 41 DF,  p-value: 0.1451
```



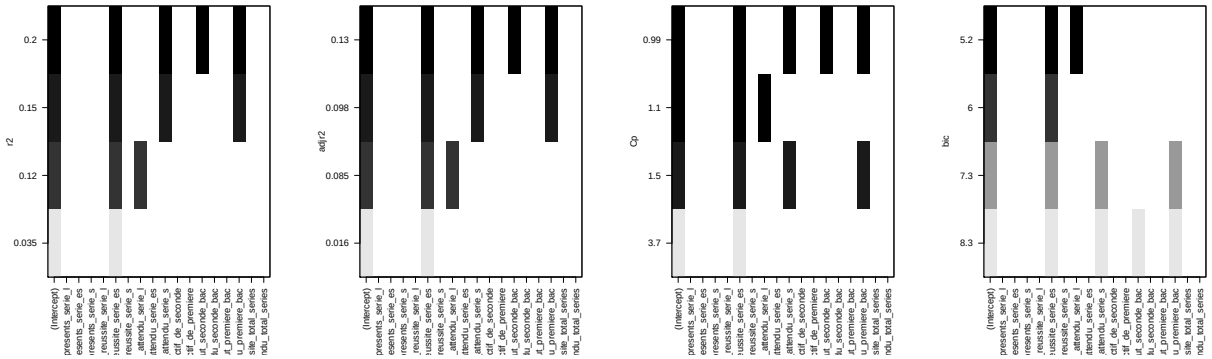

```
summary(step_mod)
```

```
##
## Call:
## lm(formula = Barre ~ effectif_presents_serie_s + taux_brut_de_reussite_serie_l +
##     taux_brut_de_reussite_serie_es + effectif_de_seconde + effectif_de_premiere +
##     taux_acces_brut_seconde_bac, data = d.math.reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -222.62  -73.59  -20.98   54.23  292.99
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -77.4081   182.6176  -0.424   0.6734
## effectif_presents_serie_s        0.7892     0.5009   1.575   0.1212
## taux_brut_de_reussite_serie_l   -3.9693     1.5624  -2.541   0.0141 *
## taux_brut_de_reussite_serie_es   3.8648     2.2119   1.747   0.0865 .
## effectif_de_seconde              0.6536     0.3883   1.683   0.0983 .
## effectif_de_premiere            -1.1044     0.4286  -2.576   0.0129 *
## taux_acces_brut_seconde_bac      4.2086     2.6145   1.610   0.1135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108.2 on 52 degrees of freedom
## Multiple R-squared:  0.3242, Adjusted R-squared:  0.2462
## F-statistic: 4.157 on 6 and 52 DF,  p-value: 0.001744
```

- Analyse fréquentiste - Mutations en Anglais

```
##
## Call:
## lm(formula = Barre ~ ., data = d.en.reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -363.03 -133.21  -24.75   97.93 1044.32
##
## Coefficients:
```

```
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -1286.4382  1860.5131  -0.691   0.4940
## effectif_presents_serie_l            -5.2435     5.6318  -0.931   0.3584
## effectif_presents_serie_es           -1.6538     2.7501  -0.601   0.5516
## effectif_presents_serie_s             0.7911     2.3383   0.338   0.7372
## taux_brut_de_reussite_serie_l        -5.1155     5.8121  -0.880   0.3850
## taux_brut_de_reussite_serie_es        5.9949     8.9053   0.673   0.5054
## taux_brut_de_reussite_serie_s         3.0029    14.7626   0.203   0.8400
## taux_reussite_attendu_serie_l        -3.7900    19.4963  -0.194   0.8470
## taux_reussite_attendu_serie_es       30.9369    17.5094   1.767   0.0862
## taux_reussite_attendu_serie_s       -24.6363    20.1117  -1.225   0.2290
## effectif_de_seconde                   1.2467     1.5277   0.816   0.4202
## effectif_de_premiere                 -1.0171     1.9638  -0.518   0.6079
## taux_acces_brut_seconde_bac           8.7706    13.6909   0.641   0.5261
## taux_acces_attendu_seconde_bac      -30.4301    23.0086  -1.323   0.1948
## taux_acces_brut_premiere_bac        -42.0258    25.2706  -1.663   0.1055
## taux_acces_attendu_premiere_bac     103.0319    52.3436   1.968   0.0572
## taux_brut_de_reussite_total_series   32.8251    31.4412   1.044   0.3038
## taux_reussite_attendu_total_series  -62.3386    54.2693  -1.149   0.2587
##
## (Intercept)
## effectif_presents_serie_l
## effectif_presents_serie_es
## effectif_presents_serie_s
## taux_brut_de_reussite_serie_l
## 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
## effectif_de_seconde
## effectif_de_premiere
## taux_acces_brut_seconde_bac
## taux_acces_attendu_seconde_bac
## taux_acces_brut_premiere_bac
## taux_acces_attendu_premiere_bac    .
## taux_brut_de_reussite_total_series
## taux_reussite_attendu_total_series
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 287.9 on 34 degrees of freedom
## Multiple R-squared:  0.3659, Adjusted R-squared:  0.0489
## F-statistic: 1.154 on 17 and 34 DF,  p-value: 0.3492
```




```
summary(step_mod)
```

```
##
## Call:
## lm(formula = Barre ~ effectif_presents_serie_l + taux_brut_de_reussite_serie_l +
##     taux_reussite_attendu_serie_es + taux_reussite_attendu_serie_s +
##     taux_acces_attendu_seconde_bac + taux_acces_brut_premiere_bac +
##     taux_acces_attendu_premiere_bac + taux_brut_de_reussite_total_series +
##     taux_reussite_attendu_total_series, data = d.en.reg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -420.1 -116.3  -33.2   88.1 1089.2
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -1696.950    822.088  -2.064  0.04521
## effectif_presents_serie_l            -3.095      2.219  -1.395  0.17038
## taux_brut_de_reussite_serie_l        -6.030      4.225  -1.427  0.16087
## taux_reussite_attendu_serie_es       37.414     13.888   2.694  0.01010
## taux_reussite_attendu_serie_s       -26.011     15.296  -1.700  0.09643
## taux_acces_attendu_seconde_bac      -26.632     15.037  -1.771  0.08381
## taux_acces_brut_premiere_bac        -29.358     16.041  -1.830  0.07432
## taux_acces_attendu_premiere_bac     101.217     34.272   2.953  0.00513
## taux_brut_de_reussite_total_series   44.714     17.511   2.553  0.01439
## taux_reussite_attendu_total_series  -74.814     32.156  -2.327  0.02488
##
## (Intercept)                        *
## effectif_presents_serie_l
## taux_brut_de_reussite_serie_l
## taux_reussite_attendu_serie_es     *
## taux_reussite_attendu_serie_s      .
## taux_acces_attendu_seconde_bac     .
## taux_acces_brut_premiere_bac       .
## taux_acces_attendu_premiere_bac    **
## taux_brut_de_reussite_total_series *
## taux_reussite_attendu_total_series *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 269.3 on 42 degrees of freedom
## Multiple R-squared:  0.3146, Adjusted R-squared:  0.1677
## F-statistic: 2.142 on 9 and 42 DF,  p-value: 0.04689
```

2.5 Conclusion

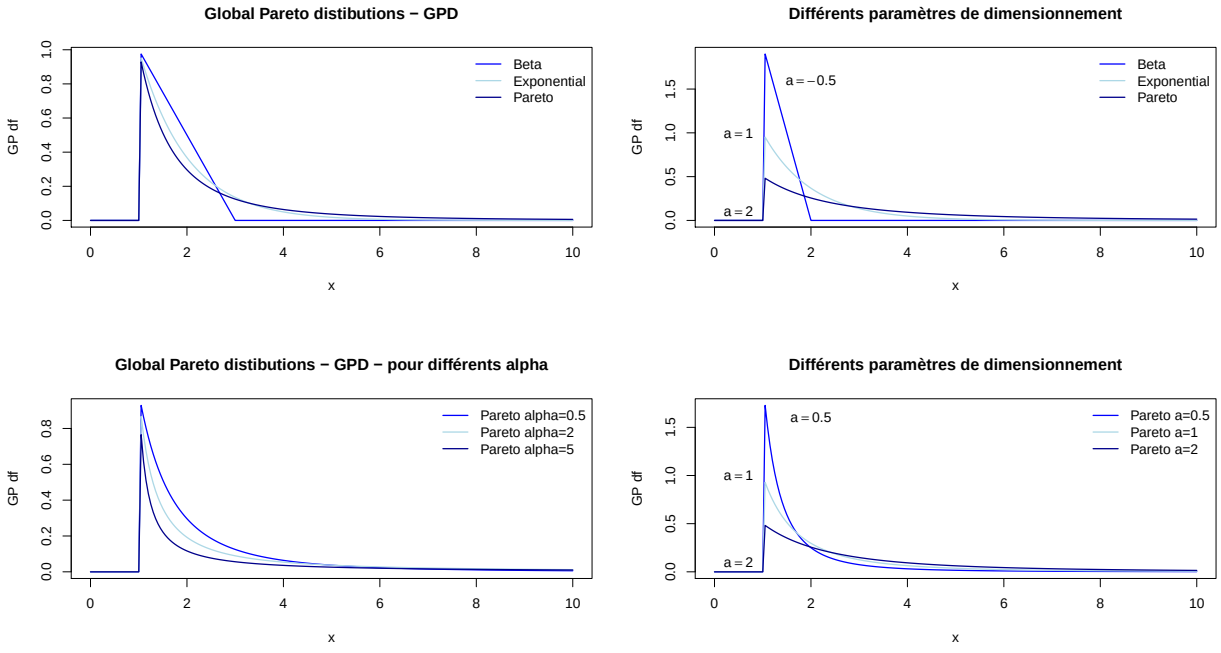
Pour les mutations en Math et en Anglais, on a plus de difficulté à sélectionner les variables dans le cas fréquentiste, alors que dans le cas bayésien une covariable ressort très nettement. Et confirme

3 Loi de Pareto

On ignore maintenant les covariables, et on s'intéresse uniquement à la loi du nombre de points nécessaire (colonne Barre). La loi gaussienne peut paraître peu pertinente pour ces données : on va plutôt proposer une loi de Pareto. Pour $m > 0$ et $\alpha > 0$, on dit que $ZPareto(m; \alpha)$ si Z est à valeurs dans $[m; +1[$ de densité:

$f(z \mid \alpha, m) = \alpha \frac{m^\alpha}{z^{\alpha+1}} \mathbb{1}_{[\geqslant, +\infty[}$

3.1 Package R pour générer des réalisation d'une loi de Paréto



3.2 Choix d'une loi à priori pour α

- Loi de paréto :

$$f(z \mid \alpha, m) = \alpha \frac{m^\alpha}{z^{\alpha+1}} \mathbb{1}_{[\geqslant, +\infty[}$$

$$f(z \mid \alpha, m) \propto \alpha e^{\alpha log(m/z)}$$

A une constante multiplicative près et après transformation en log, on reconnaît une loi exponentielle de paramètre α On peut prendre une loi a priori de type *gamma* de manière à avoir une loi conjuguée.

3.3 Loi à postériori de α

3.4 Echantillon de la loi à postériori de α

Par la méthode de votre choix, tirer un échantillon de la loi a posteriori de. Donner un intervalle de crédibilité à 95%.

3.5 8

4 Annexes

4.1 Test des méthodes BayesReg du package Bayess et BayesReg2 version modifiée

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept   3.4878      0.0304
## x1          1.0225      0.0303     Inf    (****)
##
##
## Posterior Mean of Sigma2: 0.2513
## Posterior StError of Sigma2: 0.3561

## $postmeancoeff
## [1] 3.487783 1.022509
##
## $postsqrtcoeff
## [1] 0.03039825 0.03034252
##
## $log10bf
##      [,1]
## [1,]  Inf
##
## $postmeansigma2
## [1] 0.2513425
##
## $postvarsigma2
## [1] 0.1268176

##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept   3.4878      0.2202
## x1          0.0499      0.0030     Inf    (****)
##
##
## Posterior Mean of Sigma2: 13.1889
## Posterior StError of Sigma2: 18.6866
```

```
## $postmeancoeff
## [1] 3.48778309 0.04994557
##
## $postsqrtcoeff
## [1] 0.220201027 0.003044958
##
## $log10bf
##      [,1]
## [1,]  Inf
##
## $postmeansigma2
## [1] 13.18887
##
## $postvarsigma2
## [1] 349.1907
```

```
## [1]  0.8628900  0.3852624  0.1222176 -0.1625189 -1.4271164  0.3987761
```

```
##        x1 x2 x3  x4  x5  x6  x7  x8
## [1,] 1200 22  1 4.0 1.1 5.9 1.4 1.4
## [2,] 1342 28  8 4.4 1.5 6.4 1.7 1.7
## [3,] 1231 28  5 2.4 1.6 4.3 1.5 1.4
## [4,] 1254 28 18 3.0 1.7 6.9 2.3 1.6
## [5,] 1357 32  7 3.7 1.7 6.6 1.8 1.3
## [6,] 1250 27  1 4.4 1.7 5.8 1.3 1.4
```

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept  -0.8133      0.1407
## x1         -0.5039      0.1883  0.7224      (**)
## x2         -0.3755      0.1508  0.5392      (**)
## x3          0.6225      0.3436 -0.0443
## x4         -0.2776      0.2804 -0.5422
## x5         -0.2069      0.1499 -0.3378
## x6          0.2806      0.4760 -0.6857
## x7         -1.0420      0.4178  0.5435      (**)
## x8         -0.0221      0.1531 -0.7609
##
##
## Posterior Mean of Sigma2: 0.6528
## Posterior StError of Sigma2: 0.939
```

```
## $postmeancoeff
## [1] -0.81328069 -0.50390377 -0.37548142  0.62252447 -0.27762947 -0.20688023
## [7]  0.28061938 -1.04204277 -0.02209411
##
## $postsqrtcoeff
##                  x1        x2        x3        x4        x5        x6
## 0.1406514 0.1882559 0.1508271 0.3436217 0.2803657 0.1498641 0.4759505
##        x7        x8
## 0.4178148 0.1530573
##
## $log10bf
```

```
## [1]  0.72241000  0.53918250 -0.04430805 -0.54224765 -0.33779821 -0.68568404
## [7]  0.54353138 -0.76091468
##
## $postmeansigma2
## [1] 0.6528327
##
## $postvarsigma2
## [1] 0.8817734
```

```
##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept  -0.8133      0.1407
## x1         -0.5039      0.1883  0.7224      (**)
## x2         -0.3755      0.1508  0.5392      (**)
## x3          0.6225      0.3436 -0.0443
## x4         -0.2776      0.2804 -0.5422
## x5         -0.2069      0.1499 -0.3378
## x6          0.2806      0.4760 -0.6857
## x7         -1.0420      0.4178  0.5435      (**)
## x8         -0.0221      0.1531 -0.7609
##
##
## Posterior Mean of Sigma2: 0.6528
## Posterior StError of Sigma2: 0.939
```

```
## $postmeancoeff
## [1] -0.81328069 -0.50390377 -0.37548142  0.62252447 -0.27762947 -0.20688023
## [7]  0.28061938 -1.04204277 -0.02209411
##
## $postsqrtcoeff
##                  x1        x2        x3        x4        x5        x6
## 0.1406514 0.1882559 0.1508271 0.3436217 0.2803657 0.1498641 0.4759505
##        x7        x8
## 0.4178148 0.1530573
##
## $log10bf
## [1]  0.72241000  0.53918250 -0.04430805 -0.54224765 -0.33779821 -0.68568404
## [7]  0.54353138 -0.76091468
##
## $postmeansigma2
## [1] 0.6528327
##
## $postvarsigma2
## [1] 0.8817734
```

- ModChoBayesReg