Projet de modélisation statistique

Logiciel R

CHRISTOPHE Clifford, DELCOMBEL Nicolas & LELEU Maïa

Décembre 2018

Introduction

L'aire de Broca est l'une des deux principales zones du cerveau hominidé responsables du traitement du langage. Classiquement, l'aire de Broca est la zone associée à la production des mots parlés alors qu'une zone différente du cerveau, l'aire de Wernicke, est associée à la compréhension de ces mots. Cependant, d'autres zones associées au langage ont été identifiées et les fonctions du traitement du langage sont davantage distribuées à travers toutes ces zones. De plus, une récente remise en cause de la fonction langagière de l'aire de Broca montrerait que cette aire ne serait pas l'aire de la parole.

Une analyse statistique des données d'activation (variation du signal BOLD) au cours d'une tâche de production langagière chez 124 sujets ont été récupérées dans les 6 régions cérébrales ci-dessous :

- le gyrus frontal inférieur triangulaire (ou aire de Broca, PROD_G_Frontal_Inf_Tri_1),
- le sillon supérieur temporal (ou aire de Wernicke, PROD_S_Sup_Temporal_4),
- le gyrus occipital latéral (PROD_G_Occipital_Lat_1),
- le gyrus angulaire (PROD_G_Angular_2),
- l'opercule rolandique (PROD_G_Rolandic_Oper_1),
- l'hippocampe (PROD_G_Hippocampus-1)

Pour mieux comprendre les interactions entre ces différentes régions cérébrales au cours de la production d'une phrase, nous allons expliquer les fluctuations des activations de l'aire de Broca à gauche à l'aide des autres variables présentes dans le jeu de données.

Jeu de données à traiter :

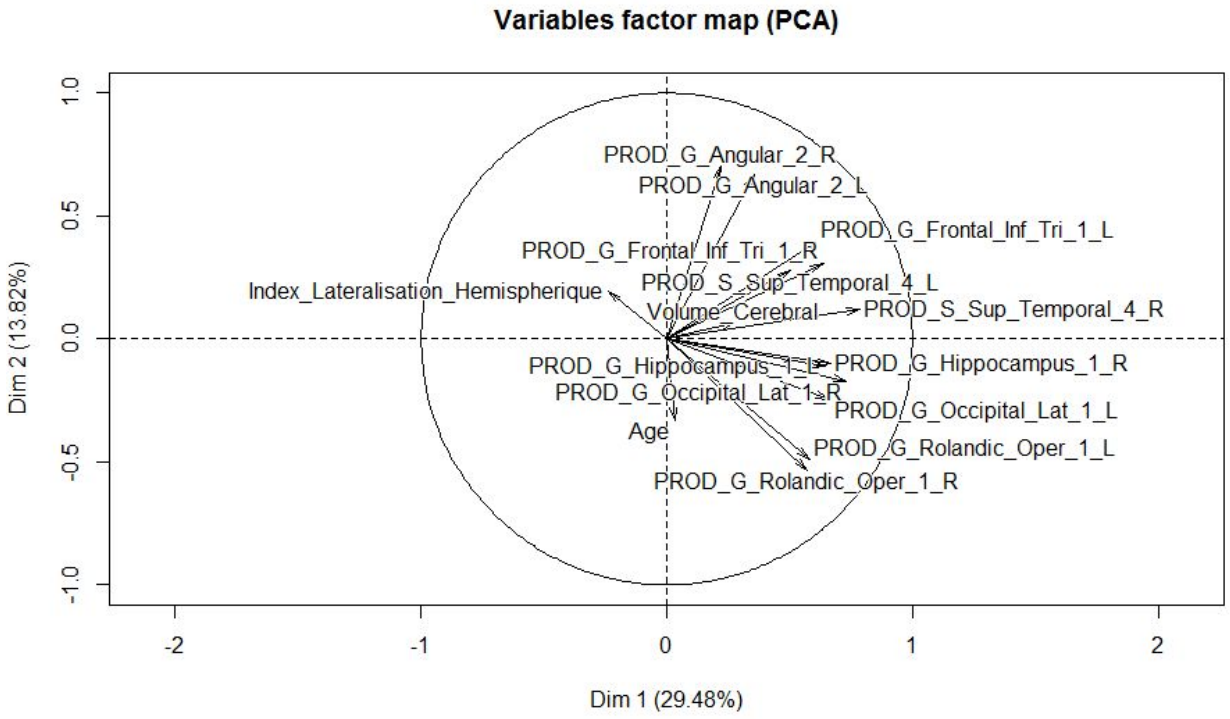
Données d'activation durant une tâche de production langagière

I. Analyse descriptive des données

On peut observer que la moyenne de l'index de latéralisation hémisphérique des individus est de 59.87097 et qu'elle positive pour chacun d'entre eux. Tous les sujets ont donc un hémisphère gauche dominant.

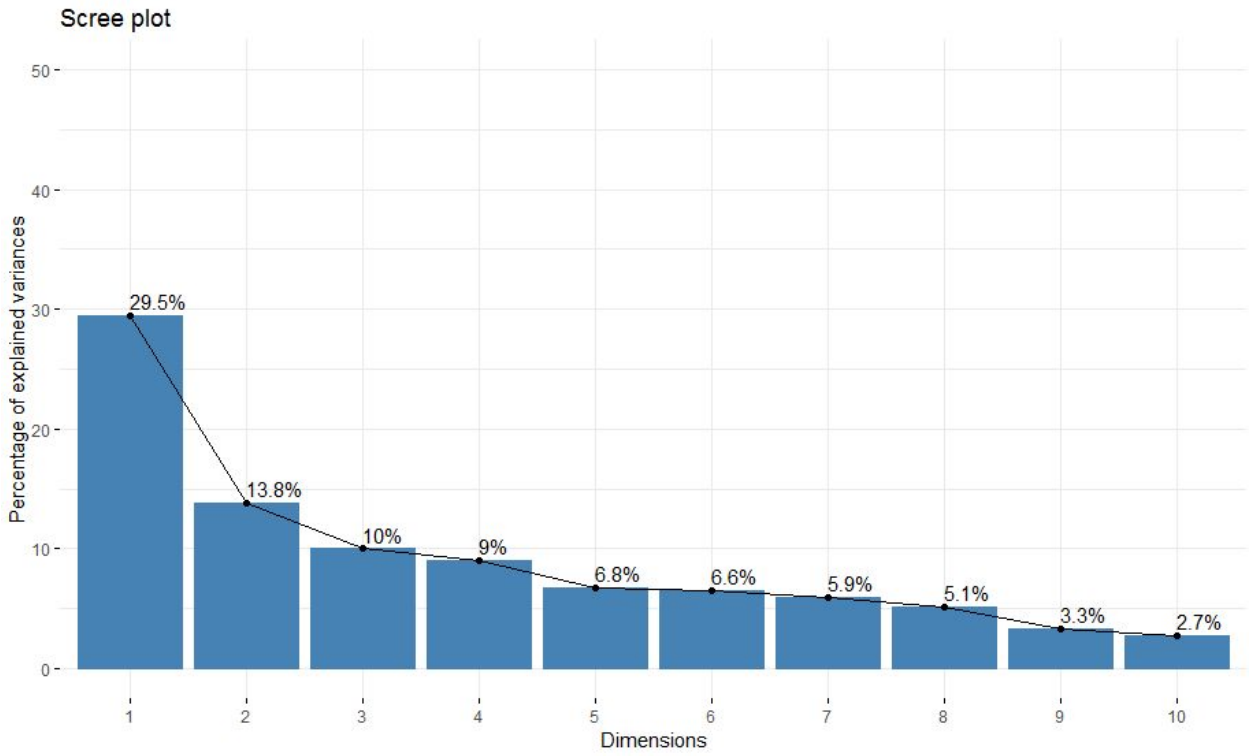
Nos données sont composées d'un ensemble de 124 individus (n=124) et de 16 variables différentes quantitatives (p=26).

Nous avons utilisé les packages FactoMineR et factoextra. Cette ligne permet d'obtenir le graphique de l'ACP des plans factoriels 1, 2 suivants : res <- PCA(data[3:17], graph = TRUE)



D'après ce graphique, on peut plus ou moins interpréter la correlations entre plusieurs variables. Notons tout que la qualité de projection des axes sur le cercle n'est pas énorme. On peut noter que l'aire de Broca ne semble pas corrélée avec l'opercule rolandique mais est plutôt bien corrélé avec le sillon supérieur temporal. On remarque aussi que les aires réparties dans deux hémisphères différents sont très corrélés, ce qui semble évident.

Les valeurs propres mesurent la quantité de variance donnée par chaque dimension : fviz_eig(res, addlabels = TRUE, ylim = c(0, 50))



On peut voir que 6 dimensions suffisent à représenter plus de 75 % de la variance totale.

II. Régression linéaire

1. Etude de l'ensemble de l'échantillon

On commence par étudier le modèle sur les hommes et les femmes, sans distinction de sexe.

Pour des raison de lisibilités, nous adopterons la légende ci-dessous pour le reste du rapport concernant la comparaison des modèles.

Points positifs (+)

Points négatifs (-)

a. Modèle M1

```
res <- lm(PROD_G_Frontal_Inf_Tri_1_L~PROD_G_Angular_2_L+PROD_G_Occipital_Lat_1_L
+PROD_G_Rolandic_Oper_1_L+PROD_S_Sup_Temporal_4_L+PROD_G_Hippocampus_1_L+PROD_G_Frontal_Inf_Tri_1_R+
+PROD_G_Angular_2_R+PROD_G_Occipital_Lat_1_R+PROD_G_Rolandic_Oper_1_R
+PROD_S_Sup_Temporal_4_R+PROD_G_Hippocampus_1_R, data=data)
```

```
summary(res)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.5845738  0.0826357   7.074 1.38e-10 ***
PROD_G_Angular_2_L        0.2654978  0.0923663   2.874  0.00485 **
PROD_G_Occipital_Lat_1_L -0.0993751  0.1605645  -0.619  0.53723
PROD_G_Rolandic_Oper_1_L  0.0789264  0.1294998   0.609  0.54345
PROD_S_Sup_Temporal_4_L   0.4994299  0.1022577   4.884 3.48e-06 ***
PROD_G_Hippocampus_1_L    0.4827193  0.1659019   2.910  0.00437 **
PROD_G_Frontal_Inf_Tri_1_R 0.6093699 0.1082234   5.631 1.35e-07 ***
PROD_G_Angular_2_R        0.0008514  0.1202390   0.007  0.99436
PROD_G_Occipital_Lat_1_R  0.3486537  0.1405364   2.481  0.01459 *
PROD_G_Rolandic_Oper_1_R -0.0079675  0.1435998  -0.055  0.95585
PROD_S_Sup_Temporal_4_R  -0.3152247  0.1205085  -2.616  0.01013 *
PROD_G_Hippocampus_1_R   -0.3972710  0.1790667  -2.219  0.02853 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3274 on 112 degrees of freedom
Multiple R-squared:  0.5917,    Adjusted R-squared:  0.5517
F-statistic: 14.76 on 11 and 112 DF,  p-value: < 2.2e-16

>
```

Le test de Fisher (F-statistic) est un test de significativité du modèle testé. On a H0 qui suppose la nullité des des paramètres (β1=β2=...=β11=0) et H1 qui suppose l'inverse. On remarque que la p-value est très inférieure à 0.05 donc on rejette fortement H0. Le modèle apparaît donc utile.

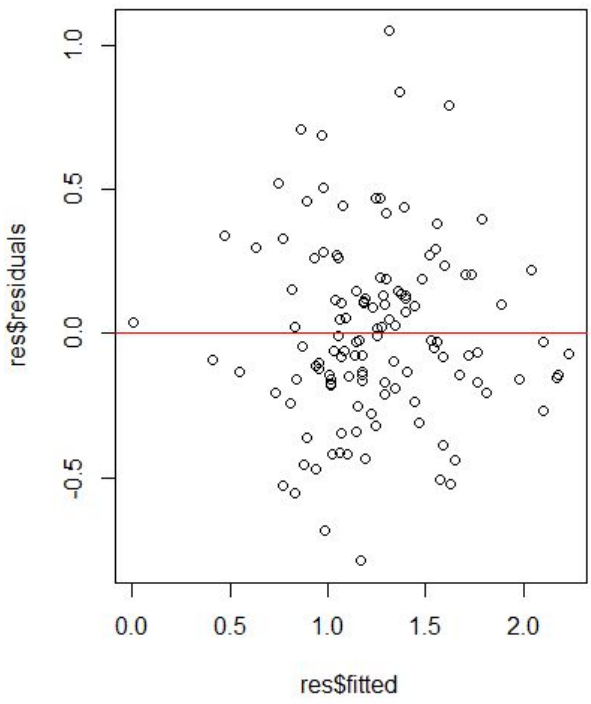
Cependant, en analysant ligne par ligne chaque variable, on peut constater que certaines apparaissent inutiles au modèle. Le test de Student indique par exemple que le gyrus occipital latéral dans l'hémisphère gauche, l'opercule rolandique dans les deux hémisphères et le gyrus angulaire dans l'hémisphère droit sont des variables potentiellement inutile au modèle. En effet, leur p-value sont toutes supérieure à 5% donc on accepte H0 qui suppose que leur paramètre β est nul. Il va donc falloir faire de la sélection de variables.

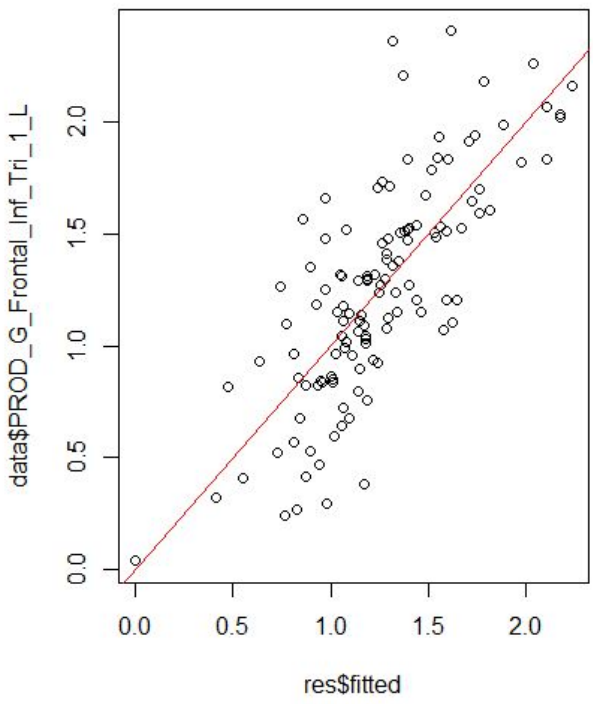
On a un R^2 environ égal à 59% ce qui indique que ce modèle explique 59% de la variance, ce qui ne semble pas énorme (R^2 ajusté environ égale à 55%).

Vérifions la normalité des résidus.

```
        Shapiro-Wilk normality test

data:  res$residuals
W = 0.98196, p-value = 0.09673
```

On constate que la p-value est supérieure à 5% donc on accepte H0 qui suppose la normalité des résidus.





Le graphique des résidus nous permet d'avoir de l'information sur les résidus. On peut remarquer qu'il n'y a pas de structure qui apparaît, ce qui est bien le reflet d'une erreur aléatoire.

On va maintenant utiliser une techniques d'estimation et d'analyse d'un modèle de régression linéaire multiple pour proposer un modèle "plus simple et meilleur", basé sur le critère AIC.

b. Modèle M1' (AIC drop)

Nous allons réaliser une sélection automatique de variables basée sur le critère AIC dans un premier temps grâce à la fonction "drop" comme suivant : `drop1(res)`

Cette fonction nous indique les variables inutiles au modèle qu'on peut supprimer pour la création du nouveau modèle.

```
resDrop <- lm(PROD_G_Frontal_Inf_Tri_1_L~PROD_G_Angular_2_L
              +PROD_S_Sup_Temporal_4_L+PROD_G_Hippocampus_1_L+PROD_G_Frontal_Inf_Tri_1_R+
                +PROD_G_Occipital_Lat_1_R
              +PROD_S_Sup_Temporal_4_R+PROD_G_Hippocampus_1_R, data=data[,6:17])


summary(resDrop)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.79898 -0.18735 -0.02331  0.17621  1.07754

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.55279    0.06921   7.987 1.12e-12 ***
PROD_G_Angular_2_L         0.25505    0.07009   3.639  0.00041 ***
PROD_S_Sup_Temporal_4_L    0.51267    0.09646   5.315 5.24e-07 ***
PROD_G_Hippocampus_1_L     0.50916    0.15955   3.191  0.00182 **
PROD_G_Frontal_Inf_Tri_1_R 0.61740    0.10376   5.950 2.91e-08 ***
PROD_G_Occipital_Lat_1_R   0.30585    0.10407   2.939  0.00398 **
PROD_S_Sup_Temporal_4_R   -0.31230    0.11761  -2.655  0.00904 **
PROD_G_Hippocampus_1_R    -0.41969    0.17153  -2.447  0.01591 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.323 on 116 degrees of freedom
Multiple R-squared:  0.5885,    Adjusted R-squared:  0.5636
F-statistic:  23.7 on 7 and 116 DF,  p-value: < 2.2e-16
```

Le test de Fisher est bon, c'est-à-dire qu'ici, la p-value est très inférieure à 0.05 donc on rejette fortement H0, c'est donc un modèle utile.

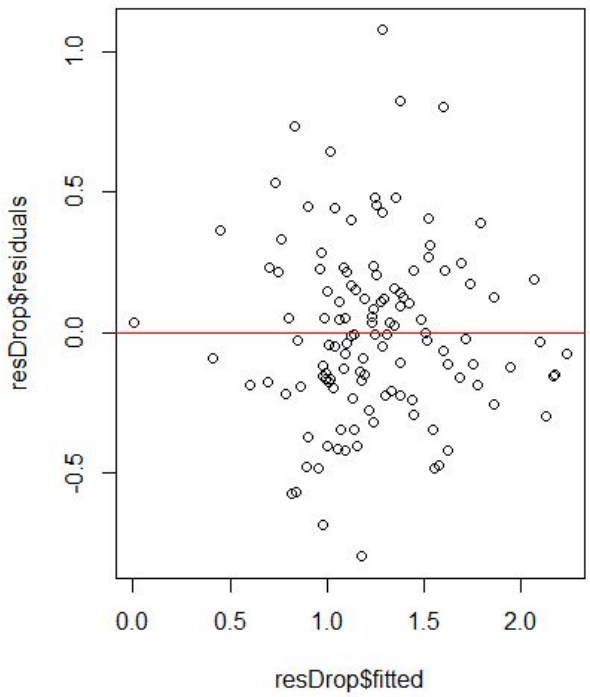
Le R^2 est environ égal à 59%, mais on a moins de variables. Le R^2 ajusté est lui, plus un peu plus élevé qu'avant (0.5636>0.5517).

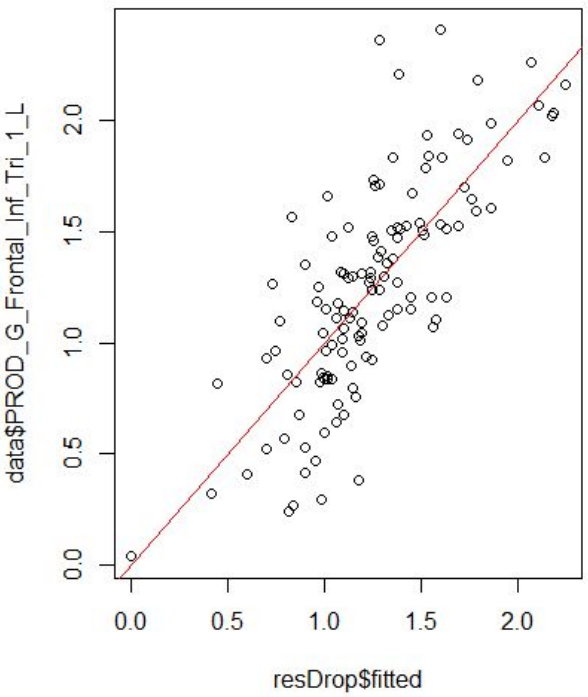
On peut aussi noter que le RSD est plus faible que celui du premier modèle (0.323<0.3274).

De plus, tous paramètres βk du modèle sont significativement différents de 0.

```
        Shapiro-Wilk normality test

data:  resDrop$residuals
W = 0.98312, p-value = 0.1248
```

Le test de Shapiro indique bien que les résidus suivent une loi normale. En effet, la p-value est supérieure à 5% donc on accepte H0.





On constate qu'il n'y pas de structure dans le graphique de résidus.

En conclusion, ce modèle est meilleur que le modèle M1.

On a y = 0.55279 + 0.25505 * PROD_G_Angular_2_L + 0.51267 * PROD_S_Sup_Temporal_4_L + 0.50916 * PROD_G_Hippocampus_1_L + 0.61740 * PROD_G_Frontal_Inf_Tri_1_R + 0.30585 * PROD_G_Occipital_Lat_1_R + (-0.31230) * PROD_S_Sup_Temporal_4_R + (-0.41969) * PROD_G_Hippocampus_1_R + ε où ε -> N(0,1).

2. Modèle spécifique aux femmes

a. Analyse du modèle M2

```
colnames(data)
data <- data[2:17]
attach(data) #Accès aux composantes directement par son nom
dataTriSexe <- split(data, Sexe)#Découpage des données en 2 (1ère partie : F, 2ème partie : H)
dataTriSexe
detach(data)
attach(dataTriSexe) #On n'a plus accès aux composantes par nom de "data" mais de "dataTriSexe"
dataF <- F #Stockage des valeurs F
dataH <- H
dataF<-dataF[2:16]
dataH<-dataH[2:16]
dataF
dataH
```

Après avoir trié préalablement les données selon le sexe grâce la fonction “split”, nous pouvons commencer par analyser le modèle de base sur un échantillon composé exactement de femmes.

```
resF <- lm(PROD_G_Frontal_Inf_Tri_1_L~PROD_G_Angular_2_L+PROD_G_Occipital_Lat_1_L
          +PROD_G_Rolandic_Oper_1_L+PROD_S_Sup_Temporal_4_L+PROD_G_Hippocampus_1_L+PROD_G_Frontal_Inf_Tri_1_R+
            +PROD_G_Angular_2_R+PROD_G_Occipital_Lat_1_R+PROD_G_Rolandic_Oper_1_R
          +PROD_S_Sup_Temporal_4_R+PROD_G_Hippocampus_1_R, data=dataF)
summary(resF)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.57089    0.12103   4.717 1.96e-05 ***
PROD_G_Angular_2_L         0.30731    0.14476   2.123 0.038743 *
PROD_G_Occipital_Lat_1_L  -0.04467    0.21588  -0.207 0.836927
PROD_G_Rolandic_Oper_1_L  -0.01248    0.19870  -0.063 0.950154
PROD_S_Sup_Temporal_4_L    0.51418    0.15577   3.301 0.001783 **
PROD_G_Hippocampus_1_L     0.56902    0.23527   2.419 0.019265 *
PROD_G_Frontal_Inf_Tri_1_R 0.64358    0.15553   4.138 0.000134 ***
PROD_G_Angular_2_R        -0.05352    0.18546  -0.289 0.774073
PROD_G_Occipital_Lat_1_R   0.21796    0.20777   1.049 0.299215
PROD_G_Rolandic_Oper_1_R   0.13927    0.20805   0.669 0.506320
PROD_S_Sup_Temporal_4_R   -0.22888    0.17441  -1.312 0.195402
PROD_G_Hippocampus_1_R    -0.56856    0.24702  -2.302 0.025557 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3162 on 50 degrees of freedom
Multiple R-squared:  0.6028,    Adjusted R-squared:  0.5154
F-statistic: 6.897 on 11 and 50 DF,  p-value: 6.482e-07
```

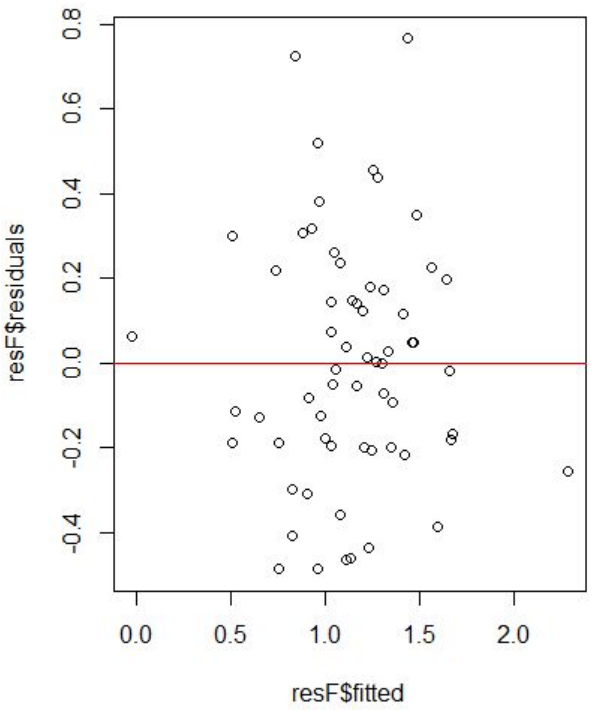
On constate que le test de Fisher indique bien que le modèle est utile.

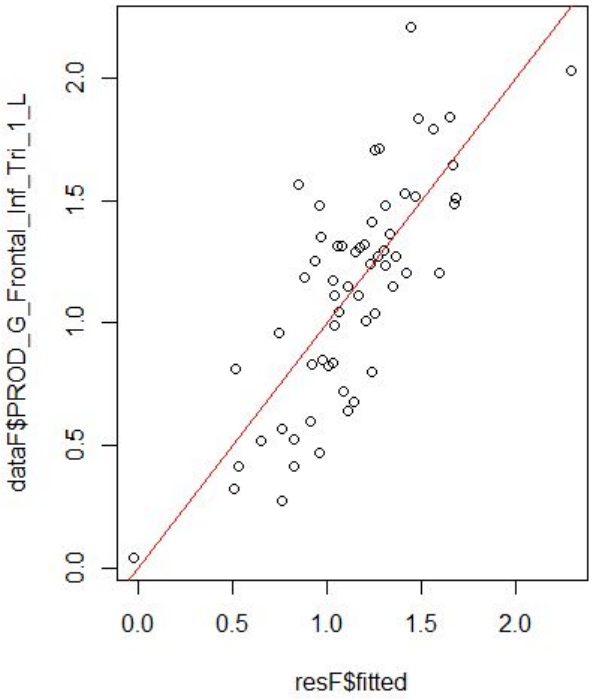
On observe un R^2 environ égal à 60% ce qui nous informe que 60% de la variabilité est expliqué par le modèle (R^2 ajusté environ égal à 52%).

Une analyse ligne par ligne de chaque variable, montre que certaines d’entre elles sont inutiles au modèle. Il apparaît ici que, potentiellement, le gyrus occipital latéral dans les deux hémisphères, l’opercule rolandique dans les deux hémisphères, le gyrus angulaire et le sillon supérieur temporal dans l’hémisphère droit sont inutiles au modèle.

```
        Shapiro-Wilk normality test

data:  resF$residuals
W = 0.97763, p-value = 0.3161
```

Les résidus suivent bien une loi normale.





Aucune structure n'apparaît dans les résidus.

b. Modèle M2' (AIC step ascendant)

Nous réalisons ici une sélection de type pas à pas, toujours sur la base du critère AIC, grâce à la fonction "step" pour avoir un compromis raisonnable entre la bonne adéquation du modèle aux données et la simplicité du modèle (avec le moins de variables possibles). Plus le critère AIC est petit, meilleur est le modèle. La fonction "step" évalue le modèle concurrent au modèle courant, les modèles concurrents étant eux privés d'une variable explicative (suppression successive ascendante ou descendante des variables inutiles, ascendant pour ce cas).

```
resF2 <- lm(PROD_G_Frontal_Inf_Tri_1_L~1, data=dataF)
resF2 <- step(resF2, as.formula(paste("~",paste(colnames(dataF)[-1],collapse="+"), sep="")), trace =TRUE)
```

Ci dessous l'étape 1 de l'AIC.

```
Start:  AIC=-96.85
PROD_G_Frontal_Inf_Tri_1_L ~ 1

                                    Df Sum of Sq     RSS      AIC
+ PROD_G_Frontal_Inf_Tri_1_R         1    4.8030  7.7848 -124.648
+ PROD_S_Sup_Temporal_4_L            1    1.7410 10.8468 -104.082
+ PROD_G_Angular_2_L                 1    1.5572 11.0307 -103.040
+ PROD_G_Occipital_Lat_1_R           1    1.5254 11.0625 -102.862
+ PROD_S_Sup_Temporal_4_R            1    1.1981 11.3898 -101.054
+ PROD_G_Rolandic_Oper_1_R           1    1.1893 11.3986 -101.006
+ PROD_G_Rolandic_Oper_1_L           1    0.8285 11.7593  -99.074
+ Index_Lateralisation_Hemispherique 1    0.7380 11.8498  -98.599
+ PROD_G_Hippocampus_1_L             1    0.7012 11.8866  -98.407
+ PROD_G_Occipital_Lat_1_L           1    0.6393 11.9486  -98.084
+ PROD_G_Angular_2_R                 1    0.6296 11.9583  -98.034
<none>                                             12.5878  -96.853
+ Volume_Cerebral                   1    0.1557 12.4322  -95.625
+ PROD_G_Hippocampus_1_R             1    0.0490 12.5389  -95.095
```

On peut comparer la première et la dernière étape de l'AIC pour constater la sélection de variables qui à eu lieu.

```
Step:  AIC=-137.8
PROD_G_Frontal_Inf_Tri_1_L ~ PROD_G_Frontal_Inf_Tri_1_R + PROD_S_Sup_Temporal_4_L +
    Index_Lateralisation_Hemispherique + PROD_G_Angular_2_L

                                    Df Sum of Sq    RSS     AIC
<none>                                           5.7160 -137.80
+ PROD_S_Sup_Temporal_4_R            1   0.13587 5.5801 -137.29
+ PROD_G_Hippocampus_1_R             1   0.10992 5.6061 -137.00
+ PROD_G_Occipital_Lat_1_R           1   0.08509 5.6309 -136.73
- PROD_G_Angular_2_L                 1   0.31130 6.0273 -136.51
+ PROD_G_Hippocampus_1_L             1   0.05518 5.6608 -136.40
+ PROD_G_Rolandic_Oper_1_R           1   0.05269 5.6633 -136.37
+ Volume_Cerebral                   1   0.03825 5.6777 -136.22
+ PROD_G_Angular_2_R                 1   0.03012 5.6858 -136.13
+ PROD_G_Rolandic_Oper_1_L           1   0.00877 5.7072 -135.90
+ PROD_G_Occipital_Lat_1_L           1   0.00178 5.7142 -135.82
- Index_Lateralisation_Hemispherique 1   0.71780 6.4338 -132.47
- PROD_S_Sup_Temporal_4_L            1   0.87066 6.5866 -131.01
- PROD_G_Frontal_Inf_Tri_1_R         1   3.11789 8.8339 -112.81
```

On peut maintenant passer à l'analyse du modèle.

```
resFStep <- lm(PROD_G_Frontal_Inf_Tri_1_L ~ PROD_G_Frontal_Inf_Tri_1_R + PROD_S_Sup_Temporal_4_L +
               Index_Lateralisation_Hemispherique + PROD_G_Angular_2_L,data=dataF)

summary(resFStep)
```

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.047835   0.249078  -0.192  0.84839
PROD_G_Frontal_Inf_Tri_1_R      0.692827   0.124252   5.576 7.04e-07 ***
PROD_S_Sup_Temporal_4_L         0.417500   0.141690   2.947  0.00465 **
Index_Lateralisation_Hemispherique 0.010332   0.003862   2.675  0.00973 **
PROD_G_Angular_2_L              0.168574   0.095677   1.762  0.08345 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3167 on 57 degrees of freedom
Multiple R-squared:  0.5459,    Adjusted R-squared:  0.514
F-statistic: 17.13 on 4 and 57 DF,  p-value: 2.802e-09
```

Pareillement, le test de Fisher est indique bien que ce modèle est utile. On rejette fortement H0.

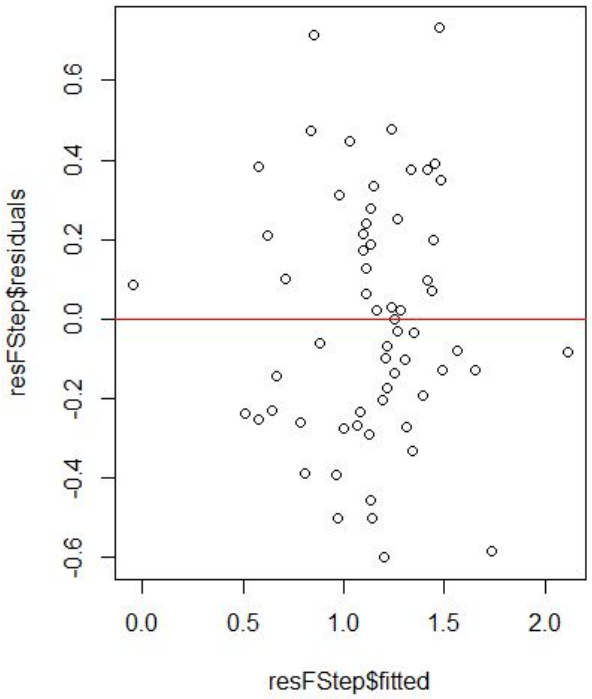
Le R^2 est environ égale à 55%, toujours pas énorme. Le R^2 ajusté est lui, est légèrement plus bas que celui du modèle M2 (0.514<0.5154). Notons qu'il y a moins de variables que dans le précédent modèle, ce qui pourrait expliquer cette baisse.

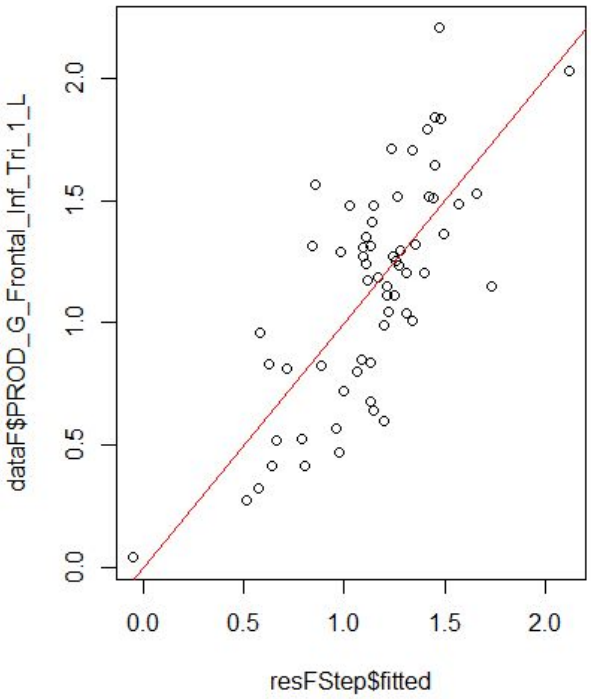
De plus, le RSD qui est légèrement supérieur à celui de M2 (0.3167>0.3162).

On peut observer qu'une variable, le gyrus angulaire dans l'hémisphère gauche, apparaît inutile au modèle (p-value>0.05).

```
        Shapiro-Wilk normality test

data:  resFStep$residuals
W = 0.98522, p-value = 0.6613
```

Les résidus suivent bien une loi normale d'après le test de Shapiro.





L'analyse du graphique des résidus ne montre à priori pas de structure.

En conclusion, le modèle M2' n'est pas franchement meilleur que le modèle M2 même s'il reste utile.

On a y = (-0.047835) + 0.692827 * PROD_G_Frontal_Inf_Tri_1_R + 0.417500 * PROD_S_Sup_Temporal_4_L + 0.010332 * Index_Lateralisation_Hemispherique + 0.168574 * PROD_G_Angular_2_L + ε où ε -> N(0,1).

3. Modèle spécifique aux hommes

a. Analyse du modèle M3

```
resH <- lm(PROD_G_Frontal_Inf_Tri_1_L~PROD_G_Angular_2_L+PROD_G_Occipital_Lat_1_L
           +PROD_G_Rolandic_Oper_1_L+PROD_S_Sup_Temporal_4_L+PROD_G_Hippocampus_1_L+PROD_G_Frontal_Inf_Tri_1_R+
             +PROD_G_Angular_2_R+PROD_G_Occipital_Lat_1_R+PROD_G_Rolandic_Oper_1_R
           +PROD_S_Sup_Temporal_4_R+PROD_G_Hippocampus_1_R, data=dataH)
summary(resH)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.56252    0.14275   3.941 0.000253 ***
PROD_G_Angular_2_L         0.24533    0.13885   1.767 0.083357 .
PROD_G_Occipital_Lat_1_L  -0.04110    0.28232  -0.146 0.884837
PROD_G_Rolandic_Oper_1_L   0.18483    0.19336   0.956 0.343724
PROD_S_Sup_Temporal_4_L    0.54773    0.16619   3.296 0.001810 **
PROD_G_Hippocampus_1_L     0.36062    0.26205   1.376 0.174907
PROD_G_Frontal_Inf_Tri_1_R 0.54465    0.17491   3.114 0.003054 **
PROD_G_Angular_2_R        -0.01084    0.19782  -0.055 0.956523
PROD_G_Occipital_Lat_1_R   0.43036    0.22041   1.953 0.056476 .
PROD_G_Rolandic_Oper_1_R  -0.21414    0.23108  -0.927 0.358531
PROD_S_Sup_Temporal_4_R   -0.41691    0.18926  -2.203 0.032239 *
PROD_G_Hippocampus_1_R    -0.17018    0.29107  -0.585 0.561398
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3625 on 50 degrees of freedom
Multiple R-squared:  0.5734,    Adjusted R-squared:  0.4796
F-statistic:  6.11 on 11 and 50 DF,  p-value: 3.125e-06
```

Le test de Fisher indique bien que ce modèle est utile, fort rejet de H0 qui suppose la nullité des βk.

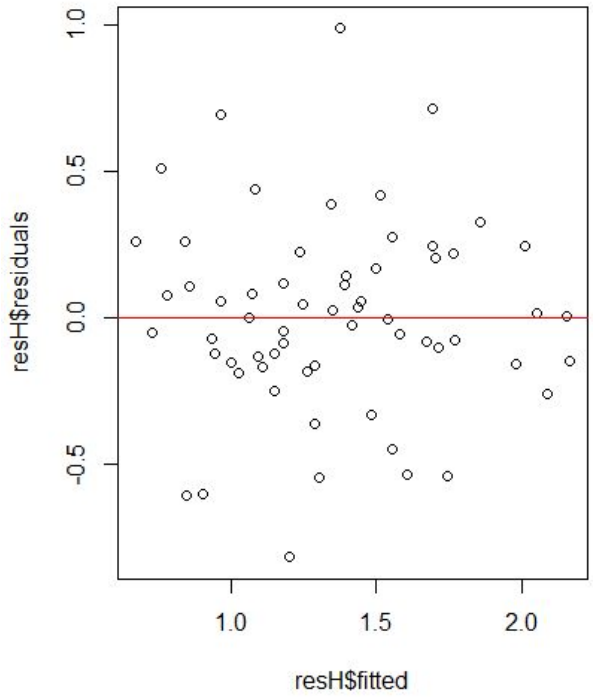
Les tests de Student pour chaque variables indiquent que certaines d'entre elles sont inutiles au modèle. 8 variables ont leur p-value supérieur à 5% ce qui laisse supposer qu'elles sont inutiles pour le modèle :

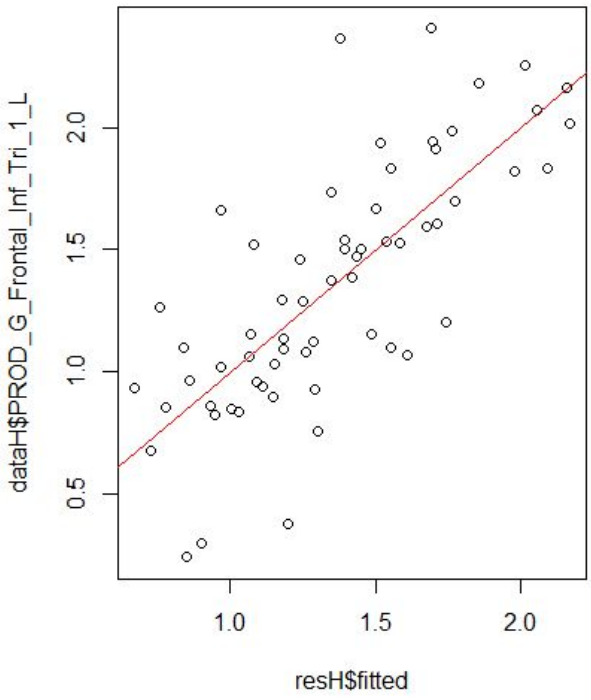
- gyrus angulaire dans l'hémisphère gauche
- gyrus occipital latéral dans l'hémisphère gauche
- opercule rolandique dans l'hémisphère gauche
- hippocampe dans l'hémisphère gauche
- gyrus angulaire dans l'hémisphère droit
- gyrus occipital latéral dans l'hémisphère droit
- opercule rolandique dans l'hémisphère droit
- hippocampe dans l'hémisphère droit

Le R^2 indique que ce modèle explique environ 57% de la variabilité avec un R^2 ajusté d'environ 48%.

```
        Shapiro-Wilk normality test

data:  resH$residuals
W = 0.97571, p-value = 0.2555
```

Les résidus ont bien distribués normalement (p-value>0.05).





Aucune structure n'apparaît dans le graphique des résidus.

b. Modèle M3' (AIC step descendant)

```
resHStep <- step(resH)
```

```
Start:  AIC=-115.16
PROD_G_Frontal_Inf_Tri_1_L ~ PROD_G_Angular_2_L + PROD_G_Occipital_Lat_1_L +
    PROD_G_Rolandic_Oper_1_L + PROD_S_Sup_Temporal_4_L + PROD_G_Hippocampus_1_L +
    PROD_G_Frontal_Inf_Tri_1_R + +PROD_G_Angular_2_R + PROD_G_Occipital_Lat_1_R +
    PROD_G_Rolandic_Oper_1_R + PROD_S_Sup_Temporal_4_R + PROD_G_Hippocampus_1_R

                          Df Sum of Sq    RSS     AIC
- PROD_G_Angular_2_R       1   0.00039 6.5706 -117.16
- PROD_G_Occipital_Lat_1_L 1   0.00278 6.5730 -117.14
- PROD_G_Hippocampus_1_R   1   0.04492 6.6151 -116.74
- PROD_G_Rolandic_Oper_1_R 1   0.11285 6.6831 -116.11
- PROD_G_Rolandic_Oper_1_L 1   0.12007 6.6903 -116.04
<none>                               6.5702 -115.16
- PROD_G_Hippocampus_1_L   1   0.24885 6.8191 -114.86
- PROD_G_Angular_2_L       1   0.41021 6.9804 -113.41
- PROD_G_Occipital_Lat_1_R 1   0.50100 7.0712 -112.61
- PROD_S_Sup_Temporal_4_R  1   0.63767 7.2079 -111.42
- PROD_G_Frontal_Inf_Tri_1_R 1 1.27405 7.8443 -106.18
- PROD_S_Sup_Temporal_4_L  1   1.42733 7.9976 -104.98
```

```
Step:  AIC=-123.05
PROD_G_Frontal_Inf_Tri_1_L ~ PROD_G_Angular_2_L + PROD_S_Sup_Temporal_4_L +
    PROD_G_Frontal_Inf_Tri_1_R + PROD_G_Occipital_Lat_1_R + PROD_S_Sup_Temporal_4_R

                            Df Sum of Sq    RSS     AIC
<none>                                    7.0205 -123.06
- PROD_S_Sup_Temporal_4_R    1   0.70599 7.7264 -119.11
- PROD_G_Angular_2_L         1   0.76122 7.7817 -118.67
- PROD_G_Occipital_Lat_1_R   1   1.12128 8.1417 -115.87
- PROD_G_Frontal_Inf_Tri_1_R 1   1.41027 8.4307 -113.71
- PROD_S_Sup_Temporal_4_L    1   1.90093 8.9214 -110.20
```

Ci dessus la première et la dernière étape du step permettant d'observer la sélection de variables utiles pour le modèle.

```
summary(resHStep)

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.5985     0.1116   5.362 1.61e-06 ***
PROD_G_Angular_2_L           0.2666     0.1082   2.464 0.016830 *
PROD_S_Sup_Temporal_4_L      0.5985     0.1537   3.894 0.000265 ***
PROD_G_Frontal_Inf_Tri_1_R   0.5193     0.1548   3.354 0.001434 **
PROD_G_Occipital_Lat_1_R     0.4494     0.1503   2.991 0.004132 **
PROD_S_Sup_Temporal_4_R     -0.4094     0.1725  -2.373 0.021098 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3541 on 56 degrees of freedom
Multiple R-squared:  0.5442,    Adjusted R-squared:  0.5035
F-statistic: 13.37 on 5 and 56 DF,  p-value: 1.393e-08
```

Le test de Fisher nous indique, d'après la p-value qui est très inférieur à 5%, qu'on rejette H0 et que ce modèle est donc bien utile.

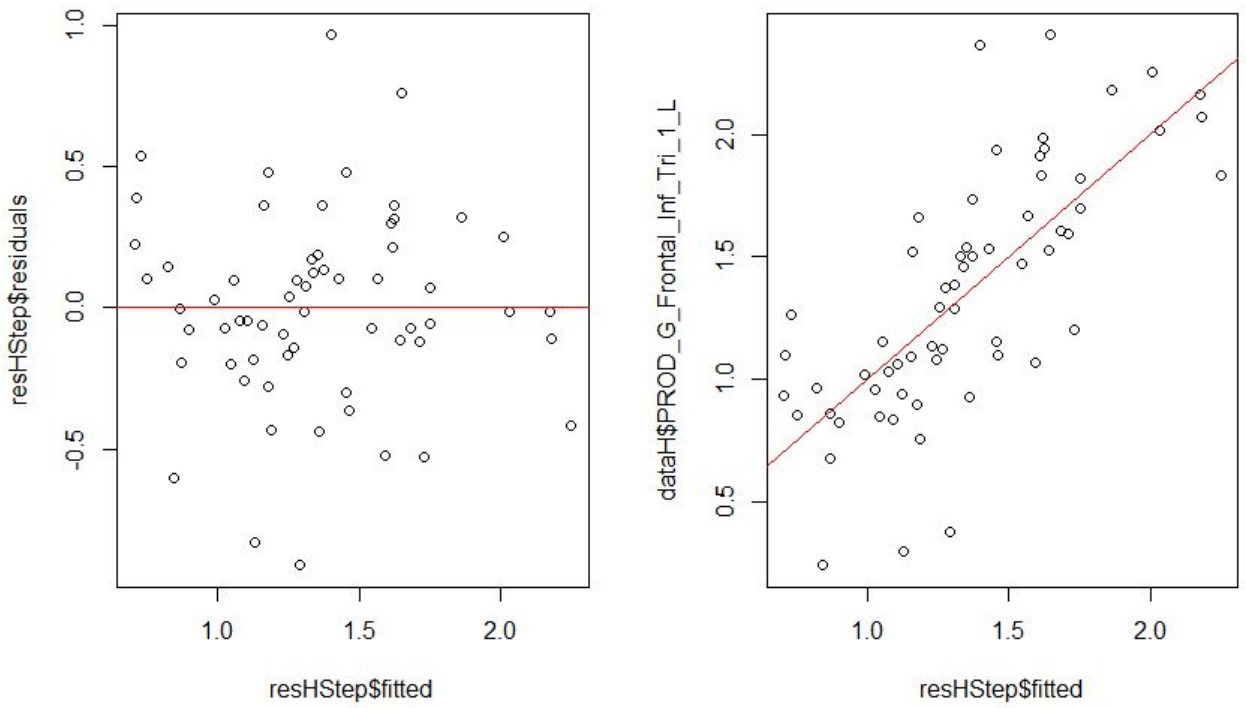
En outre, tous paramètres βk du modèle sont significativement différents de 0. Cela nous indique que toutes les variables sont utiles au modèle.

Le R^2, d'environ 54% nous indique que ce modèle explique 54% de la variabilité. Cependant, le R^2 ajusté est bien meilleur que celui du modèle M3 (0.5035>0.4796).

Le RSD est lui, plus petit que celui du modèle M3 (0.3541<0.3625).

```
        Shapiro-Wilk normality test

data:  resHStep$residuals
W = 0.98099, p-value = 0.4494
```

La p-value est très supérieure à 5% donc on accepte fortement H0. Les résidus suivent bien une loi normale.



Le graphique des résidus ne laisse apparaître aucune structure.

En conclusion, ce modèle est bien meilleur que le modèle M3.

On a y = 0.5985 + 0.2666 * PROD_G_Angular_2_L + 0.5985 * PROD_S_Sup_Temporal_4_L + 0.5193 * PROD_G_Frontal_Inf_Tri_1_R + 0.4494 * PROD_G_Occipital_Lat_1_R + (-0.4094) * PROD_S_Sup_Temporal_4_R + ε où ε -> N(0,1).

4. Tableau récapitulatif

Échantillon complet	Échantillon de femmes	Échantillon d'hommes
PROD_G_Angular_2_L PROD_S_Sup_Temporal_4_L PROD_G_Hippocampus_1_L PROD_G_Frontal_Inf_Tri_1_R PROD_G_Occipital_Lat_1_R PROD_S_Sup_Temporal_4_R PROD_G_Hippocampus_1_R	PROD_G_Frontal_Inf_Tri_1_R PROD_S_Sup_Temporal_4_L Index_Lateralisation_Hemisp PROD_G_Angular_2_L	PROD_G_Angular_2_L PROD_S_Sup_Temporal_4_L PROD_G_Frontal_Inf_Tri_1_R PROD_G_Occipital_Lat_1_R PROD_S_Sup_Temporal_4_R

On peut observer que le sexe a une influence sur la sélection des variables dans chaque modèle simplifié. Il est donc légitime de se demander l'interaction de ce facteur sur les données d'activation sur l'air de Broca. Nous allons donc procéder à une ANOVA à un facteur.

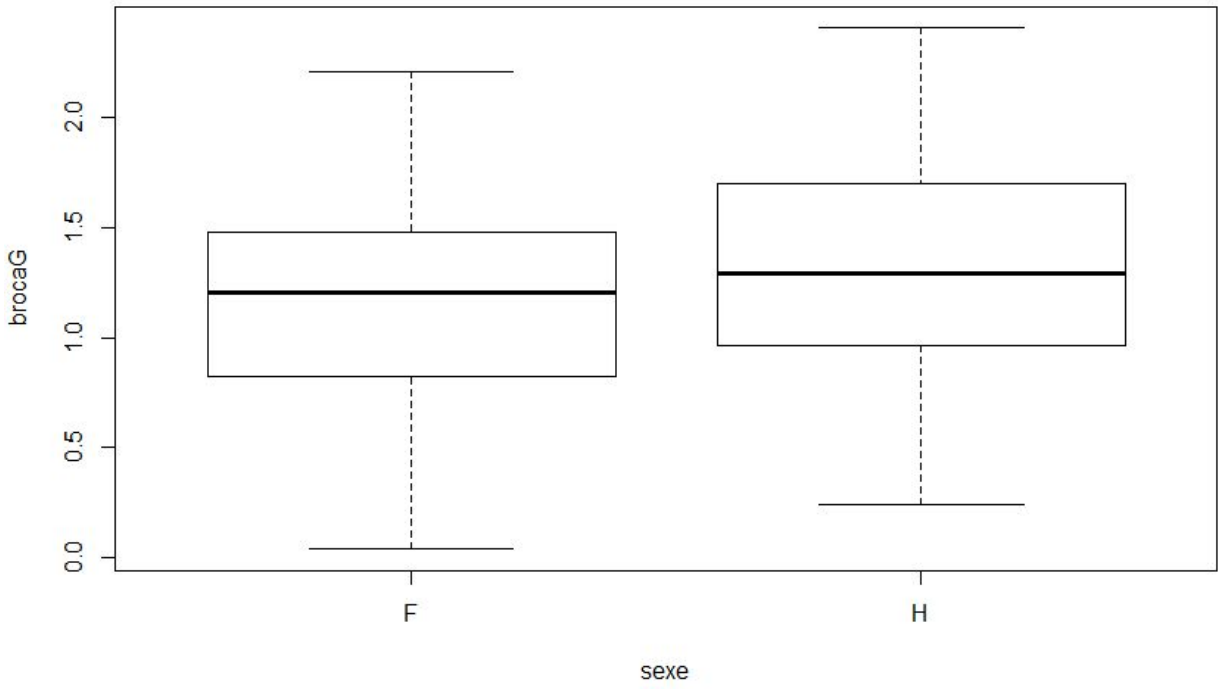
III. ANOVA

Analyse du modèle M4

On construit une ANOVA à partir des données du modèle M1'.

Nous voulions expliquer pourquoi les modèles sont différents en fonction du sexe. De plus, nous nous demandions si l'âge avait un impact sur l'activité de l'aire de Broca. C'est pourquoi nous avons réalisé une Anova à deux facteurs, elle permet de voir l'influence du sexe, de l'âge, et de leur interaction sur les variations du signal BOLD de l'aire de Broca.

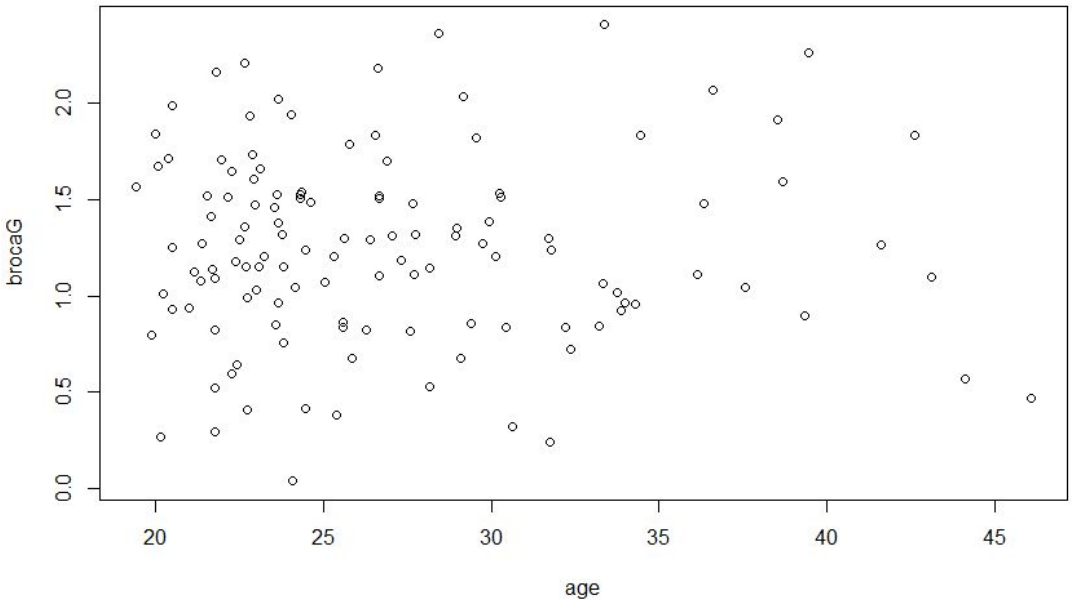
Les boxplots de l'activité de l'aire de Broca gauche en fonction du sexe semble indiquer qu'il y a une différence de moyenne et de variance entre les deux sexes.



Cela se confirme par un test de student pour vérifier l'égalité des moyennes et un test d'égalité des variances. Les variances et les moyennes des deux échantillons sont donc significativement différentes, bien que les moyennes soient proches.

```
        Welch Two Sample t-test

data:  h and f
t = 2.4911, df = 120.78, p-value = 0.01409
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04398245 0.38461755
sample estimates:
mean of x mean of y
 1.352758  1.138458

> var.test(f,h) #p-value < 0.5 ; variances significativement différentes

        F test to compare two variances

data:  f and h
F = 0.8173, num df = 61, denom df = 61, p-value = 0.433
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.492447 1.356440
sample estimates:
ratio of variances
         0.8172971
```

Le nuage de points représentant le signal BOLD de l'aire de Broca en fonction de l'âge des sujets ne montre pas de structure évidente.



Nous avons fait deux ANOVA, une avec l'interaction entre les deux facteurs et une sans.

Anova sans interaction :

```
             Df Sum Sq Mean Sq F value Pr(>F)
Sexe          1  1.424  1.4237   6.161 0.0144 *
Age           1  0.031  0.0314   0.136 0.7132
Residuals   121 27.958  0.2311
```

Anova avec interaction :

```
             Df Sum Sq Mean Sq F value Pr(>F)
Sexe          1  1.424  1.4237   6.125 0.0147 *
Age           1  0.031  0.0314   0.135 0.7140
Sexe:Age      1  0.067  0.0666   0.286 0.5935
Residuals   120 27.892  0.2324
```

L'hypothèse H0 supposant que la moyenne de l'activité de l'aire de Broca est la même pour les deux sexes est donc rejetée.

Les hypothèses H0, "la moyenne de l'activité de l'aire de broca est la même quelque soit l'âge" et "la moyenne de l'activité de l'aire de broca est la même quelque soit l'interaction entre l'âge et le sexe", sont validés.

Donc le sexe à bien un impact significatif sur le modèle de l'activité de l'aire de Broca contrairement à l'âge.

Pour vérifier les résultats de l'ANOVA, on fait une régression linéaire pour expliquer l'activité de l'aire de Broca en fonction du sexe. L'hypothèse H0 : "le coefficient de la variable sexe est nul" est rejetée au profit de H1 : non H0.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.13846    0.06083  18.715   <2e-16 ***
SexeH        0.21430    0.08603   2.491   0.0141 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.479 on 122 degrees of freedom
Multiple R-squared:  0.0484,    Adjusted R-squared:  0.0406
F-statistic: 6.205 on 1 and 122 DF,  p-value: 0.01408
```

Pour vérifier l'homogénéité du modèle, on vérifie l'égalité des variances et la normalité des résidus.

Le test de Fligner rejette l'homogénéité des variances, mais est proche de 5% donc on accepte les résultats.

```
        Fligner-Killeen test of homogeneity of variances

data:  data[, 5] and data[, 2]
Fligner-Killeen:med chi-squared = 4.4695, df = 1, p-value = 0.0345
```

Les résidus suivent bien une loi normale.

```
        Shapiro-Wilk normality test

data:  res.aov3$residuals
W = 0.99113, p-value = 0.6154
```

Conclusion

Nous gardons donc les modèles M1', M2 et M3' qui correspondent respectivement au modèle simplifié sur l'échantillon global, au modèle de base de l'échantillon des femmes et au modèle simplifié sur l'échantillon des hommes.

De plus l'ANOVA a bien montré que le sexe avait une influence significative sur l'activité BOLD de l'aire de broca, contrairement à l'âge. Ce qui explique que gardé des modèles différents en fonction des sexes est pertinent.