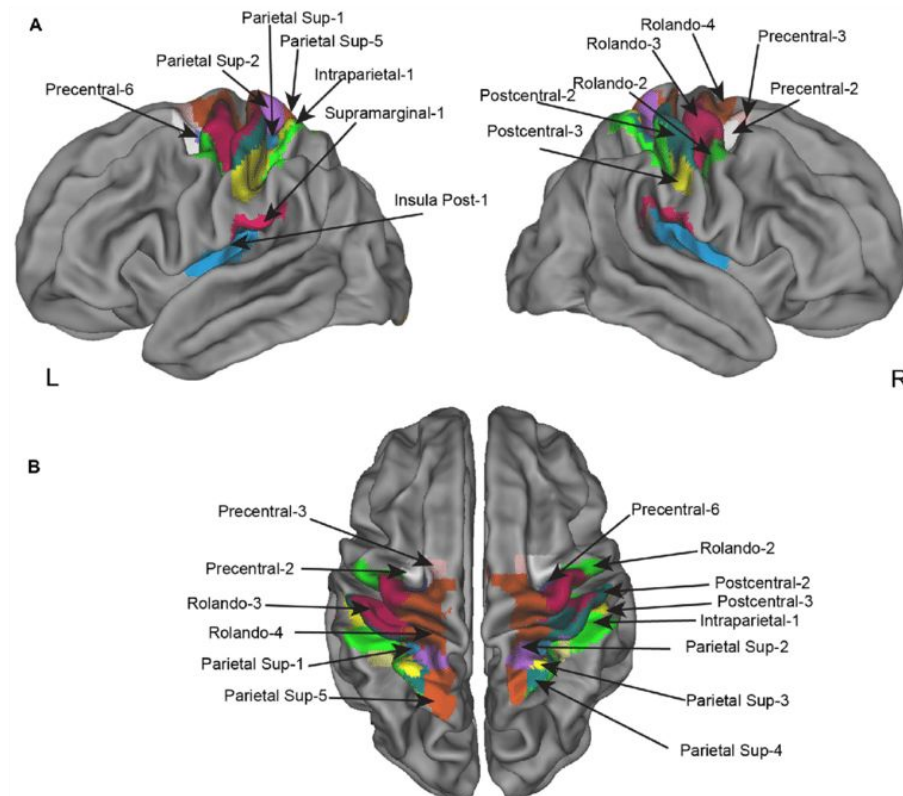


Rapport de projet

Modélisation statistique

2019-2020



Professeurs encadrants : Jérôme Saracco et Loïc Labache

Equipe projet : Lisa Bregier, Alizé Dona et Hugo Fournier

Problématique et dictionnaire des données	2
Contexte	2
Problématique	2
Dictionnaire des données	2
Analyse descriptive des données	3
Au sein des variables explicatives	3
Entre la variable à expliquer et les variables explicatives	4
ACP	4
Etude des variables qualitatives : Préférence manuelle et sexe	6
Modélisation linéaire multiple	6
Sur l'échantillon complet	6
Méthode - Step descendant	6
Méthode - Step ascendant	7
Méthode - drop1 (descendant)	7
Méthode - add1 (ascendant)	7
Etude Femme vs Homme	8
Approche computationnelle basée sur l'impact des perturbations aléatoires des covariables	9
Bilan sur les modèles	10
Modèle mixte	10
Modèles homme et femme	10
Comparaison des modèles	11

I. Problématique et dictionnaire des données

A. Contexte

Ce projet, réalisé en équipe de 3, s'inscrit dans l'UE Modélisation statistique et systèmes dynamiques et a pour objectif de nous faire utiliser les méthodes et outils de statistique vu en cours (statistique descriptive et régression linéaire multiple notamment). Ces outils sont utilisés dans une problématique de neurosciences cognitives détaillée dans la partie suivante.

B. Problématique

Nous travaillons sur le jeu de données *activation* détaillé dans la partie suivante. Il contient les données de variation du signal BOLD (blood-oxygen-level dependant) au cours d'une tâche de production de langage chez 249 sujets, qui correspondent à l'activation de différentes régions cérébrales; il provient de la base de données BIL&GIN.

Les tâches de langage associées demandent la diction d'une phrase simple lorsque les sujets voient apparaître une image à l'écran. Les activations ont été mesurées par IRMf dans 6 régions, définies à partir de l'atlas AICHA, dans les hémisphères gauche et droit.

La principale problématique de ce projet est d'identifier les réseaux du langage en analysant les aires reliées à l'aire de Broca gauche, aire connue du langage.

C. Dictionnaire des données

Nom de la variable (fichier de données)	Type de variable	Nom de la variable (scripts de modélisation linéaire)	Description
Sujet	Texte		Code d'anonymat du sujet.
Age	Quantitative	x1	Age du sujet en années.
Volume_Cerebral	Quantitative	x2	Volume occupé par le cerveau du sujet en centimètres cube.
Index_Lateralisation_Hemispherique	Quantitative	x3	Détermine l'hémisphère dominant pour le langage chez un sujet. Valeur positive \Rightarrow Hémisphère gauche dominant. Plus l'index est élevé, plus l'hémisphère est dominant.
Prod_G_Frontal_Inf_Tri_1_R	Quantitative	x4	Activation du gyrus frontal inférieur triangulaire ou aire de Broca (hémisphère droit).
Prod_G_Angular_2_R	Quantitative	x5	Activation du gyrus angulaire (hémisphère droit).
Prod_G_Occipital_Lat_1_R	Quantitative	x6	Activation du gyrus occipital latéral (hémisphère droit).
Prod_G_Rolandic_Oper_1_R	Quantitative	x7	Activation de l'opercule rolandique (hémisphère droit).
Prod_G_Hippocampus_1_R	Quantitative	x8	Activation de l'hippocampe (hémisphère droit).

Prod_S_Sup_Temporal_4_R	Quantitative	x9	Activation du sillon supérieur temporal, ou aire de Wernicke (hémisphère droit).
Prod_G_Angular_2_L	Quantitative	x10	Activation du gyrus angulaire (hémisphère gauche).
Prod_G_Occipital_Lat_1_L	Quantitative	x11	Activation du gyrus occipital latéral (hémisphère gauche).
Prod_G_Rolandic_Oper_1_L	Quantitative	x12	Activation de l'opercule rolandique (hémisphère gauche).
Prod_G_Hippocampus_1_L	Quantitative	x13	Activation de l'hippocampe (hémisphère gauche).
Prod_S_Sup_Temporal_4_L	Quantitative	x14	Activation du sillon supérieur temporal, ou aire de Wernicke (hémisphère gauche).
Prod_G_Frontal_Inf_Tri_1_L	Quantitative	y	Variable à expliquer : activation du gyrus frontal inférieur triangulaire ou aire de Broca (hémisphère droit).
Sexe	Qualitative (2 niveaux)		Sexe du sujet. H pour Homme et F pour Femme.
Preference_Manuelle	Qualitative (2 niveaux)		Préférence manuelle du sujet. G pour Gaucher et D pour Droitier.

II. Analyse descriptive des données

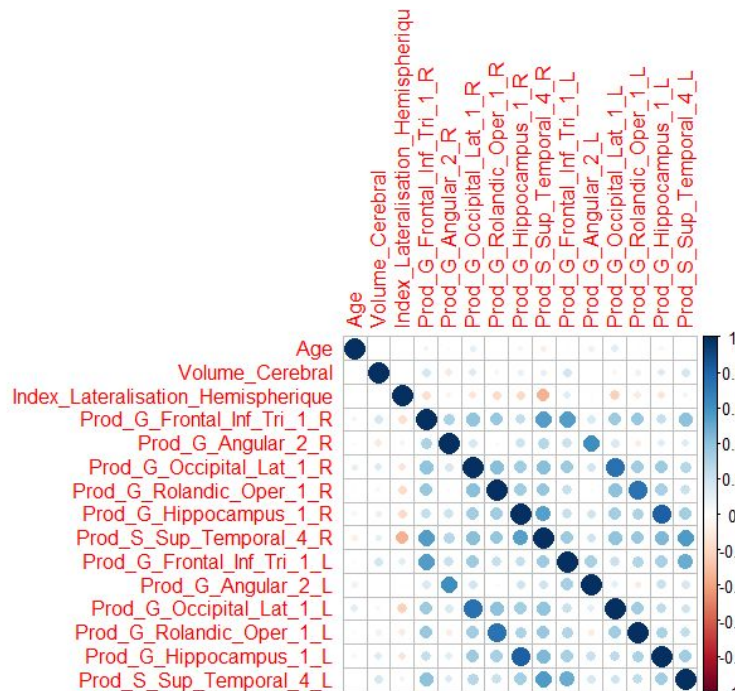
Nous avons commencé par regarder la corrélation entre les données.

A. Au sein des variables explicatives

Sur le graphique ci-dessous, on observe une forte corrélation entre Prod_G_Occipital_Lat_1_L et Prod_G_Occipital_Lat_1_R ; Prod_G_Rolandic_Oper_1_L et Prod_G_Rolandic_Oper_1_R ; Prod_G_Hippocampus_1_L et Prod_G_Hippocampus_1_R.

Ainsi, il semblerait que le gyrus Occipital Latéral, l'opercule rolandique et l'hippocampe fonctionnent de manière symétrique ou tout du moins fonctionnent ensemble lors d'une tâche langagière. Concernant leur rôle dans le langage, le gyrus occipital latéral est une zone impliquée dans l'audition, le langage et la mémoire et la vision des formes complexes. Quant à l'opercule rolandique, une lésion de cette partie du cerveau a pour conséquence des troubles de la prononciation. Enfin l'hippocampe est impliqué dans les processus de mémoire et de navigation spatiale.

Néanmoins, même s'il existe une corrélation entre les parties droites et gauches du cerveau, leur rôle n'est pas tout à fait identique : le langage est une fonction latéralisée du cerveau.



B. Entre la variable à expliquer et les variables explicatives

Il semblerait que Prod_G_Frontal_Inf_Tri_1_L soit corrélé avec Prod_G_Frontal_Inf_Tri_1_R (x4) et Prod_S_Sup_Temporal_4_L (x14). Si on calcule la corrélation entre ces variables on trouve que :

```
> cor(donnees$Prod_G_Frontal_Inf_Tri_1_L, donnees$Prod_G_Frontal_Inf_Tri_1_R)
[1] 0.5528673
> cor(donnees$Prod_G_Frontal_Inf_Tri_1_L, donnees$Prod_S_Sup_Temporal_4_L)
[1] 0.4950188
```

Par conséquent, il y a de fortes chances pour qu'on retrouve ces deux variables dans le modèle linéaire de la variable à expliquer. Concernant leur rôle dans le fonctionnement cérébral, l'aire de Wernicke est impliquée dans la compréhension du langage (oral comme écrit). L'aire de Broca d'une manière générale semble être responsable de la production de mots parlés, ou tout du moins de la formulation d'un code articulatoire mise en oeuvre ensuite par le cortex moteur.

C. ACP

```
> library(PCAmixdata)
> resACP <- PCAmix(donnees_cop, graph=FALSE)
> round(resACP$eig, digit=2)
```

	Eigenvalue	Proportion	Cumulative
dim 1	4.67	31.10	31.10
dim 2	1.86	12.37	43.48
dim 3	1.38	9.20	52.67
dim 4	1.30	8.64	61.31
dim 5	1.09	7.28	68.59
dim 6	1.00	6.65	75.24
dim 7	0.83	5.56	80.79
dim 8	0.75	4.98	85.77
dim 9	0.53	3.54	89.31
dim 10	0.45	3.02	92.33
dim 11	0.28	1.90	94.22
dim 12	0.27	1.80	96.02
dim 13	0.24	1.63	97.65
dim 14	0.21	1.42	99.07
dim 15	0.14	0.93	100.00

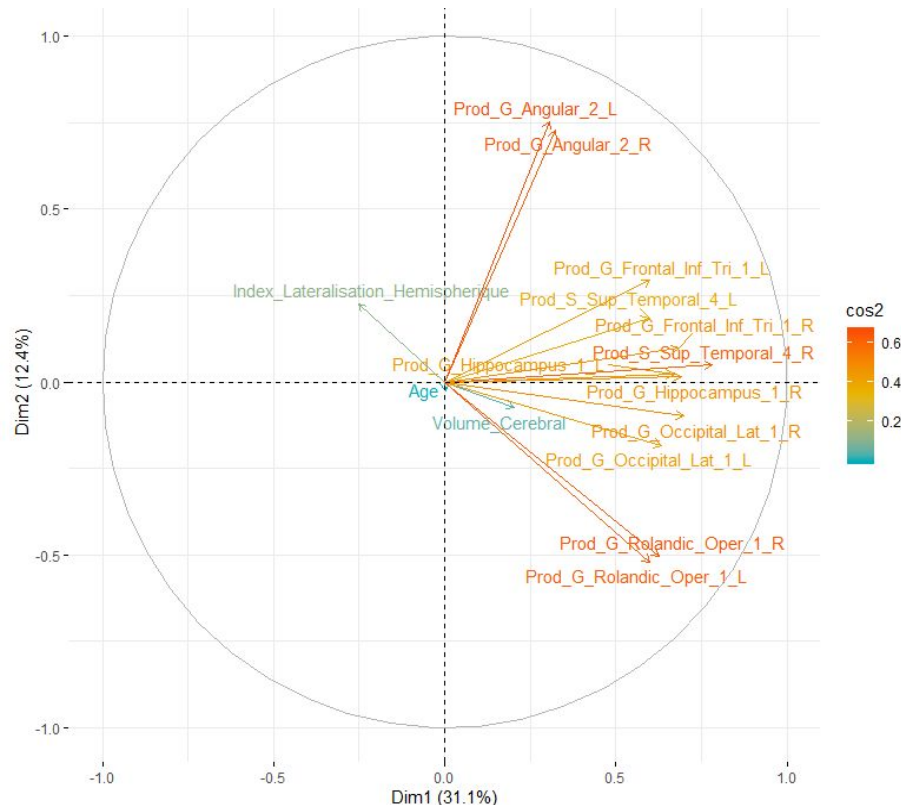
Les données ci-contre indiquent le pourcentage de variance expliquée par chacun des axes : c'est la valeur propre (aussi nommée "Eigenvalue"). Les axes à conserver sont ceux dont les échantillons présentent la plus grande variation. Ainsi, selon le critère de Kaiser (1961), les axes à conserver sont ceux déterminés par une valeur propre strictement supérieure à 1. Les 5 premiers axes expliquant 68,59% de la variance globale, on peut donc réduire l'étude aux 5 premières composantes principales.

On observe ici que 43,48% de la variance globale est expliquée par les deux premiers axes. Cela n'est pas optimal, cependant la représentation d'un graphique sur un plan à deux dimensions est la manière la plus compréhensible d'analyser des variables. On trace donc le cercle de corrélation et le graphique des individus sur le plan factoriel 1-2.

```
> round(res$quantif$contrib.pct,digit=2)
```

	dim 1	dim 2	dim 3	dim 4	dim 5
Age	0.00	0.04	0.43	22.27	31.05
Volume_Cerebral	0.89	0.30	14.53	0.42	6.56
Index_Lateralisation_Hemispherique	1.36	2.74	26.64	1.65	5.41
Prod_G_Frontal_Inf_Tri_1_R	10.09	0.51	3.04	2.41	16.19
Prod_G_Angular_2_R	2.25	28.45	7.02	0.89	3.65
Prod_G_Occipital_Lat_1_R	10.45	0.53	1.96	11.23	2.64
Prod_G_Rolandic_Oper_1_R	8.42	13.64	0.01	0.83	1.34
Prod_G_Hippocampus_1_R	10.20	0.01	2.86	19.87	9.91
Prod_S_Sup_Temporal_4_R	13.12	0.14	0.37	2.74	4.75
Prod_G_Frontal_Inf_Tri_1_L	7.70	4.63	17.75	2.31	0.10
Prod_G_Angular_2_L	2.03	30.62	1.94	1.13	1.25
Prod_G_Occipital_Lat_1_L	8.59	1.82	7.86	13.35	0.92
Prod_G_Rolandic_Oper_1_L	7.67	14.70	1.50	0.64	1.31
Prod_G_Hippocampus_1_L	9.60	0.03	0.37	19.43	14.89
Prod_S_Sup_Temporal_4_L	7.64	1.85	13.73	0.84	0.03

La première composante concentre notamment la variable "Prod_S_Sup_Temporal_4_R" à 13,12% soit la partie droite du sillon supérieur temporal. La deuxième composante quant à elle concentre à 30,62% la variable "Prod_G_Angular_2_L" soit la partie gauche du gyrus angulaire.



Cercle de corrélation sur la plan 1-2 et fidélité de la représentation

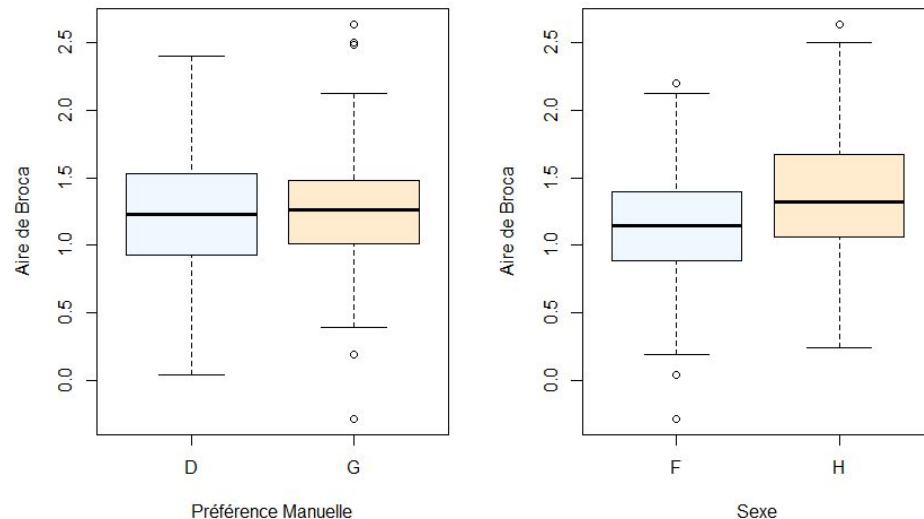
On observe sur le cercle de corrélation que les variables "Prod_G_Angular_2_L" et "Prod_G_Angular_2_R" sont fortement corrélées ce qui indique que les aires du gyrus angulaire droite et gauche s'activent de manière simultanée.

Les variables "Prod_G_Rolandic_Oper_1_L" et "Prod_G_Rolandic_Oper_1_R" sont elles aussi corrélées entre elles et sont non corrélées aux variables "Prod_G_Angular_2_L" et "Prod_G_Angular_2_R". Les parties droite et gauche de l'opercule rolandique fonctionnent aussi de manière simultanée mais leur activité n'est pas liée à l'activité des aires du gyrus angulaire.

Même si leur représentation sur le plan 1-2 n'est pas optimale, on remarque que les variables "Prod_G_Frontal_Inf_Tri_1_L", "Prod_G_Frontal_Inf_Tri_1_R", "Prod_S_Sup_Temporal_4_L", "Prod_S_Sup_Temporal_4_R", "Prod_G_Hippocampus_1_R", "Prod_G_Hippocampus_1_L", "Prod_G_Occipital_Lat_1_R" et "Prod_G_Occipital_Lat_1_L" semblent légèrement corrélées. C'est à dire que le gyrus frontal inférieur triangulaire (ou aire de Broca), le sillon supérieur temporal (ou aire de Wernicke), le gyrus Occipital Latéral et l'hippocampe ont des activations assez proches.

Ce plan 1-2 pour le cercle de corrélation ne permet pas de tirer des conclusions sur la corrélation des variables avec l'âge ou le volume cérébral. En effet, la qualité de leur représentation est moindre sur ce plan tel que l'indique le facteur cos2.

D. Etude des variables qualitatives : Préférence manuelle et sexe



Ces boxplots semblent montrer qu'il n'y a pas de lien entre la préférence manuelle et les fluctuations observées dans l'aire de Broca. Par contre, on peut se demander si le sexe n'a pas un impact sur ces fluctuations. Pour voir si les différences sont significatives, on réalise un test de Student, avec comme hypothèse H_0 : la différence est significative et H_1 : non H_0 .

Pour la préférence manuelle, $\alpha = 0.4332 > 5\%$ donc on rejette H_0 : il n'y a pas de différence significative.

welch Two Sample t-test

```
data: PM_droite$Prod_G_Frontal_Inf_Tri_1_R and PM_gauche$Prod_G_Frontal_Inf_Tri_1_R
t = -0.78499, df = 242.32, p-value = 0.4332
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.13221655 0.05686544
sample estimates:
mean of x mean of y
0.4916446 0.5293202
```

Pour le sexe, on obtient $\alpha = 0.01007 < 5\%$, on accepte H_0 : la différence est significative. Ainsi, il peut être intéressant d'étudier les sexes séparément dans le modèle de régression linéaire multiple.

welch Two Sample t-test

```
data: Sexe_F$Prod_G_Frontal_Inf_Tri_1_R and Sexe_H$Prod_G_Frontal_Inf_Tri_1_R
t = -2.5936, df = 246.21, p-value = 0.01007
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.21547982 -0.02946416
sample estimates:
mean of x mean of y
0.4476764 0.5701484
```

III. Modélisation linéaire multiple

A. Sur l'échantillon complet

Nous avons utilisé les 4 différentes méthodes afin de trouver le modèle linéaire multiple le plus pertinent selon le critère AIC.

Méthode - Step descendant

Avec cette méthode, le modèle trouvé est $y \sim x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_{10} + x_{12} + x_{13} + x_{14}$

```
Call:
lm(formula = y ~ x3 + x4 + x5 + x6 + x7 + x8 + x10 + x12 + x13 +
    x14)

Coefficients:
(Intercept)          x3          x4          x5          x6          x7          x8          x10
  0.317845    0.006128    0.500219   -0.109618    0.134727   -0.149793   -0.274940    0.262637
      x12      x13      x14
  0.194016   0.364132   0.280878
```

Méthode - Step ascendant

Avec cette méthode, le modèle trouvé est $y \sim x4 + x14 + x10 + x3 + x13 + x8 + x12 + x6 + x7 + x5$

```
Call:
lm(formula = y ~ x4 + x14 + x10 + x3 + x13 + x8 + x12 + x6 +
    x7 + x5, data = donnees)

Coefficients:
(Intercept)          x4          x14          x10          x3          x13          x8          x12
  0.317845    0.500219    0.280878    0.262637    0.006128    0.364132   -0.274940    0.194016
      x6      x7      x5
  0.134727  -0.149793  -0.109618
```

Méthode - drop1 (descendant)

On part du modèle initial : $y \sim x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14$

Valeur de l'AIC	Valeur qui minimise	Modèle résultant
-548.11	x1 : -549.96	$x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14$
-549.96	x2 : -551.60	$x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14$
-551.60	x11 : -552.60	$x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x12 + x13 + x14$
-552.60	x9 : -553.06	$x3 + x4 + x5 + x6 + x7 + x8 + x10 + x12 + x13 + x14$
-553.06	Aucune	Modèle final : $x3 + x4 + x5 + x6 + x7 + x8 + x10 + x12 + x13 + x14$

Méthode - add1 (ascendant)

On part du modèle initial : $y \sim 1$

Valeur de l'AIC	Valeur qui minimise	Modèle résultant
-391.61	x4 : -480.44	$x4$
-480.44	x14 : -512.48	$x14 + x4$
-512.48	x10 : -528.84	$x14 + x4 + x10$
-528.84	x3 : -540.16	$x14 + x4 + x10 + x3$

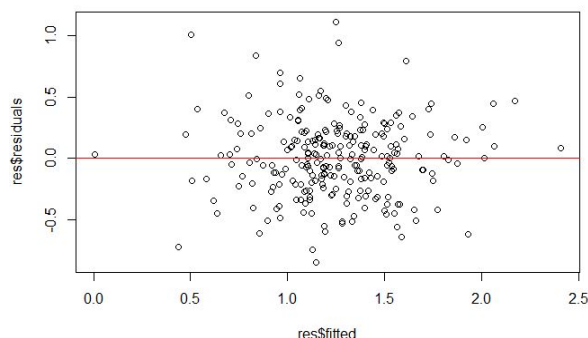
-540.16	x13 : -545.53	x14 +x4 +x10 + x3 + x13
-545.53	x8 : -550.00	x14 +x4 +x10 + x3 + x13 + x8
-550.00	x12 : -552.23	x14 +x4 +x10 + x3 + x13 + x8 + x12
-552.23	x6 : -552.64	x14 +x4 +x10 + x3 + x13 + x8 + x12 + x6
-552.64	x7 : -552.94	x14 +x4 +x10 + x3 + x13 + x8 + x12 + x6 + x7
-552.94	x5 : -553.06	x14 +x4 +x10 + x3 + x13 + x8 + x12 + x6 + x7 + x5
-553.06	Aucune	Modele final : x14+x4+x10+x3+x13+x8+x12 + x6 + x7 + x5

Ainsi, avec ces quatre méthodes on trouve le même modèle, ce qui est cohérent sachant que ces méthodes utilisent le même critère. Ce modèle est donc :

$$y = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{10} x_{10} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14}$$

Avec :

intercept = 0.317845 $\beta_3 = 0.006128$ $\beta_4 = 0.500219$ $\beta_5 = -0.109618$
 $\beta_6 = 0.134727$ $\beta_7 = -0.149793$ $\beta_8 = -0.274940$ $\beta_{10} = 0.262637$
 $\beta_{12} = 0.194016$ $\beta_{13} = 0.364132$ $\beta_{14} = 0.280878$



```
> shapiro.test(res$residuals)

Shapiro-wilk normality test

data:  res$residuals
W = 0.99054, p-value = 0.1061
```

En traçant le graphique des résidus, on n'aperçoit aucune structure. La normalité des résidus de ce modèle est vérifiée par le test de Shapiro-Wilkinson, la p-value est ici supérieure à 5% donc on accepte H0: les résidus suivent une loi normale.

B. Etude Femme vs Homme

Nous avons souhaité voir si l'on pouvait observer des différences de modèle selon le sexe du sujet. En ce qui concerne les notations, x3F et x3H correspondent respectivement à la variable x3 dans le groupe des femmes et x3 dans le groupe des hommes (idem pour toutes les autres variables).

En ce qui concerne les femmes, le modèle ayant le AIC le plus faible est :

$$y^F \sim x3^F + x4^F + x8^F + x10^F + x13^F + x14^F$$

avec :

intercept : 0.233885 $\beta_3 = 0.006871$ $\beta_4 = 0.520303$
 $\beta_8 = -0.316215$ $\beta_{10} = 0.216773$ $\beta_{13} = 0.372623$
 $\beta_{14} = 0.271208$

Les variables x5, x6, x7 et x12 n'apparaissent donc plus comme pertinentes du point de vue du critère AIC par rapport au modèle mixte.

```
> shapiro.test(res$residuals)

Shapiro-wilk normality test

data:  res$residuals
W = 0.99128, p-value = 0.6356
```

En utilisant le test de Shapiro-Wilkinson sur les résidus, on retrouve une p-value supérieure à 5%, les résidus suivent donc bien une loi normale.

Pour les hommes, le modèle le plus pertinent selon le critère AIC est :

$$y_H \sim x_{3H} + x_{4H} + x_{6H} + x_{7H} + x_{9H} + x_{10H} + x_{12H} + x_{13H} + x_{14H}$$

```
> shapiro.test(res$residuals)
```

shapiro-wilk normality test

```
data:  res$residuals  
W = 0.9915, p-value = 0.6404
```

En utilisant le test de Shapiro-Wilkinson sur les résidus, on retrouve une p-value supérieure à 5%, les résidus suivent donc bien une loi normale.

C. Approche computationnelle basée sur l'impact des perturbations aléatoires des covariables

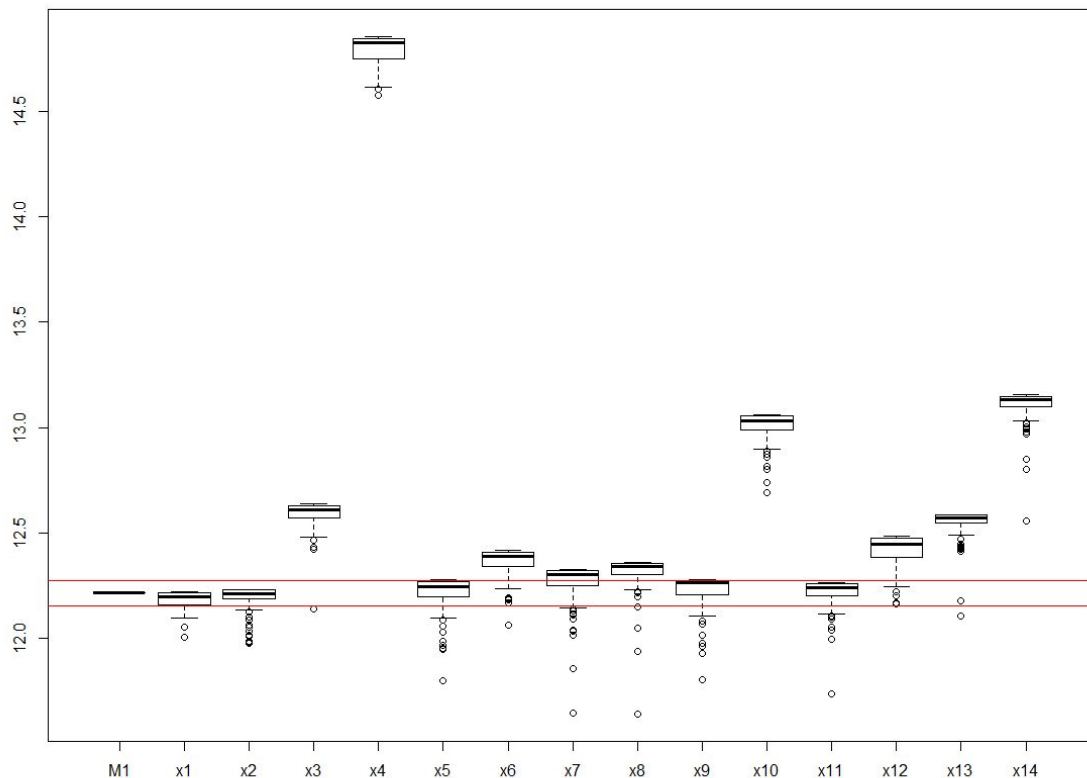
Une autre méthode pour sélectionner les variables du modèle de régression linéaire multiple consiste à perturber chacune des variables explicatives puis de comparer leur critère de qualité MSE à celui du modèle complet. Le MSE est calculé ainsi, avec ici $K=100$:

$$MSE = \sum_{k=0}^K (y - y_{estimateur})^2$$

Les variables explicatives à garder sont celles dont le boxplot correspondant est suffisamment éloigné du MSE du modèle complet (MSE_M1) car cela veut dire que des perturbations sur cette variable affectent significativement la qualité du modèle.

Ainsi, on peut établir un critère (C1) exigeant que la valeur médiane des 100 MSE correspondants à la variable soit comprise entre $MSE \pm 0.5\%$.

Boxplots des MSE obtenus par permutations des différentes variables explicatives



D'après ce critère (C1) on peut enlever du modèle les variables x_1 , x_2 , x_5 , x_9 et x_{11} . Ainsi le modèle mixte pourrait se simplifier en :

$$y = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{10} x_{10} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14}$$

```
> shapiro.test(res$residuals)
shapiro-wilk normality test
data: res$residuals
W = 0.98785, p-value = 0.03363
```

Mais un test de Shapiro-Wilkinson sur les résidus de ce modèle donne une p-value de 3,34% < 5%; on peut donc légèrement rejeter l'hypothèse H_0 correspondant à la normalité des résidus. Ceci empêche de bien appliquer un modèle de régression linéaire, ainsi nous préférons le modèle utilisant x_5 .

IV. Bilan sur les modèles

A. Modèle mixte

Avec l'approche computationnelle on retient un modèle proche M3 mais qui n'utilise pas x_5 , qui correspond à l'activation du gyrus angulaire droit et qui est non significative selon le test de Student; avec un R carré ajusté de 49,5% :

$$y = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{10} x_{10} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14}$$

Par le critère AIC on retient le modèle M2, avec un R carré ajusté de 49,7%.

Ainsi, l'hémisphère dominant (x_3) a bien un impact sur le fonctionnement de l'aire de Broca à gauche, tout comme la partie droite de l'aire de Broca (x_4). Le gyrus angulaire gauche (x_{10}), le gyrus occipital latéral droit (x_6 et x_{11}), l'opercule rolandique gauche et droit (x_7 et x_{12}), l'hippocampe gauche et droit (x_8 et x_{13}) et l'aire de Wernicke gauche (x_{14}) expliquent aussi les variations observées, comme pressenti avec l'analyse descriptive.

Il est intéressant de voir que l'âge (x_1), le volume cérébral (x_2) et l'aire de Wernicke droit (x_9) n'ont pas un impact significatif sur l'aire de Broca gauche (y).

B. Modèles homme et femme

Pour les hommes, on retient ces variables explicatives pour un modèle MH :

$$y_H \sim x_{3H} + x_{4H} + x_{6H} + x_{7H} + x_{9H} + x_{10H} + x_{12H} + x_{13H} + x_{14H}$$

Les variables x_5 (gyrus angulaire droit) et x_8 (hippocampe droit) ne font plus partie du modèle alors que d'un autre côté, la variable x_9 (aire de Wernicke droit) est considérée pertinente.

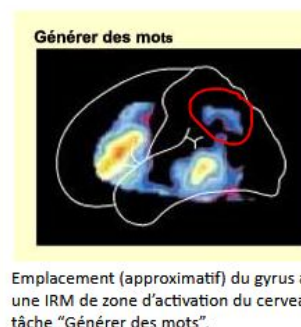
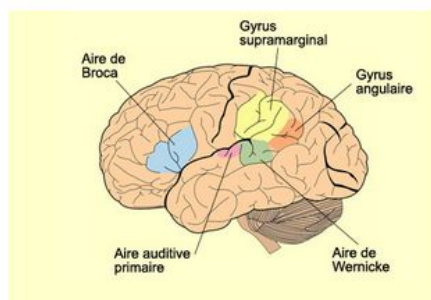
Pour les femmes, on retient ces variables pour un modèle MF:

$$y_F \sim x_{3F} + x_{4F} + x_{8F} + x_{10F} + x_{13F} + x_{14F}$$

Par rapport au modèle mixte, ce sont les variables x_5 (gyrus angulaire droit), x_6 (gyrus occipital latéral droit), x_7 (opercule rolandique droite) et x_{12} (opercule rolandique gauche) qui ne sont plus utilisées.

Bien que d'un point de vue statistique il y a des différences de modèle, on ne peut pas affirmer que l'aire de Broca fonctionne différemment selon le sexe. Il faudrait avoir plus de données et tester plusieurs aspects du langage pour vraiment l'affirmer.

La variable x_5 , l'activation du gyrus angulaire droit, est assez étonnante, elle apparaît dans le modèle mixte M2 mais pas dans le modèle homme ni dans le modèle femme. Le gyrus angulaire est impliqué dans le traitement sémantique des mots. Cette zone s'active donc lors de la génération de mots, tâche effectuée lors d'une production langagière.



C. Comparaison des modèles

Pour les tests paramétriques nous utiliserons un risque $\alpha = 5\%$. Attention, toutes les comparaisons ne sont pas à effectuer sachant que MH et MF utilisent un jeu de données plus petit et différent.

	Modèle complet M1	M2	M3 (résidus non normaux)	MH	MF
Test de Fischer avec H_0 : tous les $\beta_k = 0$	p-value < $2,2e10^{-16}$ H0 rejeté	p-value < $2,2e10^{-16}$ H0 rejeté	p-value < $2,2e10^{-16}$ H0 rejeté	p-value = $9,269e10^{-16}$ H0 rejeté	p-value = $3,562e10^{-14}$ H0 rejeté
R^2	52,34%	51,75%	51,33%	53,46%	47,14%
R^2 ajusté	49,48%	49,72%	49,5%	49,85%	44,41%
AIC	-548.11	-553,06	-552.94	-271.8	-278.63
Coefficients du modèle	β_0 : 2.269E-01; β_1 : 1.310E-03; β_2 : 8.739E-05; β_3 : 5.132E-03; β_4 : 5.223E-01; β_5 : -8.674E-02; β_6 : 2.012E-01; β_7 : -1.397E-01; β_8 : -2078E-01; β_9 : -1.033E-01; β_{10} : 2.568E-01; β_{11} : -1.12E-01; β_{12} : 2.075E-01; β_{13} : 3.243E-01; β_{14} : 3.138E-01;	β_0 : 0.317845; β_3 : 0.006128; β_4 : 0.500219; β_5 : -0.109618; β_6 : 0.134727; β_7 : -0.149793; β_8 : -0.274940; β_{10} : 0.262637; β_{12} : 0.194016; β_{13} : 0.364132; β_{14} : 0.280878;	β_0 : 0.339059; β_3 : 0.285023; β_4 : 0.210585; β_6 : 0.132807; β_7 : -0.141122; β_8 : -0.304854; β_{10} : 0.210585; β_{12} : 0.205312; β_{13} : 0.388764; β_{14} : 0.285023;	β_0 : 0.53376; β_3 : 0.003718; β_4 : 0.47936; β_6 : 0.247758; β_7 : -0.240696; β_9 : -0.307259; β_{10} : 0.191229; β_{12} : 0.333196; β_{13} : 0.226718; β_{14} : 0.397904;	β_0 : 0.233885; β_3 : 0.006871; β_4 : 0.520303; β_8 : -0.316215; β_{10} : 0.216773; β_{13} : 0.372623; β_{14} : 0.271208;;
Test de Student avec H_0 : $\beta_k = 0$ H0 rejeté H0 non rejeté	x0, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14 H0 rejeté H0 non rejeté	x0, x3, x4, x5, x6, x7, x8, x10, x12, x13, x14	x0, x14, x10, x12, x4, x3, x13, x8, x6, x7	x0, x3H, x4H, x6H, x7H, x9H, x10H, x12H, x13H, x14H	x0, x3F, x4F, x8F, x10F, x13F, x14F

Ainsi, nous n'avons pas de critère statistique déterminant pour choisir le modèle. Le modèle à utiliser dépendra d'autres critères à réfléchir avec les experts de la problématique, dont le coût de récoltes des données et la signification neurobiologique.