Projet de modélisation statistique - *Novembre 2023*

Consignes

- Le projet doit être fait avec le logiciel R.
- Le projet doit être fait **en groupe de 3 ou 4 élèves**.
- La qualité de la rédaction sera prise en compte dans la notation. Pensez à bien commenter votre démarche statistique.
- Le rapport doit être envoyé au **format PDF uniquement** à `jerome.saracco@ensc.fr`.

  Le nom du fichier PDF doit avoir la syntaxe suivante :

  `NOM1-NOM2-NOM3-NOM4-projet-stat-2A-2023-2024.pdf`
- Le rapport ne devra pas excéder 12 pages (hors courtes annexes éventuelles).
- Le rapport devra être envoyé au plus tard le **mercredi 20 décembre 2023 à 18h00**.

Jeu de données à traiter : données d'activation durant des tâches de langage

Brève description du jeu de données et de la problématique.

Les principales informations utiles concernant ce jeu de données et la problématique associées (dans le cadre de ce projet de modélisation statistique) sont indiquées ci-dessous .

Le jeu de données `activation.Rdata`[1] contient les données d'activations (variation du signal BOLD : blood-oxygen-level dependent) au cours d'une tâche de production langagière chez 124 sujets. Les données proviennent de la base de données BIL&GIN[2].

La tâches de production consiste à produire une phrase simple (sujet, verbe, complément) lorsque les sujets voient apparaître une image à l'écran.

Les activations au cours de cette tâche (variation du signal BOLD) ont été récupérées dans 6 régions cérébrales. Les régions ont été définies à partir de l'atlas AICHA[3] dans les hémisphères gauche et droit (nom de la variable suivi d'un `_L` pour l'hémisphère gauche et `_R` pour le droit) :

- 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`),
- et l'hippocampe (`Prod_G_Hippocampus-1`).

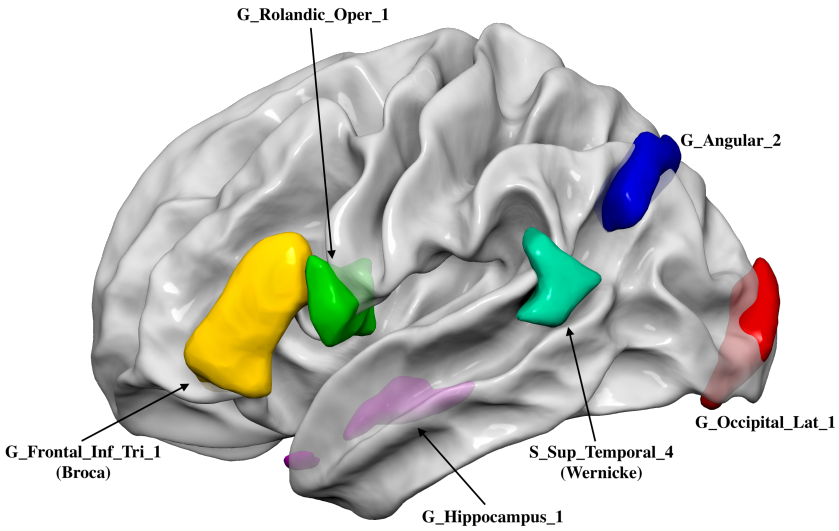
[1] Pour importer vos données, il vous faut taper la commande suivante :

```
donnees <- readRDS("activation.Rdata")
```

en spécifiant éventuellement le chemin d'accès au répertoire dans lequel se situe le fichier `activation.Rdata`.

[2] Mazoyer, B. et al. (2016). BIL&GIN: A neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *Neuroimage*, 124(Pt B), 1225-1231.

[3] Joliot, M. et al. (2015). AICHA: an atlas of intrinsic connectivity of homotopic areas. *Journal of neuroscience methods*, 254, 46-59.



En plus des activations dans les régions, vous disposez de l'âge des sujets, de leur sexe, de leur préférence manuelle, de leur volume cérébral et de leur index de latéralisation hémisphérique :

`Sexe`, `Age`, `Preference_Manuelle`, `Volume_Cerebral` et `Index_Lateralisation_Hemispherique`.

L'index de latéralisation hémisphérique permet de déterminer l'hémisphère dominant pour le langage chez un sujet :

- une valeur positive correspond à un hémisphère gauche dominant ;
- plus cet index est élevé, plus l'hémisphère sera dominant.

On sait qu'environ 90% de la population (droitier ou gaucher) ont l'hémisphère gauche dominant pour le langage.

Le but de ce projet est d'expliquer les fluctuations des activations de l'aire de Broca à gauche (variable `Prod_G_Frontal_Inf_Tri_1_L`) au cours de la tâche de production à l'aide des autres variables présentes dans le jeu de donnée et ainsi de mieux comprendre les interactions entre les différentes régions cérébrales au cours de la production d'une phrase et la notion de réseau qui se cache derrière.

Travail attendu.

- Une étape préliminaire d'analyse descriptive des données (de type statistique descriptive et Analyse en Composantes Principales) est la bienvenue.
- En utilisant les techniques d'estimation et d'analyse d'un modèle de régression linéaire multiple que vous avez vues en cours et en TP, proposez le modèle "le plus simple et le meilleur possible" (en un certain sens) de la variable `Prod_G_Frontal_Inf_Tri_1_L` en fonction des autres variables.
  - Plusieurs approches de sélection de variables peuvent être mises en œuvre (approches pas à pas ascendante ou descendante, fondées sur le critère AIC, par exemple).
  - L'étude peut être faite sur l'ensemble de l'échantillon (femmes et hommes confondus).

    Si vous le jugez utile, vous pouvez également faire un modèle spéficique à chaque sexe (femme ou homme) et comparer les différents modèles estimés.
- N'hésitez pas à commenter le modèle final obtenu à l'aide de vos connaissances en neurosciences et psychologie cognitive.