Projet de modélisation statistique - *Novembre 2018*

Consignes générales

- Le projet doit être fait avec le logiciel R.
- Le projet doit être fait **en trinôme**.
- La qualité de la rédaction sera prise en compte dans la notation. Pensez à bien commenter votre démarche statistique et les résultats obtenus.
- Le rapport doit être envoyé au **format PDF uniquement** à `loic.labache@cea.fr` avec copie à `jerome.saracco@ensc.fr`.
- Le rapport ne devra pas excéder 12 pages (annexes éventuelles comprises).
- Le rapport devra être envoyé au plus tard le **vendredi 14 décembre 2018 à 20h00**.

Jeu de données à traiter : données d'activation durant une tâche de production langagière

Brève description du jeu de données et de la problématique.

Le jeu de données et la problématique vous ont déjà été présentés en TP durant la semaine du 12 novembre 2018. Sont indiquées ci-dessous les principales informations utiles dans le cadre de ce projet de modélisation statistique.

Le jeu de données `activation`[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.

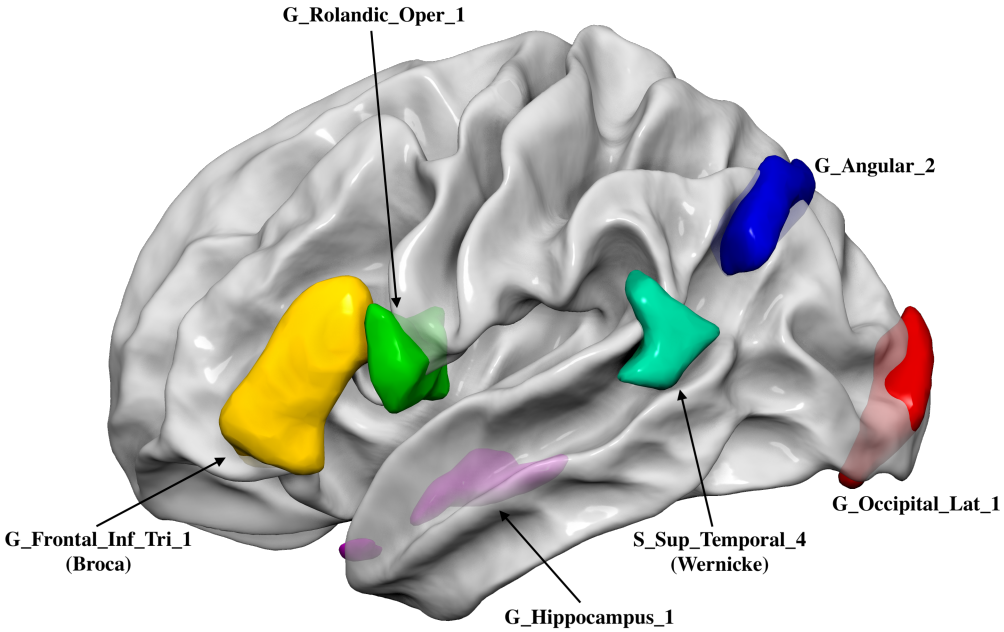
Les activations au cours de cette tâche (variation du signal BOLD) ont été récupérées dans 6 régions cérébrales. Les régions ont été définies à partir de l'atlas AICHA[3] dans les hémisphères droit et gauche (nom de la variable suivi d'un `_L` pour l'hémisphère gauche et `_R` pour le droit) :

- 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] Afin de pouvoir lire le fichier de données `activation.Rdata` dans R, il vous faut utiliser la fonction `readRDS()`.

[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 volume cérébral et de leur index de latéralisation hémisphérique :

`Sexe`, `Age`, `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'air 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.

Pour toutes questions relatives à la problématique et aux données, vous pouvez contacter (à bon escient) `loic.labache@cea.fr`.

Travail attendu.

- Une étape préliminaire d'analyse descriptive des données (de type Analyse en Composantes Principales par exemple) 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 doivent être mises en œuvre. Par exemple, on peut mentionner les approches fondées sur le critère AIC (descendante, ascendante, pas à pas ascendante, pas à pas descendante) ou encore l'approche computationnelle basée sur l'impact de perturbations aléatoires des covariables sur la qualité du modèle (la philosophie de cette approche vous a été présentée en TP).
  - L'étude peut être faite sur l'ensemble de l'échantillon (femmes et hommes).
    Vous pouvez également faire un modèle spéficique à chaque sexe (femme ou homme).
    Vous pourrez alors comparer les différents modèles sélectionnés.
- N'hésitez pas à commenter le modèle final obtenu.