GUIDE D'UTILISATION DE STATAID

Ce document vous guidera à travers les différentes étapes nécessaires à l'analyse de vos données avec StatAid. Il est volontairement simplifié sur le plan théorique afin de permettre à des médecins/chercheurs avec peu de connaissances en biostatistiques de pouvoir réaliser leurs analyses tout en en comprenant les grands principes et les limites de celles-ci. Ce guide n'étant pas un cours de biostatistiques, il ne reviendra pas sur de nombreuses notions de bases par soucis de concision. Celui-ci pourra ainsi paraitre incomplet ou manquant de justesse pour les personnes plus familières avec le domaine. N'hésitez pas à me signaler toutefois les éventuelles erreurs, imprécisions ou manques de clarté que vous auriez rencontré.

ATTENTION WORK IN PROGRESS : Ce guide est en cours de rédaction et n'est pour le moment pas finalisé. Il peut donc manquer certaines parties et conserver des problèmes de mise en forme.

1ère étape : Recueil et codage des données

Avant le recueil

Avant même de vous lancer dans le recueil, il convient de bien identifier les différentes variables que vous allez étudier. Il est ainsi important de bien définir le projet et ses limites avec les différents acteurs de l'étude, d'établir clairement les questions posées et idéalement de consulter un biostatisticien. Avoir une idée des statistiques à réaliser avant le recueil permettra de mieux définir les objectifs de votre étude et vous évitera du travail supplémentaire inutile (par exemple en recueillant des variables chronophages qui seront impossibles à analyser par la suite).

Recueil : Structure des données

Une fois les variables bien définies et les objectifs clairs, le recueil à proprement parler peut commencer. Bien qu'il existe plusieurs façons pour établir un jeu de donnée, nous aborderons ici la façon la plus classique et celle qui sera indispensable pour l'utilisation de StatAid :

- Le jeu de donnée doit être un tableau (type Excel) avec une ligne par observation (échantillon ou patient), et une colonne par variable à analyser
- La première colonne doit correspondre à la colonne d'identification des patients (exemple : Patient_1, Patient_2…)
- Les colonnes doivent toutes avoir un nom différent (pas de colonnes avec exactement le même nom)
- Les variables catégorielles (ex : Petit-moyen-grand) doivent être codées en toutes lettres et pas uniquement en chiffres (sous peine d'être considérées comme des variables numériques)
- Les variables numériques doivent quant à elles comprendre uniquement des chiffres.
- Les décimales peuvent être signalées par des virgules ou des points (bien utiliser le même codage pour tout le tableau)
- Les données manquantes doivent être signalées par une case vide ou par NA
- Evitez les caractères spéciaux (%, @, $...)

Votre jeu de donnée final bien formaté, il faudra l'enregistrer de préférence au format .txt ou .tsv (séparateur : tabulation). Des options sont disponibles si vous avez un jeu de donnée formaté différemment (par

exemple .csv, avec « ; » comme séparateur). En cas de difficultés avec le codage des variables, vous pouvez vous référez au jeu de donné fournit en exemple sur StatAid (cf « chargement des données »).

Rappel sur les variables

Nous distinguerons au cours de ce guide 3 grandes catégories de variables :

Variables continues/numériques

Ce sont les variables numériques continues, correspond à un nombre/chiffre qu'il soit entier ou décimal. Ces variables doivent bien être codées uniquement avec des chiffres.

Exemples de variables continues : Age, poids, taille, nombre de globules blancs (en nombre absolu)...

Variables catégorielles

Ce sont toutes les variables que l'on peut regrouper en catégories. Les catégories peuvent être binaires ou multimodales. Bien qu'il ne soit pas faux de coder des catégories avec des chiffres, **l'utilisation du logiciel StatAid nécessite un codage de toutes les variables catégorielles avec au minimum une lettre/un symbole afin de faciliter le traitement des données.** Les variables ne comprenant que des chiffres sont ainsi détectées comme variables numériques.

Si l'ordre des variables catégorielles n'a pas d'importance pour l'exploration des données et une analyse comparative simple, celui-ci est très important pour la suite lors de la mise en place de modèles prédictifs/explicatifs. Vous verrez en effet plus tard que dans ce type d'analyse, chacune des différentes catégories d'une variable catégorielle est comparée à la catégorie référence de la variable. Cette référence est pour le moment choisie par défaut, par ordre alphabétique. En attendant la mise en place d'un module permettant de changer directement la référence en ligne, vous devrez donc coder vos données pour changer la référence si celle si ne convient pas (en mettant par exemple : 1_Petit 2_Moyen 3_Grand ou A_Localisé B_Etendu).

Exemples de variables catégorielles : binaires (Rechute : OUI/NON ou VRAI/FAUX, Sexe : Homme/Femme), multimodales (atteinte : légère/modérée/sévère, stade: Stade_I/Stade_II/Stade_III/Stade_IV)

Variables temps-dépendant

Cette catégorie regroupe les variables dont l'apparition dépend du temps. Bien qu'elles puissent être codées de plusieurs, StatAid vous demandera de les coder via deux colonnes : une colonne numérique correspondant au temps et une colonne catégorielle/binaire correspondant à l'évènement. Une des particularités de ce type de variable est la possibilité d'inclure des données censurées dans le cas où l'événement d'intérêt n'est pas survenu.

*Exemple pratique : nous nous intéressons à la rechute des patients. Nous choisissons d'étudier le temps avant la rechute (DFS pour Disease Free Survival). Nous codons donc deux colonnes : la première pour le temps (DFS_MONTHS = variable numérique correspondant au nombre de mois jusqu'à la rechute ou bien jusqu'au dernier suivi en l'absence de rechute) et la deuxième pour l'événement (DFS_STATUS = 0 ou FALSE si aucun événement/pas de rechute, 1 ou TRUE si rechute). En l'absence d'événement (0 ou FALSE), on dit que les patients seront **censurés** de l'analyse à la fin de leur suivi (cf module Kaplan Meier).*

Cas particuliers

- Pourcentages : Si vous pouvez tout à fait explorer une variable numérique en pourcentage dans un premier temps (ex: pourcentages de lymphocytes sanguins), il est préférable d'utiliser la variable numérique de

référence en valeur absolue pour la modélisation (ex: lymphocytes en G/L ou /mm3). Un pourcentage nécessite en effet une transformation préalable pour pouvoir normaliser sa distribution, ce qui rend par la suite l'interprétation des coefficients du modèle plus difficile. Attention par ailleurs à ne pas considérer comme variable continue les pourcentages d'effectifs d'une variable catégorielle !

Chargement des données

Une fois votre jeu de donnée prêt, vous pouvez le charger sur la page de Chargement des données (Data loading, Figure 1). De base, il est considéré que votre jeu de donnée est délimité par des tabulations, que les décimales sont signalées par des virgules (comma) et les données manquantes par des cases vides. Vous pouvez modifier ces 3 paramètres dans le panneau de contrôle à droite. Après le chargement, vous pouvez contrôler que le jeu de donnée correspond bien dans le tableau du dessous. Un jeu de donnée est chargé de base dans StatAid pour vous permettre d'explorer les possibilités du logiciel. Ce jeu de donnée correspond à 151 patients issus de la base de donnée du TCGA (The Cancer Genome Atlas) et présentant une leucémie aiguë myéloïde.

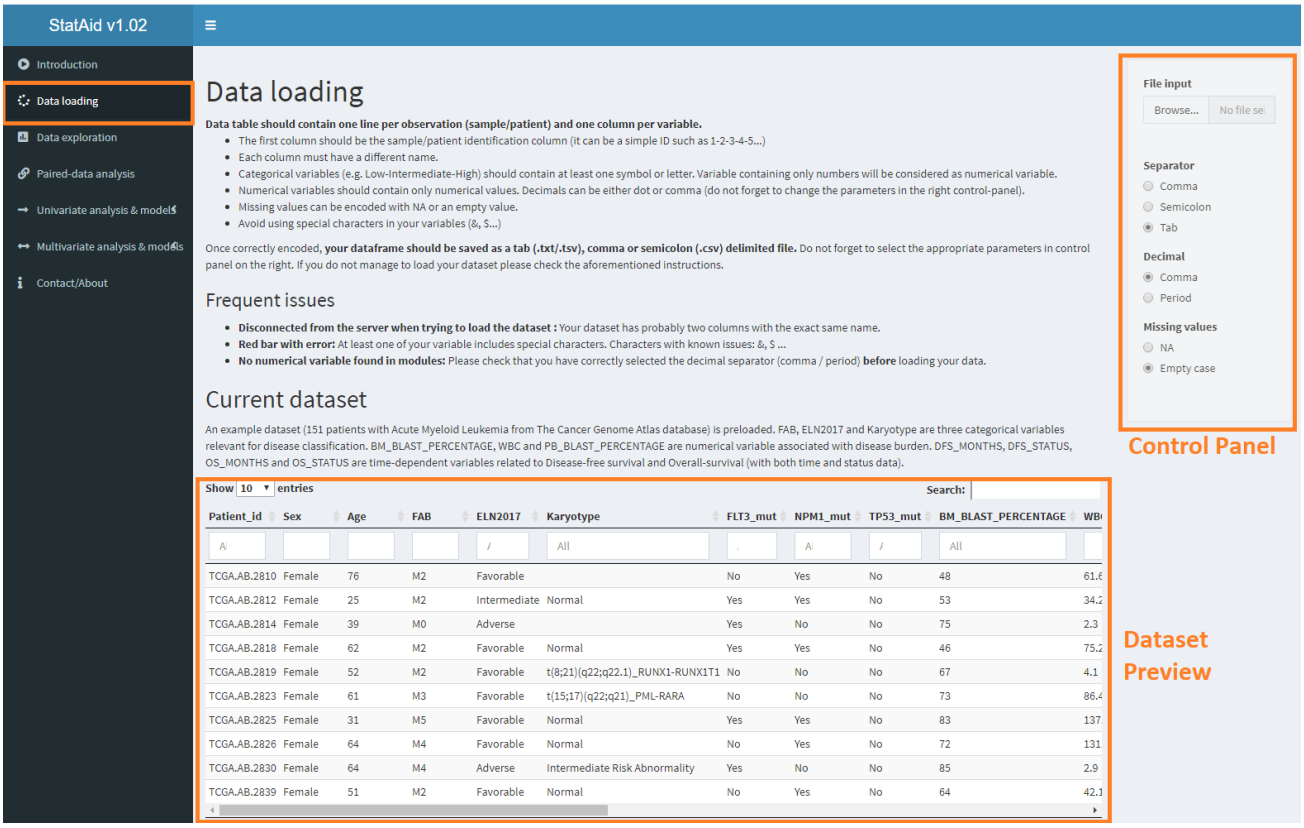

Figure 1 : Page de chargement des données.

Ne pas oublier de modifier les paramètres dans le panneau de contrôle à droite (notamment pour l'annotation des décimales, avec un point (period) ou une virgule (comma).

En cas d'erreur de chargement, vérifiez votre jeu de donnée et essayer de ne charger qu'une partie pour trouver la variable posant soucis. Si l'erreur persiste, vous pouvez me contacter pour rechercher un éventuel bug.

2ème étape : Exploration des données

Une fois votre jeu de données chargé, vous pouvez commencer l'exploration de vos données (onglet « Data Exploration »). Cette étape cruciale vous permettra de procéder à plusieurs vérifications avant de passer aux analyses comparatives :

- **Contrôle du nombre de données manquantes :** Attention à bien signaler les données manquantes de la même façon pour que le programme les détecte correctement. Le contrôle des données manquantes sera primordial si vous voulez ensuite faire une analyse multivariée. (NB: Plus de 25-30% de données manquantes pour une variable peut faire discuter son exclusion de l'étude, d'autant plus que vous avez peu de patients)
- **Détection des données aberrantes (ou outliers) :** Une donnée signalée 'Outlier' est une donnée qui s'écarte de façon importante de la distribution de la variable d'intérêt. Cette donnée doit être contrôlée car il peut s'agir d'une erreur de recueil. S'il ne s'agit pas d'une erreur de recueil, la variable peut être gardée. On peut également discuter son élimination si c'est une valeur exceptionnelle ne risquant pas de se retrouver et/ou ayant peu d'intérêt pour l'analyse.
- **Contrôle de la normalité de la distribution (variables numériques) :** La distribution de notre variable numérique va nous permettre de choisir au mieux le test statistique à utiliser (cf variables numériques)
- **Contrôle de l'équilibre des groupes comparatifs :** En règle générale, il faut éviter de multiplier les groupes, le risque étant de se retrouver avec de très faibles effectifs par groupe pour les analyses. Pour des petites cohortes (n<150), avoir des groupes comparatifs coupés en 3-4 maximum est une bonne chose.
- **Contrôle des effectifs des variables catégorielles :** De la même façon que pour les groupes comparatifs, il faut éviter de multiplier les catégories des variables d'études. Là encore, essayer de limiter à 3-4 catégories par variable max. Les effectifs par catégorie vont également nous permettre de choisir le test statistique à utiliser (cf variables catégorielles)

Lorsque vous cliquez sur l'onglet exploration des données, une fenêtre s'affiche avec un panneau sur votre droite vous permettant de sélectionner la variable à contrôler et les groupes que vous voulez comparer (ou 'cohorte totale' si vous voulez explorer la variable sur l'ensemble de la cohorte).

Variables numériques

Lorsque vous sélectionnez une variable numérique, 3 graphiques et un tableau s'affichent :

- **Boxplot :** Chaque point correspond à une observation (patient/échantillon) dans le groupe correspondant. Les données aberrantes (outliers) sont signalées en rouge. Pour rappel, la boite de couleur représente l'intervalle interquartile (IQR= intervalle entre le quartile inférieur Q1 et le quartile supérieur Q3, soit 50% des données). La ligne du milieu représente la médiane, partageant les données en 2 (50% des données sont au-dessus et 50% en dessous de cette valeur). La ligne verticale (moustaches ou whiskers) est calculée différemment selon les boxplots. Son but est généralement de représenter l'étendue de la grande majorité des données (95 voire 99% des données). Ici, les moustaches sont calculées d'une façon particulière : la ligne cherche le point le plus proche à une distance égale à environ 1,5 fois l'IQR, partant du bas et du haut de la boite. Cette méthode permet de représenter l'étendue des données en se basant sur la distribution centrale des données. Elle permet principalement de détecter les données aberrantes/outliers, situées au-delà de la ligne.
- **Barplot** : Diagramme à barres représentant la moyenne et l'écart type de la variable d'intérêt, par groupes. Les données aberrantes (outliers) sont là encore signalées en rouge.
- **Histogramme :** Un histogramme classique de la distribution par groupes. L'histogramme représente les effectifs par catégories de valeur. Il peut être utile pour apprécier la normalité d'une distribution.

- **Effectifs, NA et normalité :** Tableau récapitulant les effectifs par groupe avec le nombre de données manquantes. Pour chaque groupe, un test de Shapiro-wilk est réalisé (si effectif > 3 min).

Choix du test statistique pour les variables numériques

Après contrôle des outliers et des données manquantes, le choix du test statistique sera guidé par la distribution de la variable et le test de Shapiro-Wilk. Ce dernier est un test statistique évaluant la normalité de la distribution. Attention : une p-value > 0.05 est ici en faveur d'une distribution normale ! Le choix doit en priorité être guidé par l'effectif des groupes et l'inspection visuelle. De manière générale, si la distribution des données ressemble à une Gaussienne et que vous avez un effectif suffisant (n>30 de façon empirique) dans chacun de vos groupes, alors vous pourrez utiliser les tests paramétriques. Une p-value du test de Shapiro-Wilk supérieure à 0.05 vous confortera dans votre choix. Si vous avez des effectifs faibles (n total < 100, n<30 par groupe) ou que la distribution ne semble visuellement pas normale, privilégiez alors un test non paramétrique. Une p-value du test de Shapiro-Wilk inférieure à 0.05 vous confortera dans votre choix. En cas de doute, privilégiez les tests non paramétriques : vous gagnerez en rigueur le peu que vous perdrez en puissance.

Variables catégorielles (ou qualitatives)

Pour ce type de variables (binaires (oui/non, vrai/faux...) ou pluri-modales (Rouge/Bleu/Vert, stade T1/T2/T3/T4), le contrôle qualité peut se faire directement dans la table descriptive (étape 3). On vérifiera ici encore une fois le nombre de données manquantes, et surtout les effectifs par catégories. Lorsque les effectifs d'une catégorie sont peu nombreux (à mettre en lien avec votre nombre total de patient/échantillons), le regroupement de la variable doit se poser (par exemple pour une classification tumorale : regroupe les T1a T1b en une catégorie T1 unique).

Augmenter les effectifs par catégorie d'une variable est une très bonne façon de gagner en puissance et en fiabilité dans l'analyse statistique. Comme évoqué précédemment, se limiter à 3-4 catégories max si vous en avez le choix/la possibilité et peu de patients.

Choix du test statistique pour les variables catégorielles

Pour ce qui est des variables catégorielles, le choix du test statistique est guidé par les effectifs : s'il y a au moins n=5 dans la catégorie contenant le moins d'effectif, un test paramétrique peut être utilisé. Si n<5 dans au moins une catégorie, il faudra alors privilégier la version non paramétrique du test.

3ème étape : Décrire et comparer nos groupes d'études

Vous êtes maintenant familier avec vos données, avez contrôlé les données manquantes, les valeurs aberrantes (outliers) et avez vérifié vos variables catégorielles pour faire en sorte de ne pas multiplier les groupes de façon inutile. La prochaine étape consiste comparer les variables dans nos groupes d'études.

Tableau descriptif & comparatif

Dans ce tableau, les variables sont décrites et comparées en fonction des groupes choisis. Veuillez noter que les variables avec plus de 80% de données manquantes sont automatiquement exclues de l'analyse."),

- **Variables numériques :** La moyenne (dérivation standard) et la médiane [intervalle interquartile] sont données. Privilégier la médiane [IQR] si vous avez peu de patients et/ou des données extrêmes pouvant influer plus fortement sur la moyenne.
- **Variables catégorielles/qualitatives:** Les effectifs (pourcentage) sont donnés. Le pourcentage peut être rendu par ligne ou colonne (paramètre dans le panneau de contrôle à droite).
- **Test statistiques :** Pour chaque variable, qu'elle soit numérique ou catégorielle, deux tests statistiques sont lancés : un test paramétrique et un non paramétrique. Selon les éléments exposés dans la 2ème étape, il faudra choisir le test le plus adapté. En cas de doute, n'hésitez pas à privilégier le test non-paramétrique : vous perdrez peu de puissance et vous serez certain de la validité du résultat.
- **Filtres et paramètres :** Dans le panneau de contrôle à droite, Vous pouvez modifier vos groupes comparatifs (ou choisir 'cohorte totale' pour réaliser une description sur toute la cohorte sans comparaison), choisir les variables à analyser (choisir 'toutes' pour lancer l'analyse sur l'ensemble des variables), choisir d'inclure ou non les données manquantes et changer le type de pourcentage pour les effectifs(par ligne ou par colonne). Vous pouvez également corriger vos p-values (voir plus bas).

De nombreux résultats peuvent déjà être sortis de cette table :

- **Tableau des caractéristiques initiales (Table 1):** Ce tableau, venant généralement en premier dans les essais thérapeutiques, vise à comparer vos groupes d'études sur des caractéristiques de bases. Il ne comprend ainsi pas les variables dont vous voulez montrer qu'il existe une différence (vos variables d'études), mais plutôt des variables pouvant constituer de potentiels facteurs confondants. Une différence significative dans une de ces caractéristiques devra nous alerter sur un potentiel facteur de confusion qui devra être pris en compte pour la suite. Exemple : Nous voulons montrer que le taux de globule blanc est plus élevé chez les patients présentant une leucémie aiguë myéloïde (LAM) au diagnostic et ayant plus de 65 ans par rapport à ceux ayant moins de 65 ans. Nous faisons deux groupes (<65ans, >65ans) et comparons les caractéristiques de base de nos deux groupes (type de LAM, sexe, stade de maladie (diagnostic ou rechute), CRP...) pour faire un tableau descriptif et comparatif. Nous voyons ainsi que nos deux groupes sont comparables concernant le type de LAM, le sexe et le stade de la maladie, mais que le groupe > 65 ans a une CRP plus élevée au diagnostic. Cause ou conséquence, cela constitue un facteur de confusion important qui devra être pris en compte.
- **Comparaisons univariées :** En soit, cette analyse constitue déjà un premier niveau d'analyse univariée. Vous pouvez ainsi très bien utiliser cette table pour commencer à rechercher des différences entre vos deux groupes.

Note sur les tests statistiques

Dans le cas où vos groupes comparatifs comportent uniquement 2 niveaux, les tests correspondent à une comparaison directe entre les deux groupes.

Si vos groupes comparatifs comportent plus de deux niveaux, un test statistique significatif vous permettra de dire qu'il existe au moins une différence entre deux des groupes, mais ne vous dira pas lesquels. Si vous voulez aller plus loin, il faudra ensuite réaliser des comparaisons 2 à 2 (via le module 'graphique personnalisé' par exemple).

Vous aurez probablement remarqué que le test paramétrique comparant 2 variables numériques est systématiquement un T-test de Welch. Ce T-test est une généralisation du T-test de Student pour les situations où la variance des variables comparées n'est pas égale. Cela est très souvent le cas dans les études rétrospectives, lorsque les échantillons comparés sont de tailles différentes, c'est pourquoi il est ici laissé par défaut. Il faut noter par ailleurs que ses résultats sont très proches du test initial de Student lorsque les variances sont égales.

Correction des p-values

A ce stade de l'analyse, de nombreux tests peuvent être lancés. Il faudra ainsi prendre en compte l'inflation du risque alpha. Pour rester simple : **la p-value peut être interprétée comme la probabilité que la différence que l'on observe puisse être liée au hasard.** Ainsi au seuil alpha habituel (0.05), si vous lancez 100 tests différents, vous pouvez vous attendre à avoir 5 résultats faux positifs (p-value <0.05 par fluctuation d'échantillonnage et non différence réelle). Deux solutions principales existent pour contrôler le risque de faux positif : réduire le seuil alpha (à 0.01 par exemple) ou corriger les p-values. Notez que vous pouvez également garder des p-values non corrigées pour faire de l'exploration de données, et prendre des seuils plus restrictifs / corriger les p-values lors de la mise au point d'un modèle de régression par exemple. Vous pouvez également tout à fait interpréter vos résultats avec la valeur même de la p-value non corrigée en fonction du contexte/nombre de tests, sans fixer de seuil de significativité.

Dans tous les cas, **rappelez-vous que la p-value n'est pas l'objectif ultime de votre analyse mais simplement un indicateur de la probabilité que l'effet observé soit lié au hasard.** Un résultat doit être considéré dans sa globalité, en regardant avant tout la taille de l'effet observé, les intervalles de confiance et surtout la pertinence biologique. On aura ainsi des situations où l'on mettra en évidence une différence très importante entre deux groupes, mais où le manque de puissance (petits échantillons...) nous rendra une p-value à 0.07. Cette tendance sera importante à signaler dans les résultats. A l'inverse, une différence ridicule (modification de la pression artérielle de 5 mmHg entre deux groupes) et très significative (p<0.01) du fait d'une puissance d'étude élevée ne sera pas forcément intéressante/pertinente.

Pour ce qui est de la correction, trois méthodes sont proposées :

- **Correction de Benjamini-Hochberg:** Particulièrement adaptée aux situations où de très nombreux tests sont réalisés (bioinformatique, génomique...), ou bien aux situations où l'on cherche à screener des variables dans une première approche, je vous recommande cette méthode par défaut. Cette correction permet de contrôler le seuil de découverte de faux positif (on parle également de FDR pour False Discovery Rate) sur l'ensemble de l'expérience. L'idée globale de cette approche est de pouvoir ramener les p-values finales à un pourcentage de chance de faux positif prenant en compte l'ensemble des tests lancés. On pourra dire alors que les tests avec une p-value corrigée <0.05 après correction sont des tests significatifs avec un FDR à 5%.
- **Correction de Bonferonni:** Cette correction est simple: les p-values sont divisées par le nombre total de test effectué (ou bien le seuil alpha est lui-même divisé pour donner le nouveau seuil d'interprétration). Bien adapté pour les situations où peu de tests sont lancés (ex: corriger 3 tests de comparaison de groupes 2 à 2), cette correction est jugée trop conservatrice lorsque les tests sont multipliés.
- **Correction de Holm-Bonferonni:** Dérivée de la correction de Bonferonni, cette correction a l'avantage de mieux conserver la puissance du test (augmente moins le risque d'erreur de type II = ne pas conclure à une différence qui existe vraiment) que la méthode de Bonferonni.

4ème étape: Monter un modèle pour expliquer une variable essayer de la prédire

Afin d'aller plus loin dans l'analyse de vos données, la prochaine étape consiste à monter un modèle pour essayer soit de prédire une variable soit d'expliquer au mieux ses relations avec d'autres variables. Il existe une multitude de modèles pouvant plus ou moins convenir à vos données. Aucun modèle n'est universel, et l'avis d'un biostatisticien est vivement recommandé pour vous aider à choisir la bonne approche.

Paramètres à choisir

- **Variable Y, à prédire/à expliquer:** Ou variable indépendante, c'est comme son nom l'indique la variable que l'on cherche à expliquer/prédire. Dans la majorité des cas, on ne cherche à étudier qu'une variable à prédire à la fois.
- **Variable(s) X, prédictives/explicatives:** Ou variable(s) dépendante(s), ce sont la ou les variables que l'on utilisera pour essayer de prédire/expliquer notre variable Y.
- **Modèle :** Afin de décrire le comportement de la variable Y sous l'influence (éventuelle) des variables X, nous allons tenter de modéliser cette relation pour essayer de l'expliquer voir de la généraliser. Différents modèles existent et sont +/- adaptés selon vos objectifs. Le modèle choisit dépend en général de la variable que l'on cherche

Un des modèles les plus classiques est par exemple le modèle linéaire : pour tenter d'expliquer la relation entre deux variables continues X et Y, nous allons placer sur un graphique les différentes valeurs de X et Y et trouver une droite qui résume le mieux la relation entre les deux variables. Cette droite de régression est en fait la droite qui minimise la distance totale entre chacun des points et la droite elle-même (voir figure plus bas). Grâce à cette droite, nous pouvons ainsi quantifier la relation entre X et Y via un coefficient (= pente de la droite) mais également prédire les valeurs de Y en fonction des valeurs de X.

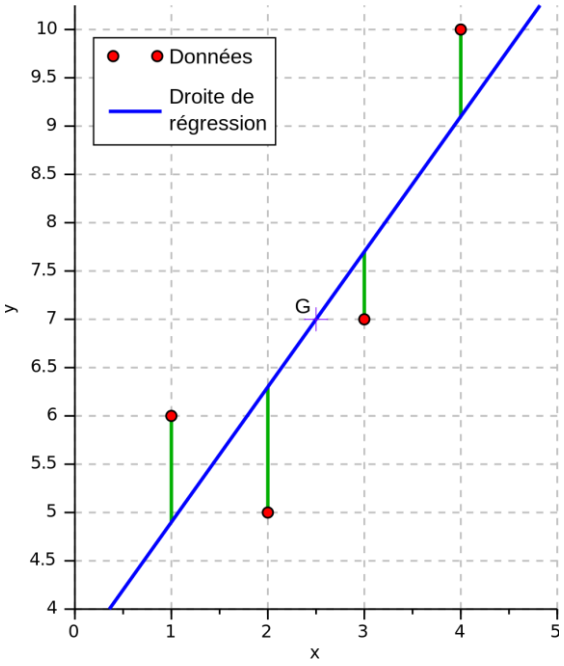

Figure: Droite de régression linéaire.

La droite de régression linéaire (bleue) obtenue par la méthode des moindres carrés est la droite qui minimise son écart (vert) avec chaque point de données (rouge). Source: https://fr.wikiversity.org