STATISTICAL MODELS FOR CLINICAL AND HEALTH DATA

V. Monbet

[1] Université de Rennes/UFR Mathématiques

Objectives

- Learn about statistical methods and models for clinical and health data
- Develop your ability in formulating a problem (from a medical description to a statistical question)
- Develop your ability in choosing a method/model for a specific (statistic) problem and use it to reply the clinical/health question
- Learn how to provide a critical analysis of the results.
- Develop your ability to read and understand biological, medical and health science articles.

- When ?
  Thuesday, from 3pm to 6 :15pm
  9 lectures + practice (If you have a laptop, please bring it with you.)
- Where ?
  Campus de Beaulieu, bat 2A

- Evaluation
  Labs + projects

- Softwares/material : R (and Python), github, google scholar, etc

- Références
  - Azaïs, J. M., Bardet, J. M. (2012). Le modèle linéaire par l'exemple-2e éd. : Régression, analyse de la variance et plans d'expérience illustrés avec R et SAS. Dunod.
  - Cornillon, P. A., Hengartner, N., Matzner-Løber, E., Rouvière, L. (2023). Régression avec R : 3ème édition. In Régression avec R. EDP sciences.
  - Delyon, B. (2023). Régression. https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf
  - Jobson, J. D. (2012). Applied multivariate data analysis : regression and experimental design. Springer Science & Business Media.

Context

Clinical studies

- What are the features characterizing patients vs controls ?
  ex : Clinical characterization of dysautonomia in long COVID-19 patients [Barizien et al., 2021].
- What is the efficiency a treatment ?
  ex : Curcumin against cancer [Guéguinou et al., 2022].
- Status prediction
  predict survival of patients with heart failure from serum creatinine and ejection fraction alone [Chicco and Jurman, 2020]

Observational studies : (Public) Health questions

- Self-poisoning by E-cigarette and E-liquids : National Reports to French Poison Control Centers from July 2019 to December 2020 : VIGIlance and VAPE : The VIGIVAPE Study [Franchitto et al., 2024]
- Biological embodiment of educational attainment and future risk of breast cancer : findings from a French prospective cohort [Berger et al., 2025]

Classical Methods

Content of the course

1. Introduction, research reproductibility, good practices
2. Statistical tests
3. ANOVA
4. Multivariate linear regression
5. Logistic regression
6. Multinomial and ordinal regression
7. Poisson regression (?)

with possibly variable selection for various datasets [1] and data quality [2].

8. Repeated data, curves, missing data
9. Mixed models

Theoretical (lectures + articles readings) and pratical aspects (R or Python o).

1. ex : longitudinal data
2. missing data may occur

Research process

- Scientific question of interest
- Decide what data to collect (and how)
- Collection and analysis of data
- Conclusions, generalizations : inference on the population
- Communication and dissemination of results

Source : Goldstein, EPFL.

Generic question : "Does a "treatment" have an effect" ?

Example

- Does smoking cause cancer, heart disease, etc ?
- Does eating oat bran lower cholesterol ?
- Does echinacea prevent illness ?
- Does exercise slow the aging process ?

Approach the question :

- One simple method for resolving this type of question is to compare two groups of study subjects :
  - Control group : gives a base level for comparison
  - Treatment group : group receiving the "treatment"

Types of studies

- A basic means to address this type of question involves comparing two groups of study subjects :
  - Control group : provides a baseline for comparison
  - Treatment group : group receiving the "treatment"

- Experimental study : subjects assigned to groups by the investigator
  - randomization : protects against bias in assignment to groups
  - "blind", "double-blind" : protects against bias in outcome assessment/measurement
  - placebo : artificial/fake treatment

- Observational study : subjects "assign" themselves to groups
  - confounder : associated with both group membership/risk factor and with the outcome of interest

A few comments

- With a well-planned and well executed controlled experiment, it is possible to infer causality
- This is not possible with observational studies due to the presence of confounders
- With confounding, it is not possible to tell whether the observed difference between groups is due to the treatment or to the confounding factor
- Not always possible to carry out an experiment, for pratical and ethical reasons

Data (examples)

- Clinical data
  A (often small) number of patients (including controls) are observed.
  Typically, two groups : patients vs controls, treated vs control, ...
  Data are obtained from humans or animal models

- (Public) Health data : observational data
  Cohorts

Essai clinique

- Écriture du protocole et du plan d'analyse statistique avant le début de l'étude
  - **Hypothèse de recherche et Objectifs de l'étude**
    - *Essai GuidAge : montrer qu'un traitement (Ginko Biloba) protège de la maladie d'Alzheimer* [Vellas et al., 2012]
  - **Déterminer le design**
    - *essai randomisé en double aveugle avec deux groupes : placebo et traitement, randomisation par block, multicentrique*
  - **Critère inclusion et exclusion**
    - *exclure les patients ayant déjà la maladie. Inclure les patients se plaignant d'une perte de mémoire*
  - **Plan d'analyse statistique** : choix du test statistique et calcul du nombre de sujets nécessaire
    - *Test du logrank. Durée de l'étude : 5 ans, on veut montrer une différence de* 20% *entre la survie du traitement et placebo à la fin de l'étude avec une survie pour le placebo à* 80%*.*
    - $\Longrightarrow n =$
    - Choix des covariables collectées
  - Effets indésirables
- **Soumission et autorisations**
  - Soumission au comité d'éthique et à l'autorité de santé (ex. : ANSM en France, EMA, FDA. . .).
  - Signature des consentements éclairés par les participants.
- **Mise en place opérationnelle, recrutement et suivi des participants**
- **Clôture et analyse, publication et communication**

Cohorte

- **Étude observationnelle** : on suit un groupe de sujets (la cohorte) dans le temps
- **Objectifs** : comparer l'incidence d'un événement entre exposés et non exposés à un facteur de risque
- **Exemple cohorte**
  - *ELFE* : suivi d' enfants de la naissance à l'âge adulte afin de mieux comprendre comment leur environnement affecte, de la période intra-utérine à l'adolescence, leur développement, leur santé et leur socialisation.
  - *PELAGIE* (IRSET Rennes) : répondre aux préoccupations de santé des enfants et adolescents dues à la présence de composés toxiques dans nos environnements quotidiens. Suivi d'environ 3500 mères-enfants réalisé en Bretagne depuis 2002.
  - *Millennium Cohort Study* (Royaume-Uni) : étude de cohorte observationnelle multidisciplinaire mise en place pour suivre la vie des enfants nés au tournant du siècle . Représentative à l'échelle nationale, 18 552 familles
- Suivi longitudinal à différents âges.
  - *Suivi de la naissance à l'âge adulte*
- Collecte d'informations avec des questionnaires, analyse de sang
- Perte de suivi possible (attrition), valeurs manquantes

Etapes

1. **Analyse descriptive** : moyenne, variance, fréquence, représentation graphique adaptée, données manquantes, valeurs aberrantes

   $\Longrightarrow$ **Il est important de regarder et comprendre les données avant toute analyse !**
2. **Analyse bivariée** : tests statistiques pour la comparaison de deux ou plusieurs groupes selon la nature de la variable)
3. **Analyse multivariée** selon la nature de la variable à expliquer
   On a besoin de modèles interprétables !
   - Continue
     - *Indice de masse corporelle (IMC)*
     - Régression linéaire
   - Binaire
     - *Maladie oui/non*
     - Regression logistique
   - Catégorielle
     - *Différents stades d'une maladie*
     - Regression polytomique
   - Comptage
     - *Nombre d'hospitalisation dans le mois*
     - Regression de poisson
   - Durée jusqu'à l'apparition d'un événement
     - *maladie d'alzheimer*
     - Modèle de Weibull, modèle de Cox
     - prise en compte de la censure : sortie d'étude ou fin d'étude. On n'observe pas l'événement mais on sait qu'il a lieu après une certaine date

Example : hibernation

- General question : How do changes in an animal's environment induce hibernation ?
- What changes should be studied ? ?
  - temperature
  - photoperiod (daylight duration)
- What measures to take ?
  - nerve enzymatic activity (Na+K+ATP-ase)
- What animal to study ?
  - golden hamster, 2 organs

Specific question

- General question : How do changes in an animal's environment induce hibernation ?
- Specific question : What is the effect of changing daylight duration on the enzyme concentration of the sodium pump in two golden hamster organs ?

Sources of variability

- Variability due to the conditions of interest (wanted)
  - Duration (long or short)
  - Organ (heart or brain)
- Variability of the response (NOT wanted) : measurement error
  - Preparation of the enzyme suspension
  - Instrument calibration/standardization
- Variability in experimental units (NOT wanted)
  - biological differences between hamsters
  - environmental differences

Types of variability

- Systematic, expected (wanted)
- Random variation (can manage this)
- Systematic, unexpected (NOT wanted)
  - biased results
  - e.g., what time the measurements are made

Questions for the hibernation study

- Long or short : Is there an effect of daylight duration on enzyme concentration ?
- Heart vs. Brain : Are the concentrations different in the 2 organs ?
- Interaction : Is the difference in enzyme concentration (long/short) different for heart and brain ?
- Hamsters : Variability between hamsters ?
- Measurement error : What is the error due to the measurement process for enzyme concentration ?

Experimental design - why do we care ?

- Poor design costs :
  - time, money, ethical considerations
- To ensure relevant data are collected, and can be analyzed to test the scientific hypothesis/ question of interest
  - Decide in advance how data will be analyzed
  - "Designing the experiment" = "Planning the analysis"
- **The design is about the biology**

Data (sources)

Clinical data

- UCI
- kaggle
- Journals

(Public) Health data

- Cohorts : Gazelle, British, NHANES...
- OCDE

Research should be reproducible

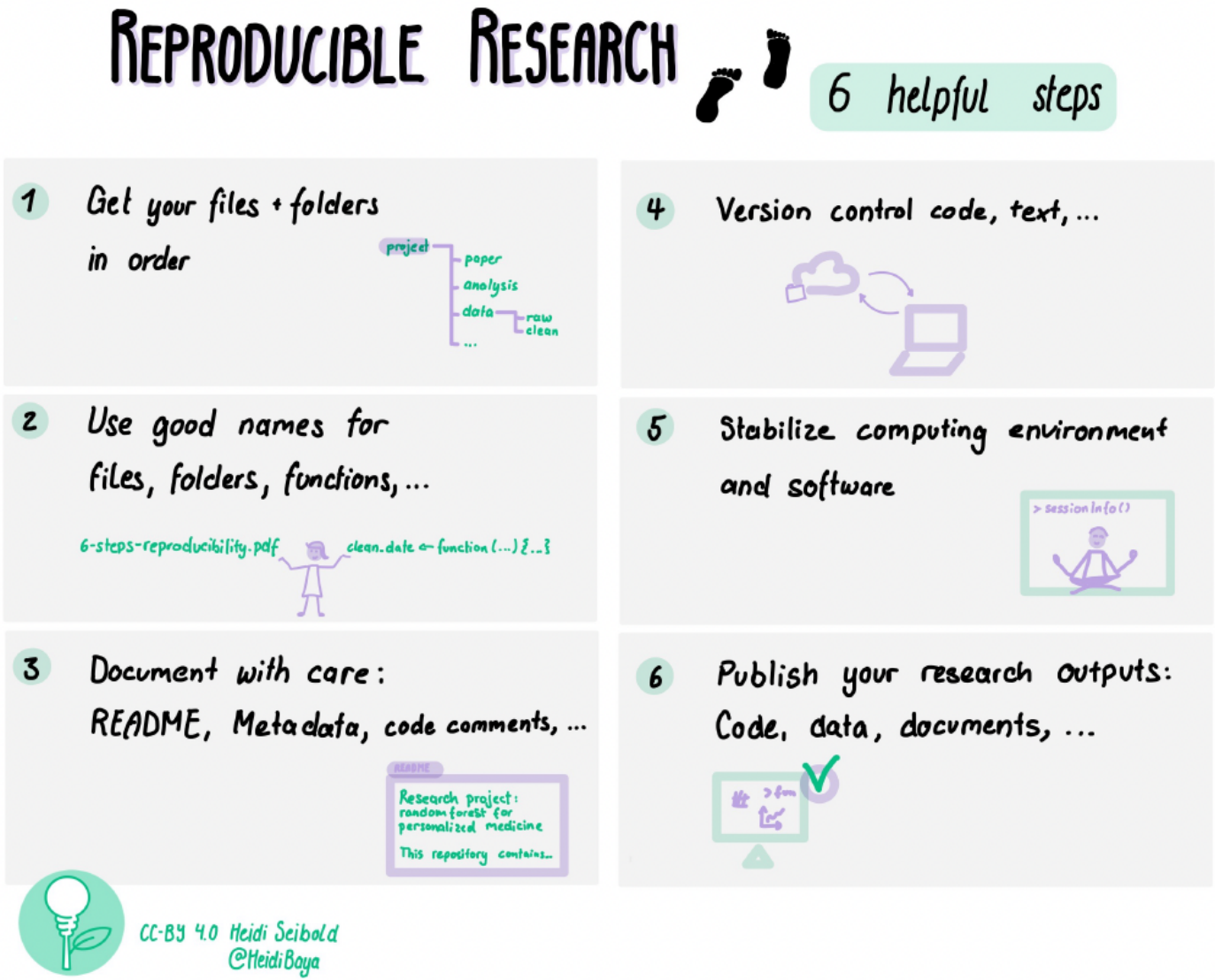
.

Two definitions from the American Statistical Association :

- **Reproducibility** : A study is reproducible if you can take the original data and the computer code used to analyze the data and reproduce all of the numerical findings from the study.

- **Replicability** : This is the act of repeating an entire study, independently of the original investigator without the use of original data (but generally using the same methods).

6 Steps Towards Reproducible Research

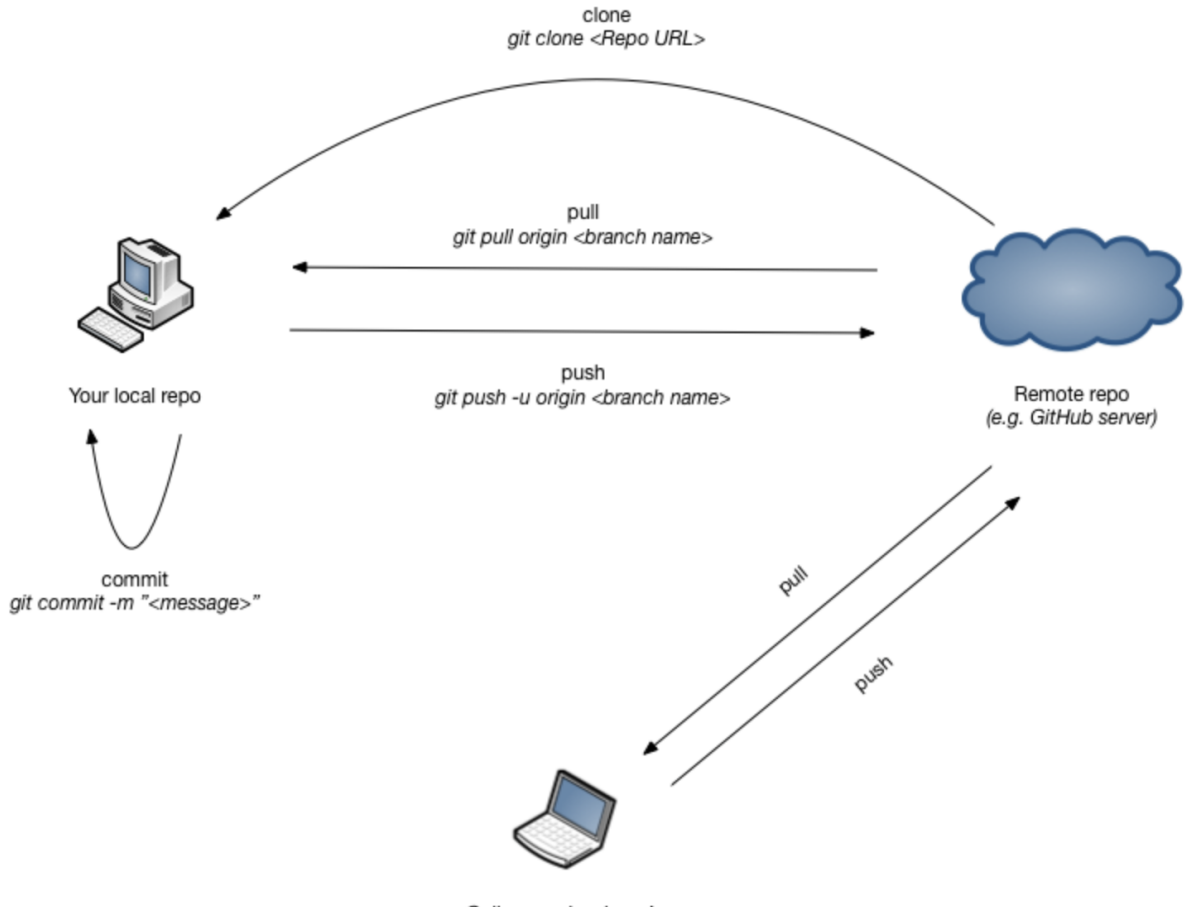


Source [Seibold, 2023]

Why should I know about Git

Some major benefits of using Git are

- Keep an archive of every version of your project
- All you and your co-authors to work at the same time
- You can easily see what changes were made and by whom
- Allows you to contribute to open source projects
- Allows you to make your project open source so others can contribute to your project



What is comming next ?

- Practice on Git
- Combine : create a github (for David's data ?), start to build a descriptive statistics table (Python or R)
- Papers reading

References


Barizien, N., Le Guen, M., Russel, S., Touche, P., Huang, F., and Vallée, A. (2021).
Clinical characterization of dysautonomia in long covid-19 patients.
Scientific reports, 11(1) :14042.


Berger, E., Dudouet, R., Dossus, L., Baglietto, L., Gelot, A., Boutron-Ruault, M.-C., Severi, G., Castagné, R., and Delpierre, C. (2025).
Biological embodiment of educational attainment and future risk of breast cancer : findings from a french prospective cohort.
BMJ open, 15(2) :e087537.


Chicco, D. and Jurman, G. (2020).
Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone.
BMC medical informatics and decision making, 20(1) :16.


Franchitto, N., Bloch, J., Solal, C., Group, F. P. R., and Pélissier, F. (2024).
Self-poisoning by e-cigarette and e-liquids : national reports to french poison control centers from july 2019 to december 2020 : Vigilance and vape : the vigivape study.
Nicotine and tobacco research, 26(3) :281–288.


Guéguinou, M., Ibrahim, S., Bourgeais, J., Robert, A., Pathak, T., Zhang, X., Crottès, D., Dupuy, J., Ternant, D., Monbet, V., et al. (2022).