
L'exploration de données avec R

Congrès annuel de l'ACL

Guilherme D. Garcia

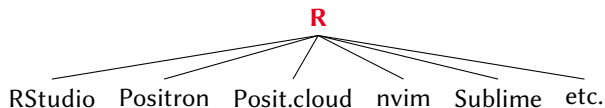
gdgarcia.ca

Juin 2025



Un langage pour l'analyse de données

- Les options les plus populaires : R, Python, Julia ; SPSS, Stata, SAS, MatLab
- R a son propre EDI (éditeur), mais RStudio/Positron est une option supérieure
- ☞ Plus de 20 000 extensions/bibliothèques disponibles (mai 2025)



Notre objectif aujourd'hui

Les étapes d'une analyse

☞ Les étapes suivantes sont typiques dans une étude quantitative (quel que soit l'outil utilisé)

1. Importer nos données
2. Nettoyer/organiser les variables pertinentes
3. Explorer les patrons d'intérêt (y compris la visualisation de données)
4. Analyser les données à partir des tests, modèles, etc.

☞ Aujourd'hui, on explore R à partir des étapes 1–3

Notre outil

posit.cloud : EDI en ligne et gratuit (25 heures par mois)

The screenshot displays the posit.cloud web interface for a workspace named 'CLA-ACL-2025'. The interface is divided into several panels:

- Top Bar:** Shows the workspace name 'Your Workspace / CLA-ACL-2025', a RAM usage indicator, settings, and the user profile 'Guilherme D. Garcia'.
- Menu Bar:** Includes 'File', 'Edit', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help'. A purple arrow points to the 'File' menu.
- Console Panel (Left):** Labeled 'Console', it shows the R version 4.4.3 (2025-02-28) and the path '/cloud/project/'. It contains the standard R startup message and a prompt '> |' at the bottom, which is circled with a purple '1'.
- Environment Panel (Right):** Labeled 'Environment', it shows 'Global Environment' and 'Environment is empty'. A purple circle with the number '2' is placed over this panel.
- Files Panel (Bottom Right):** Labeled 'Files', it shows a file explorer view for the 'project' directory. It contains a table with the following files:

	Name	Size	Modified
	..		
<input type="checkbox"/>	.Rhistory	0 B	May 16, 2025, 3:55 PM
<input type="checkbox"/>	project.Rproj	205 B	May 16, 2025, 3:56 PM

A purple circle with the number '3' is placed over this panel.

Notre outil

posit.cloud : EDI en ligne et gratuit (25 heures par mois)

The screenshot displays the posit.cloud web interface for a workspace named 'CLA-ACL-2025'. The interface includes a menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help) and a toolbar with icons for file operations and running code. The main editor shows a script with three lines of R code: `1 # Un script`, `2 hist(rnorm(1000))`, and `3`. A purple circle with the number '1' is placed over the second line. Below the script, the console shows the output of the code: `> # Un script`, `> hist(rnorm(1000))`, and `>`. A purple circle with the number '3' is placed over the last line of the console output. To the right of the script editor, the 'Environment' pane shows 'Global Environment' with 192 MiB of memory used. Below the environment pane, the 'Plots' pane displays a histogram titled 'Histogram of rnorm(1000)'. The histogram shows a normal distribution of data points, with the x-axis labeled 'rnorm(1000)' and the y-axis labeled 'Frequency'. A purple circle with the number '4' is placed over the histogram. The user's name 'Guilherme D. Garcia' is visible in the top right corner.

1

2

3

4

Un script

Notre interface avec R

- Si vous êtes familiarisé avec Excel ou SPSS, un script sera **complètement différent**
- ☞ Un script nous permet de documenter/reproduire une analyse de façon **complète** et **détaillée**

```
1 # Ceci est un commentaire
2 library(tidyverse) # Une bibliothèque
3
4 d <- c(1, 4, 5) # La création d'une variable (un vecteur)
```

- Aujourd'hui on travaillera directement dans un script **.R** pour mieux comprendre le langage
- **Des questions avant de commencer?**
- Allez sur posit.cloud et créez un compte si vous voulez reproduire notre analyse!

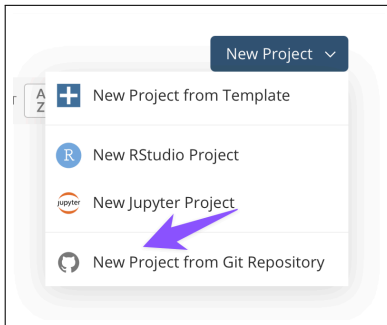
Suggestions de lecture

- Meilleure option pour R en général : Wickham *et al.* (2023)
- En linguistique : Winter (2019); Garcia (2021); Sonderegger (2023)

Comment commencer

Après avoir créé un compte sur posit.cloud

- Cliquez sur « New project » → « New Project from Git Repository »
- Entrez cette adresse : https://github.com/guilhermegarcia/CLA_data_exp_R.git



Références I

Guilherme D GARCIA : *Data visualization and analysis in second language research*. Routledge, New York NY, 2021.

Morgan SONDEREGGER : *Regression modeling for linguistic data*. MIT Press, Cambridge, MA, 2023.

Hadley WICKHAM, Mine Çetinkaya RUNDEL et Garrett GROLEMUND : *R for data science*. O'Reilly Media, Inc., Sebastopol, CA, 2023. Available at <https://r4ds.had.co.nz/index.html>.

Bodo WINTER : *Statistics for linguists : an introduction using R*. Routledge, New York, 2019.