Data Analysis in R

Wrap Up

Ken Stiller

18th January 2026

Why Do We Analyse Data?

- What we have been trying to do is to have best guesses at the value of a variable in the population based on a sample
- In the simplest case, we draw upon descriptive statistics (measures of central tendency and spread)
- If we know more variables, we can look at associations: univariate or multivariate regressions
- **Our aim is to isolate causal effects**
- Controlling for confounders aims to help us achieve this
- We always care about size of effects **and** their statistical significance

- If you suspect your relation is not the same for all units, you can add interactions
- If you suspect your relation is not linear, you can add polynomials (but be careful that this can still be extrapolated)

The Usual Process

1. Start with theory
   - What is out there?
   - Find your niche, try to contribute to debate
   - Then formulate hypotheses

2. Move to data collection
   - What is your population of interest?
   - Can you get a (random) sample?
   - If not, how much can you extrapolate?
3. Think of the right model (specification)
   - Is your treatment randomly assigned?
   - Otherwise, what confounders should you account for?
   - Can you measure them?
   - Need to re-specify model?
4. Interpret results: size *and* significance

Why Should You Care?

- Quantitative literature is everywhere: even if you don't want to do it, you will consume it
- Beyond academic literature, quantitative thinking is ubiquitous. In public discourse, people make mistakes we should be careful of:
  - Assuming $P(A|B) = P(B|A)$ (e.g., vaccines, police shootings, etc.)
  - Not thinking of possible confounders
  - Extrapolating from non-random samples

Why Should You Care? II

- Quantitative and qualitative analyses can [should] complement each other:
  - Qual. analyses help understand mechanism of quant. analyses. Example: Abdelgadir and Fouka 2020, APSR

American Political Science Review (2020) 114, 3, 707–723
doi:10.1017/S0003055420000106

Political Secularism and Muslim Integration in the West: Assessing the Effects of the French Headscarf Ban

AALA ABDELGADIR *Stanford University*
VASILIKI FOUKA *Stanford University*

In response to rising immigration flows and the fear of Islamic radicalization, several Western countries have enacted policies to restrict religious expression and emphasize secularism and Western values. Despite intense public debate, there is little systematic evidence on how such policies influence the behavior of the religious minorities they target. In this paper, we use rich quantitative and qualitative data to evaluate the effects of the 2004 French headscarf ban on the socioeconomic integration of French Muslim women. We find that the law reduces the secondary educational attainment of Muslim girls and affects their trajectory in the labor market and family composition in the long run. We provide evidence that the ban operates through increased perceptions of discrimination and that it strengthens both national and religious identities.

  - Qualitative studies can *help generate hypotheses* to test with a larger N
  - Looking at a distribution of cases can help with case selection for case studies / small-N studies

Beyond OLS: Modelling for Different Types of DV

What if my dependent variable isn't continuous?

- Logistic regression for Binary Dependent Variables
- Multinomial regressions for Categorical Dependent Variables
- Ordered logit for Ordinal Dependent Variables

  - Agresti (2018) *Statistical Methods for the Social Sciences*, Chaps. 14-15.
  - Agresti (2013) *Categorical Data Analysis.*
  - Gelman and Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Part 1).

Beyond OLS: Modelling Different Types of Data

What if my observations are not independent from each other?

- Multilevel Modelling for nested data (units are 'grouped' and we want to estimate unit- and group-level effects)
  - Gelman and Hill (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models (Part 2)

- Panel Data and Event History analysis for times-series (same units are observed at different times)
  - Woolridge (2013) *Introductory Econometrics: A Modern Approach* (Chapters 10-14)
  - Kleinbaum and Klein (2011) *Survival Analysis: A Self-Learning Text.*

Beyond OLS: Other Applications

- Measurement (Classification, PCA, Factor Analysis, IRT)
  - Lauderdale (2022) *Pragmatic Social Measurement*, available here
  - Kabacoff (2021) *R in Action*, Chaps. 14-17, available here

- Statistical Learning (aka 'Machine Learning')
  - James, Witten, Hastie and Tibshirani (2021) *An Introduction to Statistical Learning*, available here
  - Lantz (2013) *Machine Learning with R*
- Qualitative Comparative Analysis (QCA)
  - Schneider, Thomann and Oana (2021) *Qualitative Comparative Analysis Using R: A Beginner's Guide*

Beyond Base R: Making the Most out of the Software

- Data Wrangling, Working with String Variables, Creating Functions, and more...
  - Chris Hanretty's short course 'ConverRt to R', available here
  - Wickham and Grolemund (2016) *R for Data Science*, the "Bible" of `tidyverse` users, available here

- Data Visualisation
  - Healy (2019) *Data Visualization: A Practical Introduction*

- Spatial Data and Maps in R
  - Lovelace Nowosad and Muenchow (2021) *Geocomputation with R*, available here

- RMarkdown
  - Xie, Allaire and Grolemund (2023) *R Markdown: The Definitive Guide*, available here