Research Tools & Data Sources: Quick Guide

Gabriela Noemi Villalba Marecos

June 2024

1 Checklist: Starting Your Own Research Project

1. Define a clear research question and specify the estimand.
2. Map relevant literature and position your contribution.
3. Identify available data sources and assess access constraints.
4. Evaluate advantages and limitations of data types (survey, admin, big data, etc.).
5. Pre-register design or analysis plan when feasible.
6. Set up reproducible workflow (Git, container, documentation).
7. Conduct pilot data exploration and variable construction.
8. Plan identification and estimation strategy (OLS, IV, DiD, RCT, ML).
9. Anticipate robustness checks and sensitivity analyses.
10. Draft reporting structure (tables, figures, appendices).
11. Ensure compliance with ethics, data protection, and legal norms.
12. Maintain logs and README for full reproducibility.

2 Core Research Stack

Category	**Recommendations**
Statistical software	Stata (econometrics), R (tidyverse, fixest), Python (pandas, statsmodels, scikit-learn), Julia (StatsModels).
Reproducibility	Git/GitHub; `renv` (R), `pip-tools`/conda (Py); Make/`targets` (R)/`pytask`; Quarto/LaTeX for literate programming.
Workflow	VS Code; JupyterLab; tmux; containerization with Docker; CI (GitHub Actions).
Documentation	README, data dictionary, ADRs (architecture decision records).
Referencing	Zotero + Better BibTeX; `biblatex`/natbib; Citation keys in code comments.
Security	Secrets via environment variables; encrypted tokens; code review for data exfiltration.

3 Key Data Sources (Selected)

Theme	Examples (access / notes)
Macroeconomy	World Bank WDI; IMF IFS; OECD Data; FRED.
Labor markets	ILOSTAT; EU-LFS/Eurostat; IPUMS (census/ACS); UK LFS/APS (ONS Secure); US CPS/ACS (IPUMS).
Education	UNESCO UIS; OECD PISA; World Bank EdStats; national exams (e.g., Saber Pro, GCSE/A-Level microdata under secure access).
Firms/Innovation	ORBIS/BvD; Compustat/CRSP (license); WIPO Patentscope; OECD STI Microdata (restricted); enterprise surveys.
Trade	UN Comtrade; BACI (CEPII); WTO Tariff; UK Trade Info; US ITC.
Prices	CPI/PPI (national stats offices); Billion Prices (access varies).
Health/Demography	DHS; MICS; LSMS; UN Population; ONS Secure (UK) for linked health.
Geospatial	GADM; Natural Earth; OpenStreetMap (OSM); LandScan/WorldPop; EU Copernicus.
Text/Job Ads	Burning Glass/Lightcast (license); Indeed; O*NET; web archives (robots/ToS compliant).

4 Types of Data: Advantages and Limitations

Data Type	Advantages	Limitations
Survey microdata	Detailed individual/household/firms information; design-based inference possible; rich covariates.	Expensive to collect; recall/response bias; limited time coverage.
Administrative data	Large scale; often longitudinal; high accuracy on recorded variables.	Access restricted; may lack research variables; potential linkage/PII risks.
Experimental (RCTs)	Strong causal identification; transparent design.	Costly; external validity concerns; ethical constraints.
Big Data/web data	High frequency; large volume; novel phenomena.	Representativeness issues; unstable platforms; legal/ethical barriers.
Qualitative/field notes	Context-rich insights; complements quantitative analysis.	Non-replicable; limited generalizability; subjective coding.
Remote sensing/geospatial	Global coverage; fine spatial resolution; useful for exposure measurement.	Requires technical processing; may be costly; potential measurement error.

5 Data Access and Governance

- **Open vs. Restricted:** plan timelines for approvals (e.g., ONS Secure Research Service in the UK).
- **PII Handling:** minimize collection; pseudonymize; store keys separately; audit access.
- **Versioning:** track raw and processed data with DVC or Git LFS; hash files.

6 Minimal Reporting Checklist

1. Research question, estimand, and identification strategy.
2. Data provenance and construction pipeline.
3. Primary specification + robustness set; pre-trends/balance when relevant.
4. Code, environment, and artifact versions; reproducibility instructions.