Data Analysis in R

University of Bayreuth - Fall Term 2024

Course Convenor: Ken Stiller (kenneth.stiller@uni-bayreuth.de)

This course introduces students to the fundamentals of applied statistical analysis for the social sciences. We will cover the basics, starting from how we can use statistics to summarize information and describe general patterns of interest to how we can implement predictions or arrive at causal claims. To do so, this course will introduce students to coding in R, an increasingly popular statistics package. A more detailed week-by-week description of the course contents can be found below.

Students on this course are required to be familiar with basic maths and statistics but I otherwise assume no knowledge of applying statistics. The course is comprised of a lecture component and practical component every week, where you will learn to conduct your own statistical analyses. You will also be required to complete homework assignments. By the end of the class, you will be able to run your own linear regression models, present your results in publishable quality and test pre-formulated hypotheses. Please bring your own laptop for the class. If you do not have a personal laptop, please get in touch with me. Make sure R and R Studio are installed for the second session, more on this below.

Books & Readings

We will mainly be working with the following two textbooks. Moreover, the syllabus indicates the paper from which the data we work with originate, where applicable. Reading the latter is not a requirement.

- Imai, Kosuke. *Quantitative Social Science : An Introduction.* Princeton, 2017.

- Agresti, Alan., and Barbara. Finlay. *Statistical Methods for the Social Sciences.* 4th ed. Upper Saddle River, N.J.: Pearson Prentice Hall, 2009

Assessment

- Contribution to discussions in class (15%)

- Homework (10%)

- In January, you'll be completing a Take-Home Exam (75%). Further details of its content and timeframe will be announced in class.

Whenever you are asked to submit your homework, please do so via email.

Collaboration on Homework & the Take-Home Exam

Collaboration is not allowed. You may discuss general approaches to solving tasks as well as ask for clarification in case you are not sure you fully understand the tasks. **However, you are strongly encouraged to reach out to me rather than to contact your classmates.** In any case, each of you is required to produce his or her own final code and write-up - and, if you did reach out to classmates, to indicate in your output-file which classmates you discussed with as well as the contents of your discussion. Discussion of your homework; take-home exam; or your responses on online channels is strictly forbidden. Students will be penalized for violating this rule.

Course Website

Slides, course materials, data, and lab scripts we will be working with can be found on the course website: https://bayreuth-politics.github.io/R24/.

Drop-In Sessions

Instead of office hours, a weekly drop-in session held will take place (tbc, probably on Wednesdays). Feel free to pop in to ask any questions on or discuss the lecture materials, homework or coding issues. You are also welcome to reach out to me to discuss any questions, including those related to your dissertation or other projects involving data analysis.

About R

Below you can find instructions for the installation of R and R Studio as well as some useful information and tutorials :

• To download R, go to: https://cran.rstudio.com/

• Many people like R-Studio as a way of managing your work in R. Like R, the basic version of R-Studio is free. You can download it here.

• There are a few tutorials which you might deem useful. Feel free to have a look at these websites if you want to improve your R skills (this is entirely voluntary and not a course requirement):
– Try R: http://tryr.codeschool.com/levels/1/challenges/1
– swirl: http://swirlstats.com
– Jared Knowles R bootcamp: https://www.jaredknowles.com/r-bootcamp/

Syllabus

Week 1: Introduction

Brief overview of the course and discussion of how quantitative methods are useful in answering problems in the social sciences. No readings required.

Week 2: Causality & The Basics of Statistics

Introduction to causality, the fundamental problem of causal inference and counterfactuals. The crucial role of randomization and statistical approaches using observational data. We'll discuss basic descriptive statistics: types of variables (interval, ordinal, categorical, and ratio); centres of distributions (median and mean); spread of distributions (quantiles, variance and standard deviation).

Readings:
• Imai Ch. 2
• Agresti and Finley Ch. 2.1 & 3

Data Paper:
• Dupas, Pascaline and Jonathan Robinson. 2013. *Why Don't the Poor Save More? Evidence from Health Savings Experiments.* American Economic Review, 103(4): 1138-1171, http://dx.doi.org/10.1257/aer.103.4.1138
• Jones, Benjamin and Benjamin Olken. 2009. *Hit or Miss? The Effect of Assassinations on Institutions and War.* American Economic Journal: Macroeconomics 1(2): 55–87, http://dx.doi.org/10.1257/mac.1.2.55.

Week 3: Sampling & Measurement

Measurement is key - and not a purely technical exercise as it involves important decisions. We'll also start visualising data and distributions. We'll also discuss the role of sampling, covering: simple random sampling; problems of non-probability sampling; cluster and stratified random sampling; sampling error and non-sampling biases. We'll also learn various ways to summarise bivariate relationshipsand visualise them using R.

Readings:
- Imai Ch. 3;
- Agresti and Finley Ch. 2

Data:
- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2022. *The Quality of Government Standard Dataset*, version Jan22. University of Gothenburg: The Quality of Government Institute, https://www.gu.se/en/quality-government doi:10.18157/qogstdjan22

Week 4: Prediction

Bivariate linear regression & short explanation of the idea of dependent and independent variables. We move from scatter-plots and fitting a line to discussion of the line equation (e.g. the slope and intercept) and OLS regression (e.g. least squares criterion and the error term). Regression Anatomy: TSS, ESS, RSS and goodness of fit. Interpretation of regression coefficients. Inference with bivariate regressions.

Readings:
- Imai Ch. 4.1 & 4.2;
- Agresti and Finley Ch. 9

Data:
- Brexit Referendum Data. Available here.

Week 5: Multivariate Regression

There are various reason to include more than one independent variable in a regression model. We will discuss the basic principle of multivariate regression from a statistical perspective; specify models; and learn how to interpret regression coefficients in multivariate regression models. We'll also learn how to make inference in the multivariate case.

Readings:
- Imai Ch. 4.3.2;
- Agresti and Finley Ch. 10 & 11

Data Paper:
• Ana de la O. (2013). *Do Conditional Cash Transfers Affect Voting Behavior? Evidence from a Randomized Experiment in Mexico.* American Journal of Political Science, 57:1, pp.1-14.

Week 6: Probability & Uncertainty

This week we'll address probability and probability distributions, beginning to consider the accuracy of estimates - and understand how regressions critically hinge on probability theor. We will discuss basic probability models; probability distributions; large sample theorems; standard errors; confidence intervals and learn to how to estimate and evaluate them in R.

Readings:
• Imai Ch. 6;
• Agresti and Finley Ch. 4

Data Paper:
• Levin, Dov H. (2016). *When the Great Power Gets a Vote: The Effects of Great Power Electoral Interventions on Election Results.* International Studies Quarterly Vol. 60, No.1, pp. 189-202.

Week 7: Hypothesis Testing

This week builds upon the issue of accuracy of estimates discussed last week and expands on it with a focus on hypothesis testing. We will learn more about standard errors and confidence intervals for means and proportions before covering several aspects relevant to hypothesis testing, such as null and alternative hypotheses; p-values; type I and II errors.

Readings:
• Imai Ch. 7.1 & 7.2;
• Agresti and Finley Ch. 5 & 6

Data Paper:
• Levin, Dov H. (2016). *When the Great Power Gets a Vote: The Effects of Great Power Electoral Interventions on Election Results.* International Studies Quarterly Vol. 60, No.1, pp. 189-202.

Week 8: Assumptions & Limits of OLS

This week considers uncertainty in relation to the linear regression model, with a particular focus on OLS assumptions and the limitations of simple OLS models. We'll cover bias; Gauss Markov assumptions; re-assessment of OLS inference. Practically, we will use regression diagnostics in R to assess the validity of OLS models.

Readings:
• Imai Ch. 7.3;
• Agresti and Finley Ch. 14

Data:
• Brexit Referendum Data. Available here.

Week 9: Interactions & Non-Linear Effects

Looking beyond simple OLS models, we'll discuss interactions using both binary and continuous moderators. You'll learn how to interpret main effects and the interaction terms. Moreover, we'll address non-linearities using simple higher-order polynomials. We implement these estimators in R and learn to interpret them. Final assessment of what has been learnt and what else is out there.

Readings:
• Imai Ch. 4.3 & 4.4;
• Agresti and Finley Ch. 11.5

Data Paper:
• Timothy Hellwig and David Samuels (2007), *Voting in Open Economies: The Electoral Consequences of Globalization.* Comparative Political Studies 40 (3): 283-306.