APPLIED STATISTICS FOR SOCIAL AND POLITICAL SCIENCES

Year 2025/2026

Spring term, Year 1

Instructors Lecture (20 hours): Daniel Auer (coord.), Francesco Colombo, Kasia Nalewajko, Aron Szekely
Instructors R Lab (20 hours + 4 voluntary): Arturo Bertero, Gonzalo Franetovic, Fabio Torreggiani

Timetable

Starting date: 13.02. (Thursday – Lab)
Lectures: Thursdays, 14:00 – 16:00, Classroom 1
Labs: Thursdays, 16:30 – 18:30, Classroom 1

Course Aim

The aim of this course is to provide you with the core statistical and conceptual tools needed to understand and conduct reliable empirical research in the social and political sciences. At the end of the course, you should be able to:

1) display and explore data, compute and graph linear relations, understand basic probability distributions and statistical inferences, and simulate random processes to forecast uncertainty
2) build, fit, understand, use, and assess the fit of linear regression models and be have a basic understanding of logistic regression models.
3) understand the assumptions underlying causal inference and perform causal inference in simple experimental settings using regression to estimate treatment effects.

Readings

Recommended preliminary readings:

- Agresti, A. (2018). *Statistical Methods for the Social Sciences*, (fifth edition) Boston: Pearson. Chapters 1-3.
- Kellstedt, Paul M., and Guy D. Whitten (2018). *The fundamentals of political science research* (third edition). Cambridge: Cambridge University Press. Chapter 6

Reference textbooks:

- Gelman, Andrew, Jennifer Hill, Aki Vehtari (2020). *Regression and Other Stories*. Cambridge: Cambridge University Press.
- Kellstedt, Paul M., and Guy D. Whitten (2018). *The fundamentals of political science research* (third edition). Cambridge: Cambridge University Press.

Additional reading material can be assigned in some weeks.

Evaluation

- Individual assignments: Students will be asked to hand in two assignments (after week 5 and after week 8, respectively).
- Essay: After the end of the course, students will be assigned a research question and a dataset to work with. They are expected to develop a (statistical) strategy to address the research question, apply it to the data, and report the results in a short essay.

Schedule

Click on the titles for more details

Each week consists of a 2-hour lecture, where topics will be presented from a theoretical and intuitive way, and a 2-hour hands-on tutorial, which will guide you to the application of each topic with the statistical software R (Rstudio). In the lab sessions, we will create and continuously feed an RMarkdown file that collects all coding information. You are expected to read the assigned material before each class.

Week		Lecture (DA / FC / KN / AS)		Lab (AB / FT / GF)	
14.02.		Rstudio + RMarkdown 101			AB
20.02.	**1**	Basic concepts	DA	Getting started with R	FT
27.02.	**2**	Knowing your data	KN	Data visualization	FT
03.03. tbd		Voluntary Recap 1			GF
06.03.	**3**	Statistical inference	AS	Statistical analysis in practice	FT
13.03.	**4**	Bivariate relations	DA	Bivariate hypothesis testing	FT
20.03.	**5**	Linear Regression	KN	Linear regression with a single predictor	AB
		Assignment 1			
27.03.	**6**	Multiple Regression	FC	Linear regression with multiple predictors	AB
03.04.	**7**	Assumptions and diagnostics	AS	Regression diagnostics and evaluation	AB
07.04. tbd		Voluntary Recap 2			GF
10.04.	**8**	Modeling probabilities	FC	Logistic regressions	AB
		Assignment 2			
17.04.	**9**	From correlation to causation	DA	Regression models for causal inference	GF
08.05.	**10**	A statistical western	AS	Statistical analysis in practice II	GF
		Final Assignment: Essay			

Week 1: Review of descriptive statistics

Lecture 1: Review of basic concepts in descriptive statistics *Instructor: Daniel Auer*

Outline:

- Variable measurement: quantitative vs categorical (nominal, ordinal, interval); discrete and continuous
- Description of data: relative frequency distribution; types of distribution (u-shape, bell-shape, skewed)
- Central tendency: Mean; median; mode
- Examples of data visualization of univariate statistics with tables and graphs; Frequency tables; Graph bars; Histograms

Required readings:

- Agresti (2018): Chapters 1-3
- Kellstedt & Whitten (2018): Chapter 6

Lab 1: Basic workflow in R *Instructor: Fabio Torreggiani*

Outline:

- Getting familiar with R & Rstudio – Interface
- Loading data
- Generating new variables, renaming, …
- Fundamental commands for data inspection

Before class:

- Download and install R & Rstudio on your laptop

Week 2: Understanding probability, understanding your data

Lecture 2: Knowing your data *Instructor: Kasia Nalewajko*

Outline:

- Definitions: statistical science (methods for design, description, and inference from data); population and sample -> statistics and parameters (sample mean and population mean); descriptive vs inferential statistics
- Variability of the data: Range; standard deviation: formulation, properties (scaling issue), empirical rule; interquartile range; outliers; z-score
- Challenges of statistics: generalize from sample to population; generalize from treatment to control group; generalize from measures to constructs of interest
- Association between variables: dependent and independent variables; cross-tabulation; correlation: covariance and formulation
- Probability distribution: normal distribution; lognormal distribution; binomial distribution; Poisson distribution; real data

Required readings:

- Gelman et al. (2020): Preface, Overview (only pp. 3-13), Chapter 3

Lab 2: Data visualization in R *Instructor: Fabio Torreggiani*

Outline:

- Describing data: univariate statistics; relative frequencies
- Visualizing data: bar charts, scatterplots, histograms

Required readings:

- Gelman et al. (2020): Chapter 2

Voluntary Recap 1 *Instructor: Gonzalo Franetovic*

During this voluntary 2-hour session students have the opportunity to recap concepts and their application in R, ask general questions, and get their statistical skills up to speed.

Week 3: Statistical inference

Lecture 3: Statistical inference *Instructor: Aron Szekely*

Outline:

- Statistical inference: what is it and why do we need it?
- Samples and populations, sampling distribution
- Quantifying uncertainty in sampling: standard error
- Confidence intervals
- Hypothesis testing and Type 1 and Type 2 errors
- *p*-values

Required readings:

- Gelman et al. (2020): Chapter 4 (only pp. 49-60)
- Kellstedt, Paul M., and Guy D. Whitten (2018). Chapters 7 and 8

Lab 3: Statistical analysis in practice *Instructor: Fabio Torreggiani*

Outline:

- Samples and population: Possible sources of data; Random sampling; Normal distribution; Kernel density; Standard deviation and z score
- Confidence intervals and p-value
- Hypothesis testing in practice: Plot the distribution and test H0 and H1; T-test; P-value

Required readings:

- Gelman et al. (2020): Chapter 4 (only pp. 63-65)

Week 4: Bivariate hypothesis testing

Lecture 4: Concepts and tools to test bivariate relations *Instructor: Daniel Auer*

Outline:

- The Null Hypothesis and p-values (reminder): Strength and statistical significance of an association between X and Y
- Choosing the right bivariate hypothesis test
- Tabular Analysis: Getting cross-tabulation right (reminder); Chi-Squared Test of Independence; Strong vs. weak association in a contingency table
- Difference of means: Two-sample t-test
- Correlation coefficient: Assumptions of Pearson's correlation; Pearson's r and Spearman's rho

Required readings:

- Kellstedt & Whitten (2018): Chapter 8

Lab 4: Bivariate hypothesis testing *Instructor: Fabio Torreggiani*

Outline:

- Group comparisons with real data: Create a treatment variable and identify treated and control groups; Cross tabulation; Pearson/Spearman correlation coefficient
- Graphical tools with real data: Histogram of two groups on one graph; Overlapping density plots for two groups; Box plot for a continuous variable in two groups; Dot plot for a continuous variable comparing two groups

Week 5: Linear regression with one predictor

Lecture 5: Concepts and tools for linear regression modeling *Instructor: Kasia Nalewajko*

Outline:

- Introduction to regression models: Historical origins of regression; Using regressions for prediction or comparison; Linear and nonlinear regression models
- Which line fits best? Population and sample regression models; Ordinary least-squares
- Measuring uncertainty about the regression line: Goodness-of-Fit: R-Squared Statistic; Confidence Intervals about Parameter Estimates
- Descriptive and causal interpretations of regression: coefficients as comparisons, not effects
- The paradox of regression to the mean

Required readings:

- Kellstedt & Whitten (2018): Chapter 9 (until pag. 205)
- Gelman et al. (2020): Chapter 6

Supplementary readings

- Kahneman, Daniel (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux. Chapter 17 (Regression to the mean)

Lab 5: Linear regression modeling with a single predictor *Instructor: Arturo Bertero*

Outline:

- Preliminary steps of data inspection: using real data
- Outliers: Graphically detect outliers and assigning labels to recognize them
- Bivariate regression
- Regression output: how to read it; Interpreting coefficients
- Different dependent and independent variables (continuous, categorical, dummy)
- Export regression results: Tables and graphs

Required readings:

- Gelman et al. (2020): Chapter 6

Assignment 1

The first assignment will reflect concepts and tools discussed during weeks 1 to 5. Assignments must be done individually (no group work) and will be graded. Hand-in via email.

Week 6: Linear regression with multiple predictors

Lecture 6: Introducing multiple predictors in your regression model *Instructor: Francesco Colombo*

Outline:

- Dealing with more than two variables: Control variables, omitted variables and spurious associations
- Regressions with multiple predictors; Understanding the fitted model
- Interpreting regression coefficients in regressions with multiple predictors: Comparing models A, B and C; Interpreting the R squared; Which predictor is stronger? Standardized and unstandardized predictors
- Standardizing predictors for comparing regression models (using z scores, using an external population, using reasonable scales)
- Testing interactive hypotheses: The logic of interactions; Interactions between predictors (categorical-categorical, categorical-continuous, continuous-continuous interactions)

Required readings:

- Gelman et al. (2020): Chapter 10 (only pp. 131-139)
- Kellstedt & Whitten (2018): Chapter 10

Lab 6: Linear regression modeling with multiple predictors *Instructor: Arturo Bertero*

Outline:

- Spurious associations in real data
- Multiple regression with real data: Add lurking variables; Compare regression outputs with and without lurking variables; Comparing regression models F test
- Interactions with real data: How to interpret interactions in the regression output; Compare regression outputs with and without interactions; Plot the results

Week 7: Behind regression: assumptions, diagnostics, and evaluation

Lecture 7: Regression assumptions, diagnostics, and evaluation *Instructor: Aron Szekely*

Outline:

- The assumptions: Validity; Representativeness; Linearity; Independence; Normality; Equality of variance
- Diagnostics: Residuals vs. fitted; Histogram of residuals
- Other issues: Influential points; Multicollinearity
- Solutions: Adding independent variables; Transforming variables; Using robust estimation procedures; Excluding influence data points; Removing or combining independent variables

Required readings:

- Gelman Chapter 11.
- Kellstedt and Whitten Chapter 9, Section 9.5 and the last parts of Chapter 11 (sections 11.4 and 11.5)

Lab 7: Regression diagnostics and evaluation *Instructor: Arturo Bertero*

Outline:

- Testing for regression assumptions with real data: Detect heteroskedasticity and fix it; Detect Non-Normality and fix it; Test for Linearity
- Influential points in real data: How to check for influential points; Calculate the Variance Inflation Factor (VIF)
- Multicollinearity in real data: Inspect the correlation table

Voluntary Recap 2

Instructor: Gonzalo Franetovic

During this voluntary 2-hour session students have the opportunity to recap concepts and their application in R, ask general questions, and get their statistical skills up to speed.

Week 8: Introduction to logistic regression

Lecture 8: Modelling probabilities *Instructor: Francesco Colombo*

Outline:

- From continuous to dichotomous outcomes
- Probabilities, odds, and odds ratios
- From linear to logistic regression
- Interpreting the logistic regression
- Predictions and comparisons: Odds-ratios; Marginal effects
- Model fit, assessment, and assumptions

Required readings:

- Gelman et al. (2020): Chapter 13 (only pp. 217-226)
- Agresti (2018): Chapter 15

Lab 8: Logistic regression *Instructor: Arturo Bertero*

Outline:

- Binary outcomes using real data: Analyze associations with categorical variables
- Odds and odds ratios: How to calculate them using real data
- The linear probability model with real data: How to estimate and plot the results
- Logistic regression: How to estimate effects in logged odds
- How to interpret the results: Odds ratios; Marginal effects; Predicted probabilities
- Differences between predicted probabilities and OLS: Marginal effects

Assignment 2

The second assignment will reflect concepts and tools discussed during weeks 6 to 8. Assignments must be done individually (no group work) and will be graded. Hand-in via email.

Week 9: Causal inference

Lecture 9: Using regression models for causal inference *Instructor: Daniel Auer*

Outline:

- Basics of causal inference: Potential outcomes, counterfactuals, and causal effects; The fundamental problem of causal inference; The problem of selection bias; Assumptions underlying causal inference
- The logic of randomized experiments: Sample, Conditional, and Population average treatment effects; Problems with self-selection into treatment groups
- Using regressions to estimate average treatment effects in experimental settings: Interpreting regression coefficients as causal effects; Pre-treatment covariates, treatments, and potential outcomes; Including pre-treatment predictors

Required readings:

- Gelman et al. (2020): Chapter 18 (only pp. 339-346)
- Gelman et al. (2020): Chapter 19 (only pp. 363-367)

Supplementary readings

- Kellstedt & Whitten (2018): Chapter 3

Lab 9: Statistical analysis in practice *Instructor: Gonzalo Franetovic*

Outline:

- Aggregating and transforming the data
- Data cleaning and merging
- General recap

Week 10: Statistical analysis in practice

Lecture 10: The good, the bad, and the ugly: a statistical western *Instructor: Aron Szekely*

Outline:

- The puzzle: are we wrong about time and causality? (Bem, 2011).
- Consequences
- We must do better: transparency, honesty, accuracy, and good data
- Suggestions for better practice: Exploratory vs. confirmatory research; Pre-registration; Substantive and statistically significance; Reducing flawed statistical reasoning; The importance of data
- *p*-values and Bayes' theorem

Required readings:

- Gelman et al. (2020): Chapter 4 (only pp. 49-60)
- Bem (2011), *Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect*

Supplementary readings:

- Tim Harford - How to Make the World Add Up (2020)
- https://www.smbc-comics.com/comic/science-fictions
- Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability

Lab 10: Statistical analysis in practice *Instructor: Gonzalo Franetovic*

Outline:

- How to analyze data: Replication of a published paper (Acemoglu (2011); Simulation of Simmons et al. (2011)
    - Good and bad practices in data analysis:

Required readings:

- Gelman et al. (2020): Chapter 4 (only pp. 60-67)

Final Assignment: Essay