APPLIED STATISTICS FOR SOCIAL AND POLITICAL SCIENCES

Year 2021/2022

Spring term, Year 1

Instructors: Krzysztof Krakowski, Camilla Borgna, Elena Pisanelli, Aron Szekely

Hours: 40 (20 hrs of lectures, 20 hrs of STATA lab)

Timetable

Starting date: Thursday, February, 10
Lectures: Thursday, 17.30-19.30
Labs: Friday, 17.30-19.30
Classroom: 1 [except February 18: on Zoom only]
All lectures and labs will be streamed online (Zoom) and recorded

Requirements

No formal requirements in the Allievi Program.

Recommended preliminary readings:

- Agresti, A. (2018). *Statistical Methods for the Social Sciences*, (fifth edition) Boston: Pearson. Chapters 1-3 (covering the basics of descriptive statistics (types of variables, measures of central tendency, measures of dispersion, cross-tabulation, scatter plots, bivariate descriptive statistics e.g. correlation coefficient).

- Kellstedt, Paul M., and Guy D. Whitten (2018). *The fundamentals of political science research* (third edition). Cambridge: Cambridge University Press. Chapter 6

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

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

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.

Outline

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 STATA.
You are expected to read the assigned material before each class.

Week 1: Review of descriptive statistics

During this week, we will review the main concepts and tools of descriptive statistics.

Lecture 1: Review of basic concepts in descriptive statistics
February, 10
Instructor: Krzysztof Krakowski

Outline:

- Definitions
  - statistical science (methods for design, description, and inference from data)
  - population and sample -> statistics and parameters
  - descriptive vs inferential statistics
- Variables
  - measurement scale:
    - 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
- Variability of the data
  - range
  - standard deviation: formulation, properties (scaling issue), empirical rule
  - interquartile range
  - outliers
  - z-score
- 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 analysis in STATA
February, 11
Instructor: Krzysztof Krakowski

Outline:
- Getting familiar with STATA
    - Output window
    - Command window
    - Do files
- General commands: help, display, cd, use, insheet, edit, save
    - Inspecting your data: describe, sort, order, summarize
- Data manipulation: generate, egen, drop, keep, rename, recode, encode

- Describing data: univariate statistics
    - Relative frequencies: tab, graph bar, histogram
    -

Before class:
- Get access to STATA on your laptop
- Check you can open the data file sent to you by the instructor

Week 2: Understanding probability, understanding your data

During this week, we will introduce you to the logic of regression and why it is a fundamental statistical tool for social science. In the lab session, we will dig deeper into data visualization, a key step to understanding your data.

Lecture 2: Why learning regression?
February, 17
Instructor: Krzysztof Krakowski

Outline:
- 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
- Regression—definition & examples
- Regression—uses
    - predictions (forecasting new data using existing observations)
    - exploring associations

    - extrapolation (adjusting for a mismatch between the sample and the population of interest)
    - causal inference
- Regression—interpretation
    - interpreting coefficients
- Building blocks of regression modelling
    - weighted mean
    - vectors & matrices
    - graphing a line
    - logarithmic transformation
    - linear transformation (standardizing)
- Probability distribution
    - normal distribution
    - lognormal distribution
    - binomial distribution
    - Poisson distribution
    - real data

Required readings:

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

Lab 2: Data visualization in STATA
February, 18
Instructor: Krzysztof Krakowski

Outline:

- Recap
    - data manipulations
    - descriptive statistics

- Plotting relations between variables
    - scatter plots -- twoway scatter –
    - linear fit
    - using labels to mark observations

Required readings:

- Gelman et al. (2020): Chapter 2

***
Homework for sessions 1-2
***

Week 3: Statistical inference

This week we will look at the fundamentals of statistical inference. We will address what statistical inference is, why you do it, and some issues with it. We will do this both by hand and using STATA.

Lecture 3: Statistical inference by hand
February, 24
Instructor: Aron Szekely

Outline:
- Statistical inference: what is it and why do we need it?
- Samples and populations
- The 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 inference in STATA
February, 25
Instructor: Elena Pisanelli

Outline:
- Correction of homework
- Samples and population:
    - Possible sources of data
    - Random sampling
    - Normal distribution
    - Kernel density
    - Standard deviation and z score by hand and with STATA
- Random sampling from simulated data
    - How to create simulated data
    - How to random sample from simulated data
    - How to plot the distribution of the data
- Sample mean and population mean
- Standard error by hand and with STATA
- Confidence intervals and p-value
- Hypothesis testing in practice
    - Plot the distribution and test H0 and H1
    - Ttest
    - P-value
- New commands in the lab: sum, kdensity, gen, egen, clear, set obs, sort, hist, ttest

Required readings:

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

Week 4: Statistical analysis in practice

In 2011 Daryl Bem proved that people can perceive the future. Is this really possible? What does Bem's paper really show? We begin with this paper (which is important albeit for entirely unintended reasons) and use it as a stepping stone to examine bad practice, its prevalence and consequences. This will help us learn about good practice.

Lecture 4: The good, the bad, and the ugly: a statistical western
March 3
Instructor: Aron Szekely

Outline:

- The puzzle: are we wrong about time and causality? (Bem, 2011).
- An answer…
- Is this an exception?
- 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 4: Statistical analysis in practice using STATA
March 4
Instructor: Elena Pisanelli

- Recap of the previous lab
- Tip of the day: How to organize the STATA window and use a dofile
- How to analyze data in STATA

    - Replication of a published paper (example: Acemoglu (2011))
    - Simulation of Simmons et al. (2011)
- Good and bad practices in data analysis
    - Put in practice the advice you received in the theoretical class
        - Show variables' distribution
        - Show summaries
        - Plot the distribution across groups
- Interpreting findings in STATA: how to interpret the table of a ttest
- New commands in the lab: local, cd, use, ssc install, keep, drop, use of if, use of by, graph box

Required readings:

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

Week 5: Bivariate hypothesis testing

This week will delve into inferential statistical analysis starting from the simplest scenario, where you want to establish whether a difference that you observe in your sample can be taken as indicative of an actual association between two variables in the population.

Lecture 5: Concepts and tools to test bivariate relations
March 10
Instructor: Camilla Borgna

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 5: Bivariate hypothesis testing in STATA
March 11
Instructor: Elena Pisanelli

Outline:

- Recap of last lab
- Tip of the day: How to edit graphs in STATA (change labels, …)
- Group comparisons in STATA 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 in STATA
    - Overlapping density plots for two groups in STATA
    - Box plot for a continuous variable in two groups in STATA
    - Dot plot for a continuous variable comparing two groups in STATA
- New commands in the lab: tab, row, col, chi, pwcorr, spearman, hist…by, twoway kdensity… || kdensity…, ciplot

No assigned readings to do before class.

***
Homework for sessions 3-4-5
***

Week 6: Linear regression with one predictor

This week will present regression analysis as a key tool of statistical inference. It will focus on linear regression with one predictor, considering how it works, what it can be used for, and how to apply it in practice.

Lecture 6: Concepts and tools for linear regression modelling
March 17
Instructor: Camilla Borgna

Outline:

- Introduction to regression models
    - Historical origins of regression
    - Using regressions for prediction or comparison
    - Linear and non linear regression models

- Which line fits best?
    - Population and sample regression models
    - Fitting a line: Ordinary least-square (OLS) regressions

- Measuring uncertainty about the regression line

    - Goodness-of-Fit: R-Squared Statistic
    - Confidence Intervals about Parameter Estimates

- Descriptive and causal interpretations of regression
    - Interpret 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 6: Linear regression modelling with a single predictor in STATA)
March 18
Instructor: Elena Pisanelli

Outline:

- Correction of homework
- Tip of the day: exporting tables and saving graphs in STATA
- Preliminary steps of data inspection: using real data
    - Outliers
        - Graphically detect outliers and assigning labels to recognize them
- OLS with real data
    - Bivariate regression
    - Regression output: how to read it
    - Interpreting coefficients
        - Different dependent variables (continuous, categorical, dummy)
        - Different independent variables (continuous, categorical, dummy)
- Export regression results
    - Tables and graphs

STATA

- New commands in the lab: asdoc…save, tabout…using, graph save, graph export, reg, twoway scatter, mlab(variable), when to use = or ==, outreg2, margins, marginsplot

Required readings:

- Gelman et al. (2020): Chapter 6

Week 7: Linear regression with multiple predictors

This week will extend the treatment of regression analysis considering how you can include multiple predictors in a linear model and how to test for their possible interactions.

Lecture 7: Introducing multiple predictors in your regression model
March 24
Instructor: Camilla Borgna

Outline:

- Dealing with more than two variables
    - Control variables, omitted variables and spurious associations

- Regressions with multiple predictors: an example with two predictors
    - Model A: A binary predictor
    - Model B: A continuous predictor
    - Model C: Including both 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

- Standardization
    - 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 7: Linear regression modelling with multiple predictors in STATA
March 25
Instructor: Elena Pisanelli

Outline:

- Recap of last lab
- Spurious associations in real data
- Multiple regression with real data
    - Add lurking variables

    - Compare regression outputs with and without lurking variables
    - Logarithmic transformation and how to interpret them (Lin-log, Log-lin, Log-log)
    - 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
- Standardization of the coefficients
    - How to obtain regression coefficients that are comparable in magnitude
- New commands in the lab: est store, ftest, ##, option beta

No required readings to do before class.

Week 8: Behind regression: assumptions, diagnostics, and evaluation

Can you really trust your model? During this week, we will consider the assumptions behind regression analysis, learn how to check for these, and understand some approaches to solve or minimise the issues.

Lecture 8: Regression assumptions, diagnostics, and evaluation
March 31
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 8: Regression diagnostics and evaluation in STATA
April 1
Instructor: Elena Pisanelli

Outline:

-
- Recap of last lab
- 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
- New commands in the lab: rvfplot, rob option, predict residuals, dfbeta, list, corrtable, vif

No required readings to do before class.

***
Homework for sessions 6-7-8
***

Week 9: Introduction to logistic regression

Often, in the social sciences our dependent variable is not continuous but dichotomous. You may need to model the probability that a given event occurs, or that your analytical units display a given state. As linear regressions are not well suited to model probabilities, you need to resort to other modelling strategies. This week, you will learn the basics of logistic regression and its applications.

Lecture 9: Modelling probabilities
April 7
Instructor: Aron Szekely

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 9: Logistic regression in STATA
Instructor: Elena Pisanelli
April 8

Outline:

- Correction of homework
- 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 with real data
    - How to estimate effects in logged odds
    - How to interpret the results
        - Odds ratios
        - Marginal effects
        - Predicted probabilities
    - Differences between predicted probabilities and OLS
        - Compare marginsplots
        - Using variables at the mean in marginsplot
- STATA New commands in the lab: logit, margins dydx

No assigned readings to do before class.

Week 10: Statistical and causal inference

This final week discusses the difference between statistical inference and causal inference and highlights under which circumstances and how statistical analysis can be used when you want to address a causal research question. Next year, you will pick these topics up again from the perspective of research design in the course "Advanced Research Design" and from the perspective of econometrics in "Applied economics".

Lecture 10: Using regression models for causal inference
Instructor: Camilla Borgna
April 14

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 10: Regression models for causal inference: applications in STATA
Instructor: Elena Pisanelli
April 15

- Catchup

If time allows:

- Aggregating and transforming the data: reshape, merge, append, collapse, bysort
- Q&A session
- Lab Notes on the steps needed when analyzing the data