Analysis of cross-sectional data

Michael Kühhirt
michael.kuehhirt@uni-koeln.de

Winter 2019/20

Office Hours: Wed 10:00-11:30am
Office: Universitätsstr. 24, 0.05
Web: bit.ly/2oRhzZ9

Class Hours: Wed 12:00-13:30pm
Class Room: 211-H114, Herbert-Lewin-Str. 2
Teaching Assistants: Charlotte Becker, Daria Tisch

Class content

How are math skills distributed in school children? Are there differences between boys and girls? Can election outcomes be predicted based on politicians' looks? Does having a university degree reduce xenophobia? This course provides an introduction to basic quantitative methods for answering these and other social science questions using cross-sectional data, that is, data which include one observation per unit of analysis (e.g., persons, families, firms, countries). The follow-up course next semester focuses on methods to analyse data with multiple observations per unit (i.e., longitudinal data).

The course is divided into four main parts: (1) cross-sectional data, their exploration and description, (2) statistical tools to model cross-sectional data (i.e., Regression models and algorithms to estimate their parameters), (3) statistical uncertainty and model selection, (4) causal inference. The goal is to begin preparing students to analyze data to answer their own research questions and to enable them to critically evaluate quantitative analyses portrayed in the media, in politics, and in academic research.

Each week, key concepts are presented and discussed in a 90min lecture. Every lecture is accompanied by an exercise, which is used to review course contents and to practice statistical techniques in `Stata` labs. The exercise is offered in three alternative time slots, which are allocated on a first-come-first-serve basis:

- Wed 5:45-7:15pm, 107b PC-Pool B III
- Thu 2:00-3:30pm, 107b PC-Pool B III
- Thu 4:00-5:30pm, 107b PC-Pool B III

Version: 2019-12-31

Goals and learning objectives

The course has three broad goals, which I enumerate below. Each goal is associated with a number of specific learning objectives, which you should be able to perform as the class concludes.

1. Understanding methods to describe and model cross-sectional data. Upon course completion students
   - explain common descriptive statistics and regression techniques,
   - interpret their results, and
   - discuss their underlying assumptions.
2. Implementing these methods in the statistical software Stata. This includes that students
   - prepare data for analysis,
   - select appropriate `Stata` commands, and
   - use analytical scripts to ensure reproducibility of their analyses.
3. Applying these methods to answer substantive research questions. In this process, students
   - choose an appropriate method depending on the research question,
   - infer a (tentative) answer to the research question, and
   - identify limitations of their analyses.

Overall, the course should equip you with a solid understanding of common statistical techniques, which provides a basis for learning about and applying more advanced methods. Ideally, the course not only helps you advancing your studies but also prepares you to conduct data analyses for industrial applications and to evaluate statistical claims in the media and elsewhere.

Requirements and grading

The requirements to successfully complete the course are as follows:

- **Written exam:** There will be a 60-minute exam, which takes place at two alternative dates, one at the beginning and one at the end of the semester break. You can achieve a maximum of 60 points in the exam. A minimum of 30 points is required to pass the exam and thus to successfully complete the class.
- **Take-home assignments:** There will be two assignments in which you apply the course contents to answer specific research questions by analyzing real-world data, one due on December 23, 2019, and one due on February 3, 2020. The assignments are passed (but not graded) by completing them at least 50% successfully. You must pass both assignments in order to pass the course. *While I strongly encourage you to discuss the assignments (and class contents in general) with other participants, submission of assignments that are identical in full or in part will result in a failing grade for all participants involved.*

For successful course completion you're awarded 9 credit points. Please be aware that this corresponds to a workload of 270SWS (i.e., 13.5h/week). Only 3h/week are allotted for in-class learning (i.e., lecture and exercise). Therefore, course completion requires a substantial time investment outside class (i.e., up to 10.5h/week), including preparation and review of class contents and labs, assignments, and reading.

The final grade is awarded according to the following rules:

Grade	Points
1.0	$60 \geq P \leq 58$
1.3	$58 > P \leq 55$
1.7	$55 > P \leq 51$
2.0	$51 > P \leq 48$
2.3	$48 > P \leq 45$
2.7	$45 > P \leq 42$
3.0	$42 > P \leq 39$
3.3	$39 > P \leq 36$
3.7	$36 > P \leq 33$
4.0	$33 > P \leq 30$
n.p	$30 > P \leq 0$

Given you pass the exam and both assignments, you may improve your grade by a maximum of 0.7 by collecting up to 6 **activity points** (AP). To unlock the ability to collect AP, you must correctly complete at least 80% of an online quiz (Quiz 0) scheduled for the second course week. There are three ways to collect AP:

- **Review quizzes:** There are 10 quizzes scheduled for the first 15min of the exercise time slots (see class schedule below). These quizzes contain questions on class contents of the previous week that are specific to each of the three exercise groups. Because quizzes are completed online, physical attendance of the exercise isn't required. However, you can only participate in the quiz for the exercise group that you were initially assigned to. You can collect up to 5 points per quiz, and thus 50 points in total. Total quiz points are divided by 10 upon course completion, yielding a maximum of 5 AP.
- **Asking "good questions":** You can also collect AP by asking questions on Piazza (see class resources below). A "good question" pertains to a specific course content (but not course organization etc.), is clear and concise, and is subsequently marked as "good question" by the instructors. You'll receive 1 AP for each 5 good questions.
- **Providing "good answers":** The last way to collect AP is to give "good answers" to questions on Piazza that were previously marked as "good" by an instructor. A "good answer" is the first to adequately address a "good question" on Piazza within 24h, is well-structured, and is endorsed by an instructor. For each 3 good answers, you are awarded 1 AP.

Attempts at cheating may lead to partial or complete termination of the AP program at any time.

Class resources

Textbook

- Fox, J. 2016. *Applied Regression Analysis and Generalized Linear Models. Third Edition.* Thousand Oaks: Sage. Errata | Appendices

Ilias:

- Course materials (slides, readings, data, etc.)
- Quizzes
- Assignment submission
  https://www.ilias.uni-koeln.de/ilias/goto_uk_crs_2994321.html

Piazza: This is an online platform for fast and efficient class Q&A. Rather than emailing instructors your questions, please post them on Piazza. Doing so will also enable you to collect activity points to boost your grade. Piazza is also available as a cost-free app for Android and iOS. You can find the course page here: https://piazza.com/uni-koeln.de/fall2019/bmi/home

SoFS: Free online storage at University of Cologne. Find information on how to use it here: https://rrzk.uni-koeln.de/sofs.html?&L=1

Other online resources

Statalist (Stata online forum): https://www.statalist.org/

StackExchange (Online forums for statistics etc.): https://stackexchange.com/

StatQuest (Youtube channel with helpful statistics videos):
https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw

Further reading

Quantitative social research:

- Firebaugh, G. 2008. *Seven Rules for Social Research.* Princeton: Princeton University Press.
- Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks: Sage.
- Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data. Second Edition.* Cambridge, MA: MIT Press.

Causal inference:

- Hernán, M. A. and Robins, J. M. 2019. *Causal Inference.* Boca Raton, FL: Chapman & Hall/CRC.
- Pearl, J., Glymour, M., and Jewell, N. P. 2016. *Causal Inference in Statistics: A Primer.* West Sussex, UK: Wiley.
- Pearl, J. and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect.* New York: Basic Books.

Stata:

- Kohler, U. and Kreuter, F. 2012. *Data Analysis Using Stata. Third Edition.* College Station, TX: Stata Press.
- Long, J. S. 2009. *The Workflow of Data Analysis Using Stata.* College Station, TX: Stata Press.

Class schedule (subject to change)

Introduction

Week 1: October 9/10

- Lecture: Overview and course mechanics
- Lab: Course setup, Introduction to Stata
- Reading: Fox Ch. 1, Syllabus, Stata Cheat Sheets 1+2

Cross-sectional data

Week 2: October 16/17

- Lecture: Measurement, research design, data
- Lab: Data exploration and processing
- Reading: Fox Ch. 3, Stata Cheat Sheets 2-4
- Quiz 0: Course mechanics (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Week 3: October 23/24

- Lecture and lab: Summarizing data
- Reading: Fox Ch. 3, Stata Cheat Sheets 2-5
- Quiz 1: Week 2 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Statistical models for cross-sectional data

Week 4: October 30/31

- Lecture and lab: Linear regression I
- Reading: Fox Ch. 5.1 + 7.1/2, Stata Cheat Sheet 6
- Quiz 2: Week 3 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)
- Assignment 1: Statistical models and uncertainty (due December 23)

Week 5: November 6/7

- Lecture: Linear regression II
- Lab: Recap and Assignment 1
- Reading: Fox Ch. 5.1
- Quiz 3: Week 4 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Week 6: November 13/14

- Lecture and lab: Probability models I
- Reading: Fox Ch. 14.1.1/3
- Quiz 4: Week 5 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Week 7: November 20/21

- Lecture and lab: Probability models II
- Reading: 14.1.1/3, StatQuest video
- Quiz 5: Week 6 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Week 8: November 27/28

- Lecture and lab: Modelling nonlinearities
- Reading: Statistics 101 video 1 + Statistics 101 video 2
- Quiz 6: Week 7 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Week 9: December 4/5

- Lecture and lab: Multiple predictors and interactions
- Reading: Fox Ch. 5.2+7.3+14.1.4
- Quiz 7: Week 8 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Statistical uncertainty

Week 10: December 11/12

- Lecture and lab: Quantifying uncertainty
- Reading: Fox Ch. 6
- Quiz 8: Week 9 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Week 11: December 18/19

- Lecture and lab: Statistical models with uncertainty
- Reading: Fox Ch. 6
- Quiz 9: Week 10 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Week 12: January 8/9

- Lecture: Further aspects of statistical models for sample data
- No exercise/lab
- Reading: Fox Ch. 22.1+3, StatQuest video

Causal inference

Week 13: January 15/16

- Lecture and lab: Correlation and causation
- Reading: Elwert pp. 245-252
- Assignment 2: Causal inference (due February 3)

Week 14: January 22/23

- Lecture and lab: Covariate selection
- Reading: Hernan & Robins Ch. 18.1/2 + Elwert pp. 256-261
- Quiz 10: Week 13 review (Wed 5:45pm/Thu 2:00pm/Thu 4:00pm)

Wrap up

Week 15: January 29/30

- Lecture: Recap, evaluation, Q&A
- Lab: Assignment, Q&A