Syllabus: Data Analysis
Spring 2018

Last updated: April 23, 2018

Course information

- Course title: Data Analysis
- Course number: POLSC-UH 2211
- Credits: 4
- Term: Spring 2018 (14 weeks)
- Lecture location: A2-004
- Lecture hours: MoWe 1:15PM - 2:30PM
- Office hours: We 230-330PM, Office A5 147
- Course prerequisites: Statistics (SOCSC-UH 1010Q Statistics for the Social and Behavioral Sciences). Motivated students can take it simultaneously with this course. A knowledge of basic statistical ideas and inference (t-tests, standard error, p-value, etc.) is recommended.

Instructor information

- Instructor: Dr. Peter van der Windt
- Email: petervanderwindt@nyu.edu
- Website: www.petervanderwindt.com
- Office: Building A5, Office 147
- Dr. Jonathan Rogers will help us during the course:
  - Email: jdr7@nyu.edu
  - Office hours: Flexible by appointment

Course description & learning objectives

This course introduces concepts and methods necessary to facilitate sophisticated data analyses. Participants who successfully complete this course will be able to:

- Explain and implement various statistical tests and regression techniques.
- Consume and criticize complicated methodological research.
- Write computer programs in the R language.
- Diagnose problematic results in a regression framework.
- Distinguish causal and descriptive research designs.
- Create custom data visualizations.
- Present and interpret quantitative analyses.
- Make arguments using data and analyses.

Books

I teach data analysis in a very hands-on way. As such, I try to avoid standard mathematical treatments of statistics found in many books. That said, I have picked out several books that will help you deepen your understanding of data analysis and statistics (the last three should be in the bookstore):

- "OpenIntro Statistics" textbook is freely available at `http://www.openintro.org/stat/textbook.php`. This is a no-nonsense book.
- Andy Fields' "Discovering Statistics Using R" is a bit more reader friendly.
- "Interpreting and Using Regression" by Achen will also be available from the bookstore to round out our discussions.
- "Quantitative Social Science: An Introduction" by Kosuke Imai will be especially useful when we discuss causal inference.

In addition, I will provide supplementary slides and code as the course progresses. Additional readings will be available on reserve in the library, or available on Newclasses.

Software

I strongly recommend that students use R for the course, but students may use whatever statistical software they choose. The two most prominent choices are STATA and R. We will discuss the trade-offs between these two statistical tools in order to find the one that best suits your needs. Class demonstrations will be in R, given its power, breadth, and flexibility. R is versatile, fast, and open-source. You can download it at `http://www.r-project.org/`. R has a steep learning curve, and may be frustrating at first. The upside of learning R is flexibility and popularity. Many different disciplines use R. Because R is open-source, vast libraries of R code are out there for you to use. R is excellent for data visualization, GIS/mapping, and is a skill in high demand outside of academia. Rstudio (`http://www.rstudio.com/`) is an excellent workspace for organizing and implementing your analyses in R. Install it after installing R.

I recommend that you learn to use R markdown (`http://rmarkdown.rstudio.com/`) or LaTeXto prepare homeworks and exams (this syllabus is written in LaTeX). This allows you to streamline data analysis, creation of tables and visualizations, and written responses to home work or practicum questions. I will accept homeworks composed in Word or another word processor, though. LaTeXis a type-setting program that greatly simplifies document layout, mathematical notation, and bibliography management. To install LaTeX, Windows users should try `http://tug.org/protext/`; Mac users can try `http://tug.org/mactex/`. If there is demand/interest, I will schedule a introduction to R markdown and/or LaTeXearly in the semester.[1]

Requirements and grading

Your final grade will consist out of four components: four homeworks (5% each), one practicum (30% each), final paper (40%), and class participation (10%).

- **Homework:** There will be 4 homework assignments building on material covered during class or in readings. Homework will usually be due one week from the day it is assigned. You are allowed to work in groups, but you are required to write-up your own results and explanations individually. Bottom line: what you submit for your homework should be your

[1] If you plan on going to graduate school in the social sciences, definitely learn R and LaTeX. If you would like to work with me as a research assistant or co-author, definitely learn R and LaTeX.

own work. If you do work in a group, you are required to list the members of your group at the beginning of your homework. Homeworks will be graded on a binary scale (0/1). In the event that you receive a zero, there is a one-week rewrite policy, assuming that you handed the assignment in on time and made a good-faith effort to complete the assignment.[2] That is, you will have one week to revise and resubmit the assignment in the event you receive a zero. Late assignments will not be accepted except in the case of a documented emergency. For grading purposes, a 1 will represent a grade of 95%. Especially well-organized, lucid, or creative homework responses may be give a 1+, corresponding to 100%. I highly recommend that you use LaTeXor Rmarkdown to prepare your homework. This will allow you to make reproducible documents that include both your code and your write-up in one convenient place. Moreover, these systems make documents that look far superior to Word, and manage your bibliography and formatting for you.

- **Practicums:** Once during the semester, I will provide you with a data set and a series of questions about the dataset. You will then write a polished two-page report answering the questions at hand in simple, non-technical terms. The practicum challenges you to apply the material covered in class to a real-world dataset, and write up your analysis and findings. That is, the practicum is an opportunity to actually do data analysis. We will devote one session to understanding what makes a good practicum response. The practicum exercise will be sent to you via email and responses should be submitted via email within two weeks. Students are not allowed to work together or discuss practicums under any circumstances. If you have a question about the practicum scenario, send an email. I — or the instructor — will then anonymize your question and respond to the entire class. The goal of the practicum is not to display technical expertise alone, but rather to demonstrate that you can use data and statistical analysis to make an argument regarding a set a questions.
- **Final paper:** The final project is due on the last day of class and should be approximately 10 pages long. The paper should employ at least one of the methodological tools covered in class to answer a social science question of substantive interest using data assembled by the student. Students should submit a one paragraph description of the final project and data to be used to the instructor by the sixth week of class. The proposal should describe the substantive question of interest, a (very) preliminary research design, and discuss the sources of data. The final draft of your project will be due the last week of class. If you choose to turn in your paper late, the highest grade for which it will be eligible will be reduced by one step (e.g., A to A-) for every 12 hours or part thereof. NB: If you turn in a rough draft of your final project by week 11, I will read your draft and provide comments to help improve your final draft. You are expected to work consistently on the final project throughout the semester, meeting with me or Jonathan Rogers as needed. Please see the appendix for an example.
- **Participation:** Class participation makes up 10% of your final grade. Participation will be observed two main ways. First, class attendance and active participation: be sure to ask questions during class and help your classmates. Second, you will give a 10 minute presentation in mid-April on your final project topic in order to get feedback.

Course schedule

The following schedule is a partial list of topics that we will cover during the semester. The first weeks cover basic material that you should (but probably don't) remember from your introductory statistics class. The next set of weeks cover linear and logit regression – workhorse models in the social sciences. Finally, we will cover specialized topics of interest to the class. Possibilities include: causal inference, missing data, advanced data visualization, spatial data/mapping in R, etc. The

[2] As judged by the instructor.

topics for these last set of weeks remain flexible, allowing us to decide to explore other topics that may be of interest to the class.

Week 1: jan22, jan24

- Introductions
- Topic: Course overview; review of syllabus; research question; hypotheses; etc. Introduction to R.
- Reading: OpenStatistics (3rd edition), sections 1.1 - 1.8; Field, chapter 2.

Week 2: jan29, jan31

- Topic: Distribution; mean; standard deviation; sampling; estimating a sample mean; thinking about uncertainty.
- Reading: OpenStatistics, section 3 (mainly 3.1 and 3.2); section 4 (mainly 4.1, 4.4).

Week 3: feb5, feb7

- Topic: Standard error; confidence intervals; ggplot2; hypotheses; z-scores; p-value.
- Reading: OpenStatistics 1.3; 1.6; 4.2.
- Task: HW1 distributed

Week 4: feb12, feb14

- Topic: Hypothesis testing; categorical variables; estimating mean of a proportion.
- Reading: OpenStatistics 4.1, 4.2, 4.3, 5.1, 5.3, 6.1, 6.2.

Week 5: feb19, feb21

- Topic: More hypothesis testing; one and two sample t-test; relationships between variables; conditional expectation.
- Reading: OpenStatistics 4.1, 4.2, 4.3, 5.1, 5.3, 6.1, 6.2.
- Task: HW 2 distributed

Week 6: feb26, feb28

- Topic: Linear regression with dummy variables and continuous variables; Ordinary Least Squares; assumptions.
- Reading: OpenStatistics 7.1, 7.2.

Week 7: mar5, mar7

- Topic: More linear regression; interpreting regression output; predicting outcomes; quick and easy data manipulation.
- Reading: OpenStatistics 7.2, 7.4. Imai chapter 4.

Springbreak: *mar12*, *mar14*

Week 8: *mar19*, mar21

- Topic: Even more linear regression; multivariate regression; statistical and substantive significance; interaction effects.
- Reading: OpenStatistics 7.4, 8.1, 8.2., 8.3. Plus scan: **?**.
- Task: Practicum. Distributed: mar21. Due: apr4 before midnight.

Week 9: mar26, mar28

- Topic: Presentations of research paper ideas
- Eating: This will be difficult work and we will get hungry. What kind of food should I bring? Email me.

Week 10: apr2, apr4

- Topic: Outliers; missing data.
- Reading: OpenStatistics 7.3.
- Task: HW3. Distributed: apr4. Due: apr11 before midnight.

Week 11: apr9, apr11

- Topic: What to do when your outcome isn't continuous; logistic regression.
- Reading: OpenStatistics 8.4.

Week 12: apr16, apr18

- Topic: Correlation is not causation; causal inference; random assignment to treatment.
- Reading: Imai, Chapter 2.
- Task: HW4. Distributed: apr18. Due: apr25 before midnight.

Week 13: apr23, apr25

- Topic: Matching; regression discontinuity.
- Reading: Imai, Chapter 4.3.

Week 14: apr30, may2

- Topic: Spatial data; mapping with R.

Last class: may7

- Topic: What next? Where do you go from here? Ideas on how to continue learning about R and statistics once you leave this course.
- Task: Final project due may7 before midnight.
- Wrapping things up

Acknowledgement

Thanks to Andy Harris who taught this course before me and on which the course material (including this syllabus) relies heavily.

Appendix: example final paper

- For this final paper, please use a linear model to answer a social science question of substantive interest using data assembled by the student.
- Document is due before midnight on the last day of class.
- Approximately 10 pages of main text. Appendices are allowed (but please not many pages).
- Please write the document in Rmd: i.e. have both the text and the code in your document.
- Please hand in both your .Rmd and a PDF. For the PDF we do not have to see the code, just the 'final product'. We'll check the .Rmd to see if the code you wrote make sense.
- In this exercise you can show how much you learned in class.
- Writing a document like this is not easy, but an important skill to have. It follows the same rules that you would use for writing an academic paper. Below I list several of them. Please stick to this. The Do's:
  - Use writing that people understand. No need to write more complicated than necessary. "I have set out to argue" can also be written as "I will argue".
  - Be academic in your writing. Do not use words like "fascinating", "amazing", etc. That's your opinion. Readers don't care about your opinion. The data will provide evidence.
  - It also means that you have to be precise. Avoid words like "a lot" or "in the beginning" if you can be more precise.
  - Use references. If you use a number somewhere (e.g. in Congo unemployment is 95%), you need to add a source.
  - Use the same tense throughout your document. If you start writing in the present tense, do that throughout your document.
  - Hypotheses should follow from the literature/motivation. Your hypotheses should not come out of the blue. If your hypothesis is "Women in the Congo work more than men in the Congo", your literature/motivation section should not (only) contain information about poverty in the Congo, but also about differences between men and women in the Congo.
  - All pages should have a page number.

  The Dont's:

  - Do not have hypotheses that are a combination of multiple hypotheses. E.g. "X leads to increases in Y and does not decrease Y." These are two hypotheses. Write them out as two sentences.
  - Do not make value judgements. We are academics.
  - Avoid contractions: write "they've" as "they have", "can't" "as cannot", etc.
  - When writing up a reference, avoid writing the full name of the authors, the title of their book or article, etc. Use referencing as is given in the example on the next page.
  - No need for a coverpage and pictures. That's for secondary school.
- Below is an example document with pointers. It is only a subset of potential things you can write down, but it serves as an idea. Needless to say, your document will have more meat and is more detailed. You can add sections and take sections out, if you want to. The page numbers are just just indications.

Title
Name: Peter van der Windt[3]

1 Introduction

- About 1 page
- Write why we should care about the topic
- Give a few references to other people working on this topic and what they found, and how you are going to improve or add to this literature
- State your research question

2 Hypotheses

- About 1/2 page
- Please explicitely state the hypotheses that you'll test
- Don't have too many hypotheses

3 Data

Given this is a course about data, this is an important section.

3.1 Data source

- Can be as short as one or two sentences
- Please state the source of the data

3.2 Variables

- About 2 to 3 pages
- Add a summary table about your variables, at the very least your independent and dependent variables (following from your hypotheses), and any control variables you might want to use. This table should have at least the following information: mean, standard deviation, minimum, maximum, number of observations.
- Discuss the variables that you'll use. Discuss the mean, the standard deviation, etc.
- Maybe add a figure (e.g. histogram) if you think it is interesting or illuminating to the reader.
- How many observations do you have? Any missing observations?
- Do you have any outliers? What are you going to do with it?

[3] Email: petervanderwindt@nyu.edu

4 Estimating equation

- About 1 to 2 pages
- Please write down the estimating equation(s) that you'll estimate (i.e. your model)
- Please discuss your model(s). What are the estimates that you're going to estimate: e.g. "Beta 1 gives the impact of X on Y".
- Are you standardizing your variables? Why, why not?
- Are you using logs, or do any other transformation with your data? Please discuss.
- If you do an interaction effect, please discuss this in detail.
- Say a few words about causality.

5 Results

- About 2 pages
- Think about how to best show your results. A table (you can use stargazer)? A figure (e.g. for interaction effects)?
- Discuss your results.
  - Use/explore terms that we used in class, like the R^2.
  - The magnitudes of the coefficients: "if X goes up with A then Y goes up with B". And whether this is a lot or little.
  - The significance of your results (discuss this in such a way that I can see you understand what e.g. p<0.05 means).
  - Etc.

6 Conclusion

- About 1 page
- Just briefly say one more time what we learned and why we should care about it
- Please give 1 or 2 policy recommendations

7 Appendix

- You can have an appendix but only do it if it is necessary