Quantitative Module

Statistics: "science dealing with data about the condition of a state or community"

Gottfried Aschenwall, 1770

University of Turku
Faculty of Social Sciences
Turku, Finland

Last updated: January 18, 2022.
Download last version here.

General Overview

Professor: Héctor Bahamonde, PhD.
e:hibano@utu.fi
w:www.HectorBahamonde.com
Office Hours: Schedule time with me here.

Place: TBA.
Time: TBA.

Course website: TBA.

TA: Valtteri Pulkkinen.
e: valtteri.s.pulkkinen@utu.fi
TA Bio: TBA.

Program: Master of Social Sciences, University of Turku.
Semester: Spring.
Credits: 2.
Timing: 4 modules of 3 hours each.

Motivation: Why take this course?

What's the effect of education on income? How can we evaluate the effectiveness of a public policy? Does legalizing some drug increase its consumption? Which political candidate will win the election? All these questions entail some kind of relationship between two social phenomenon (a.k.a. *correlation*). In this course we will learn (1) how to answer these questions from a quantitative perspective, (2) how to select the "right" quantitative method depending on the type of data we have (a.k.a. *functional form*), (3) how to quantify the amount of *uncertainty* we have (and why it matters), and (4) how to communicate quantitative results effectively (even for non-specialized audiences).

Public entities guide their strategic decisions based on quantifiable information, i.e., data. This decision-making process has taken even much more relevance nowadays where there has been a wave of data digitization, making available much more high-quality data. I believe this is a great opportunity for social scientists like ourselves, putting heavy pressures on us to learn how to analyze those data. If ten years ago we complained that there was not enough data, our problem now is different: there are so much data that we need to learn how to analyze it.

Though what we will learn this semester is highly mathematical and numeric, and thus, it might seem to you as "very scientific" or "irrefutable," don't get confused, these methods are *not* "bullet-proof:" they will *never* "proof" anything at all. During the semester, we will learn *inferential* statistics, that means that everything we do will come with some degree of **uncertainty**. All the time, we will also rely on *untestable* **assumptions**. **As we shall see, inferential statistics is more art than science**.

Depending on our progress during the semester, we will pay special attention to an issue that is absolutely relevant nowadays in applied social sciences: *causal inference*. Experiments are the gold-standard for making causal claims. However, often times conducting experiments is either too expensive, unethical, or impossible. Under the "right" circumstances, though, some times it is possible to get quasi-experimental statistical designs that might get us closer to the gold-standard. We will also discuss why the methods we will learn this semester are *not* causal, i.e., *correlation is not causation*.

Honestly, I hope this course captivates your enthusiasm, and gets you interested and curious about ways to study different social phenomena from a quantitative perspective, *tervetuloa*!

Description

Enrolled students will acquire a basic inferential statistics toolkit. The course will pay special attention to Ordinary Least Square regression (OLS) and a selection of Generalised Linear Models (such as logit/probit, multinomial, ordered models and/or rare events data generating processes)—the workhorses of quantitative social sciences. This course is very hands-on, and while some statistical theory will be covered, the core of it will be on data analyses and programming in R.

Overall, this course is an opportunity for students to make progress on their Master theses, particularly on the data analyses portion of it. For those matters, the actual content of the course will follow the students' research questions and data structure. Thus, during the course, students will perform real analyses on their own data (if they have those data already available), otherwise students will perform replications.

Organization

- ☞ I need you to **email/meet me two weeks before this module begins** so we can discuss about your Master thesis, particularly, (1) your research question and (2) your data structure. This will help me tailoring the actual contents of this course to your research project.
- The course will be taught in English in the computer lab as 4 sessions of 3 hours each.
- The course will be organized in different "Lessons." Each class we will try to cover as many lessons as we can.
- We will meet between early March and early May at the end of the week (Wed, Thu and/or Fri) starting at 11.00 AM. Exact dates will be confirmed later.

In terms of contents, this course will address four general topics.

1. Basic functions in R.

2. Descriptive statistics in R.
3. Introduction to lineal models in R.
4. Causal inference in R.

Programming

We will learn to program in R, the most-used programming language in social sciences. There are several advantages. R is free and runs on all platforms. Second, it's an object-oriented language. This implies—third—that R forces the student to think hard about what s/he is doing. Unlike other statistical packages such as Stata or SPSS, where the user "clicks and points," you have to tell R specifically what you need and how you need it. Fourth, if you know R, you can easily learn about other pieces of software.

Installing R. First, download R. Click on "CRAN" (upper-left corner), then select any "mirror" you want. Then select which version of R you will need depending on your OS (i.e., Windows, Mac, Ubuntu). R will start downloading. Once it's all done, install R. Now, download R Studio, the most-used interphase to "talk" to R. Click on *Download R-Studio* and make sure you select *FREE*. Also, select the version that works according to your OS.

Academic Integrity

I expect nothing but the best out of my students.

- I expect students to do their reading *before* class.
- Practical exercises should also be done *before* class.
- If you need to see me, plan your time accordingly. It's best to assume that my office hours will get busier before tests and submissions. Ask your TA or myself when in doubt.
- I usually don't answer emails during weekends.

☞ Plagiarism will not be tolerated. Make sure you follow the University's rules and definitions of plagiarism. Also, make sure you know how to cite your work.

☞ I won't accept late work.

Policy About Collaborative Work and External Resources

I do recommend collaborative work. It's good that you work with your classmates. However, I will grade individual work.

Another advantage of R is that it has a really engaged community of software developers and Internet bloggers. They are your best friends whenever R gives you trouble. Maybe the best website to look for answers *before* start asking online, is StackOverflow. In any case, feel free to contact your TA or myself.

Evaluations

1. ***Quizzes***: 5% each, 10% in total.

   There will be two quizzes. **The first one is due at the beginning of our first session!** (i.e., first day of class with me). Both are due in hard copy (i.e. paper).

2. ***Problem Sets***: 10% each, 40% in total.

   The *problem sets* are hands-on programming exercises. Some times, you may expect some epistemology-type questions for which I expect epistemology-type answers (one well-crafted paragraph will do). For the programming ones, I will give you an `R` script with a dataset. You'll have to answer the questions within the same `R` script, and then turn that file in. The TA and myself will be available to answer questions if needed.

   - ◇ While it shouldn't be necessary, you *may* use resources on the Internet to answer the questions.
   - ◇ It is important that the code runs without issues. In other words, your coding shouldn't get stuck.
   - ◇ It is important to guide and explain your reasoning. For that use the # symbol.

3. **Research Project *or* Replication (30%) plus a Final Presentation (20%)**: 50% in total.

   This is the core of course. You'll have to give a 20 minutes presentation, very much like a professional conference. Using your (1) actual dataset and (2) research question of your actual Masters thesis, you will have to:

   - ◇ **Motivate the problem**: Why should we care about your research question? (1 minute).
   - ◇ **Short lit review**: What's the main gap your work intends to bridge? (1 minute).
   - ◇ **Hypotheses**: Tell us about your hypotheses.
   - ◇ **Data**: What's your dependent and independent variables? Use plots, summary statistics, etc. (2 minutes).
   - ◇ **Data analyses and hypotheses testing**: What's the relationship between your IV and DV's? Why are you performing the analyses you're performing? Are there any alternative ways to analyze your data? *Convince us you've done all that there is to be done with your data!* The audience, the TA and myself are going to ask you questions about this. Thus, anticipate those questions and make sure you include in your presentation a large Appendix covering, at least: extra diagnostics, plots, tables, alternative variable re-coding, etc. (rest of the time)
   - ◇ **Conclusion**: What are the main conclusions you can derive from your work? What are your suggestions for future research? (1 minute).

   Life isn't perfect. Thus, you might not have the data ready by the time we need them for the class. In those cases, you will conduct a replication. A replication consists of obtaining someone else's data and performing the exact same data analyses published in the paper (models, tables and plots). *The idea is to obtain the exact same results!* The data must come from a well-published paper. To obtain those data, you will need to contact the actual authors by sending them a *very* nice email. I'll ask you to see me before sending that email. Since replications take time, you will need to gather the data of your replication ASAP. In any case, **you will need to have the data ready from our second meeting**. **Failure to obtain the replication data will negatively impact your grade tremendously**.

   Regardless, either you have your own data ready or you're conducting a replication, you'll have to:

   - ◇ **Email/meet me two weeks before this module begins** so we work out a plan. We will discuss whether you having your Master thesis dataset is possible at all, otherwise, which paper you will replicate. Hopefully is a paper that either addresses a similar problem of your thesis, or it's a paper you consider the best paper in your area.
   - ◇ Turn in an `R` script with *all* data manipulations.
   - ◇ Give the 20-minutes presentation at the end of the module.

To summarize:

	Percentage	Cumulative Percentage
Quizzes	10%	10%
Problem Set #1	10%	20%
Problem Set #2	10%	30%
Problem Set #3	10%	40%
Problem Set #4	10%	50%
Research Project or Replication	30%	80%
Final Presentation	20%	100%

Basic Bibliography

- Guido Imbens and Donald Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.
- Joshua Angrist and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.
- Jeffrey Wooldridge. 2002. *Introductory Econometrics: A Modern Approach.* 2nd. South-Western College Pub.
- Krishnan Namboodiri. 1984. *Matrix Algebra: An Introduction.* Sage.
- Jeff Gill. 2006. *Essential Mathematics for Political and Social Research.* Cambridge University Press.
- John Fox and Sanford Weisberg. 2010. *An R Companion to Applied Regression.* 2nd. Sage

Suggested Bibliography

- Paul Rosenbaum. 2010. *Design of Observational Studies.* Springer Series in Statistics. Springer New York.
- James Monogan. 2015. *Political Analysis Using R.* Springer.

☞ We will also read some papers.

Schedule

1. **Basic Functions in `R`**
   - **Lesson #1**
     - Introduction: syllabus, requirements, expectations, etc.
     - *What's `R`?* Installing `R` and `RStudio`.
     - **Reading(s)**:
       - Wooldridge (2002): 1.
   - **Lesson #2**
     - Basic functions: mean, `help()`, operators, objects (character, arrays, dates, lists, data.frames).
     - Working with data.frames (I): format, labels, types of variables, basic description.
     - **Reading(s)**:
       - Fox and Weisberg (2010): 1.1.

- **Lesson #3**
  - Transformations, generating new variables.
  - Working with data.frames (II): generating matrices and data.frames, `merge`, `append`. Logs.
  - **Reading(s)**:
    - ◇ Gill (2006): 1.7.
    - ◇ Fox and Weisberg (2010): 2.3 and 3.4.

- **Lesson #4**
  - Data visualization (I): bar plots, scatter plots, histograms, time series plots.
  - **Reading(s)**:
    - ◇ Fox and Weisberg (2010): 3.2.

- **Lesson #5**
  - Data visualization (II): more complex plots, maps.
  - **Reading(s)**:
    - ◇ Fox and Weisberg (2010): 7.3.

2. **Descriptive Statistics in R**

- **Lesson #6**
  - Descriptive Statistics (I). Measures of central tendency (mean, median, mode) and dispersion (variance and standard deviation).
  - **Reading(s)**:
    - ◇ Gill (2006): 8.4—8.5.

- **Lesson #7**
  - Estadística descriptiva (II). Probability Theory and distributions: binomial, normal, others; simulation.
  - **Reading(s)**:
    - ◇ Gill (2006): 8.3.

☞ Problem set #1. Turn it in next class.

3. **Introduction to Linear Regression in R**

- **Lesson #8**
  - *What's OLS?*
  - **Reading(s)**:
    - ◇ Wooldridge (2002): 2.1—2.2.

- **Lesson #9**
  - The mechanics behind OLS (II): matrices in R.
  - **Reading(s)**:
    - ◇ Namboodiri (1984): 1—2.

- **Lesson #10**
  - Coefficients.
  - **Reading(s)**:

- ◇ Wooldridge (2002): 3.1—3.2.
- ◇ Fox and Weisberg (2010): 4.3 till p. 177.

- **Lesson #11**
  - Error, residual and ϵ_i: Statistical, practical and philosophical differences (and why it matters).
- **Lesson #12**
  - Confidence intervals, standard error and covariance matrix.
  - **Reading(s)**:
    - ◇ Wooldridge (2002): 4.3.
    - ◇ Fox and Weisberg (2010): 4.3.1.
- **Lesson #13**
  - Hypothesis testing (t test), Type I and II Errors, statistical significance (p-values).
  - **Reading(s)**:
    - ◇ Wooldridge (2002): 4.2.
- **Lesson #14**
  - Interaction terms. Motivation. Estimation. Interpretation.
  - **Reading(s)**:
    - ◇ Wooldridge (2002): 7.4.
    - ◇ Thomas Brambor, William Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (01): 63–82.

☞ Problem set #2. Turn it in next class.

- **Lesson #15**
  - Numeric properties of OLS, Gauss-Markov, omitted variable bias.
  - **Reading(s)**:
    - ◇ Wooldridge (2002): pp. 89—94, 102—104.
- **Lesson #16**
  - Goodness of fit, coefficient of determination (r^2), prediction ($\hat{y}$).
  - **Reading(s)**:
    - ◇ Wooldridge (2002): pp. 40—41, 6.3.
    - ◇ Gary King. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30 (3): 666–687.
- **Lesson #17**
  - Issues and post-estimation: perfect multicollinearity and variance inflation, heteroskedasticity, non-linearity, outliers, non-normality of residuals and auto-correlation.
  - **Reading(s)**:
    - ◇ Wooldridge (2002): 8 y 9.5.

☞ Problem set #3. Turn it in next class.

4. **Causal Inference in R**

- **Lesson #18**

- Causal Inference: The Fundamental Problem of Causal Inference, the ignorability assumption, and the Potential Outcomes Framework.
- **Reading(s)**:
  - Imbens and Rubin (2015): 1.

- **Lesson #19**
  - Instrumental Variables and the Two-Stage Least Squares Regression.
  - **Reading(s)**:
    - Angrist and Pischke (2009): 4.1—4.2.

☞ Problem set #4. Turn it in next class.

- **Lesson #20**
  - Regression Discontinuity Designs (RDD): *Sharp Designs*.
  - **Reading(s)**:
    - Angrist and Pischke (2009): 6—6.1.

- **Lesson #21**
  - Regression Discontinuity Designs (RDD): *Fuzzy Designs*.
  - **Reading(s)**:
    - Angrist and Pischke (2009): 6.2.

- **Lesson #22**
  - Incorporating the Time Element: Fixed Effects (FE), Differences-in-Differences (DID).
  - **Reading(s)**:
    - Angrist and Pischke (2009): 5.

- **Last Class**
  - Presentations. Format is conference.

References

Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Brambor, Thomas, William Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (01): 63–82.

Fox, John, and Sanford Weisberg. 2010. *An R Companion to Applied Regression.* 2nd. Sage.

Gill, Jeff. 2006. *Essential Mathematics for Political and Social Research.* Cambridge University Press.

Imbens, Guido, and Donald Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30 (3): 666–687.

Monogan, James. 2015. *Political Analysis Using R.* Springer.

Namboodiri, Krishnan. 1984. *Matrix Algebra: An Introduction.* Sage.

Rosenbaum, Paul. 2010. *Design of Observational Studies.* Springer Series in Statistics. Springer New York.

Wooldridge, Jeffrey. 2002. *Introductory Econometrics: A Modern Approach.* 2nd. South-Western College Pub.