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 11, 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): TBA.
e:: TBA.
TA Bio: TBA.
Hora de ayudantía: *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 a public policy? Does legalizing some drug increase consumption? Which political candidate will win the election?

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 quality data. This is fundamental for social scientists like us, putting heavy pressures for 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, don't get confused: these methods are not "bullet-proof," they will *never* "proof" anything at all. During the semester I need you to exercise your criticism skills all the time. As we will see, you will need to *suspect* everything we do, particularly, because we will learn *inferential statistics*, that is, we will do *inferences*, not "proofs." On top of that, everything we do will come with some degree of **uncertainty** and hard-to-check **assumptions**, probably the most-used words throughout the semester.

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*. For those matters, we will discuss why the methods we will learn this semester are not causal, i.e., we cannot derive a causal statement from our results. However, when certain conditions apply, we might get *quasi-causal* statements. What are those "conditions"? We shall see...

Honestly, I hope this course captivates your enthusiasm, and gets you interested and curious about ways to study social phenomena, *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

- The course will be taught in English in the computer lab as 4 sessions of 3 hours each.
- 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 use "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, install `R` from the official Website. Click on "CRAN" (upper-left corner), then select any "mirror" you want. `R` will start downloading. Once it's all done, install `R`. Now, download `R Studio`, the most-used interphase to "talk" to `R`. For those purposes, download this piece of software from its official website. Click on *Download R-Studio* and make sure you select *FREE*. Also, select the version that works according to your OS (i.e., Windows, Mac, Ubuntu).

Academic Integrity

I expect nothing but the best out of my students, in particular, it's necessary to mention the following:

- I expect students to do their reading *before* class. Participation is not only encouraged but *graded*.
- Practical exercises should also be done *before* class.
- If you need to see me, plan your time accordingly. It's good to assume that it will get busier before tests and submissions.
- 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

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

Recitation

TBA.

Evaluations

1. **Lecturas y Participación** : 10%.

   El TA y yo asumiremos durante todo el semestre que haz leído. Nosotros empleamos un método de clases interactivo, pero este método necesita de tu participación activa en clases.

   Si no puedes asistir a la clase sincrónica, existirán opciones para dejar entradas en la sección *Foro* de `uCampus`.

2. ***Problem Sets***: 10% cada uno, 40% en total.

   Estos *problem sets* son ejercicios prácticos. Nosotros te entregaremos un *script* de `R` junto a una base de datos. Tú tendrás que resolver las preguntas dentro de `R` y devolvernos ese *script*. El ayudante y el profesor estaran disponibles para resolver preguntas vía email o Zoom.

   - **Aunque no es necesario, sí puedes ocupar recursos externos, como Internet**.
   - Es importante que estas líneas corran bien: el usuario (yo) tiene que ser capáz de ver cómo `R` ejecuta cada linea, sin estancarse.

- ◇ Es importante que vayas guiando al usuario (yo) sobre tu raciocinio. Asegúrate de comentar (usando el simbolo #).

3. **Un trabajo final obligatorio/no-eximible (30%) y una presentación final (20%, vía Zoom)**: 50% en total.

   En este curso, la actividad final es un trabajo final (30%) que tiene formato de trabajo grupal. Usando una base de datos que nosotros te daremos, tú y tu grupo deberán responder una serie de preguntas. El producto final (i.e. lo que debes entregar) consiste en un *script* de `R`. La nota es grupal (i.e. todo el grupo recibirá la misma nota). **Los grupos serán de 2 personas**. La formacion del grupo es endógena.

   El paper (*script*) se puede entregar antes, pero una vez cerrado el plazo, no se recibirán trabajos. Los *scripts* que se entreguen tarde o vía *email* tendrán un 1 (sin opción a reclamo). **No hay excepciones**.

   En un formato muy parecido a una conferencia académica (virtual, no presencial), tendrás (junto a tu grupo) que presentar los principales hallazgos (20%). Todos/as presentan. Cada presentación debe durar no menos de 15 minutos, pero nunca más de 20 minutos. Las presentaciones se realizarán virtualmente (i.e. vía Zoom) el último día de clases. Tendrás que ocupar *slides* ("Power Point"). Para tales efectos, tendrás que compartir pantalla desde tu casa, y hacer tu presentación de esa manera.

   Les recomiendo "verme" (vía Zoom) en mis office hours *antes* del plazo de entrega. Si quieres, SEND ME AN EMAIL con tu borrador, y yo te devolveré comentarios. Vélo como una pre-corrección. Esto es voluntario. También puedes contactar al/la TA. **No se procesarán preguntas durante fines de semana, y/o festivos**.

En resumen:

	Porcentaje	**Porcentaje Acumulado**
Participación (cátedra, foro `uCampus` y ayudantía)	10%	10%
Problem Set #1	10%	20%
Problem Set #2	10%	30%
Problem Set #3	10%	40%
Problem Set #4	10%	50%
Trabajo final grupal	30%	80%
Presentación grupal	20%	100%

Textos Mínimos

- Guido Imbens and Donald Rubin (1998). *Causal Inference for Statistics, Social, and Biomedical Sciences*.
- Joshua Angrist and Jorn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*.
- Jeffrey Wooldridge (2010). *Introducción a la Econometría. Un Enfoque Moderno*.
- Urdinez y Cruz (2019). *AnalizaR Datos Políticos*.
- Krishnan Namboodiri (1984). *Matrix Algebra, an Introduction*.

Textos Recomendados

- Paul Rosenbaum (2010). *Design of Observational Studies*.
- James Monogan (2015). *Political Analysis Using R*.

☞ También se considerarán algunos *papers*. Estos estarán señalados en las fechas indicadas y en la sección de Bibliografía.

Calendario

1. **Funciones básicas en `R`**

   - **Clase #1**
     - Introducciones: programa de curso, requerimientos, expectativas, etc.
     - *Qué es `R`?* Instalación de `R` y `RStudio`.
     - **Lecturas**:
       - ◇ Wooldridge (2010): Cap. 1.
       - ◇ Urdinez and Cruz (2019): Cap. 2.

   - **Clase #2**
     - Funciones básicas: promedio, `help()`, operadores, tipos de objetos (*character*, *arrays*, fechas, listas, *dataframes*).
     - Cargando bases de datos (I): formatos, etiquetas, tipos de variables, descripción básica.
     - **Lecturas**:
       - ◇ Urdinez and Cruz (2019): Cap. 5.

   - **Clase #3**
     - Cargando bases de datos (II): transformaciones, creación de nuevas variables.
     - Manipulando bases de datos: generación de matrices y *dataframes*, `merge`, `append`. Logs.

   - **Clase #4**
     - Visualización de datos (I): *bar plots* (variable categórica/continua, categórica/categórica), *scatter plots*, histogramas, *time series plots*.
     - **Lecturas**:
       - ◇ Urdinez and Cruz (2019): Cap. 4.

   - **Clase #5**
     - Visualización de datos (II): *plots* más complejos (por categorías), mapas.
     - **Lecturas**:
       - ◇ Urdinez and Cruz (2019): Cap. 15.

2. **Estadística descriptiva en `R`**

   - **Clase #6**
     - Estadística descriptiva (I): Teoría de probabilidades: distribuciones, varianza.

   - **Clase #7**
     - Estadística descriptiva (II): binomial, normal, otras; simulación.

☞ Entrega temario del *Problem set* #1. Una semana de plazo.

3. **Introducción a modelos lineales en `R`**

- **Clase #8**
  - Introducción a modelos lineales: *Qué es OLS?*
  - **Lecturas**:
    - ◇ Wooldridge (2010): 2.1—2.2.
- **Clase #9**
  - La mecánica detrás del OLS (II): matrices en `R`.
  - **Lecturas**:
    - ◇ Namboodiri (1984): Caps. 1 y 2.
- **Clase #10**
  - Coeficientes.
  - **Lecturas**:
    - ◇ Wooldridge (2010): Caps. 3.1—3.2.
- **Clase #11**
  - Error, residual y ϵ_i.
- **Clase #12**
  - Intervalos de confianza, error estándar y *variance-covariance matrix*.
  - **Lecturas**:
    - ◇ Wooldridge (2010): Cap. 4.3.
- **Clase #13**
  - Test de hipótesis (*t test*), errores Tipo I y II, significancia estadística (*p-values*).
  - **Lecturas**:
    - ◇ Wooldridge (2010): Cap. 4.2.
- **Clase #14**
  - Términos de interacción. Motivación. Estimación. Interpretación.
  - **Lecturas**:
    - ◇ Wooldridge (2010): Cap. 7.4.
    - ◇ Thomas Brambor, William Clark and Matt Golder (2006). *Understanding Interaction Models: Improving Empirical Analyses*. Political Analysis, 14(1): 63—82.

☞ Entrega temario del *Problem set* #2. Una semana de plazo.

- **Clase #15**
  - Propiedades numéricas del OLS, Gauss-Markov, sesgo de variable omitida.
  - **Lecturas**:
    - ◇ Wooldridge (2010): pp. 89—94, 102—104.
- **Clase #16**
  - *Goodness of fit*, "coeficiente de determinación" (r^2), predicción.

- **Lecturas**:
  - Wooldridge (2010): pp. 40—41, Cap. 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.

- **Clase #17**
  - Problemas y *post-estimation*: multicolinealidad perfecta, heteroskedasticidad, no linearidad, *outliers*, no normalidad de residuos, auto-correlación.
  - **Lecturas**:
    - Wooldridge (2010): Caps. 8 y 9.5.

☞ Entrega temario del *Problem set* #3. Una semana de plazo.

4. **Inferencia causal en** `R`

- **Clase #18**
  - Inferencia Causal: El *Problema Fundamental* en Inferencia Causal, el Supuesto de la "Ignorabilidad" y el "*Potential Outcomes Framework*".
  - **Lecturas**:
    - Imbens and Rubin (2015): Ch. 1.

- **Clase #19**
  - Variables instrumentales y *two-stage least squares*.
  - **Lecturas**:
    - Angrist and Pischke (2009): 4.1—4.2.

☞ Entrega temario del *Problem set* #4. Una semana de plazo.

- **Clase #20**
  - Regression discontinuity designs: *Sharp Designs*.
  - **Lecturas**:
    - Angrist and Pischke (2009): 6—6.1.

- **Clase #21**
  - Regression discontinuity designs: *Fuzzy Designs*.
  - **Lecturas**:
    - Angrist and Pischke (2009): 6.2.

- **Clase #22**
  - Incorporando el elemento *tiempo*: fixed effects, differences-in-differences.
  - **Lecturas**:
    - Angrist and Pischke (2009): Ch. 5.

☞ Entrega temario del trabajo final.

- **Última Clase**
  - Presentaciones Grupales. Formato "conferencia online".

References

Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* 392. Princeton University Press.

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

Imbens, Guido, and Donald Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences.* 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,* 1–99. Sage.

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

Urdinez, Francisco, and Andrés Cruz. 2019. *AnalizaR Datos Políticos.* Edited by Francisco Urdinez and Andrés Cruz. https://arcruzo.github.io/libroadp/.

Wooldridge, Jeffrey. 2010. *Introducción a la Econometría. Un Enfoque Moderno.* 4th. Cengage Learning.