ECO 541: Healthcare Data Analytics

Spring 2023 Syllabus

Instructor: Peter Bondarenko
Email: pbondare@depaul.edu

Course Time: The course follows an **online async** format which involves pre-recorded mini lectures posted weekly on D2L.

Meet Up Time: 7:30-9:00pm Thursdays. Optional live Zoom meeting (see D2L for link) for a chance to meet online (in virtual person) and review material and Q&A.

Discussion/Quick Comms: ECO 541 – Healthcare Data Analytics – Spring 2023 Slack workspace. All of you on the class-list should have received an invite. If not, please click here to enroll.

While this is an online async class, I encourage you to utilize the weekly meet up time on Thursday evenings to bring your questions about the material covered. I will come to those with a pre-determined set of material to review, so there will always be some content we go over during this time. While this meeting is optional, I encourage you to utilize it as this will be our only chance to engage in real-time via live Zoom during this course.

We will also utilize our Slack workspace for quick questions and weekly discussions on the material. When you accept the invite and join the workspace you will be automatically placed in the #health-data-analytics channel. We will use this channel for general inquiries/questions. I will be creating weekly topic channels to help keep our discussions organized and easier to search.

Course Description and Objectives

This course introduces analytical methods using healthcare data. While we continue to build on our health care knowledge from ECO 540: Business of Health, we also learn to adopt a statistician's perspective to approach key health problems. We use our economics intuition to help build better models that describe the existing data, which we can then utilize for various purposes such as helping evaluate interventions or coming up with accurate predictions.

The purpose of this course is to equip you with a solid analytics foundation that you can build on in your future roles as health care managers and practitioners. In other words, the goal is to teach you "how to fish" in the land of healthcare data analytics, as opposed to memorization of concepts.

The topics follow a systematic progression from simple to complex: we start with small data sets and learn the basics of data loading and manipulations (aka data wrangling), calculating summary statistics, identifying basic correlations, and estimating simple linear regression models to help

answer some interesting questions. We gradually build towards more advanced techniques, such as multiple regression, logistic regression, and survival analysis. Time permitting, we explore clustering, random forest, and natural language processing algorithms.

While almost every analysis we perform in this course can be (albeit with much difficulty) done in MS Excel, we use R as our statistical software of choice. R is an open-source software that has a lot of widespread use in healthcare data analytics. The idea is to add R to your toolkit and spend minimal time in figuring out how to perform an analysis to free up more time to discuss model building and the intuition behind the results.

Prerequisites

This course builds on the material from ECO 519: Business Analytics Tools (formerly GSB 420) and ECO 540: Business of Health. The former is an official pre-requisite for this class: while we will do a quick review of the ECO 519 material during the first week of class, the general expectation will be a working knowledge of the basic statistical concepts.

The learning objectives for the course can be summarized as follows:

- Build a solid understanding of various statistical techniques widely used in healthcare data analytics.
- Develop a good working knowledge of R.
- Gain familiarity with major publicly available health datasets.
- Obtain ability to build models and empirically analyze a health related question of interest.

Assessment

- Weekly Assignments: 40% - these build directly on the weekly material.
- Midterm: 30% - this is a fully guided health data analytics exercise (due week 6)
- Final Project: 30% - this is a health data analytics project (due on Sunday 6/12)

Grading A = 93-100, A- = 90-92, B+ = 87-89, B = 83-86, B- = 80-82, C+ = 77-79, C = 73-76, C- = 70-72, D+ = 67-69, D = 60-66, F = <60

Required Text

- None

Recommended Texts

- Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Harrell, FE. New York, NY: Springer, 2001

- Healthcare Data Analytics, Edited by Chandan K. Reddy and Charu C. Aggarwal (2015), CRC Preess, ISBN: 978-1-482-23212-7
- R for Health Data Science, Ewen Harrison and Riinu Pius (2021), CRC Preess, ISBN: 978-0-367-85542-0, available from https://argoshare.is.ed.ac.uk/healthyr_book/

Other Information

- Our class D2L site will include course material and supplementary readings throughout the term.
- I will only send email to you using your DePaul email.

Final Project

The final project is an empirical research paper. I would like you each to write a simple research paper (8-10 pages max, double spaced) to tackle a health question of interest to you. This could build on a previous paper you wrote - for example the one you wrote in your ECO 540: Business of Health class - or it could be on a new topic you want to investigate. The only restriction is for it to have a health-related problem statement and make use of the empirical methods we learned in this class. Being an empirical paper, it should involve analysis using a health data set (could be the ones we will introduce in class or other) and utilize the R statistical software.

At the end of the 8th week of the course, I will ask you to provide a short description of your intended paper topic, the data set you plan to use, as well as the analytics methods you plan to utilize. This will provide a useful checkpoint that you are on-track and an opportunity for me to provide some feedback.

(Tentative) Course Outline

Week 1, March 27-31

- Review of course requirements
- Review of statistical concepts from ECO 519
- Introduction to R

Week 2, April 3-7

- Simple linear regression
- Interpreting R regression output
- Statistical significance

Week 3, April 10-14

- Introduction to multiple regression
- Model building in R

Week 4, April 17-21

- Multiple regression – continued
- Introduction to logistic regression
- Further exercises in R

Week 5, April 24-28

- Logistic regression – continued
- Interpreting R logistic regression output

Week 6, May 1-5

- Introduction to survival analysis
- R exercises and examples

MIDTERM DUE – May 7th, Sunday, 5pm

Week 7, May 8-12

- Survival analysis - continued
- Intro to major public health datasets - ACS, NHIS, HCUP-NIS, and BRFSS

Week 8, May 15-19

- Exploration of public health datasets - continued
- Introduction to Machine Learning (ML): clustering and random forests

DRAFT PAPER IDEAS DUE – May 21st, Sunday, 5pm

Week 9, May 22-26

- ML: Clustering and random forests – continued
- Introduction to Natural Language Processing (NPL)

Week 10, May 29-June 2

- Natural Language Processing (NLP) – continued
- Analyzing Electronic Health Records (EHR) using NLP

PAPERS DUE – June 12, Sunday, 5pm