BUS 41100: Applied Regression Analysis

Autumn 2021

(version 0.1, subject to change)

Instructor: Max H. Farrell
Email: max.farrell@chicagobooth.edu
Office hours (zoom): TBA: some in person, some zoom if needed

TA: TBA
TA Email: max.farrell.ta@gmail.com

Course websites:

1. All material like slides, homework, and data sets, is on my website: https://maxhfarrell.com/bus41100/
2. Piazza Q&A website, where you ask any/all questions: http://piazza.com/chicagobooth/fall2021/41100/home
3. **No** Canvas site

Course Summary

This course is about regression, a powerful and widely used data analysis technique. Students will learn how to use regression by analyzing a variety of real world problems. Heavy emphasis will be placed on analysis of actual datasets. Topics covered include: simple and multiple regression, prediction, variable selection, causal inference, residual diagnostics, classification (logistic regression), and time series (auto-regression), and introductory machine learning.

Course Format

For Autumn 2021–2022 this course will be **in person** for Sections 01 (Mon AM) and 02 (Wed AM)and **dual-modality** for Section 81 (Wed night).

Contents

1 Course Schedule & Topics Covered

This schedule is subject to change, depending on how well remote learning progresses.

Week	What's Due	Material
1	Nothing!	Introduction, simple linear regression model
2	HW 1	Inference for linear regression
3	HW 2	Finish SLR, start multiple linear regression
4	HW 3	Causation & Other variables types in MLR
5	HW 4	MLR Pitfalls & Some Fixes, Clusters and Panels
6	Nothing!	**Midterm Exam** (in class) & Time series
7	Project Proposal	Logistic regression & classification
8	HW 5	Model Building
9	HW 6	More on Discrete Outcomes (if time)
10	Project write-up	**Final Exam** (in class)

2 Prerequisite

You should be familiar with basic statistics, such as BUS 41000 or its equivalent, including:

- random variables and their properties (mean, variance, distribution, etc),
- expectation and conditional expectation,
- the Normal distribution,
- hypothesis testing,
- confidence intervals, and
- sampling distributions.

A review can be found in any decent statistics textbook, e.g. *Statistics for Business: Decision Making and Analysis*, by Stine and Foster). There is a "Homework 0" here: https://maxhfarrell.com/bus41100_remote/homework0.pdf. If you find this homework challenging or the concepts unfamiliar, you should either review the material or consider a different course.

3 Textbook

I **strongly** recommend you use *A Modern Approach to Regression with* R, by Sheather. You can get the PDF through the U of C library here (must authenticate): http://link.springer.com/978-0-387-09608-7.

This book is excellent for both conceptual material and computing help. It gives a careful introduction to the concepts of simple and multiple linear regression, overlapping with my material a great deal. The book also worked-through examples in R so you can learn that as

well. The author's website (http://gattonweb.uky.edu/sheather/book/) has R, STATA, and SAS code to accompany the text, and tutorial videos.

This book is not required in the sense that I will not assign readings or homework from it. But it will help you succeed in the course.

4 Computing

Statistical computing is a key part of the class. In-class analysis will be conducted in R and all course material (code and data) is in R format. R is free (gratis and libre) and available for download at http://www.r-project.org. For help, a quick google search is your best friend. After that post on Piazza. You can use whatever platform/language (Matlab, Stata, SAS, Python, Minitab, etc) you like, but I cannot help with anything but R.

R has a command line interface. I **strongly encourage** you to install the software as soon as possible and get familiar with simple operations; you should do this **before** the course starts. There many GUIs available, such as R Studio.

Some resources:

- A good introduction/tutorial to R: https://data.princeton.edu/R
- UCLA has a fantastic help page for R (and statistics/regression in general) with everything from installation/basic help, worked-through examples, books, and link to more resources: http://www.ats.ucla.edu/stat/r/
- E-Books available from U of C: https://catalog.lib.uchicago.edu/vufind/Search/Results?lookfor=%22R+%28Computer+program+language%29%22&type=TopicBrowse&filter[]=format%3A%22E-Resource%22
- The University offers R workshops in the Research Computing Center, see schedule here: https://rcc.uchicago.edu/support-and-services/workshops-and-training
- The resources out there are continually changing, so you may find other options. Please let me know if you find something helpful that isn't listed here.

5 Assignments / Exams / Grading

Grades. The course grade is determined by: the seven homeworks (25%), a course project (25%), and midterm (20%) and final (30%) exams. Class participation is not graded. The due date schedule is above.

Homework. You will be *randomly* assigned a homework group at the start of the quarter that you will work with for homeworks 1–4. You will be assigned a second group for homeworks 5 & 6 and the course project. Turn in only one copy of the assignment per group, with everyone's name on it.

Homework reinforces material introduced in lecture and *extends* it, introducing new topics and ideas that are relevant to real-world regression analysis but can not be covered in lecture.

Such topics are always explained as needed.

Homework assignments should have a clear and professional presentation. Please do not submit unedited computer output; cut and paste the relevant portions of the output into your homework document. Late homework is not accepted.

Midterm Exam. The midterm will be take home. You will have from Saturday evening through Thursday to complete it, during week 6. You are required to complete it individually, but otherwise it is open book, notes, Internet, ... The only thing you may not use is each other. Exams are designed to test understanding of the course material. Questions are more concrete than the homework, but still may be conceptual in nature. Exam questions never ask for computer code, but are based on R output from regression analyses. Sample/practice exams are available.

Re-grades. Clerical errors will be corrected without hassle, but other requests must be submitted **in writing** within **one week** of the homework/exam return. The entire homework/exam will be regraded from scratch (by me!), and as such your score may go up, down, or stay the same.

Students must adhere to *Booth Honor Code*. But you do not need to include the honor code, signature, etc., on your work.

6 Course Project

The goal of the project is to produce an essay that addresses a realistic empirical question by conducting a thorough regression analysis. Students will work in groups of up to four people. These groups will be assigned during the course. You may investigate any empirical question you choose: you will find the data, decide on the analyses to perform, and draw all the conclusions. I can help you, or course, but will deliberately avoid explicit guidance. The project is meant to be an open-ended exercise.

The project write-up is due on the last day of the academic quarter, even if this is after the final day for your section.

Project Goals & Requirements. The most important thing is to demonstrate conceptual mastery of the course material and its implementation. Your essay should clearly state your empirical question and write down your regression model(s) as we did in class, e.g. $Y = \beta_0 + \beta_1 X + \beta_2 Z + \cdots + \varepsilon$, $\mathbb{E}[\varepsilon|X, Z, \ldots] = 0$. You should provide a motivation for which variables you include in your regressions and clear definitions of any constructed variables like indicator variables for categories. You should carefully address diagnostic issues and model selection and testing.

There is no formal requirement on the length or format of the essay. The goal is to write something that is clear, readable, and thorough; however you feel you can best accomplish those goals is fine. Most write-ups are around 15 pages including tables and graphs (no need to include a ton of computer code/output).

Proposal. The project proposal due in week 7 should include the following five things:

1. a description of your essay's empirical question,
2. why it is important,
3. the data sources you will use and how you will get access to them,
4. the methods you will use,
5. preliminary results you have obtained, and
6. any specific questions you have.

Send one email per group with the subject heading "41100 Project Proposal" to me (not the TA) by class time in week 7. Make sure the name of each group member is in the email.

I will provide feedback via email on each proposal, discussing your preliminary results, details of your approach, additional strategies in addressing the empirical question, and any problems you bring up. In-person meetings can of course be scheduled. The proposal isn't part of the grade, but the more you have done, the better the feedback you will get.

Of course, feel free to talk to me (or email me) before/after the proposal is due about your idea. You can always get informal feedback.

Data. Some sources you might find useful (only a few ideas):

- Macroeconomic Data from Federal Reserve Economic Database (FRED): http://research.stlouisfed.org/fred2/
- Wharton Research Data Services (WRDS): http://wrds.wharton.upenn.edu/
- Center for Research on Securities Prices (CRSP): detailed securities data
- Global Insight: financial and economic data
- Compustat: firm level data for publicly traded firms
- IPUMS: U.S. census data: http://www.ipums.org
- Prediction and data mining competitions (all sorts of application areas): http://www.kaggle.com/
- City of Chicago data: http://data.cityofchicago.org/
- Harvard's open source research data repository (not all of these are rich enough for a thorough project): http://dataverse.harvard.edu/
- A bunch of data sets in R format (not all of these are rich enough for a thorough project): http://vincentarelbundock.github.io/Rdatasets/datasets.html

Some data sets are interesting, but are not high quality in one way or another. That's fine. If your data set is limited in one way, think about exploring/expanding your project in a different direction. Are there different ways of using those variables? Different outcomes you could predict? Different ways to evaluate model quality? Interactions that are interesting? Diagnostics and transformations that are useful? If you have only 5 variables, then you'll want to explore these issues carefully. If you have 5000, you are going to be more worried about variable selection methods. Different techniques for different projects and none are a priori better or worse. Don't feel that you must use each and every tool discussed in class.

The only real limitation is that you have enough data for a thorough analysis, to demonstrate mastery of the techniques from class. For example, a state-level data set ($n = 50$) will usually not be sufficient.

Project Grading. People often ask how the project and the proposal are graded. The proposal is not formally graded, beyond you turning it in on time. It's a way for you to get feedback to make your project better. Only under extreme circumstances will I "reject" a proposed project. As for the project itself, the goal of the project is to demonstrate that you can do a thoughtful, thorough job of investing a real-world empirical question using the material from class. Doing so is a "good" project. So you do *not* have to: *(i)* use every single technique from class; some will apply to your project and some won't; *(ii)* come up with an earth-shattering question or result; *(iii)* have the best data.

The project is deliberately open ended, and to reflect that spirit, so is the grading. I will not post a "sample A+" project, or anything of that sort. There's no specific structure or content required, so different projects can look *very* different.

7 Accommodations for Disabilities

The University of Chicago is committed to ensuring the full participation of all students in its programs. If you have a documented disability (or think you may have a disability) and, as a result, need a reasonable accommodation to participate in class, complete course requirements, or benefit from the University's programs or services, please contact Student Disability Services as soon as possible. To receive a reasonable accommodation, you must be appropriately registered with Student Disability Services. Please contact the office at 773-702- 6000/TTY 773-795-1186 or `disabilities@uchicago.edu`, or visit the website at `http://disabilities.uchicago.edu`. Student Disability Services is located at 5501 S. Ellis Avenue.

If you have an approved accommodation from Student Disability Services that you plan to use in this course, please contact Academic Services (`AcademicServices@lists.chicagobooth.edu`) as soon as possible. Academic Services will provide support to you and your instructor and coordinate the details of your accommodations on your behalf.

8 Other Recommended Books

- Nontechnical treatment of why statistics and uncertainty matter for decision making and what to do about it: *Public Policy in an Uncertain World*, by Manski. The book is in the context of public policy/health, but all the ideas apply to decision making in any area, including business.
- Introduction to basic statistics: *Statistics for Business: Decision Making and Analysis*, by Stine and Foster.
- *Applied Regression Analysis*, by Dielman, is a more traditional textbook than Sheather,

covering the classical regression material from the course in some detail. The most recent few editions are probably fine for this course.
- Taddy's *Business Data Science* is an accessible introduction to "big data" regression-type problems and why they are useful, as well as some introduction to some machine learning methods.

Students sometimes ask for references that cover certain topics in greater depth. Here are a few suggestions out of many. I have tried to single out references that are at a manageable level conceptually, but by necessity some of these books are more advanced.

- Basic Statistics (week 2): *Statistics for Business: Decision Making and Analysis*, by Robert Stine and Dean Foster.
- Time series (weeks 5): *Analysis of Financial Time Series* and/or *An Introduction to Analysis of Financial Data with R*, both by Tsay.
- Panel data & clustering (week 6): *Introductory Econometrics*, by Wooldridge.
- Causal inference: *Causal Inference for Statistics, Social, and Biomedical Sciences*, by Imbens and Rubin and/or Manski's book above.
- Data mining and Logistic Regression: Taddy's book above and/or *An Introduction to Statistical Learning*, by James, Witten, Hastie, and Tibshirani or Taddy's *Business Data Science*.