BUS 41100: Applied Regression Analysis

Autumn 2019–2020

Instructor: Max H. Farrell
Email: max.farrell@chicagobooth.edu
Office hours: By appointment

TA: Mauricio Chikitani, Gustavo Gonzalez, Min Park
Email: max.farrell.ta@gmail.com
Office hours:
Monday 4:30–5:30pm — Harper 3A
Tuesday 5:00–6:00pm — Harper 100-9A

Course website: https://maxhfarrell.com/bus41100/
Piazza Q&A website: http://piazza.com/chicagobooth/fall2019/41100/home

There is **no** blackboard/chalk/canvas/whatever site for this class!

Course Summary

BUS 41100 is a course 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, residual diagnostics, classification (logistic regression), time series (auto-regression), and panel data methods.

Contents

1 Prerequisite

This prerequisite is taken seriously in this course, please read carefully!

The prerequisite is familiarity with the topics covered in BUS 41000, or 41000 itself, or its equivalent. This prerequisite is not enforced in the registration system (anyone may take the class), but lecture material, assignments, and exams will assume you have this knowledge and additional help on these topics will not be provided. This course moves quickly and covers a great deal of material, some is quite advanced. To see for yourself if this course is appropriate, there is a "Homework 0" on the class website containing problems on basic statistic topics at the level expected. If you find this homework challenging or the concepts unfamiliar, you should *very* carefully consider your options.

Ensure that you have a complete command of these concepts (available for review in any decent statistics textbook, e.g. *Statistics for Business: Decision Making and Analysis*, by Stine and Foster):

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

It is also useful to start getting familiar with R before the course begins. There are links below and on the course webpage.

2 Schedule & Topics Covered

Week	What's Due	Material
1	Nothing!	Introduction, simple linear regression model
2	HW 1	Inference for linear regression
3	HW 2	Multiple linear regression
4	HW 3	Logistic regression & classification
5	HW 4	Time series models and autoregression
6	HW 5	**Midterm Exam** (in class) & Panel Data
7	HW 6	Regression issues & diagnostics
8	Proposal	Model building 1: testing and prediction
9	HW 7	Model building 2: data mining & causality
10	HW 8	Advanced discrete outcomes
11	Project (due Dec 14)	**Final Exam** (in class)

3 Website & Piazza

There is no blackboard/chalk/canvas/whatever site. All course material (slides, code, data sets, assignments, and solutions) is on the course website:

https://maxhfarrell.com/bus41100.

Note well that the lecture slides *accompany* the lecture. They are not designed to be a complete exposition of the material and will not serve as a self-contained reference.

There is also a Piazza Q&A website:

http://piazza.com/chicagobooth/fall2019/41100/home.

Students are **strongly** encouraged to post any and all questions there, regarding material, homework, R, computing, etc. Use Piazza before you directly email me or the TAs. You should be able to self-enroll at the link above.

4 Computing

Statistical computing is a key part of the class. You can use whatever platform you like. In-class analysis will be conducted in R and all course material (code and data) is in R format. R is free (as in speech) and available for download at http://www.r-project.org, and you can find manuals and installation guidelines on this site.

R has a command line interface (you type commands to get what you want). 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.

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.

Other languages (Matlab, Stata, SAS, Python, Minitab, etc) are allowed, but not supported by me or the TAs.

5 Optional Textbooks

There is no course pack or required textbook, but it is a good idea to have a textbook in case you are unclear about anything from class. Two suggestions are:

- *A Modern Approach to Regression with* R, by Sheather, closely follows the classical regression material from this course, and is a very good, careful introduction to the concepts and computing, with worked-through examples in R. The author's website (http://gattonweb.uky.edu/sheather/book/) has R, STATA, and SAS code to accompany the text, and tutorial videos. PDF here (must authenticate): http://link.springer.com/978-0-387-09608-7.
- *Applied Regression Analysis*, by Dielman, is a more traditional textbook, covering the classical regression material from the course in some detail. The most recent few editions are probably fine for this course.

Other useful books:

- Introduction to basic statistics: *Statistics for Business: Decision Making and Analysis*, by Stine and Foster.
- Nontechnical treatment of why statistics and uncertainty matter and what to do about it: *Public Policy in an Uncertain World*, by Manski. The book is in the context of public policy, but all the ideas apply to decision making in any area, including business.
- Taddy's *Business Data Science* is a fair accessible treatment to "big data" regression-type problems and why they are useful, as well as some introduction to some machine learning methods.

In addition, students often ask for references that cover 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 these books are sometimes more advanced.

- Causal inference (week 5): *Causal Inference for Statistics, Social, and Biomedical Sciences*, by Imbens and Rubin.
- Data mining (week 7) and Logistic Regression (week 8): *An Introduction to Statistical Learning*, by James, Witten, Hastie, and Tibshirani.
- Count data (week 8): *Regression Analysis of Count Data*, by Cameron and Trivedi.
- Time series (weeks 9 & 10): *Analysis of Financial Time Series* and/or *An Introduction to Analysis of Financial Data with R*, both by Tsay.
- Panel data & clustering (week 10): *Introductory Econometrics* or *Analysis of Cross Section and Panel Data*, both by Wooldridge; *Microeconometrics: Methods and Applications*, by Cameron and Trivedi.

6 Assignments / Exams / Grading

Grades. The course grade is determined by: the seven **homeworks** (15%), a **course project** (30%), and midterm and final **exams** (25% & 30%). Class participation is not

graded. The due-date schedule is above.

Homework. These reinforce material introduced in lecture and *extend* 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.

Homeworks are intentionally *not* in the style of exams. There are sample exams available for that. Instead, homework is where you learn/practice skills closer to the real world. The problems are more open-ended, designed to explore and struggle with issues that often manifest in everyday practice.

Students may work in groups of up to 4 people. Please turn in only one copy of the assignment per group, with each group member's name on it. Turn in your homework via email attachment to `max.farrell.ta@gmail.com` before the start of class according to the schedule above. 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.

Exams. The midterm and the final will be in class. The exams are closed book and closed notes, with the exception of **one** 8.5x11 "cheat sheet" (both sides). You may use a calculator (but you don't need anything fancy). The final exam is cumulative.

Exams are designed to test understand 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 exams are available on the course website.

If you need to reschedule an exam, and have a good reason, talk to me in person and email me as soon as possible. No request for rescheduling will be entertained on or after the date of an exam.

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.

7 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 are encouraged to work in groups of up to 4 people. 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 (or the TA) can help you, or course, but we will deliberately avoid explicit guidance. The project is meant to be an open-ended exercise.

The project is due on the last day of the academic quarter, that is, the final day of week 11, even if this is after the final exam 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 8 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, and
4. the methods you will use, and
5. preliminary results you have obtained.

Send one email per group with the subject heading "41100 Project Proposal" to me (not the TA) by class time in week 8. 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.

8 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.

9 FAQs

1. I have a question about {class material, computing, class logistics, class policies}, should I email you or the TA?

   Post it on Piazza! The website is: http://piazza.com/chicagobooth/fall2019/41100/home *You should be able to self-enroll using that link.*

2. I can't find/access the blackboard/chalk/canvas site for the course, am I appropriately registered for it?

   There is no blackboard/chalk/canvas/marker/whatever-system-we-bought-this-year *site for the class. All material will be at the url at the start of this syllabus. Announcements will be via email.*

3. I took such-and-such statistics class (using the book by blah-blah-blah) and did very well, am I ready for this class?

   If "such-and-such" was BUS 41000 here at Booth, then from a statistical knowledge point of view, yes, otherwise, take Homework Zero and decide for yourself!

4. Is there a first-class assignment?

   Yes, Homework Zero should be completed before the first class (see above under "Pre-requisite"), but is not to be turned in.

5. But where are the solutions for Homework Zero?

   They aren't available. What ends up happening is that people look at the answers and think, "oh yeah, I knew that" without really trying and get a false sense confidence. Either you feel confident in your understanding of the material or you don't; a solution set won't change that.

6. Follow-up: I tried to complete Homework Zero and I felt very comfortable with about X% of it, am I ready for this class?

   If X=0, then definitely not. If X=100, then at least from a statistical knowledge point of view, yes. For anything in between, you have to decide for yourself. For the benefit of both of us, I will not put myself in the position of having said "yes" and then you performing poorly in the class.

7. Can I audit the course?

   I discourage auditing. The course material is taught in a hands-on way, and without doing homeworks or the project, and without any group interaction, you will not get much from the course.

8. I need to miss class, what should I do?

   If you need to miss a lecture*, no need to tell me or ask permission. Attending an alternate section always fine. You are responsible for the material (the lecture slides are* not *a self-contained reference) ensuring timely delivery of any assignments.*

   If you need to reschedule an exam *talk to me as soon as you can.*

9. I need to reschedule an exam and I have a good reason, what should I do?

   Talk to me in person and be sure *to email me. We will work out an alternate exam time. When possible, taking the exam with another section is the easiest way to go. You must have a good reason to reschedule an exam.*

10. Can I have a provisional grade?

    Sure. But I need a rough draft of your project before I can give you a provisional grade. The final exam and project together make up 60% of your overall grade, and so I need a rough draft of the project in order to have something to base your provisional grade upon. All provisional grades are conditional on completing the project and sitting for the final exam.

11. Can we have 5 people in our homework/project group?

    No, 4 is a hard limit. I have found that with more than 4 people per group, at least one person (usually more!) ends up free-riding.

12. Can I be in a group with people from other sections?

    Sure, the only caveat is that all homework must be turned in during the earliest *class from any member of your group. The limit of 4 still applies.*

13. Will you post a "sample A+" project, so we know what to do?

    Definitely not! There's no specific structure or content required, so different projects look very *different. In fact, part of the point of the project is for your group to decide how best to present your analysis and conclusions.*

14. I want to do _____ for my project, is that OK?

    Yes, as long as you apply the methods from class. Of course, feel free to talk to me (or email me) before the proposal is due about your idea. You can always get informal feedback.