Syllabus

Ec240a - Second Half, Fall 2020

Course Description

After introducing some basic optimization tools, this course begins with an analysis of a basic prediction problem. A decision maker obtains a random sample of covariates (features) and outcomes. She wishes to use her sample to forecast the outcomes of new units on the basis of their covariates. We motivate this problem and provide a canonical representation of it (the K Normal means problem). We use this problem to introduce some elements of (i) statistical decision theory and (ii) modern regression methods.

We then develop some properties of regression functions. The iteration properties of mean and linear regression will receive special emphasis.

Finally, we will develop methods for conducting inference on linear regression coefficients estimated by the method of least squares under random sampling. We will develop two approaches. The first is a nonparametric Bayesian method. The second, frequentist approach, is based on large sample (i.e., asymptotic) approximations. Methods of hypothesis testing and confidence interval construction will be reviewed. If time permits we will introduce some methods for causal analysis using instrumental variables.

Instructor: Bryan Graham, 665 Evans Hall, email: `bgraham@econ.berkeley.edu`

Time and Location: Monday and Wednesday, 10:00AM to 12:00PM (On Zoom)

Office Hours: Thursdays 2 to 3:40PM (sign up online here).

Graduate Student Instructor: Mingduo Zhao, e-mail: mingduo@berkeley.edu

Prerequisites: linear algebra, multivariate calculus, basic probability and inference theory.

Course Webpage: Various instructional resources, including occasional lecture notes and Jupyter Notebooks, can be found on GitHub in the following repository

https://github.com/bryangraham/Ec240a

The GSI may make additional resources available on bCourses.

Due to the COVID-19 pandemic Ec240a lectures will be delivered virtually this semester using ZOOM. Details for accessing virtual lectures will be shared via bCourses to officially enrolled students. In lieu of the traditional live lecture some combination of pre-recorded asynchronous videos and live Q&A sessions may be offered. Totally weekly instructional hours will equal four hours (plus section time with the GSI).

Textbook: There is no mandatory text. Material will be delivered primarily through lecture and assigned papers. Good note taking is essential for successful performance in the class. Nevertheless I do recommend the following book as useful supplement to the material presented in lecture.

1. Wooldridge, Jeffrey M. (2011). *Econometric Analysis of Cross Section and Panel Data, 2^{nd} Ed*. Cambridge, MA: The MIT Press.

This is a useful long term reference for anyone who anticipates undertaking empirical research. An excellent textbook by Bruce Hansen, which I will assign readings from, is available online here. Much of the material covered in this class has an analog in the Hansen textbook (however I will not be "lecturing" from this book).

Additional books which you may find helpful include Ferguson (1996), Wasserman (2004), Wasserman (2006) and Manski (2007). Ferguson (1996) is a compact introduction to large sample theory. My treatment of the K Normal means problem draws from Wasserman (2006). Wasserman (2004) is a nice introductory mathematical statistics reference. Manski (2007) provides a textbook treatment of identification with applications of interest to economists.

Grading: Grades for *this half of the course* will equal a weighted average of homework (40%) and mid-term performance (60%). The mid-term will be held on the last day of class (**December 2nd, 2020**). There will be 5 homework assignments (plus a review sheet). Homeworks are due at 5PM on the assigned due date (the GSI may elect to make small modifications to all things homework related). Homeworks are graded on a ten point scale with one point off per day late. You are free, indeed encouraged, to work in groups but each student must submit an individual write-up and accompanying Jupyter Notebook (when required; see below). Your lowest homework grade will be dropped, with the average of the remaining scores counting toward your final grade. There will be no 'make-up' midterms. I will add 5 points to homework aggregates for students who make serious efforts to complete all five problem sets. *Concretely this means that students may amass up to 45 homework points; it also means that if you only do four problem sets you can earn no more than 40 out of 45 homework points.* Problem Set 5 is formally the course final assessment.

The due dates for the five problem sets are:

Problem Set	Due Date
1	October 26th
2	November 9th
3	November 23rd
4	December 4th
5	December 18th

Computation: All computational work should be completed in Python. Python is a widely used general purpose programming language with good functionality for scientific computing. There are lots of ways of accessing Python (EML, on the web etc.). For those wishing to manage a Python environment on their personal computer, the Anaconda distribution, which is available for download at https://www.anaconda.com/distribution/, is a convenient way to get started. Some basic tutorials on installing and using Python, with a focus on economic applications, can be found online at `http://quant-econ.net`. Good books for learning Python, with some coverage of statistical applications, are Guttag (2013), VanderPlas (2017), and McKinney (2017).

The code I will provide will execute properly in Python 3.6, which is (close to) the latest Python release. Python is also available on the EML workstations (which are remotely accessible). Students wishing to work with another technical computing environment (e.g., MATLAB, Julia, Fortran 2008, C++, R, etc.) should speak with the GSI. This will be allowed at the GSI's discretion. There are a large number of useful resources available for learning Python (including classes at the D-Lab).

While issues of computation may arise from time to time during lecture, I will not teach Python programming. *This is something you will need to learn outside of class.* I do not expect this to be easy. I ask that those students with strong backgrounds in technical computing to assist classmates with less experience.

Extensions: Extensions for assignments will not be granted. The penalty for lateness is relatively minor and I also drop the lowest homework grade.

Accommodations: Any students requiring academic accommodations should request a 'Letter of Accommodation' from the Disabled Students Program at http://dsp.berkeley.edu/ *immediately.* I will make a good faith effort to accommodate any special needs conditional on certification. Please plan well in advance as I may not be able accommodate last minute requests.

Academic Integrity: Please read the Center for Student Conduct's statement on Academic Integrity at http://sa.berkeley.edu/conduct/integrity. I take issues of intellectual honest *very* seriously.

Cooperation/Civility: I remember graduate school as a period of immense intellectual excitement, punctuated by periods of equally intense frustration and stress. My classmates were an important source of intellectual support and encouragement. Please be open to helping one another learn the material. Don't be afraid to ask classmates for help and, if asked for help, be generous and gracious in providing it. Everyone will learn more in this class if they work together. *Class is not a competition.* We are here to learn together. I also encourage each of you to familiarize yourself with Berkeley's Principles of Community (available here). This is not an easy time, so the above advice is all the more important.

E-mail and office hours: I prefer to avoid having substantive communications by e-mail. Please limit e-mail use to short yes/no queries. I am unlikely to read or respond to a long/complex e-mail. Please make use of my office hours. This is time specifically allocated for your use; please come by (virtually!) I look forward to getting to know all of you. You can sign-up for office hour slots online here.

Course Outline

Date	Topic	Readings/Notes
W 10/14	**Projection Theorem**	Hansen (2018, Appendix A)
M 10/19	**Probability Distributions**	Hansen (2018, Appendix B)
W 10/21	**Conditional Expectation Functions**	Hansen (2018, Ch. 2.1-2.17, 2.30-2.32) Wooldridge (2010, Ch. 2)
M 10/26	**K-Normal Means**	Wasserman (2006, Ch. 7)
W 10/28	**K-Normal Means**	Wasserman (2006, Ch. 7) Stein (1981)
M 11/2	**Linear Regression**	Hansen (2018, Ch. 2.18-2.27) Wooldridge (2010, Ch. 2) Card (1995); Card & Krueger (1996)
W 11/4	**Bayesian Bootstrap**	Chamberlain & Imbens (2003)
M 11/9	**Large Sample Theory for OLS**	Hansen (2018, Ch. 4-7) Wooldridge (2010, Ch. 3)
W 11/11	**No Class (Veteran's Day)**	
M 11/16	**Applications of OLS**	Deaton (1989); Altonji & Pierret (2001) Ashenfelter et al. (2006); Goldberger (1984)
W 11/18	**Instrumental Variables**	Angrist et al. (1996) Angrist et al. (2000)
M 11/23	**Instrumental Variables** **& EM Algorithm**	Imbens & Rubin (1997) Gupta & Chen (2010)
W 11/25	*No Class*	*Thanksgiving recess*
M 11/30	**Review Session**	
W 12/2	**2nd Midterm Exam**	Good luck!

References

Altonji, J. G. & Pierret, C. R. (2001). Employer learning and statistical discrimination. *Quarterly Journal of Economics*, 116(1), 313 – 350.

Angrist, J. D., Graddy, K., & Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, 67(3), 499 – 527.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444 – 455.

Ashenfelter, O., Ashmore, D., Baker, J. B., Gleason, S., & Hosken, D. S. (2006). Empirical methods in merger analysis: econometric analysis of pricing in ftc v. staples. *International Journal of the Economics of Business*, 13(2), 265 – 279.

Card, D. (1995). Earnings, schooling, and ability revisted. *Research in Labor Economics*, 14(23 - 48).

Card, D. & Krueger, A. (1996). *Does Money Matter?*, chapter Labor market effects of school quality: theory and evidence, (pp. 97 – 140). Brookings Institution Press: Washington D.C.

Chamberlain, G. & Imbens, G. W. (2003). Nonparametric applications of bayesian inference. *Journal of Business and Economic Statistics*, 21(1), 12 – 18.

Deaton, A. (1989). Rice prices and income distribution in thailand: a non-parametric analysis. *Economic Journal*, 99(395), 1 – 37.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. London: Chapman & Hall.

Goldberger, A. S. (1984). Reverse regression and salary discrimination. *Journal of Human Resources*, 19(3), 293 – 318.

Gupta, M. R. & Chen, Y. (2010). Theory and use of the em algorithm. *Foundations and Trends in Signal Processing*, 4(3), 223 – 296.

Guttag, J. V. (2013). *Introduction to Computation and Programming Using Python*. Cambridge, MA: The MIT Press.

Hansen, B. (2018). Econometrics.

Imbens, G. W. & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variable models. *Review of Economic Studies*, 64(4), 555 – 574.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

McKinney, W. (2017). *Python for Data Analysis*. Cambridge: O'Reilly.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6), 1135 – 1151.

VanderPlas, J. (2017). *Python Data Science Handbook*. Boston: O'Reilly.

Wasserman, L. (2004). *All of Statistics*. New York: Springer.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2nd edition.