Syllabus

Ec240a - Second Half, Fall 2017

Course Description

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 decision theory.

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 predictor 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 partial means and quantile regression.

Instructor: Bryan Graham, 665 Evans Hall, email: `bgraham@econ.berkeley.edu`
Time and Location: Monday and Wednesday, 10:00AM to 12:00PM, Barrows 170
Office Hours: Monday and Wednesday 9:00 AM - 10 AM (by Appointment)
Graduate Student Instructor: Ingrid Haegele (or Hägele), e-mail: inha@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.

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 also 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 at http://www.ssc.wisc.edu/∼bhansen/econometrics/. 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:

1. Ferguson, Thomas S. (1996). *A Course in Large Sample Theory.* London: Chapman & Hall.
2. Freedman, David A. (2009). *Statistical Models: Theory and Practice.* Cambridge: Cambridge University Press.
3. Manski, Charles F. (2007). *Identification for Prediction and Decision.* Cambridge, MA: Harvard University Press.
4. Wasserman, Larry. (2004). *All of Statistics.* New York: Springer.
5. Wasserman, Larry. (2006). *All of Nonparametric Statistics.* New York: Springer.

Ferguson (1996) is a compact introduction to large sample theory. Freedman (2009) is a book on applied data analysis and provides a useful complement to the material covered in lecture. This book has been used as a text for STAT 215 for many years here at Berkeley. 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 (**November 29th, 2017**). There will be 4 homework assignments (plus a review sheet). Homeworks are due at 5PM on the assigned due date. They are graded on a ten point scale with one point off per day late. You are free to work in groups but each student must submit an individual write-up and accompanying Jupyter Notebook (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 four problem sets.

The due dates for the four problem sets are:

Problem Set	Due Date
1	October 23rd
2	November 6th
3	November 20th
4	December 8th

Note problem sets 1 to 3 are due on Mondays, while problem set 4 is due the Friday of Reading/Review/Recitation Week.

Computation: All computational work should be completed in Python. Python is a widely used general purpose programming language with good functionality for scientific computing. I highly recommend the Anaconda distribution, which is available for download at http://continuum.io/downloads. Some basic tutorials on installing and using Python, with a focus on economic applications, can be found online at `http://quant-econ.net`. You may also wish to install Rodeo, which is an integrated development environment (IDE) tailored to statistics or "data science" applications. Rodeo makes working in Python look and feel similar to working in Stata or MATLAB. Rodeo is also free and available at `https://www.yhat.com/`.

Good books for learning Python, with some coverage of statistical applications, are Guttag (2013), VanderPlas (2017), and McKinney (2013). The last of these is now somewhat dated, but still useful (particularly for the pandas module). The first is an excellent introduction to computer science as well as Python.

The code I will provide will execute properly in Python 3.6, which is the latest Python release. Python is also available on the EML workstations. 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 his/her 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.

Additional notes: 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. Do feel free to chat with me immediately before class. For longer questions please make use of my office hours. This is time specifically allocated for your use; please come by! I look forward to getting to know all of you.

Course Outline

Date	Topic	Readings/Notes
M 10/16	**Probability Distributions**	Hansen (2015, Appendix B)
W 10/18	**Conditional Expectation Functions**	Hansen (2015, Ch. 2.1-2.17, 2.31-2.32) Wooldridge (2011, Ch. 2)
M 10/23	**K-Normal Means**	Wasserman (2006, Ch. 7)
W 10/25	**K-Normal Means**	Wasserman (2006, Ch. 7) Stein (1981)
M 10/30	**Linear predictors**	Hansen (2015, Ch. 2.18-2.28) Wooldridge (2011, Ch. 2) Card (1995), Card & Krueger (1996) Case & Paxson (2008)
W 11/1	**Bayesian Bootstrap**	Chamberlain & Imbens (2003)
M 11/6	**OLS**	Hansen (2015, Ch. 4.1-4.17, 5.1-5.15) Wooldridge (2011, Ch. 3)
W 11/8	**OLS**	Hansen (2015, Ch. 6.1-6.16) Wooldridge (2011, Ch. 4)
M 11/13	**Quantiles**	Mood, Graybill & Boes (1974, Ch. 11.3)
W 11/15	**Conditional Quantiles**	Chamberlain (1994) Hahn (1997)
M 11/20	**Partial Means**	Blundell & Powell (2003) Wooldridge (2005) Imbens (2004)
W 11/22	**No Class**	Thanksgiving recess
M 11/27	**Partial Means**	Robins, Mark & Newey (1992) Olley & Pakes (1996) Griliches & Mairesse (1998)
W 11/29	**2nd Midterm Exam**	Good luck!

Readings

1. Blundell, Richard W. and James L. Powell. (2003). "*Endogeneity in nonparametric and semi- parametric regression models*," Advances in Economics and Econometrics: Theory and Applications II: 312 - 357. (M. Dewatripont, L.P. Hansen, S. J. Turnovsky, Eds.). Cambridge: Cambridge University Press.

2. Card, David. (1995). "Earnings, school, and ability revisited," *Research in Labor Economics* 14(1): 23 - 48 (S.W. Polachek, Ed.). Greenwich, CT: JAI Press Inc.

3. Card, David and Alan B. Krueger. (1996). "Labor market effects of school quality: theory and evidence," *Does Money Matter: The Effect of School Resources on Student Achievement and Adult Success*: 97 - 140 (G. Burtless, Ed.). Washington D.C.: Brookings Institution Press.

4. Case, Anne and Christina Paxson. (2008). "Stature and status: height, ability, and labor market outcomes," *Journal of Political Economy* 116 (3): 499- 523.

5. Chamberlain, Gary. (1994). "Quantile regression, censoring, and the structure of wages," *Advances in Econometrics, Sixth World Congress I*: 171 - 209. (C. Sims, Ed.). Cambridge: Cambridge University Press.

6. Chamberlain, Gary and Guido W. Imbens. (2003). "Nonparametric applications of Bayesian inference," *Journal of Business and Economic Statistics* 21 (1): 12 - 18.

7. Griliches, Zvi and Jacques Mairesse. (1998). "Production functions: the search for identification," *Econometrics and Economic Theory in the 20th Century: The Ragner Frisch Memorial Symposium:* 169 - 203 (S. Strom, Ed.). Cambridge: Cambridge University Press.

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

9. Hahn, Jinyong. (1997). "Bayesian bootstrap of the quantile regression estimator: a large sample study," *International Economic Review* 38 (4): 795 - 808.

10. Imbens, Guido W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics* 86 (1): 4 - 29.

11. McKinney, Wes. (2013). *Python for Data Analysis.* Cambridge: O'Reilly Media, Inc.

12. Mood, Alexander M., Franklin A. Graybill and Duane C. Boes. (1974). *Introduction to the Theory of Statistics.* New York: McGraw-Hill, Inc.

13. Olley, G. Steven and Ariel Pakes. (1996). "The dynamics of productivity in the telecommunications equipment industry," *Econometrica* 64 (6): 1263 - 1297.

14. Robins, James M., Steven D. Mark and Whitney K. Newey. (1992). "Estimating exposure effects by modeling the expectation of exposure conditional on confounders," *Biometrics* 48 (2): 479 - 495.

15. Stein, Charles M. (1981). "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics* 9 (6): 1135 - 1151.

16. VanderPlas, Jake. (2017). *Python Data Science Handbook.* Cambridge: O'Reilly Media, Inc.

17. Wooldridge, Jeffrey M. (2005). "Unobserved heterogeneity and estimation of average partial effects," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 27 - 55 (D.W.K. Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.