Advanced Quantitative Methods

Jos Elkink

School of Politics and International Relations
University College Dublin

jos.elkink@ucd.ie
Newman Building, Rm F304
http://www.joselkink.net/teaching

Spring 2011

Introduction

This course extends the analytical and theoretical background developed in Quantitative Methods I. It focuses on building a greater understanding of the methods introduced in that course, such as the workings of multiple regression and problems that arise in applying it, as well as going deeper into the theory of inference underlying regression and most other statistical methods. The contents are in line with what would be covered in any modern introduction to econometrics book, but with applications in political science rather than economics.

This course is primarily about data analysis and developing a deeper understanding of the generalized linear model. The focus is on practice, and this focus is reflected in the choice of texts and in the emphasis on applied coursework. While this course deals to some degree with the generalized linear model on a mathematical and theoretical level, its main focus is practical, the ability to use the techniques when faced with the need in practical research. Consequently the learning method combines lectures and reading with hands-on statistical programming exercises using real datasets.

The statistical package being used is R[1] (see also Verzani 2005: Appendix A). You should download and install this at home, so you can get as much hands-on practice as possible.[2]

[1] Freely available at http://www.r-project.org.

[2] A very nice short video tutorial of R can be found here: http://www.decisionsciencenews.com/?p=261.

The learning outcomes associated with this twelve-week course are aimed at students being able to:

- Develop a deeper understanding of the linear regression model and its limitations;
- Know how to diagnose and apply corrections to some problems found in real data when applying the generalized linear model;
- Use and understand generalizations of the linear model to limited dependent variables (primarily binary data);
- Understand basic concepts of time series analysis;
- Develop a greater familiarity with a range of techniques and methods through a diverse set of theoretical and applied readings;
- Know where to go to learn more about the techniques in this class and those called for that were not covered in this class.

Prerequisites

Quantitative Methods I for Political Science or an equivalent course. A basic knowledge of mathematics, in particular algebra and simple calculus, is beneficial but not assumed. Also since the practice component is done in the R statistical package, it is assumed that students have already used this program in Quantitative Methods I.

Texts

This course will assign a variety of reading materials, some essential and some supplementary. Readings are absolutely *central* to this course and you will learn very little if you attempt to rely on the lectures alone.

As the main textbook you can use either Kennedy (2008) or Gujarati (2009). Kennedy (2008) provides a three-level discussion of each topic: first a general discussion, then a technical discussion, and then a very technical discussion. Most students find this quite useful since it permits them to dig as deep as their abilities let them or as their need allows. Gujarati (2009) uses a slightly more conventional approach, but contains a very clear exposition of basic econometrics. Where one of the two is unclear, it might help to check the equivalent chapter in the other. In addition to the main text, Faraway (2005) is a useful manual for linear regression in R, with brief introductions to each topic and clear demonstrations in R.

A text that is a very basic, very accessible but thorough introduction to statistics, written by statisticians, is Wonnacott and Wonnacott (1990). I have listed chapters on the recommended

lists for the first several weeks that will provide a very useful counterpoint to the applied, social science-oriented readings represented by the Kennedy and Faraway texts.

You may find some of the readings difficult or uncomfortable. This is completely normal. Your response should not be avoidance but rather a renewed effort to understand the material by (1) reading it with even greater care, (2) rereading it several times, (3) seeking other readings that might make the primary texts more comprehensible, and (4) working with other students in study groups. It is also perfectly normal in methods classes that you do not absorb all a text has to offer upon the first reading, but rather return to it several times over the years and learn new things as your knowledge accumulates.

There are two texts I would recommend that you purchase, namely either Kennedy (2008) or Gujarati (2009), and Faraway (2005). The other texts which you should consider purchasing, are Wonnacott and Wonnacott (1990); King (1998); Verzani (2005). The most useful references for working in R are probably Faraway (2005); Verzani (2005); Maindonald and Braun (2007); Gelman and Hill (2007). If you enjoy this course and consider applying it in your research, the next main step should be to read Gelman and Hill (2007), which is a slightly more advanced, but much more applied introduction to limited dependent variable and multilevel models, using R and additional software for Bayesian estimation.

Classes

Classes take place once a week, Wednesday from ... at ... . This is a computer lab (?), so we will be able to do practical exercises in class. The amount of material and the short duration of each session, however, will probably mean that 80% of the time will be taken up by lectures.

Grading

The only way to properly learn statistics is by hands-on training. You will need to work with actual data and produce your own statistical analyses - just the theory will never be sufficient. For that reason, a substantial part of the grading will be based on regular homework assignments. The assignments will be available online. For late submissions the standard policies apply, i.e. losing one point of a grade per day and a NG (no grade) for more than one week late. Exemptions will be granted only on the basis of illness or bereavement, documented in all cases.

Grading will be based on three components.

1. **Problem sets: 50%**. Problem sets will be handed out each Tuesday and must be submitted electronically before class the following Tuesday. Each problem set will consist of a number of problems combining computer analysis with interpretation and analytical

problems. Computer output, when supplied, should include both the commands used as well as results. Computer results should be indicated clearly. You are encouraged to work in groups on the problem sets, although work should be submitted individually. If you have any answers to problems that you wrote by hand, then you can use our department's excellent scanner to convert them easily to pdf. The grade will be determined by the average of the ten best grades.

2. **Replication project: 50%**. This project will be quantative reanalysis of a published quantitative work. Your job will be to obtain the data from the original author (or obtain the same data he or she used for the original piece), replicate his or her analysis, and extend the analysis using a new model or new variables whenever possible. If done properly this replication may be suitable for publication, which should be your objective. This project will require you to *begin searching immediately for an article to replicate*, including contacting the author or taking equivalent steps to obtain the data for your replication. The article you replicate may be from any field in political science, but must be an empirical application using inferential statistics. The only other restriction is that you may not replicate any article assigned for class reading or exercises. Your project must be submitted with your own replication dataset, so that someone else could replicate your analysis. You must also include a copy of the article whose analysis you have replicated. Examples of published replications may be seen in volumes 41 and 42 of the *American Journal of Political Science*, available for browsing through JSTOR. See also King (2006) for some helpful advice regarding an assignment very similar to this one.
   The deadline for this final assignment is **29/4, 2010, 5 pm**.

3. **Examination: 0%.** There is neither an exam nor a traditional research paper for this course. The problem sets substitute for the exam and the replication project replaces the research paper.

Plagiarism

Although this should be obvious, plagiarism - copying someone else's text without acknowledgement or beyond "fair use" quantities - is not allowed. UCD policies concerning plagiarism can be found online.[3] A more extensive description of what is plagiarism and what is not can be found at the UCD Library website.[4]

Contact

If you need to contact me outside class hours, you can find me in room F304 in the Newman Building at UCD. I do not have fixed office hours, so if you want to make sure I am present,

[3] http://www.ucd.ie/regist/documents/plagiarism_policy_and_procedures.pdf
[4] http://www.ucd.ie/library/students/information_skills/plagiari.html

you can make an appointment by email. If a personal visit is not necessary, the easiest way to reach me is by email (jos.elkink@ucd.ie).

Schedule overview

1	19/1	Mathematics review
2	26/1	Statistical estimators
3	2/2	Ordinary Least Squares
4	9/2	Hypothesis testing
5	16/2	Specification, multicollinearity and heteroscedasticity
6	23/2	Autocorrelation
7	2/3	Time-series analysis
		Study break
8	23/3	Maximum Likelihood
9	30/3	Limited dependent variables
10	6/4	Bootstrap and simulation
11	13/4	Multilevel data
12	20/4	Panel data

Schedule details

Week 1: Mathematics review

Introduction to the course; introduction to matrix and vector algebra; probabilities and probability distributions; non-technical discussion of derivatives.

The readings are quite substantial and mathematical, but primarily important as a reference throughout the course. It is not necessary to study this thoroughly for the first class.

required	Wonnacott and Wonnacott (1990: ch 3-4)
	Searle (1982: ch 1-3)
remedial	Gujarati (2009: Appendix B)
recommended	Namboodiri (1984)
recommended (R)	Maindonald and Braun (2007: ch 1, 14)
further	Hammer (1971: ch 1-2)
	Searle (1982)
	Davidson and MacKinnon (1993: Appendix A)
	Harville (1997)

Week 2: Statistical estimators

Discussing evaluation criteria of statistical estimators, including bias, consistency, asymptotics, errors.

required	Kennedy (2008: ch 1-2)
	Faraway (2005: ch 1)
remedial	Verzani (2005: ch 10)
	Gujarati (2009: Appendix A)
recommended	Wonnacott and Wonnacott (1990: ch 7)

Week 3: Ordinary Least Squares

Calculating the OLS estimates; properties of OLS estimators; assumptions underlying OLS properties.

required	Kennedy (2008: ch 3)
	or Gujarati (2009: ch 3)
	Faraway (2005: ch 2)
remedial	Verzani (2005: ch 7-8, 10)
	Wonnacott and Wonnacott (1990: ch 11-13)
	Gujarati (2009: ch 1-2)
	Maindonald and Braun (2007: ch 5-6)
recommended	Berry (1993)
	Gelman and Hill (2007: ch 2-4)
	Searle (1982: ch 14)
further	Verbeek (2008: ch 2)
	Gujarati (2009: ch 2, 4, 7, Appendix C)
	Kutner et al. (2005: ch 1, 5-6)
	Venables and Ripley (2002: ch 6)
	Maddala (2001: ch 3-4)
advanced	Davidson and MacKinnon (1993: ch 1)
	Greene (2003: ch 1-5)
	Amemiya (1985: ch 1)

Week 4: Hypothesis testing

Various testing methods in regression analysis, including t-test and F-test.

required	Kennedy (2008: ch 4)
	or Gujarati (2009: ch 5, 8)
	Faraway (2005: ch 3)
recommended	Wonnacott and Wonnacott (1990: ch 8-9)
	Maindonald and Braun (2007: ch 4)
further	Kutner et al. (2005: ch 2, 7)
	Maddala (2001: ch 2, 4, §3.5)
advanced	Greene (2003: ch 6)
	Davidson and MacKinnon (1993: ch 13)

Week 5: Specification, multicollinearity and heteroscedasticity

Detection of problems with model specification and multicollinearity in the independent variables. Detecting heteroscedasticity and discussion of consequences.

required	Kennedy (2008: ch 6-8, 12, 21)
	or Gujarati (2009: ch 10-11, 13)
	Faraway (2005: ch 4-5)
recommended	Berry and Feldman (1993)
further	Fox (1993)
	Verbeek (2008: ch 3, §4.1-4.5)
	Maddala (2001: ch 5, 12)
	Kennedy (2008: ch 21)
advanced	Greene (2003: ch 7-9, 11)
	Davidson and MacKinnon (1993: ch 16)

Week 6: Autocorrelation

Consequences of and testing for autocorrelation.

required	Gujarati (2009: ch 12)
further	Verbeek (2008: §4.6-4.11)
	Maddala (2001: ch 6)
advanced	Davidson and MacKinnon (1993: ch 10)
	Greene (2003: ch 12)

Week 7: Time-series analysis

Basic introduction to time-series data.

required	Kennedy (2008: ch 19)
	or Gujarati (2009: ch 17, 21)
	Verbeek (2008: ch 8-9)
recommended	King (1998: ch 7)
further	Venables and Ripley (2002: ch 14)
	Kutner et al. (2005: ch 12)
	Maddala (2001: ch 13-14)
	Maindonald and Braun (2007: ch 9)
advanced	Amemiya (1985: ch 5)
	Hamilton (1994)
	Davidson and MacKinnon (1993: ch 19-20)
	Greene (2003: ch 20)

Week 8: Maximum Likelihood

Introduction to and implementation of maximum likelihood estimators.

required	King (1998: ch 4)
	Wonnacott and Wonnacott (1990: ch 18)
further	Davidson and MacKinnon (1993: ch 8)
	King (1998: ch 3)
	Gelman and Hill (2007: ch 18)
	Verbeek (2008: ch 6)
	Maddala (2001: ch 16)
advanced	Greene (2003: ch 17)

Week 9: Limited dependent variables

Estimating and interpreting logistic and probit regressions, including ordinal and multinomial models.

required	King (1998: ch 5)
	Kennedy (2008: ch 16)
	or Gujarati (2009: ch 15)
recommended	Gelman and Hill (2007: ch 5-6)
further	Verzani (2005: ch 12)
	Long (1997)
	Maindonald and Braun (2007: ch 8)
	Aldrich and Nelson (1984)
	Kutner et al. (2005: ch 14)
	Verbeek (2008: ch 7)
	Venables and Ripley (2002: ch 7)
	Kennedy (2008: ch 17)
	Maddala (2001: ch 8)
	King and Zeng (2001*b*,*a*)
advanced	Davidson and MacKinnon (1993: ch 15)
	Greene (2003: ch 21, 22)
	Amemiya (1985: ch 9)
	Maddala (2001: ch 17)

Week 10: Bootstrap and simulation

Presenting logistic and probit regression results.

required	Verzani (2005: ch 6)
	King, Tomz and Wittenberg (2000)
further	Maddala (2001: ch 15)
	Kennedy (2008: ch 23)
	Davison and Hinkley (1997)

Week 11: Multilevel data

Conceptual discussion of multilevel data structures; fixed and random effects models.

required	Gelman and Hill (2007: ch 11-12)
recommended	Gelman and Hill (2007: ch 13-17)
further	Verbeek (2008: ch 10)
	Maindonald and Braun (2007: ch 10)
	Snijders and Bosker (1999)

Week 12: Panel data

Discussion of panel data analysis, having multiple observed units over multiple time periods.

required	Kennedy (2008: ch 18)
	or Gujarati (2009: ch 16)
advanced	Greene (2003: ch 13)
	Baltagi (2005)
	Hsiao (2003)
	Wooldridge (2002)

Additional readings

Not required reading, but potentially very useful additional references are:
King (2006) on how to write a publishable replication paper;
Kastellec and Leoni (2007) on how to use graphs instead of tables (I strongly support this approach!);
Brambor, Clark and Golder (2006) on interpreting interaction models;
Gelman and Hill (2007) is generally a very useful and accessible textbook on modern regression and multilevel analysis.

If you are interested in this course, there are a number of major topics that, due to time limitations, have been left out of this course. Topics you want to familiarize yourself with at least conceptually are (roughly in order of importance):

- missing data analysis, using statistical procedures to fill in the blanks in your datasets in such a way that the precision of your estimation increases, without the data in the blanks determining the results: King et al. (2001); Honaker and King (2010);
- estimating causal effects, thinking more carefully about the precise causal effect that is being estimated: Gelman and Hill (2007: ch 9-10); King, Keohane and Verba (1994: ch 3); Morgan and Winship (2007); Pearl (2000, 2009); jae Lee (2005);
- measurement error / error-in-variables, statistical models to deal with measurement error in the independent variables: Kennedy (2008: ch 10); Maddala (2001: ch 11);
- instrumental variable models: Kennedy (2008: ch 9-10); Davidson and MacKinnon (1993: ch 7); Verbeek (2008§5.3-5.5);
- survival analysis / duration models, statistical models that deal with explaining different survival or duration rates: Kennedy (2008§17.4);
- Bayesian analysis, a more flexible alternative to maximum likelihood, but based on a fundamentally different philosophical position towards statistical inference: Wonnacott and Wonnacott (1990: ch 19-20); Kennedy (2008: ch 14); Gelman and Hill (2007); Lancaster (2004); Gelman et al. (2004);
- simultaneous equation models, statistical models for dealing with complicated, endogenous models: Kennedy (2008: ch 11); Gujarati (2009: ch 18-20);

- spatial autocorrelation, regression models when observations are interdependent across space (or social network): Ward and O'Loughlin (2002); Beck, Gleditsch and Beardsley (2006); Franzese and Hays (2007); Anselin (1988, 2002).

References

Aldrich, John H. and Forrest D. Nelson. 1984. *Linear probability, logit, and probit models*. Beverly Hills: Sage.

Amemiya, Takeshi. 1985. *Advanced econometrics*. Cambridge: Harvard University Press.

Anselin, Luc. 1988. *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic Publishers.

Anselin, Luc. 2002. "Under the hood. Issues in the specification and interpretation of spatial regression models." *Agricultural Economics* 27:247–267.

Baltagi, Badi H. 2005. *Econometric analysis of panel data*. 3rd ed. Chichester: John Wiley & Sons.

Beck, Nathaniel, Kristian Skrede Gleditsch and Kyle Beardsley. 2006. "Space is more than geography: using spatial econometrics in the study of political economy." *International Studies Quarterly* 50(1):27–44.

Berry, William D. 1993. Understanding regression assumptions. In *Regression analysis*, ed. Michael Lewis-Beck. London: Sage.

Berry, William D. and Stanley Feldman. 1993. Multiple regression in practice. In *Regression analysis*, ed. Michael Lewis-Beck. London: Sage.

Brambor, Thomas, William Roberts Clark and Matt Golder. 2006. "Understanding interaction models: improving empirical analyses." *Political Analysis* 14(1):63–82.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and inference in econometrics*. Oxford: Oxford University Press.

Davison, Anthony and David Hinkley. 1997. *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Faraway, Julian J. 2005. *Linear models with R*. Boca Raton: Chapman & Hall.

Fox, John. 1993. Regression diagnostics. In *Regression analysis*, ed. Michael Lewis-Beck. London: Sage.

Franzese, Robert J. and Jude C. Hays. 2007. "Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data." *Political Analysis* 15(2):140–164.

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Analytical Methods for Social Research Cambridge: Cambridge University Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. *Bayesian data analysis*. 2nd ed. Boca Raton: Chapman & Hall.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River: Prentice Hall.

Gujarati, Damodar N. 2009. *Basic econometrics*. 5th ed. Boston: McGraw-Hill.

Hamilton, James D. 1994. *Time series analysis*. Princeton, NJ: Princeton University Press.

Hammer, A.G. 1971. *Elementary matrix algebra for psychologists and social scientists*. Rushcutters Bay: Pergamon Press.

Harville, David A. 1997. *Matrix algebra from a statistician's perspective*. New York: Springer-Verlag.

Honaker, James and Gary King. 2010. "What to do about missing values in time series cross-section data." *American Journal of Political Science* .
http://gking.harvard.edu/files/pr.pdf

Hsiao, Cheng. 2003. *Analysis of panel data*. 2nd ed. Cambridge: Cambridge University Press.

jae Lee, Myoung. 2005. *Micro-econometrics for policy, program, and treatment effects*. Oxford: Oxford University Press.

Kastellec, Jonathan P. and Eduardo L. Leoni. 2007. "Using graphs instead of tables in political science." *Perspectives on Politics* 5(4).
http://www.tables2graphs.com

Kennedy, Peter. 2008. *A guide to econometrics*. 6th ed. Malden, MA: Blackwell.

King, Gary. 1998. *Unifying political methodology. The likelihood theory of statistical inference*. University of Michigan Press.

King, Gary. 2006. "Publication, publication." *Political Science and Politics* 39(1):119–125.
http://gking.harvard.edu/files/paperspub.pdf

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American Political Science Review* 95(1):49–69.
http://gking.harvard.edu/files/evil.pdf

King, Gary and Langche Zeng. 2001*a*. "Explaining Rare Events in International Relations." *International Organization* 55:693–715.

King, Gary and Langche Zeng. 2001*b*. "Logistic Regression in Rare Events Data." *Political Analysis* 9:137–163.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the most of statistical analyses: improving interpretation and presentation." *American Journal of Political Science* 44(2):341–355.

King, Gary, Robert Keohane and Sidney Verba. 1994. *Designing social inquiry*. Princeton: Princeton University Press.

Kutner, Michael H., Christopher J. Nachtsheim, John Neter and William Li. 2005. *Applied linear statistical models*. 5th ed. McGraw-Hill.

Lancaster, Tony. 2004. *An introduction to modern Bayesian econometrics*. Malden, MA: Blackwell.

Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Maddala, G.S. 2001. *Introduction to econometrics*. 3rd ed. Chichester: Wiley.

Maindonald, John and John Braun. 2007. *Data analysis and graphics using R. An example-based approach*. 2nd ed. Cambridge: Cambridge University Press.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and causal inference. Methods and principles for social research*. New York: Cambridge University Press.

Namboodiri, Krishnan. 1984. *Matrix algebra: an introduction*. Quantitative Applications in the Social Sciences London: Sage.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearl, Judea. 2009. "Causal inference in statistics: an overview." *Statistics Surveys* 3:96–146.

Searle, Shayle R. 1982. *Matrix algebra useful for statistics*. New York: John Wiley & Sons.

Snijders, Tom and Roel Bosker. 1999. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage.

Venables, W.N. and B.D. Ripley. 2002. *Modern applied statistics with S*. 4th ed. Springer.

Verbeek, Marno. 2008. *A guide to modern econometrics*. Chichester: John Wiley & Sons.

Verzani, John. 2005. *Using R for introductory statistics*. Boca Raton, FL: Chapman & Hall/CRC.

Ward, Michael D. and John O'Loughlin. 2002. "Spatial processes and political methodology: introduction to the special issue." *Political Analysis* 10(3):211–216.

Wonnacott, Thomas H. and Ronald J. Wonnacott. 1990. *Introductory statistics*. 5th ed. New York: Wiley.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. MIT Press.