Institut for Political Science—University of Zürich

Quantitative Methods 1

Lucas Leemann

December 15-22, 2010

Contact Information:
Lucas Leemann, Department of Political Science, Columbia University.
Email: `ltl2108@columbia.edu`. Office hours: Daily, TBA.

Course Information:
Lecture times: Every day 10.15-12.00 and 13.30-14.30. Lecture location: Cityport Oerlikon. Recitation/Lab times: Every day 15.00-17.00. Recitation location: Cityport Oerlikon. Language of instruction: English or German (depends on students), slides will be in English.

Overview

In this class we will focus on three main topics; probability theory, maximum likelihood theory, and binary models. First, we will revisit the concept of a random variable, moments of a distribution, see how to derive them, and gain familiarity with a number of common distributions. Within this theoretical part we will also talk about estimation and how to evaluate estimators. Second, building on the statistical theory we will see how this can be applied to relevant problems. Maximum likelihood estimation is the corner stone of quantitative methods and a thorough understanding will be extremely helpful in the future. We will see how one derives an ML estimator for various problems, how one can analytically derive the estimator and the Hessian. In addition we will cover test theory and you will be introduced to the likelihood ratio test and the Wald test. After this theoretical part the focus will shift to programing estimators in R. We will rely heavily on R for various purposes during the entire class.

Finally, we will use the theory and apply it to regression models in which the outcome variable is binary. Such binary models, e.g. logit and probit, can be derived and programed (from scratch in R) by the students by the end of the

course. In this latter applied section, we will also spend time on interpretation (marginal effects, predicted probabilities, ...) and presentation (i.e. graphical) of statistical results.

We will rely on R during the entire class for illustrations, exercises, and coding estimators.[1] In the evening of the first day of class I will provide a 120 minute introduction for those unfamiliar with R. Nevertheless, I expect students being familiar with Stata.[2]

Basic Idea

Seven days is not a lot of time to cover the material. Students are expected to participate in all lectures and labs. In addition students are expected to be prepared and have done the relevant readings. Make sure you reserve the week for this course exclusively. You can expect to be occupied from 10am till 6pm with lecture, preparing and going to the labs. Afterwards you will need to read for the next day.

Since there is only limited time for the PhD quantitative methods classes the foci are slightly different. Instead of teaching students how to apply models and cover practical issues the emphasis will be on theory. If you understand the basic theories you will be able to learn from books on your own.

Books - Manuscripts

The main book will be Scott Long's *Regression Models for Categorical and Limited Dependent Variables*. Several other books are very helpful for this topic and will prove helpful during class.

- **Buy - or borrow:**

  Wooldridge, Jeffrey. 2006. *Introductory Econometrics: A Modern Approach*. Mason, OH: Thompson.

  Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

  Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice*. Newbury Park, CA: Sage.

[1] This program is free and available for download from `http://www.r-project.org/`. It is a fairly easy tool with which you will become very familiar during the class. Make sure that you download and install it. In addition, please make sure you also download these packages: `foreign`, `xtable`, `mvtnorm`, `MASS`, and `arm` .

[2] If anybody is not familiar with Stata I recommend doing the homework in the BA class *Vertiefung Methoden* which are carried out in Stata. Although Stata is a very important program and every PhD student has to know it, I will not spend any time on introducing the program.

- **You get as pdf or hard copy:**

  Steenbergen, Marco R. 2008. *Maximum Likelihood for Political Analysis: Theory and Applications.* Bern: Department of Political Science, University Bern. (pdf)

- **You might want to take a look at:**

  King, Gary. 1989. *Unifying Political Methodology. The Likelihood Theory of Statistical Inference.* New York: Cambridge University Press.

  Faraway, Julian. 2006. *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Boca Raton: Chapman & Hall/CRC.

- **Refresher mathematics:**

  Chiang, Alpha C. and Kevin Wainwright. 2005. *Fundamental Methods of Mathematical Economics.* New York: McGraw Hill.

- **Statistics refresher and background:**

  Wooldridge, Jeffrey. 2006. *Introductory Econometrics: A Modern Approach.* Mason, OH: Thompson. **(Appendix B)**

  DeGroot, Morris H. and Mark J. Schervish. 2002. *Probability and Statistics.* Reading, MA: Addison-Wesley, 2nd edition.

- **Software:**

  R: Fox, John. 2002. *An R and S-PLUS Companion to Applied Regression.* New York: Sage Publication.

  R: Venables, William N. and Brian D. Ripley. 2002. *Modern Applied Statistics with S-PLUS.* Fourth ed. New York: Springer-Verlag.

  R: Kuhnert, Petra and Bill Venables. 2005. *An Introduction to R: Software for Statistical Modelling & Computing.* CSIRO Mathematical and Information Sciences Cleveland, Australia. (pdf)

  Stata: Long, J. Scott, and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata.* College Station: Stata Press, 2nd ed.

Grading:

- *Take-Home Exam:* The first take-home exam will cover all the material we discuss in the first few days (W, TH, F). You will receive a set of questions and you will have a certain amount of time (probably 12 hours) to answer these questions and submit your results. You can work wherever you feel comfortable. You can use any sources (books, scripts, class notes, web, ...) – but you must not ask other people, whether these are friends or fellow class mates. The exam is likely to take place on the one of the weekend days (18/19).

- *Application with Real Data:* A short three to four page paper showing the students ability to use the learned skills. Any social science data may be used – either a binary model or a more advanced model should be employed. Special emphasis is on interpretation and illustration of the results. Students are encouraged to use either Stata or R. The deadline for this final assignment will be mid January. I will make myself available in early January for office hours. I expect every student to discuss her/his project with me before starting it.

Class and Lab:

The class is a straight lecture. The quality of the class will depend heavily on your participation. Make sure to be prepared for every class and ask whenever anything remains unclear. On most days there will be two lectures and one lab. On certain days we will convert the second lecture into a lab and work on problem sets.

The labs will be structured around a problem set. You will receive this problem set in advance. I expect you to work through it in advance. This will allow us to skip all the pieces which are clear to everybody and we can focus on the tricky parts.

December 15, 2010 (Lll)

Class:
Repetition of statistical theory and the linear regression model. We will cover expectation, variance, discrete and continuous distributions. What is estimation and how are estimators evaluated? In the second part we focus on the linear regression model – the assumptions, the violations thereof, and possible remedies.
Lab:
The first part is an introduction to R for those students who do not know R yet. In the second part we will see how one can use simulations to test estimation performance. This procedure is known as *Monte Carlo simulation*.

Readings:

- Gujarati, Damodar. 2003. *Basic Econometrics*. Boston, MA: McGraw-Hill. **(read 3.8)**
- Wooldridge, Jeffrey. 2006. *Introductory Econometrics: A Modern Approach*. Mason, OH: Thompson. **(skim Appendix B)**
- Wooldridge, Jeffrey. 2006. *Introductory Econometrics: A Modern Approach*. Mason, OH: Thompson. **(skim Ch. 2, 3, and 4)**

December 16, 2010 (Lll)

Class:
Theory of maximum likelihood estimation. Estimation of a parameter, estimation of the variance of the parameter estimate. Simple hypothesis tests. Derivation of simple parameter estimates for Bernoulli and normal distribution.
Lab:
We will see how one can write a function in R. We will use a numerical optimizer and find maxima and minima of different functions. We will carry out a maximum likelihood estimation of a single parameter for a Bernoulli distribution and the normal distribution.

Readings:

- Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice*. Newbury Park, CA: Sage. **p. 1-25**
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. **(Ch. 3.1-3.6)**
- Steenbergen, Marco R. 2007. *Maximum Likelihood for Political Analysis: Theory and Applications*. Chapel Hill: Department of Political Science, University of North Carolina. **(tba)**
- * Train, Kenneth E. 2003. *Discrete Choice Methods with Simulation*. New York: Cambridge UNiversity Press.
- *Long, J. Scott, and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. College Station: Stata Press.

December 17, 2010 (LLl)

Class:
If the outcome variable is binary (e.g. yes or no vote) we are usually reluctant to estimate a linear probability model. We will briefly revisit the LPM (linear probability model) and see how one would apply it. We will then turn to Logit and Probit. Both models will be derived and thereby we will cover stochastic utility theory. Participants will be introduced to the latent variable framework.

Lab: We write/program are own Logit and Probit functions in R. For the lab we will use data on the Titanic passengers and estimate different models – we will see that the survival probability depends on gender, boarding class and age.

Readings:

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. **(Ch. 3.1-3.6)**
- Steenbergen, Marco R. 2007. *Maximum Likelihood for Political Analysis: Theory and Applications*. Chapel Hill: Department of Political Science, University of North Carolina. **(tba)**

December 18/19, 2010

Take Home Exam: There will be a take-home exam covering all the material up to the weekend. You will have a given amount of time to solve the exercises. You may use any non-human resources (class notes, slides, books, webpages,...). You must not ask other class mates or colleagues for support. I will make myself available during the exam period and will assist if necessary (via e-mail). On the remaining day we will not meet for class and there will be no homework.

December 20, 2010 (Lll)

Class:
Second part of maximum likelihood theory. We will discuss alternative hypothesis tests, such as the likelihood ratio test and the Wald test. Several measures of model fit will be introduced. This will allow students to carry out most of the hypotheses test they will be interested in.

Lab:
I will discuss the solutions of the take-home exam. Hypothesis tests in R. We will focus on the Wald test as it is the more complicated of the two tests. Most exercises will be based on the binary models from Wednesday.

Readings:

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. **(Ch. 3.7, 4.1 & 4.3)**
- Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice*. Newbury Park, CA: Sage. **p. 28-34**
- Steenbergen, Marco R. 2007. *Maximum Likelihood for Political Analysis: Theory and Applications*. Chapel Hill: Department of Political Science, University of North Carolina. **tba**

December 21, 2010 (Lll)

Class:
Third part of maximum likelihood theory. We will discuss ways of interpreting estimation results. We will especially focus on change in predicted probabilities and marginal effects. We will use these tools to see how one can effectively communicate the key results of a statistical model.

Lab:
How can we generate the desired quantities, i.e. predicted probabilities and marginal effects, in R? How do we do that in Stata?

Readings:

- Long, J. Scott, and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. College Station: Stata Press, 2nd ed.
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. **(Ch. 3.7, 4.1 & 4.3)**
- Steenbergen, Marco R. 2007. *Maximum Likelihood for Political Analysis: Theory and Applications*. Chapel Hill: Department of Political Science, University of North Carolina. **tba**

December 22, 2010 (Lll)

Class:

Many possible outcome variables in the social sciences are ordered. Typically these variables take on three, five or seven categories and we know the ordering but not the distance between two answer categories. A typical example for such a variable is the answer to how much one favors policy A over policy B where the answer possibilities are 'less', 'equal', or 'more'. The correct model for such an outcome variable may be the ordinal Logit or ordinal Probit model.

Lab:

We will write/program our own ordered Logit and ordered Probit model. We will see what the crucial assumption (PRA) is and what its implications are.

Readings:

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. **(Ch. 5.1-5.7)**
- Steenbergen, Marco R. 2007. *Maximum Likelihood for Political Analysis: Theory and Applications*. Chapel Hill: Department of Political Science, University of North Carolina. **(tba)**
- * McKelvey RD, Zavoina W (1975). "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," *Journal of Mathematical Sociology* 4: 103-120.