10th ECPR Summer School in Methods and Techniques, 23 July - 8 August
University of Ljubljana, Slovenia
Course Description Form[1] - Week 1 course, 15 hrs (27 July - 31 July)

Course title

SD105. Multiple Regression Analysis: Estimation, Diagnostics, and Modeling

Instructor details

First name, last name : Prof. Dr. Bernhard Kittel
Department/Unit : Institute of Economic Sociology
Institution : University of Vienna
Full postal address for ECPR correspondence : Oskar-Morgenstern-Platz 1, 1090 Wien
Phone : 0043-1-4277-38311
E-mail : bernhard.kittel@univie.ac.at

Short Bio

Bernhard Kittel is professor of Economic Sociology at the University of Vienna. Previous engagements include the University of Oldenburg, the University of Amsterdam, the University of Bremen, and the Max Planck Institute for the Study of Societies in Cologne. His research interests cover collective decision-making, political economy, and comparative research methodology.

Prerequisite knowledge

a) The course makes use of the freeware statistical package R, which can be downloaded from http://www.r-project.org, and assumes basic skills in using the package. For students unfamiliar with R, a preparatory course will be offered prior to the course. A good introduction into using R for statistics is Field et al (2012).

b) Students are expected to understand the logic of inferential statistics. Prior knowledge of multivariate statistics is useful but not necessary. Remedial reading: any good statistics textbook, for example: Alan Agresti & Barbara Finlay (2008). Students familiar with R but in need of a refresher in basic statistics are encouraged to take part in the preparatory course on statistics.

c) Furthermore, although a brief session on the use of matrix algebra in regression analysis is scheduled at the beginning of the course, it is assumed that the logic of matrix algebra is understood at the level presented in Gujarati.

Short outline

The course starts with a discussion of the logic of the multivariate regression model and the central assumptions underlying the ordinary least squares approach. This part includes an extensive discussion of the derivation of the ordinary least squares estimator, its standard error and summary statistics. Then it proceeds with testing for the adequacy of the assumptions and suitable corrections and extensions to the estimation techniques in the context of cross-sectional data. Particular emphasis will be laid on influential cases, multicollinearity, heteroskedasticity, and autocorrelation. Finally, categorical predictors, interactions, and nonlinearities will be considered.

Long outline

Multiple regression analysis has become the workhorse of quantitative analysis in many social sciences. In the process of learning statistics, a thorough discussion of the multiple regression model

[1] *Disclaimer: the information contained in this course description form may be subject to subsequent adaptations (e.g. taking into account new developments in the field, specific participant demands, group size etc.). Registered participants will be informed in due time in case of adaptations.*

constitutes the bridge between basic statistics and advanced statistical and econometric techniques, which are all in one way or another extensions or alternatives to the linear ordinary least squares approach of multiple regression. Hence mastering multiple regression opens the avenue to a wide variety of more sophisticated techniques for quantitative social science data analysis.

The course starts by discussing the derivation of the ordinary least squares principle in a bivariate model and then extends the idea to the matrix representation of the multivariate model. A brief introduction to the matrix algebra of regression analysis is included. Then, the standard errors are derived and fundamental criteria for model adequacy (t-test, F-test, coefficient of determination) are discussed. It then proceeds to diagnosing model robustness and fit in more detail, covering multicollinearity, including tests and remedial actions, as well as measures for leverage and influence of observations. Furthermore, residual structures will be discussed. The assumption that standard errors are identically distributed across observations (homoscedasticity) and the assumption that residuals are uncorrelated (no autocorrelation) will be covered along with test procedures and remedial treatments. Finally, some deviations from the linearity assumption will be considered. We will first discuss situations with categorical independent, then explore situations in which the effect of one variable depends on the level of another variables, and end by studying nonlinear associations between the independent variables and the dependent variable.

The course is a direct precursor to "Generalized Linear Modeling" offered in the second week of the summer school. It is recommended to take both courses as two modules of a full two-week course on "Multiple Regression Analysis and Generalized Linear Modeling".

In the lab exercises, a variety of data sets will be used, in particular individual-level data from the European Social Survey will be used. This dataset is freely available at www.europeansocialsurvey.org. Some topics will be discussed on the basis of a dataset containing aggregate political economy data. All lab sessions will use the software package R (www.r-project.org). The course is set up as follows: First a topic is presented in the lecture, then students are asked to work on a set of tasks in which they stepwise develop and test a regression model, using the ideas and formulae presented in the lecture.

Note on readings: No book is fully satisfactory in the details of presentation or the emphases laid. Therefore, I work with two textbooks which I like for different reasons, Gujarati and Fox. I expect participants to have worked through at least one of the indicated chapters in the literature when we discuss the topic in class.

It is highly recommended to bring a copy of the books as well as a laptop computer with a recent version of R installed.

The course assumes proficiency with descriptive and inferential statistics at the level of test theory and bivariate regression analysis. It is highly recommended bring along a basic understanding of the logic of matrix algebra.

Day-to-day schedule

	Topic(s)	**Details**
Day 1 (90')	The Logic of Regression Analysis: Coefficients	90' lecture - Topic of course and course goals - Derivation of regression coefficients 90' Lab: Matrix algebra for coefficients
Day 2 (3 hours)	Interpreting Regression Results: Standard Errors and Summary Statistics	90' Lecture: Derivation of standard errors, hypothesis testing 90' Lab: Matrix algebra for standard errors, basic tests and summary statistics model assessment

Day 3 (3 hours)	Regression Diagnostics 1: Influential Cases and Multicollinearity	90' Lecture:, OLS assumptions, outlying and influential observations, multicollinearity 90' Lab: Diagnostics and interpretation of results
Day 4 (3 hours)	Regression Diagnostics 2: Heteroskedasticity and Autocorrelation	90' Lecture: Heteroskedasticity and Autocorrelation: problem, tests, remedial treatments 90' Lab: Testing and correcting for heteroskedasticity and autocorrelation
Day 5 (3 hours)	Regression Modeling	90' Lecture: Dummies, Interactions, Nonlinearities 90' Lab: Model specification

Day-to-day reading list

	Readings
Day 1	Gujarati, Ch. 1, 2, 3 Fox, Ch. 5
Day 2	Gujarati, Appendix B, C, Ch. 4
Day 3	Gujarati, Ch. 5, 7, 8, 10, 13.1-13.10 Fox, Ch. 6, 11, 12.1, 12.3, 13
Day 4	Gujarati, Ch. 11, 12 Fox, Ch. 12.2, 16
Day 5	Gujarati, Ch. 6 Fox, Ch. 4, 7 Brambor et al 2006

Software and hardware requirements

Software programme
R, http://www.r-project.org

Hardware requirements
None.

Literature
Readings
Fox, J., 2008. Applied Regression Analysis and Generalized Linear Models, London: Sage.
Gujarati, D.N. and D. Porter, 2009. Basic Econometrics, 5th edition, New York: McGraw-Hill.
Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. Understanding Interaction Models: Improving Empirical Analysis. *Political Analysis* 14 (1): 63-82.

Further Literature
Agresti, A. and B. Finlay, 2008. Statistical Methods for the Social Sciences, 4th edition, Pearson.
Field, A., J. Miles and Z. Field, 2012. Discovering Statistics Using R, Sage.
Fox, J. and S. Weisberg, 2010. An R Companion to Applied Regression, London: Sage.
Fox, J., 2003. Effect Displays in R for Generalised Linear Models, Journal of Statistical Software, 2003, Vol. 8, No. 15

Lecture room requirement
Ordinary lecture hall with beamer facilities, computer lab