Statistical methods for register-based research

During the academic year of 2017-2018 the Graduate School in Population Dynamics and Public Policy offer in collaboration with the Department of Statistics at Umeå School of Business and Economics at Umeå University a package of four courses on the theme *Statistical methods for register based research*.

The package is open to PhD students from all faculties working with register data. The aim of the course package is to provide the PhD student with adequate statistical tools and knowledge to analyse register-based data. The package consists of four courses:

1. Introduction to R
2. Analysis of longitudinal and hierarchical/multilevel data with R
3. Causal inference in register studies with R
4. Statistical analysis of big and complex data with R

The student is free to take all four courses or to cherry-pick from them. If not familiar with the statistical software package R, it is highly recommended to take the first course, “Introduction to R”, since R will be used during the subsequent courses.

**Required prior knowledge:**
At least 15 Credits in Statistics, e.g., *Quantitative Research Methods for the Social Sciences* (master level/post-graduate level) or *Statistics A1* (basic level).

**Registration:**
Register to Johan Lundberg by august 1st 2017.

**Contact persons:**
Director of The Graduate School in Population Dynamics and Public Policy:
Johan Lundberg, Department of Economics, USBE, johan.lundberg@umu.se, 090-786 95 40

Director of Studies at The Graduate School in Population Dynamics and Public Policy:
Maria Karlsson, Department of Statistics, USBE, maria.karlsson@umu.se, 090-786 55 96
Course 1: Introduction to R

When? September 2017

How? 2-3 half days of lectures and computer lessons. The examination consists of written assignments.

What? R is a free software environment for statistical computing and graphics. R and its libraries implement a wide variety of statistical and graphical techniques. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages.

In this course, your will learn the basics in R and R-studio, which is a practical and more user-friendly environment to run R in. You will learn to import data into R/R-studio, “data manipulation” (i.e., selecting, inserting, deleting and updating data), some tools for visualizing data, but also to analyse data with some of the methods that you already should be familiar to (e.g., regression analysis), using the built-in functions in R. You will also learn strategies to learn more about R yourself and strategies to search for and use (contributed) packages with a specific topical function.

Why should you learn to use R? Here are two strong arguments:

- R is a popular (maybe the most popular) data programming language, and ranks highly in several popularity surveys, see e.g., http://blog.revolutionanalytics.com/2016/07/r-moves-up-to-5th-place-in-ieee-language-rankings.html
- Learning R is a great way of learning statistics (or data science) too. There are many R-based books and resources for statistics, data visualization, machine learning and more.
Course 2: Analysis of longitudinal and hierarchical/multilevel data with R

When? October 2017

How? 2-3 half days of lectures and computer lessons. The examination consists of written assignments.

What? Longitudinal data consists of measurements of explanatory and response variables at several points in time for many individuals. This means that the usual assumption of independent data is not valid, observations are typically correlated when the same individuals are measured repeatedly. In this course, you will get familiarized with methods that can deal with this type of data.

Many of the methods suitable for longitudinal data are also useful for analysing data with a hierarchical/multilevel structure also making observations not statistically independent. Hierarchical data can, e.g., be patients nested (grouped) within hospitals, employees nested within firms, students nested within schools, and so on.

Drop-out is a very common complication in the analysis of data from longitudinal studies which may bias the results. Methods for dealing with drop-out is therefore an important topic that covered in this course.

In the course, you will also acquire knowledge about methods for so called sequence analysis, i.e., analysis of state (or event) sequences describing life trajectories such as family life courses or employment history.

---

1 Repeated measurements data and panel data are common synonyms to longitudinal data.
Course 3: Causal inference in register studies with R

**When?**
November 2017

**How?**
3-4 half days of lectures and computer lessons. The examination consists of written assignments.

**What?**
In many empirical applications of statistics, the central aim of a study is a question about causality rather than questions about description or association. To move from association to causation is however not trivial. Several schools of thought have been developed over the past decades that have expanded the collection of methods available for causal inference.

In this course, you will acquire awareness of the common threads across methods for estimating the (causal) effect of a treatment/intervention on some outcome variable/-s of interest. An important matter that will be discussed during the course is the different methods dependence on (untestable) assumptions. You will also acquire competence in applying these methods in simple settings.

Register studies commonly involve drawing causal conclusions about investigated relationships. A key feature of register data with respect to causality is that the data is observational and not experimental and thereby rely on strong assumptions for causal analysis.
Course 4: Statistical analysis of big and complex data with R

When? February – March 2018

How? Customize your own course by choosing one or more of the modules listed below. The course is taught with “flipped classroom”-pedagogics, i.e., you will see video lectures on your own and then meet the teacher and your fellow students in the classroom where you will discuss questions related to the course content and solve exercises (by hand or in R), together.

The course is coordinated with parts of the course “Statistik C1”, which is the second to last course before receiving a Bachelor’s degree in Statistics (only the Bachelor’s thesis course comes after this course).

The examination consists of written assignments.

What? Statistical methods aimed at solving the problems involved in analysing big and complex data sets, with emphases on methods for model selection and dimension reduction. The methods considered can be divided in two main groups: “supervised (statistical) learning” methods and “unsupervised (statistical) learning” methods. Supervised learning involves building statistical models for predicting a response variable based on one or more explanatory variables (predictors, covariates), while unsupervised learning is used to find relationships and structures in data where none of the variables is a response variable. Supervised learning methods consists of regression methods, in a broad sense, and classification methods, i.e., methods for prediction of qualitative response variables.

Module 1: Methods for classification. (1.5 weeks fulltime studies)

Module 2: Advanced regression methods: model selection, regularization and moving beyond linearity. (3.5 weeks fulltime studies)

Module 3: Introduction to unsupervised learning methods (1.5 weeks fulltime studies)

---

2 If two or more modules are selected, you can subtract half a week from the total length of the course since there is an overlap of course content between the modules.