Scientific report of the EHPS-Net International Summer School 'Demography and Event History Analysis using R', 1-12 June 2015, Centre for Demography and Ageing Research, Umeå University, Umeå, Sweden

Summary

This course is supposed to be advanced, by which we mean that participants should have some knowledge about basic statistical concepts, like sample, mean, estimate, standard error, hypothesis testing. Knowledge about survival analysis and Cox regression is desirable but not required.
The work horse *RStudio* is used for producing dynamic statistical reports, a basis for documentation and reproducible research. The plan is to introduce and utilize the ideas of literate programming in this way, thus making statistical programming more intuitive and easy to comprehend. The R markdown tools (Allaire et al., 2015), included in RStudio, are used.
We start by providing a basic overview of R (R Core Team, 2015) as a statistical environment. It is shown by example how R is used for solving traditional statistical tasks like displaying data by graphs and tables, summary statistics, linear regression, and survival analysis (Broström, 2012, 2015; Therneau, 2015). Then RStudio is introduced, together with the knitr (Xie, 2015) package which is used for data analysis and report writing in R markdown. Output is possible to get in three flavors: HTML, pdf, and (not really desirable) as a Word document.
The statistical content focuses on survival and event history analysis. A brief introduction to the statistical theory and its demographic applications is given, and hands-on instructions of how to perform the research and at the same time write a research report will be emphasized throughout the course. Most of the course material is publicly available on GitHub (http://github.com/goranbrostrom/cedar) and on the course home page, http://capa.ddb.umu.se/cedar15.
Presentation of the result of research project, given during the first week, concludes the course. There is no formal examination, but participants are given a certificate stating their participation in the course and in the research project. Participants are working on the projects in small groups.

Scientific content and discussions at the event

The main course themes were *reproducible research* and good habits in report writing, *event history and survival analysis* (Broström, 2012), *version control*, and *R programming* with the lexical approach provided by *R markdown* in *RStudio*.
The more scientific-demographic content was concentrating on fertility and mortality, and less migration, as application fields of survival and event history analysis. The main topics were *the life table and its connection to the survivor function, non-parametric estimation of proportional-hazards models, parametric proportional-hazards and accelerated-failure-time models*, with the *Gompertz* distribution for the study of adult mortality.
One session was devoted to *causality and statistics* (including *Simpson's paradox*), with *matching as a main theme*. The concept *local independence* was introduced and exemplified by a study on infant and maternal mortality (Broström, 1987).
Topics touched upon include *choice of time scales*, *stratification*, and *frailty models*, the latter an important ingredient in *multiple-events models*, frequently present in *fertility studies*.
The seven guest lectures (see Programme) were much appreciated.

Discussions at the event

There were many discussions regarding *interactions* and their interpretation, the superiority of the *likelihood ratio test* over the *Wald test*, and how to do everything in Rand *RStudio*.

The very sensitivity of the proportional hazards assumption was discussed in connection with omitting important covariates (don't do that!). It was emphasized that *Cox regression* is much overused. Often there are better alternatives, like parametric proportional hazards models and/or accelerated failure time models.

References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., and Hyndman, R. (2015). *rmarkdown: Dynamic Documents for R*. R package version 0.7, http://CRAN.R-project.org/package=rmarkdown.

Broström, G. (1987). The influence of mother's mortality on infant mortality. a case study in matched data survival analysis. *Scandinavian Journal of Statistics*, 14:113–123.

Broström, G. (2012). *Event History Analysis with R*. CRC Press/Chapman & Hall, Boca Raton.

Broström, G. (2015). *eha: Event History Analysis*. R package version 2.4-3.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Therneau, T. (2015). *survival: Survival Analysis in R*. R package version 2.38-2, http://CRAN.R-project.org/package=survival.

Xie, Y. (2015). *knitr: A General-Purpose Package for Dynamic Report Generation*. R package version 1.10.5, http://CRAN.R-project.org/package=knitr.

Schedule

Monday 1 June

9:00–12:00, room 1007

Maria Josefsson, Erling Häggström Lundevaller, Göran Broström
- Course introduction
- EHPS-Net and CEDAR presentations
- Course presentation
  - Why *R* and *RStudio*?

14:00–15:15, room 1007

Maria, Erling, Göran: Data sets and Facilities

15:30–18:00, room 1031

Invited lectures:
- Wolfgang Lutz, Vienna Institute of Demography: *World population and human capital in the 21st century*
- James Vaupel, Max Planck Institute for Demographic Research, Rostock: *Life expectancy and lifespan equality*

19:00 Welcome dinner, restaurant Rex

Tuesday 2 June

9:00–12:00, Aula Nordica

Population Aging and Longer Lives (KVA Conference):
- Wolfgang Lutz: *The effects of education over the life course: Macro level implications*
- James W. Vaupel: *The remarkable rise of longevity*

- Coffee break
- Tom Kirkwood: *Why and how are we living longer?*

14:00–17:00, room 1006

Erling, Maria
- *R* and *RStudio* basics
- Data types
- Data input/output
- Basic statistical tasks
  - summary statistics
  - tables
  - regression analysis and the formula interface
  - graphics

Wednesday 3 June

9:00–12:00, room 1007

Göran: Introduction to event history analysis
- *Censoring*
- *Length-biased sampling and left truncation*
- Characterization of survival distributions
  - The *hazard* function
  - The *survival* function
- The Lexis diagram
- The *eha* package

14:00–17:00, room 1007

Maria, Erling, Göran
- Introducing data from the Skellefteå region
- Distribution of projects
- Introduction to *R* markdown

Thursday 4 June

9:00–12:00, room 1007

Göran: Cox regression
- proportional hazards
- the log-rank test
- explanatory variables
- Doing it in *R*

14:00–17:00, room 2005

Maria, Göran: Practising
- The *survival* package
- Cox regression on
- mortality studies

Friday 5 June

9:00–12:00, room 1007

Erling
- R markdown, *knitr*

- R programming
- report writing

14:00–17:00, room 1007

Maria, Erling
- Practicing today's lecture
- Start working on projects

Saturday 6 June

National Day of Sweden

Evening: Shrimps boat on the Umeå river

Sunday 7 June

Own time.

Monday 8 June

9:00–12:00, room 1006

Göran
- *Parametric* proportional hazards models
- pros and cons (visa vi Cox regression)
- The *Gompertz* distribution and adult mortality
- The *Accelerated Failure Time* (AFT) model

14:00–17:00, room 1006

Göran: Theme *mortality:*
- Guest lecture: Professor Sören Edvinsson talks about adult mortality and socio-economic status
- Old age mortality
- Practicing today's lecture

Tuesday 9 June

9:00–12:00, room 1006

Göran: Multiple events data

- Fertility:
  - Choice of time scales
  - Stratification
  - Frailty models
- Competing risks models

14:00–17:00, room 1006

Göran, Erling
- Guest lecture: Dr Glenn Sandström talks about fertility and sex preferences
- Practicing, project work

Wednesday 10 June

9:00–12:00, room 1006

Göran
- *Causality*, what do we mean?
- Simpson's paradox
- Matching

14:00–17:00, room 1006

Göran, Erling: Practicing matching and project work

19:00 Guitars museum with dinner

Thursday 11 June

9:00–12:00, room 1006

Erling, Göran, Johan Junkka
- Reproducible research
- Version control
  - Git
    - GitHub

14:00–17:00, room 1006

Erling, Göran: Project work, report writing

Friday 12 June

9:00–12:00+, room 1006

Göran, Erling
- Project presentations,
- Course evaluation.
- Conclusion.

Assessments of the results and impact of the event on the future directions of the field

The participants worked hard on their projects, that were distributed early, maybe too early. However, their project presentations (some available at http://rpubs.com/cedar) on the last day showed that they had learned a lot.

Impact on future directions
We, both the participants and the teachers, learned a lot during the course. The teachers learned that they were not coherent in the teaching of "how to do it in **R**": The power of the R package *dplyr* was underestimated (by me) and was introduced too late in order to be useful in the project work phase. The projects were distributed *too* early, which had as a consequence that, quite naturally, the students focused on them rather than practice on the exercises connected to the lectures.

Conclusion
The next time the course will be straightened up, more focus on early learning of basic **R**/RStudio skills. The scientific content (event history analysis) will be more concentrated to fewer topics, but deeper.
The teaching staff was under-dimensioned: One or two lab assistants more had made a large difference. This will be taken care of for the next course opportunity, in, say, two years time.

Please refer to the Annexes – Course evaluation for more information on this subject.

Annexes

Full list of instructors

Göran Broström professor, PhD of Mathematical Statistics, Umeå University: Event History Analysis, the eha and survival packages in R. Assisted under exercises and project work.

Erling Häggström Lundevaller assistant professor, PhD of Statistics, Umeå University: Basics in R and RStudio. Assisted under exercises and project work.

Johan Junkka PhD candidate of History, Umeå University: The dplyr package for easy data management.

Maria Josefsson post doc, PhD of Statistics, Umeå University: Assisted under exercises and project work (first week).

Guest lecturers

Wolfgang Lutz, Vienna: "World population and human capital in the 21st century", and "The effects of education over the life course: Macro level implications". Two hours.

James Vaupel, Rostock: "Life expectancy and lifespan equality", and "The remarkable rise of longevity". Two hours.

Tom Kirkwood, Newcastle: "Why and how are we living longer?"

Sören Edvinsson, Umeå: "The Demographic Data Base", and "Socio- economic status and adult mortality". Two hours.

Glenn Sandström, Umeå: "Fertility and sex preferences". Two hours.

Full list of students

In total eighteen participants came from nine different countries. The distribution is shown in Figure 1.

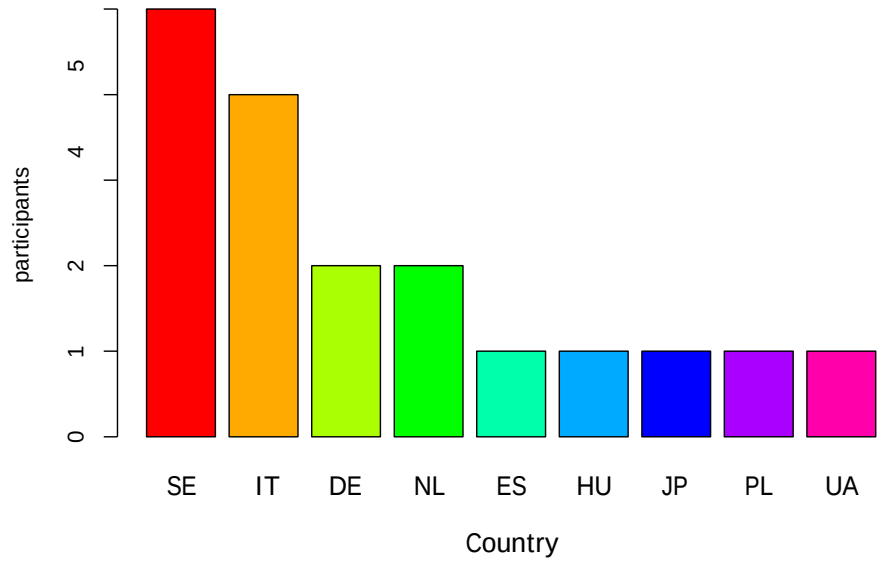

Figure 1: Distribution of participants over countries (ISO 3166-1 alpha-2).

Germany

Stephanie Klages
Contact: s.klages@uni-muenster.de
Department: Institute of Economic and Social History
University: University of Münster

Position: PhD Candidate
Research interest: The impact of economic stress on marriage behavior and fertility in Wittgenstein, 1750–1875.

Benjamin Matuzak
Contact: bmatuzak@gmx.net
Department: History
University: Martin-Luther-Universität, Halle
Position: PhD Candidate
Research interest: Life under Economic Pressure in the 19th Century: Borders and Demographic Systems.

Hungary

Gábor Koloh
Contact: koloh.gabor@gmail.com
Department: History
University: Eötvös Loránd University, Budapest
Position: PhD Candidate
Research interest: The fertility transition in southern Transdanubia.

Italy

Jessica Busco
Contact: jessica.busco@gmail.com
Department: Statistical Sciences
University: Bologna
Position: Assistant researcher
Research interest: Determinants of mortality and health in Italy, smooth- ing of life tables (Lee-Carter), Share data.

Paolo Cardone
Contact: paoloemilio.cardone@gmail.com
Department: Statistical Sciences
University: Sapienza University of Rome
Position: PhD Candidate
Research interest: Historical demography and event history and survival analysis.

Chiara Sanna
Contact: chiara.sanna1986@tiscali.it
Department: Science Economics
University: Cagliari and Sassari
Position: PhD Candidate
Research interest: Woman's role, family formation and reproductive behaviors in Sardinia in the last decades.

Nicoletta Signoretti
Contact: nicoletta.signoretti@uniroma1.it (nicoletta.signoretti@libero.it)
Department: Demography
University: La Sapienza, Rome (PhD 2012, Demography)
Position: Statistical researcher, Provincia di Roma
Research interest: New forms of family in Italy (and Europe) and the correlation with demographic aspects.

Japan

Ryohei Mogi
Contact: rmomniv@gmail.com

Department: Political Science and Economics
University: Graduate School of Meiji University, Tokyo
Position: PhD Candidate
Research interest: Do woman's work places and styles change in Spain from 20 century?, if so how?

Poland

Marília Nepomuceno
Contact: mariliare@gmail.com
Department: European Doctoral School of Demography
University: Warsaw School of Economics
Position: PhD Candidate
Research interest: Adult and elderly mortality.

Spain

Francisco Marco
Contact: fran.marco.gracia@gmail.com
Department: Economics
University: University of Zaragoza
Position: PhD Candidate
Research interest: Family reconstitution.

Sweden

Evans Korang Adjei
Contact: evans.korang.adjei@umu.se
Department: Geography
University: Umeå University
Position: PhD candidate
Research interest: The role of family networks in the Swedish labor market.

Helena Haage
Contact: helena.haage@umu.se
Department: CEDAR
University: Umeå University
Position: PhD Candidate
Research interest: Life-courses for disabled people in the nineteenth century.

Finn Hedefalk
Contact: Finn.Hedefalk@nateko.lu.se
Department: Physical Geography and Ecosystem Science
University: Lund University
Position: PhD Candidate
Research interest: Integration of and analysis of longitudinal demographic and geographic data on the micro-level.

Emilie Hane-Weijman Jansson
Contact: emilie.haneweijman@umu.se
Department: Geography
University: Umeå University
Position: PhD Candidate
Research interest: Labor market dynamics and the effects on people and regions.

Johan Junkka
Contact: johan.junkka@umu.se
Department: History
University: Umeå University

Position: PhD Candidate
Research interest: Demographic history, religious history, gender studies.

The Netherlands

Ingrid van Dijk
Contact: mail@ingridvandijk.com
Department: History
University: Radboud University Nijmegen
Position: PhD Candidate
Research interest: Clustering in infant and child mortality in The Nether- lands 1812–1912.

Tim Riswick
Contact: t.riswick@let.ru.nl
Department: History
University: Radboud University Nijmegen
Position: PhD Candidate
Research interest: Differences in the Mortality Changes of Brothers and Sisters in Taiwan and the Netherlands, 1860–1945.

Ukraine

Yuliia Bezsmertna
Contact: jbezsmertna@gmail.com
Department: History
University: Poltava National Pedagogical University named after V.G. Korolenko
Position: PhD Candidate
Research interest: Servants and domestic servants in regimental towns of Hetmanshchyna in the 60s–70s of the 18th century.

Course evalution

We asked the students the first day to start thinking of evaluation, and discuss issues at any time during the course. Nothing much of that happened. The second week we distributed an evaluation form (it is also on the course home page), which the students filled in anonymously and delivered the last day. Thirteen students have so far given their verdicts.
The questions, with a summary of the answers, were

1. *Gender*
   Nine women and nine men participated in the course.

2. *Did you get enough information at the start of the course about the content and readings?*
   "Yes" was the dominant answer, but a few wanted more reading advise.

3. *Do you think the course developed as could be expected from the information given?*
   Again, "Yes" was the dominant answer, but one wrote "Expected exercise on R . . . and not only in the project." Another comment: "Yes. The program was very good and we followed that."

4. *How do you rate the course in terms of level of difficulty?*

   - *Theoretical sessions*
     Dominant answer "About right", two thought too easy, and three too difficult.

   - *Applied R/RStudio sessions*
     Same.

- *Exercises/Project work*
  Even between 'too difficult' and 'about right'. The main criticism was that it was too little of Exercises and too much Project work.

5. *How do you think the exercises and projects reflected the course content?*
   Answers varied between "Good" and "Very good." Again complaints about to few exercises and too much project work. (I agree.)

6. *How well did the theoretical part (Event History/Survival analysis) cover what you expected from the course?*
   All but two "Well" or "Very well". Two comments: "I was expecting deeper discussions" and "I'd preferred more focused concepts".

7. *How well did the RStudio/R part (R markdown, programming, knitr, etc.) cover what you expected from the course?*
   A majority was very satisfied, but a minority complained about too few exercises (rightly so).

8. *What is your overall assessment of the course structure?*
   "Good" or "Very good". One comment: "Too much information and not enough time for reflection."

9. *In summary, how do you rate this course?*
   Eight "Very good" and five "Good". One comment: "Very difficult, but very good."

10. *How do you assess the service provided by the* Centre for Demography and Ageing Research *concerning practical matters?*
    "Perfect" and "Very good" were the dominant answers. No complaints.

11. *What did you like most about the course?*
    The answers were evenly divided between statistics, R/RStudio, and Project work. Some also mentioned the good infrastructure and the social events.

12. *Suggestions for further improvements?*
    There should have been more focus on basic R/RStudio skills from the beginning. *dplyr* lecture much earlier. Some comments: "I think that the course was especially designed for people who has a historical background", "Less seminar and conference, *more applied* work !", "It should have been better if we worked in a lab with computers of the University", "Stick people together with the same knowledge of R or slightly different and adjust the group work to their skills."

13. *Other comments?*
    "Please let people choose their own group", "No laptop", "No, it's OK", "Most people did the project work in an automatic way, without understand well what they did and why."

What do we learn for the future?

It is obvious that we need to have more focus on the basic skills of data handling in R/RStudio from the beginning. The projects should have been distributed later, on the last day of the first week. The *reproducible research* part was hidden too far away.

In all, better and tighter structure is needed. Exactly where this will lead us in the future will depend on future directions for the basic courses in the network, primarily the introduction of R/RStudio.