SPE-2019: postmortem

Next year's SPE: in Tartu, from Friday 14 to Wednesday 19 August, 2020.

General experiences from the course

Great heterogeneity was again observed in the participants' preknowledge and skills in R. What to do in order not to bore those who already are familiar with basics in R?

Exercises of the first two days in particular need rethinking from that viewpoint.

General points for improvement

Recruit external reviewers (Merli and Nele?) for the exercises, especially those of the first two days. Other exercises good to be reviewed among us.

Simplify the code as much as possible. Don't use negative indices. Develop a wrapper for `cut()` having defaults that are conventional in epidemiology.

Develop a robust method for delivering to the students the data that is not in R packages. Problems with Mac users related to the somewhat less transparent directory structure of the operating system needs some thought.

Continue encouraging the use of explicit naming of arguments. Emphasize this also in the exercises – modify them as necessary. Avoid using excessively dense code which would be readable only to seasoned R user.

Change `T` → `TRUE` and `F` → `FALSE` when these logical constants are needed.

Remember to have some motivational lines in the beginning of lectures and practicals; why we think these things are important.

Put all `library()` statements needed in a practical to the beginning.

Use `round()` to cut unnecessary decimal points – or put `options(digits=3)`, or even `digits=2`. Encourage students to do this as well.

Continue to instruct, how to efficiently write, check and run the R code.

Continue to develop the idea of having enough explanation about what the next lines of cryptic looking script are supposed to do for you.

Choice of colors: Try to develop the exercises on graphics to comply with recent recommendations on colours. Avoid at least red and green in the same plot. Consult package `colorspace` and the paper by Zeileis et al. in *Computational Statistics & Data Analysis*, 53, 3259-3270.

Day 1

Lecture on history and ecology of R (Martyn) was OK.

Exercises 1.1: Emphasize at the very beginning: Always start with a clean workspace and never save a workspace! Use R interactively but turn to use a script in subsequent exercises. In section 1.1.10 reduce the number of variants for selecting variables and observations from a data frame. For the latter, recommend just `subset()`. At least sections 1.1.8 on creating data frames and 1.1.16 on saving and loading data could be moved towards the end. Sections 1.1.17 on search path and 1.1.18 on attaching deserved to be removed completely. – Martyn develops this exercise.

Exercise 1.2: It continues to be important. We should make sure that students have time to do this, which means that Ex 1.1 must be shorter. – Martyn takes care of this.

Lecture on language and basic data (Krista): Abandon the lecture, review the corresponding exercises so that the lecture material would be covered by the exercises. Optional exercises for those who are already familiar with the basics. Recap of those exercises may be a bit longer.

Lecture on `dplyr/tidyverse` (Damien) – extend to at least 30 minutes.

Exercises 1.3 on tabulation generally OK. – Krista continues be in charge.

Exercise 1.5 on graphics quite good, too. Introduce `segments()` explicitly. The program must be planned such that people have time to do 1.5. – Krista works with this.

Exercise 1.6 on simple simulation not needed: remove

Advanced participants: Exercise 1.4 on `dplyr` and `tidyverse` tools for those already familiar with basic R: Expand substantially – Damien prepares.

Day 2

Lecture on Poisson & logistic regression (Janne): OK but shift emphasis on using `family=poisreg`. Even then it is important to start from the classical approach of using `family=poisson` with the number of cases as the response and log-person-times as offset. Drop `D/Y` with `weights`. Change "logistic" to "binary" as appropriate. Drop the `D/N` with `weights`. Motivation for the different confidence intervals and references to relevant literature (e.g. Clayton & Hills).

Exercise 1.7 on Poisson rates: Drop the optional part on `weights`. Motivation for the variants of confidence intervals?

Exercise 1.8 on logistic regression: Better be in the afternoon after the lecture on LM and GLM. Must be shortened and simplified so that basic things come early on and that everybody would have time to do them. Keep `effx()` to some extent. Remove housekeeping; provide the data set already with pertinent categorical variables. Some rearrangements of the different items required. Martyn In logical expressions change `T` and `F` to be `TRUE` and `FALSE`, respectively. – Janne works on this.

Exercise 1.9 on estimation of effects: Better to have this already in the morning, also such that everybody would have time to do the essential parts. – Esa improves.

Lecture on linear and generalized linear models (Esa): from 30 to 45 min. Add things from GLMs, interactions, and review items from the morning practical

Lecture on splines (Martyn): shorten from 45 to 30 min. Remove slides on Brownian motion. Bring closer to the needs of the course: use of package `mgcv` and function `gam()`.

Exercise 1.10 on estimation and reporting curved effects: Drop the section on natural splines. Simplify the code; at least in section 6 on categorical age and period. – Esa develops further.

Day 3

Lecture on graphics (Martyn): Nearly perfect as it is now.

Exercise 1.11: Encourage people to have a look at the contents of the vectors and matrices resulting from the housekeeping script, but not the script itself. – Martyn will do.

Day 4

Lecture on survival (Janne): Somewhat dense because competing risks are introduced quite soon after a very brief introduction to basic concepts for censored data.

Exercise 1.12 on survival methods: Add legends to some plots. Change the color palette to be more color-blind neutral (see above). – Janne.

Lecture on representation of follow-up, Lexis & SMR (Bendix): OK.

Exercise 1.13: Refer to the specific slides in the lecture handouts when asking to compute and tabulate summary measures like `D, Y, rate, E, SMR`. Pay attention to the execssive numerical precision when printing results. Mark SMR modeling as optional. – Bendix

Day 5

Lecture on NCC & CC (Esa): perhaps add a comment that with heavy censoring case-cohort design may become less efficient than NCC.

Exercise 1.14 on NCC & CC: `cal.yr()` works for data frames, so simplify the code accordingly.

Lecture on causal inference (Krista): Add the example on coffee drinking to the slides. Introduce the term "collider".

Exercise 1.15 on causal inference: Pay attention to how the results are printed; more concise and less decimal points. Use `ci.lin()`; focus on confidence interval.

Day 6

Lecture on multistate models (Bendix): Is developed.

Exercise 1.16 on multistate model: Drop Cox models. Drop decimal points. Consider moving from natural splines to penalized splines here, too.