ST: Advanced R (2 Credits) —

Stat 6910, Section 004, Fall 2013

Instructor: Dr. Jürgen Symanzik
Office: AnSc 313
Phone: 797–0696
FAX: 797–1822
e–mail: symanzik@math.usu.edu
WWW: `http://www.math.usu.edu/~symanzik/`
`http://www.math.usu.edu/~symanzik/teaching/2013_stat6910_004_fall/stat6910.html`

Office Hours: MWF 4:00pm – 4:55pm and by appointment.

Classes & Rooms:
MWF 6:00pm – 6:50pm, Mo 9/30 – Fr 12/6, 2013: AnSc 320.

Please note that there are no classes on Fr 10/18 (Fall Break). However, Fr classes will be held on Th 10/17 (i.e., no regular Th classes will be held during that week). There are also no classes during the entire Thanksgiving week (Mo 11/25 – Fr 11/29).

Please visit the course Web page listed above and/or the Canvas page for this course frequently for lecture notes, data sets, R code, etc. — in particular if you miss class for any reason.

Course Objectives:

Recall what Andreas Buja, the Liem Sioe Liong/First Pacific Company Professor in the Statistics Department, The Wharton School, at the University of Pennsylvania in Philadelphia, USA, stated in a 2006 interview: *"I think in education we still have some ways to go to find a balance of things that we want to teach students. There is the traditional curriculum that gives Ph.D. students a solid foundation in theory, but then they also should acquire computational skills, they should become good applied statisticians who have good sense, good data sense. Many of these things are really hard to teach. You can teach them details, but ultimately they have to pick up the high level of thinking, the creative way of thinking, on their own or by being thrown like fish into the water, either they swim or they don't. That is hard to teach. Something that I see lacking right now is, especially if you are interested in education with industry in mind, what you need out there in industry I think is not specifics of modeling, it is good data sense, it is data skills, data literacy. I think that was a term used by one of the earlier interviewees. Data literacy, in general, is the ability to get data and start doing something sensible, and that is of utmost importance. Part of that is that we cannot assume that other people are doing data cleaning for us. We have to do that ourselves. So here I see a gap actually in our education. I don't think most statistics programs teach something like a scripting language and practice data cleaning, reshaping of data, basic tabulations, mild aggregations, getting subsamples, systematic ones and random ones, and so on. These are very important activities, and we still need to get better at teaching them."* (see Computational Statistics (2008) 23:177–184 for the full interview).

In Stat 6910, Section 003 ("Introduction to R"), you have learned basic data skills and data literacy via R to achieve this goal. You were thrown like fish into the water (without

any prior swimming course) — and you learned to swim! Now, we will extend these basic skills to do something really meaningful with a variety of data sets. Eventually, this two–course sequence may be the most valuable courses for a future career in research (assuming you are working with any kind of data) or in industry.

Prerequisites:

The most important prerequisite is Stat 6910, Section 003 ("Introduction to R"), or knowledge from a similar in–depth introductory R course. Just having used R in previous classes likely will not be sufficient. If you are in doubt, talk to me first.

Moreover, I expect basic "operational" knowledge from an introductory stats course such as Stat 2000, Stat 3000, or higher. "Operational" means that you still recall sufficient details from regression, ANOVA, hypothesis tests, etc. (it is not sufficient that you have taken such a course several years ago and have forgotten almost all details).

IDEA Center Learning Objectives:

Objective 1) Gaining factual knowledge (terminology, classifications, methods, trends).

Objective 2) Learning fundamental principles, generalizations, or theories.

Objective 3) Learning to apply course material (to improve thinking, problem solving, and decisions).

Objective 5) Acquiring skills in working with others as a member of a team.

Objective 11) Learning to analyze and critically evaluate ideas, arguments, and points of view.

Topics: (subject to change)

This course will continue where Stat 6910, Section 003 ("Introduction to R") ended:

1. Writing functions, loops, if/else–statements, and debugging.
2. Reproducible research.
3. LaTeX and Sweave via RStudio.
4. More about graphics.
5. Basics of simulation.
6. Representation of information.
7. Regular expressions.
8. XML.
9. Data bases and SQL.
10. Resampling/bootstrap.
11. Others (as time permits).

We will work with real "messy" data that have not been preprocessed nor analyzed so far. These data will contain surprises — for you and for me. Do not expect that someone is going to give you the final answer or model. We jointly will have to work towards such an answer or model.

For MS and PhD students majoring in Statistics, it is important to learn LaTeX — from basic document preparation, over the inclusion of R graphics into your LaTeX documents to advanced topics such as Sweave (`http://www.statistik.lmu.de/~leisch/Sweave/`)

and the LaTeX bibliography BibTeX (`http://www.bibtex.org/`). LaTeX is essential for graduate work (at the MS and PhD level) and will be used for many theses, dissertations, and scientific publications.

Quizzes:

There will be a few (perhaps 3 or 4) quizzes throughout the semester. The length of each quiz will be about 15 to 20min. These quizzes will determine whether you have understood the basic ideas discussed in the previous lectures. You have to interpret the output of some R code, spot the errors in the R code provided to you, or write short segments of R code (often just 1 or 2 lines) that accomplish a particular task. Quizzes will be closed book/closed computer. In particular, you won't have access to R during the quizzes. Quizzes will account for about 15% of your final grade.

In case you miss a quiz, there will be no makeup quiz offered! However, a score of 0 points will be replaced with your **lowest** score from the other quizzes. If you are doing well on the earlier quizzes, you may want to consider to miss one of the final quizzes. ☺ If you miss 2 quizzes (or more), your score will be 0 for all of the quizzes you have missed!

Quizzes will be announced in class and on Canvas the lecture before a quiz is going to take place.

Homework Assignments:

There will be about 3 to 4 homework assignments throughout the semester. Each assignment will include a value that it will be scored out of. The value of each assignment will be roughly proportional to its importance and the amount of work involved. Regular homework assignments will be done individually or in groups of 3 to 5 students. Homework assignments will account for about 35% of your final grade.

Projects:

There will be (very likely) 2 major projects throughout the semester. This will include the preparation of a final project report and possibly a short presentation of your work for the other students in this course. Projects most likely will be done in groups of 3 to 5 students. These projects will account for about 50% of your final grade.

There will be no regular (in–class or take–home) midterm or final exams. Nevertheless, this will be a very challenging course that requires a lot of individual time to work on the assignments and projects. Just attending classes will not be enough to pass this course! In addition, you will have to do a lot of individual reading of textbooks, online documentation, and help pages, and search for available information on the web.

Textbooks:

J. Adler (2010), R in a Nutshell, O'Reilly.

P. Murrell (2009), Introduction to Data Technologies, Chapman and Hall/CRC. [Note that the entire book is available online from `http://www.stat.auckland.ac.nz/~paul/ItDT/` under a Creative Commons licence.]

If you want to use a printed reference book for R, I would suggest that you obtain a copy of the book by Adler (2010). Murrell (2009) provides a very good insight on many of the advanced topics discussed in this 2–credit course. There exist numerous other books and tutorials (in print and on the web) that can be used as well.

Software:

We will primarily be using R (`http://cran.r-project.org/`), a GNU–license statistical

package and clone of S–Plus. Please install the most recent version of R, i.e., 3.0.1, on your own computer so we can exchange R code. R and all R packages we are going to use are available on the web for free download from the URL listed above. To create R code, debug your code, or look at the results of your computations in a convenient way, you should also install and use RStudio from `http://www.rstudio.com/`. This can also be downloaded for free.

Credits:

This course uses some of the course materials provided by Dr. Duncan Temple Lang (UC Davis: `http://www.stat.ucdavis.edu/~duncan/`) and Dr. Deborah Nolan (UC Berkeley: `http://www.stat.berkeley.edu/~nolan/`). We are likely to include parts from additional web sources that will be specified later on.

Courtesy:

Please turn off pagers and cell phones before class, and please keep conversations to a minimum during lectures. Please do not read/reply to your e–mails or browse other Web pages than the ones discussed during class.

I will not keep track if you come to class or not. However, I would highly recommend to attend all lectures.

Americans with Disabilities Act:

If a student has a disability that will likely require some accommodation by the instructor, the student must contact the instructor and document the disability through the Disability Resource Center, during the first week of the course. Any requests for special considerations relating to attendance, pedagogy, taking of examination, etc. must be discussed with and approved by the instructor. In cooperation with the Disability Resource Center, course materials can be provided in alternative formats — large print, audio, diskette or Braille.

Grading System:

Quizzes	about 15%
Homework	about 35%
Projects	about 50%
Total	100%

Due to the considerable weight of the projects, I expect that everyone who puts enough efforts into this course will do well. **There may be a few opportunities for individual extra credit assignments, in particular for students who did not do so well on the quizzes.**

This course is intended to make you even more proficient in R. Thus, even if your R code is correct, but not very efficient, you are going to lose points. You are also going to lose points if your R code is not properly documented or does not follow formatting recommendations from Google's R Style Guide (see `http://google-styleguide.googlecode.com/svn/trunk/Rguide.xml`). Carefully check your graded homework assignments from Stat 6910, Section 003 ("Introduction to R"), and note where you lost points — and avoid losing "easy" points in this Advanced R course.

Note:

The above schedule and procedures in this course are subject to change in the event of extenuating circumstances.