Introduction to Statistical Computing —

Stat 5810, Section 004 / Stat 6810, Section 003, Fall 2012

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/2012_stat5810/stat5810.html

Office Hours: MWF 11:00am – 12:00noon and by appointment.

Classes & Rooms:
MWF 9:30am – 10:20am, Mo 8/27 – Fr 12/7, 2012: AnSc 320.

Please note that there are no classes on Mo 9/3 (Labor Day) and on Fr 10/19 (Fall Break). However, Fr classes will be held on Th 10/18 (i.e., no regular Th classes will be held during that week). There are also no classes during Thanksgiving Break (We 11/21 – Fr 11/23).

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

Course Objectives:

In a 2006 interview, 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: *"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 this course, we will learn basic data skills and data literacy. Be prepared to be thrown like fish into the water (without any prior swimming course)! This will be by far the most unusual course you have taken so far — and it may remain the most unusual course for

many semesters to follow. But, this course will also be a required prerequisite for most (if not all) applied statistics courses at the 6000 and 7000 level you may take in the future. Finally, this course may be the most valuable course for a future career in industry.

Prerequisites:

I expect basic "operational" knowledge from an introductory stats course such as Stat 2000 or Stat 3000. "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).

Prior R and LaTeX knowledge is not required, but you will rather learn these in this course.

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)

1. Introduction of R (basic R usage, loops, graphical concepts, etc.).
2. Introduction to Data Technologies (including data input, data cleaning).
3. Data Bases (general and in R).
4. Others (as time permits)

We will work with real "messy" data that have not been preprocessed nor analyzed so far. Examples likely will be taken from iPod log files, web logs, GPS tracks, and data from data mining and visualization competitions. 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.

If you take this course at the 6000 level, you will also 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, but it is not that essential for a job in industry.

Assignments:

There will be a variety of assignments throughout the semester. Each assignment will include a value (typically 20–100 points) that it will be scored out of. Your final grade will be determined by the sum of your points in all assignments. 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 2 or 3 students. Major assignments such as seminar–like presentations or a major project will be done individually or also in groups of 2 or 3 students.

There will be no regular (in–class or take–home) midterm or final exams. Nevertheless, this will be a very challenging class 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.

If you take this course at the 6000 level, you will also obtain points for your LaTeX performance in the assignments, projects, and presentations.

Textbooks:

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

W. J. Braun and D. J. Murdoch (2007), A First Course in Statistical Programming with R, Cambridge University Press.

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.]

Every student should have access to each of these three books, but it is not necessary that every student buys all of these books. Perhaps you can make arrangements with some of the other students in class (or your group members) who purchases which book(s).

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 current version of R, i.e., 2.15.1, on your own computer so we can exchange code. R and all R packages we are going to use are available on the web for free download from the URL listed above.

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.

Note:

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