ST: Introduction to R (1 Credit) —

Stat 6910, Section 003, Spring 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/stat6910.html`

Office Hours: TBA and by appointment.

Classes & Rooms:
TBA — In a regular semester, there are 15 weeks with about 3 weekly 50min–lectures or 2 weekly 75min–lectures, i.e., about 2250 lecture minutes in total for a 3–credit course. Thus, we have to meet for about 750 lecture minutes for our 1–credit course. After our organizational meeting on Mo 1/14 at 5pm in AnSc 314, there will be 3 weeks with 3 75min–lectures each week, thus a total of 9 lectures (= about 675 lecture minutes, plus about 60min for the organizational meeting, i.e., about 735min in total).
Meeting dates, times, and locations will be determined during or after the organizational meeting. I expect that our last lecture will be held on (or before) Fr 2/1.
Please visit the course Web page listed above frequently for updates on meeting dates, times & locations, 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 via R to achieve this goal (a 2–credit follow–up course likely will be offered in the next semester). Be prepared to be thrown like fish into the water (without any prior swimming course)! Note that this 1–credit course will be a required prerequisite for most (if not all) applied statistics courses at the 6000 and 7000 level you may take in the future that work with R. Finally, this course may be the most valuable course for a future career in research (assuming you are working with any kind of data) or 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 knowledge is not required, but you will rather learn the basics of R 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 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, writing functions, etc.).
2. Introduction to Data Technologies (including data input, data cleaning).
3. Others (as time permits)

We will work with real "messy" data that have not been preprocessed nor analyzed so far. Examples include temperature data, family data, housing data, and data from the 2012 Olympic Games. These data sets 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.

Assignments:

There will be 3 homework assignments in our compact course. Each assignment will include a value (around 40–60 points) 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. Answers to homework assignments have to be submitted about 7 to 10 days after an assignment has been handed out. Homework assignments will account for about 66% of your final grade.

You are allowed to discuss questions on the assignments with other students, but each student has to submit his/her individual set of answers.

Weekly Quizzes:

There will be 3 weekly quizzes at the end of the 3rd (final) lecture in each week. The length of each quiz will be about 15 to 20min. These quizzes will determine whether you have understood the basic ideas discussed earlier in the week. 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 33% of your final grade.

Be aware that this will be a very challenging class that requires a lot of individual time to work on the assignments and prepare for the quizzes. Just attending classes will not be enough to pass this course! In addition, you will have to do a lot of individual reading in textbooks, online documentation, and help pages, and search for available information on the web. Most importantly, you should spend as much time as possible working with R.

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) goes far beyond our course and provides a very good insight what to expect in the 2–credit follow–up 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 a recent version of R, i.e., 2.15.1 or 2.15.2, 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 33%	about 75 pts
Homework	about 66%	about 150 pts
Total	100%	about 225 pts

Due to the short length of this course, course grades will very likely be much lower than grades assigned in other graduate statistics courses. When you lose points in one of the 3 homework assignments or the 3 quizzes, there will be hardly any opportunity to compensate for this point–loss in the remaining HWs and quizzes. **There will also be no opportunity for individual extra credit assignments.**

This course is intended to introduce you to R and make you proficient in its use. 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/google-r-style.html`). Overall, you have to learn quickly how to write good, correct, efficient, and well–formatted R code. This knowledge will be useful for your future classes, for your research at the university (so that your correct R code can be shared and understood by your friends and advisors), and possibly outside the university when you plan to publish your R code that was developed as part of your MS or PhD research as an R package later on. Obtaining this knowledge quickly won't be easy!

Therefore, I would strongly suggest to take this course on a pass/fail (P/F) basis, rather than on the usual letter grade basis, given your departmental requirements allow you to do so!!

Note:

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