Data Technologies —

Stat 5080, Section 001 & Stat 6080, Section 001

Fall 2024 (2 Credits)

Instructor: Dr. Jürgen Symanzik
Office: AnSc 313
Phone: 435–797–0696
e–mail: juergen.symanzik@usu.edu
Web: `https://www.usu.edu/math/symanzik/`
`https://www.usu.edu/math/symanzik/teaching/2024_stat5080/stat5080.html`

Office Hours: Tuesday (T) & Thursday (H) 1:30pm – 2:30pm in AnSc 313,
Friday (F) 9:00am – 10:00am (via Zoom), and by appointment.

Classes & Rooms:
TH 12:00noon – 1:15pm, T 9/10 – H 11/14, 2024 (tentatively): ENGR 238 (face–to–face).

Please visit the course Web page listed above for emergency announcements, e.g., when Canvas is unavailable. Otherwise, visit Canvas frequently for lecture notes, data sets, R code, etc. — in particular if you miss our face–to–face lectures for any reason. All (additional and updated) materials, announcements, discussions, recordings, etc. from Canvas are part of the course materials. Not seeing one of these in time does not serve as an excuse for not getting point deductions for the course. Deadlines may change or unexpected new *Coronavirus/Covid–19* regulations and requirements may occur. It is your responsibility to make sure to receive all announcements in time.

Detailed Class Schedule:
For a 2–credit course, we need 20 lectures/lecture days (in contrast to 29 or 30 lectures/lecture days for a 3–credit course). Those days are marked as "Lecture 01" to "Lecture 20" in the overview below:

Week	Tuesday	Thursday
1	8/27 No class	8/29: No class
2	9/3: No class	9/5: No class
3	9/10: Lecture 01	9/12: Lecture 02
4	9/17: Lecture 03	9/19: Lecture 04
5	9/24: Lecture 05	9/26: Lecture 06
6	10/1: Lecture 07	10/3: Lecture 08
7	10/8: Lecture 09	10/10: Lecture 10
8	10/15: Lecture 11	10/17: Lecture 12
9	10/22: Lecture 13	10/24: Lecture 14
10	10/29: Lecture 15	10/31: Lecture 16
11	11/5: Lecture 17	11/7: Lecture 18
12	11/12: Lecture 19	11/14: Lecture 20
13	11/19: Backup	11/21: Backup
14	11/26: Backup	11/28: No class
15	12/3: Backup	12/5: Backup

Note: "No class" means guaranteed no class that day. I have marked the dates in the three last weeks of the semester as "Backup", e.g., in case we miss lectures because of weather conditions, sickness, or any other reason. If nothing goes wrong, our tentative last lecture date will be on H 11/14/2024. Backup lectures (if any) will be held via Zoom (and not face–to–face). There is no need to stay in Logan after H 11/14/2024 for a backup lecture in this course.

Course Objectives:

Note that 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, already 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, `https://doi.org/10.1007/s00180-008-0113-0` for the full interview).

In the "Introduction to R" course, you have learned basic data skills and data literacy via R to achieve this goal. You were (likely) 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 course will

be one of your most valuable courses for a future career in research (assuming you are working with any kind of data) or in industry.

Prerequisites:

STAT 5050 ("Introduction to R") with a C− or better. STAT 3000 ("Statistics for Scientists") or STAT 5100 ("Modern Regression Methods") with a C− or better. STAT 5550 ("Statistical Visualization I") is recommended. Moreover, you should be familiar with a tool such as R Markdown, knitr, or Sweave that allows you to combine text, R code, graphics, and numerical results in high–quality documents. LaTeX is a plus, but is not formally required at the 5000 level, but it will be required at the 6000 level of this course.

Moreover, I expect that you still have "operational" knowledge from your STAT 3000 or STAT 5100 (or higher) course. "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 a basic understanding of the subject (e.g., factual knowledge, methods, principles, generalizations, theories).

Objective 4) Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course.

Objective 13) Learning appropriate methods for collecting, analyzing, and interpreting numerical information.

Topics: (subject to change)

1. Data.
2. Basics of simulation.
3. Representation of information.
4. Regular expressions.
5. HyperText Markup Language (HTML) and Web scraping.
6. Text data from Portable Document Format (pdf) files and via Optical Character Recognition (OCR).
7. The *Tidyverse* (R packages for Data Science).
8. Data bases and Structured Query Language (SQL).
9. Extensible Markup Language (XML).
10. Resampling/bootstrap/others (Outlook only).

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 (`https://stat.ethz.ch/R-manual/R-devel/library/utils/doc/Sweave.pdf`), knitr (`https://yihui.org/knitr/`), 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. **Therefore, LaTeX will have to be used for all homeworks, projects, presentations, etc. at the 6000 level of this course.**

Homework Assignments:

There will be 3 HW assignments for this course, roughly one every three weeks. Each HW assignment will include a value (typically 20–100 points) that it will be scored out of. HW assignments will contribute to 100% (for Stat 5080), repectivley 70% (for Stat 6080), of your course grade. The value of each HW assignment will be roughly proportional to its importance and the amount of work involved.

You will be allowed to discuss general approaches to questions on the HW assignments with other students, but each student must write and submit their own R code and comments. Any students caught sharing R code or other parts of their homework submissions will fail the class.

Unless otherwise stated on the HW assignment sheet, all homework assignments have to be submitted electronically via Canvas. **You will have about 2 or 3 weeks after the last lecture to finalize and submit the last HW assignment.**

There will be no (in–class or take–home) quizzes, midterm exams, or final exams. We will have a few worksheets for training purposes only. Nevertheless, this will be a very challenging course that requires a lot of individual time to work on the assignments (and project for Stat 6080). 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.**

No Excuse Needed Late Homework Submission Policy:

Each student has **3 tokens** for late homework submissions. These tokens can be used in multiple ways, e.g., one token at a time for each of the 3 HWs or all 3 tokens can be used for a single late HW submission. You will need 1 token if your HW is 1 min – 24 hours late; you will need 2 tokens if your HW is $>$ 24 hours – 48 hours late; and you will need all 3 tokens if your HW is $>$ 48 hours – 72 hours late. Once you have used up all 3 tokens, late HWs will count as 0 points, even if such a HW is only one additional minute late. As an example, if you used 2 tokens early on, you will have only 1 token left and can submit your next HW only up to 24 hours late. If you submit that HW more then 24 hours late, it will count as 0 points. All times are based on the Canvas time stamp that is created when a HW gets submitted — and not on your local computer clock.

It is your responsibility to keep track of the number of tokens you have used. Record that information on your phone or check how you can extract the submission dates and times in Canvas on your side. I will not be able to tell you shortly before the next submission deadline how many tokens you have left. In fact, I only plan to use that information when I determine final course grades. You may therefore get assigned a score $>$ 0 points for a certain HW submission initially, but that score may get adjusted to 0 points if it turns out that this was for a late submission and you had already used up all your tokens.

You will not get any credit for any unused tokens, so you could use the remaining ones for proofreading, checking that all figures and R code are included, etc., in particular for the later HWs — and if you still have tokens left at that time!

As the name indicates, there is no excuse needed when you use a token, be it for a minor sickness, personal travel due to family events, or just needing more time as you couldn't finish that HW by the deadline. If you are officially absent from USU when a HW is due, e.g., due to travel with a USU sports team or band, attending a conference, doing fieldwork, etc., I need an official statement from a coach or supervisor — and no token will be used. However, we then need to determine on a case–by–case basis when that particular HW will be due for you — and when you would have to use additional token(s) to further extend your extended deadline.

You can resubmit your HW as often as you want in Canvas prior to the deadline. Apparently, if you resubmit it after the deadline, one (or more) tokens will be used. In that case, also let me know via e–mail at the time of the deadline that you are still working on your HW beyond the original deadline and plan to resubmit a new version later on. Once a HW has been graded on my side, you will no longer be allowed to resubmit a revision of that HW (or in case you did, you still would get the points based on the last submission prior to the original deadline).

Project (Stat 6080 only):

There will be one major project towards the end of the semester. This will include the preparation of a final project report and a short presentation of your work for the other students in this course. The project will be done in a small group of students. The project will account for 30% of your course grade.

Textbooks:

Murrell, Paul (2009) *Introduction to Data Technologies*, Boca Raton, FL: Chapman and Hall/CRC.
Note that the entire book is available online from `https://www.stat.auckland.ac.nz/~paul/ItDT/` under a Creative Commons licence.

Nolan, Deborah, and Temple Lang, Duncan (2015) *Data Science in R — A Case Studies Approach to Computational Reasoning and Problem Solving*, Boca Raton, FL: CRC Press/Taylor & Francis.

Every student should have access to each of these books, but it is not necessary that every student buys all of these books. The USU library holds several of these books or provides online access. If you plan to work in the area of Data Science for your MS or PhD degree, you should consider to purchase these books for an ongoing use beyond this course.

Software:

We will primarily be using R (`https://cran.r-project.org/`), a free software environment for statistical computing and graphics. Please install a recent version of R, i.e., 4.4.1 or later, on your own computer so we can exchange code. Working with version 4.4.0 should still be OK. I would discourage the use of version 4.3.3 or older. Also install the most recent version (2024.04.2 + 764, released on 2024-06-10) of RStudio (`https://posit.co/`) as a front end to R. **Note that some versions of RStudio released earlier in 2024 contained bugs that prevented them from opening certain Rnw files, similar to the ones used in class. If you get error messages when you try to open some of the files from class, check whether you have the most recent version of RStudio installed — and if not, upgrade to the most recent version.** For Windows, also install Rtools 4.4 (`https://cran.r-project.org/bin/windows/Rtools/`).

For most work in this course, using the *tinytex* R package is all that is needed, but if you plan to work on more complicated LaTeX documents later on (such as your MS thesis or your PhD dissertation), you should also install a full version of LaTeX that can be used in connection with RStudio. My personal recommendation is TeX Live (`https://www.tug.org/texlive/`).

Credits:

This course uses some of the course materials provided by Dr. Paul Murrell (University of Auckland: `https://www.stat.auckland.ac.nz/~paul/`), Dr. Duncan Temple Lang (UC Davis: `https://www.stat.ucdavis.edu/~duncan/`) and Dr. Deborah Nolan (UC Berkeley: `https://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 cell phones and similar devices 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. If you have to miss a lecture, there will be a Kaltura recording of the lecture available in Canvas later in the day (if the technology doesn't fail).

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 (DRC – `https://www.usu.edu/drc/`), preferably 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, or Braille.

Note:

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