Week 1: Introduction & Software

MATH-517 Statistical Computation and Visualization

Tomas Masak

September 23rd 2022

Classes

Lectures

- Teacher: Tomas Masak
- Time: Friday 10:15-12:00
- Place: GC D0 386

Exercises

- Teacher: Almond Stoecker
  - available only from Week 3 on
- Time: Friday 13:15-15:00
- Place: GC D0 386

Statistical **Computation** and Visualization

Offered the choice between mastery of a five-foot shelf of analytical statistics books and middling ability at performing statistical Monte Carlo simulations, we would surely choose to have the latter skill.

– Press et al., *Numerical Recipes*

Statistical **Computation** and Visualization

Offered the choice between mastery of a five-foot shelf of analytical statistics books and middling ability at performing statistical Monte Carlo simulations, we would surely choose to have the latter skill.

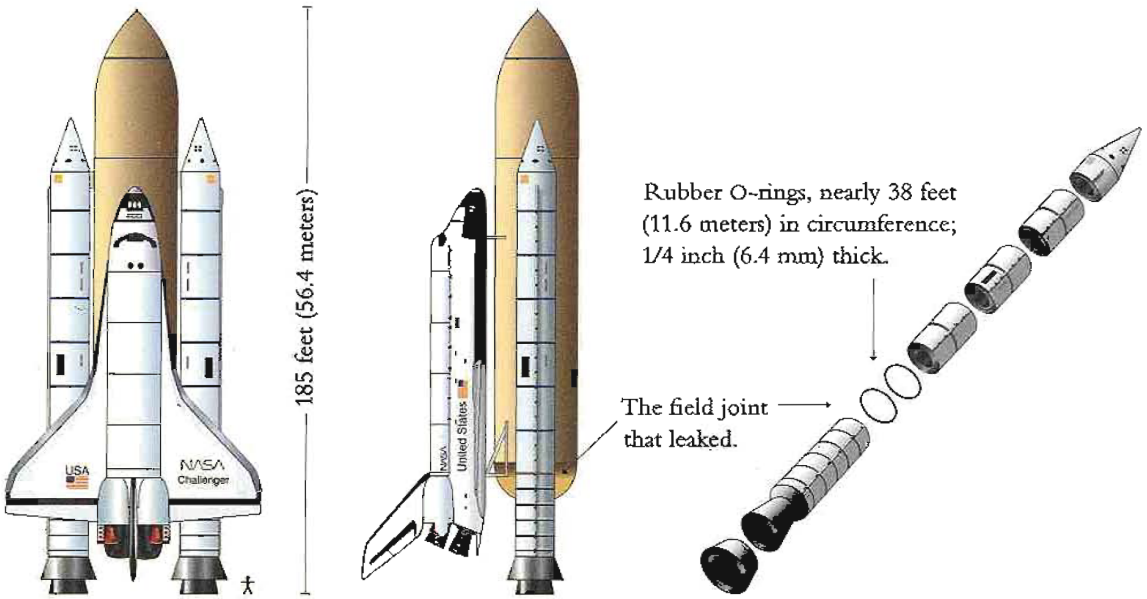
– Press et al., *Numerical Recipes*

Apart from Monte Carlo (MC), we will cover (re-)sampling methods such as

- cross-validation
- bootstrap
- jackknife
- Bayesian MC extensions

The Three Pillars of Science

Statistical Computation and **Visualization**



Bad Vizualization

BLOW BY HISTORY

SRM-15 WORST BLOW-BY

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY

- 2 CASE JOINTS (30-40°)

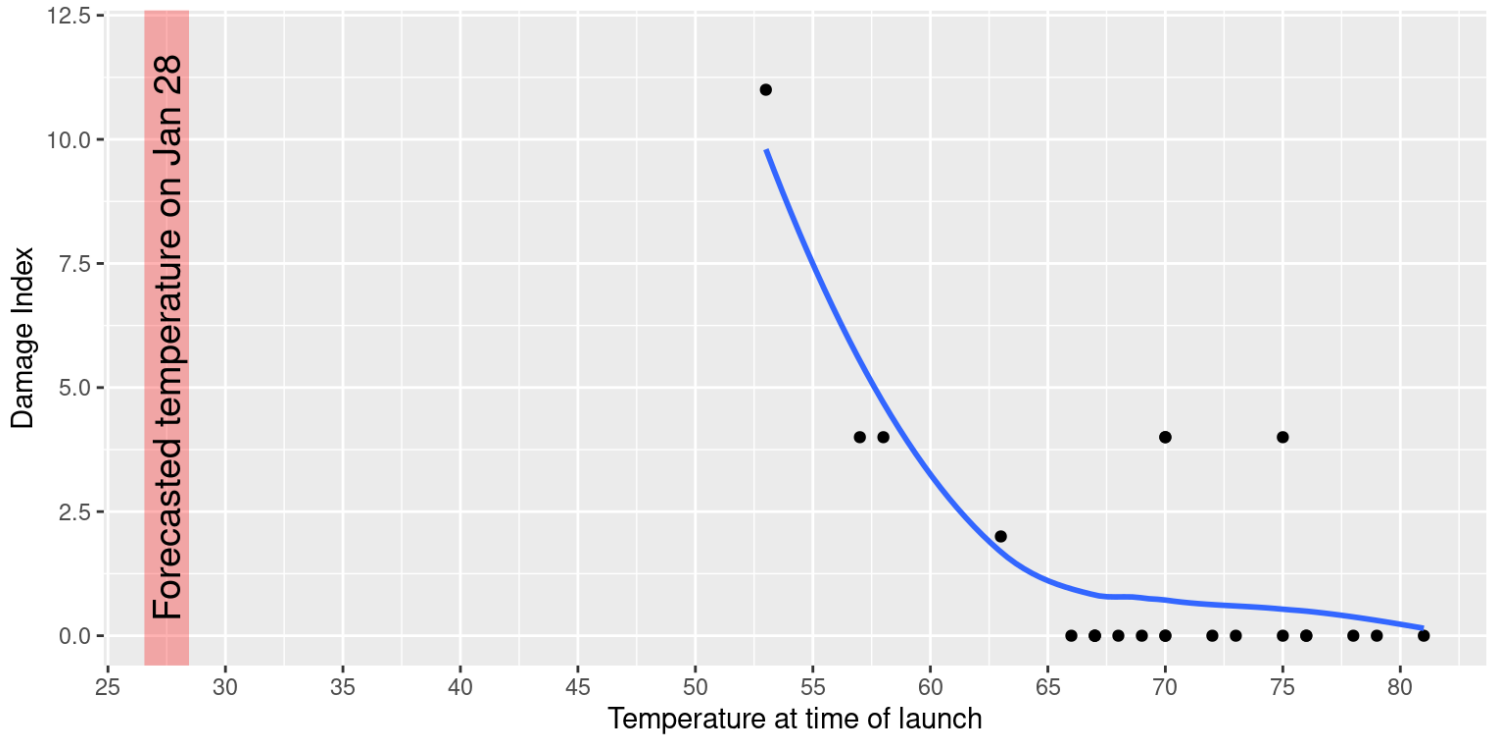
SRM-13A, 15, 16A, 18, 23A 24A

- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

Statistical Computation and **Visualization**



Course Content (Chronologically)

- Soft Start
  - R and other software
  - reproducibility and ethics
  - data wrangling and visualization
- Course Core
  - kernel density estimation
  - non-parametric regression
  - cross-validation
  - simulations
  - Monte Carlo (MC)
  - bootstrap
  - EM algorithm
- Bayesian Dessert
  - basic thinking
  - Markov Chain Monte Carlo (MCMC)

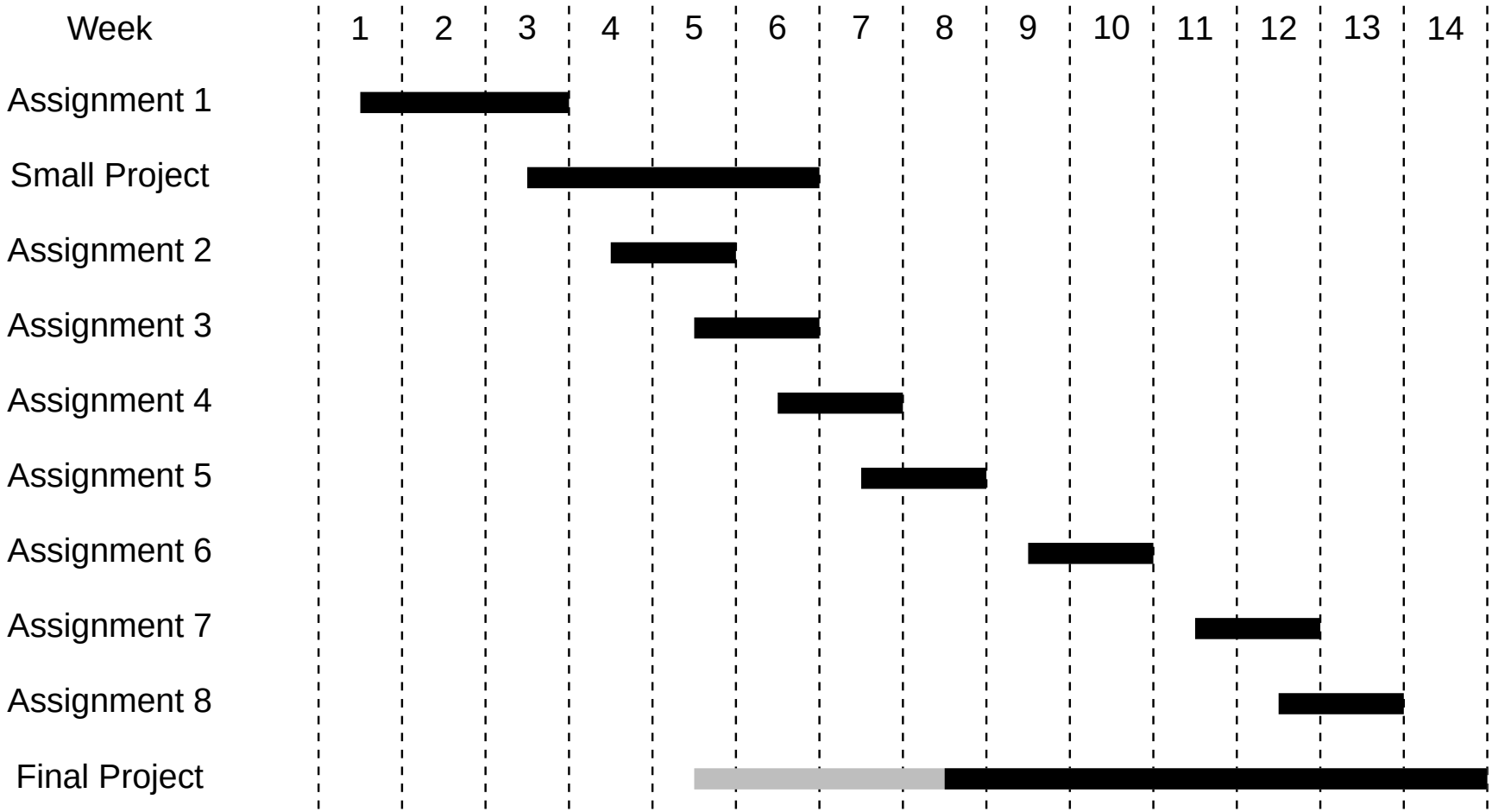
Polls

- Have you ever written a for-loop and if-else statement?
- Have you ever worked with R?
- Have you ever worked with Python, Matlab, etc.?
- Have you taken a class dedicated to linear models?
  - prerequisite
- Can you define the p-value?

Course Requirements

- assignments
  - 40 % of the grade (say 8 assignments of 5 %)
  - to be solved during the exercise classes
  - graded on the binary scale
  - collaboration (and questions) encouraged, but individual submissions required (avoid perfect copies!)
- data exploration – small project
  - 20 % of the grade
  - if the chosen data set too simple, can be composed of multiple data sets
  - in groups of 2-3 students
- project: data exploration+analysis *or* simulation study
  - 30+10 % of the grade
  - the 10 % for added value
  - in groups of 2-3 students

Expected Progress



Course Requirements

- 1 assignment = 5 % of the grade = 0.25 on the 1-6 grade scale
  - missing all assignments ⇒ final grade 4.0 at best!
- R, Markdown and GitHub for the assignments and projects will be needed
  - submission of GitHub links to the Moodle needed for grading purposes
  - this is not a programming course, learn by doing!
- 2 hours of lecture per week
  - going through the course content
- 2 hours of exercises per week
  - working on assignments and projects
  - keeping up with the lecture (e.g. with R)

active participation = success in this course

Questions and feedback are always appreciated.

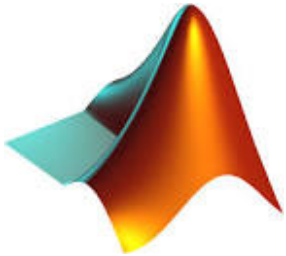
Evaluation starts right away!

Business Software

- all commercial
  - it has pros and cons
  - all (claim to) offer free academic versions
- popular in different fields
  - SAS: biomedicine, clinical research, etc.
  - SPSS: psychology, social sciences, etc.
  - STATA: econometrics, finance, etc.

Academic Software

Python

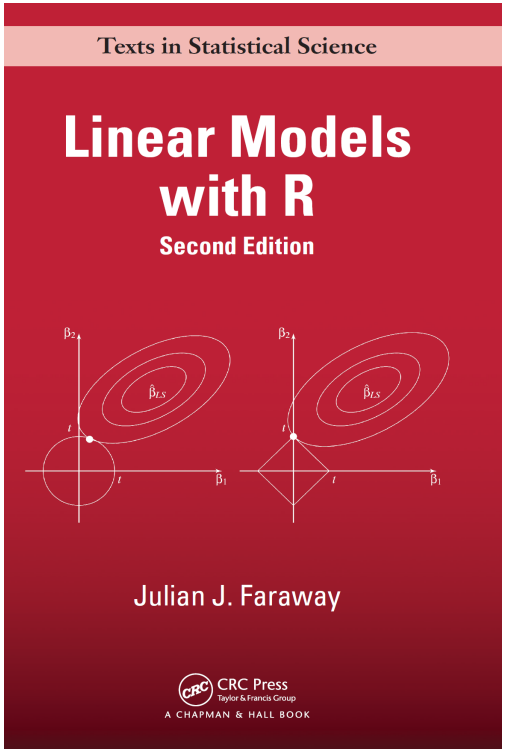


Matlab

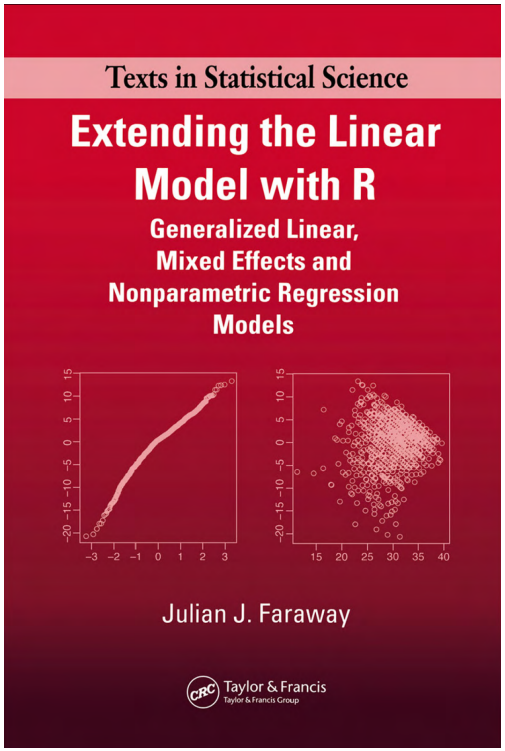
- all well documented, easy to use, with lots of examples and extensive community support
- each has its strengths and weaknesses, none is perfect
- we will use R!
- software packages are our **tools**, not skills!

free open source	free (mostly) open source (mostly)	paid (accessible) closed source
Statistics	Machine Learning	Numerical Math
Data Science		Optimization

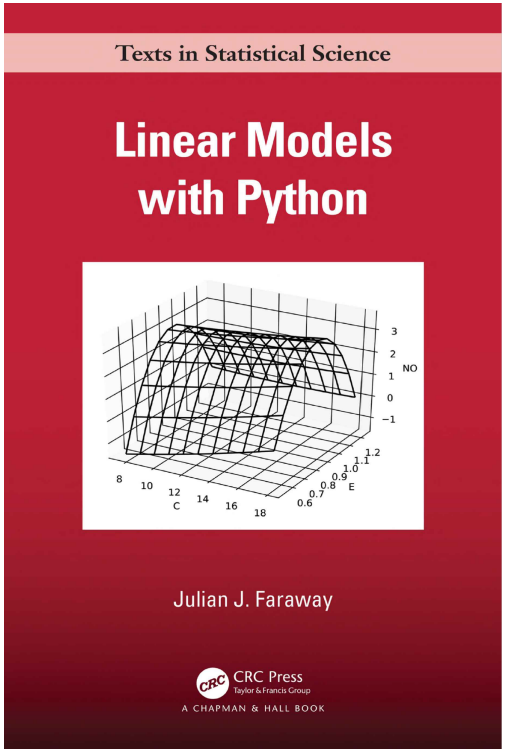
Statistics is done in R!



2004 – 1st Edition
2015 – 2nd Edition



2006



2020

Linear Models Recap

Data: $(Y_1, X_1)^\top, \ldots, (Y_N, X_N)^\top$, $Y_n \in \mathbb{R}$ is the response

Gaussian model: $Y = (Y_1, \ldots, Y_N)^\top \sim \mathcal{N}(X\beta, \sigma^2 I_{N\times N})$
- $X \in \mathbb{R}^{N\times p}$ is the model matrix containing (transformations of) $X_1, \ldots, X_N$

Model: $Y \sim (X\beta, \sigma^2 I_{N\times N})$ meaning that
- $\mathbb{E}\, Y_n = x_n^\top \beta$
- $\mathrm{var}(Y) = \sigma^2 I_{N\times N}$

Least Squares: $\min_\beta \|Y - X\beta\|_2^2$ or $\min_\beta \sum_n (Y_n - x_n^\top \beta)^2$

Fit: $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$ assuming X is full-rank
$\widehat{Y} = X(X^\top X)^{-1} X^\top Y$ are called fitted values

lm Example Takeaways

- doing statistics is more than just "massaging" data and taking whatever comes out
- linear models are the bread and butter of statistics, and as all statistical methods they are most easy to work with in R
- effective visualization is important beyond reporting results (code debugging, model verification, etc.)

lm Example Takeaways

- Jupyter Notebook (Python) or Live Script (Matlab) seem more interactive than R Markdown
  - not necessarily a good thing, e.g. reproducibility issues
- one can run R in Jupyter Notebook or conversely Python in R Markdown
  - I generally do *not* recommend either, except if you, say, work in R Markdown and want to use some Python packages (e.g. for deep learning)
- instead of R Markdown, one can use R Notebook (also in RStudio) to get some of the Jupyter Notebook interactivity

RStudio

Assignment 1 [5 %]

"**Assignment**" = **mandatory**, counts towards the final grade [5 %]

Go to Manual 02 and set up R, RStudio and GitHub for yourself as described there.

Make your first push to GitHub and submission to Moodle (push and submit whatever you want) as described in the course requirements.

Exercise

"**Exercise**" = does not count towards the grade

1. Familiarize yourselves with Markdown, and R itself by following the respective Manuals available in the GitHub repo.
2. We have done most of the analysis of the `chredlin` data in R, but we have done some pieces in Python and Matlab. Do the full analysis in R, i.e. complete the Markdown script corresponding to R translating the pieces of code from Python and Matlab. Also try to improve the model by considering transformations of (some of the variables.

References

JJ Faraway (2015) Linear Models with R (2nd Edition)

JJ Faraway (2020) Linear Models with Python

Poldrack (2019) Statistical Thinking for the 21st Century (online)

Tufte (1997) Visual Explanations