36-617: Applied Linear Models
Fall 2020

MW 1:30 -- 2:50 pm Pittsburgh Time
Zoom Meeting ID[1]: 954 6049 2939 / Passcode: 042583
Class Materials: https://canvas.cmu.edu/

"[I]t makes sense to base inferences or conclusions only on valid models" – *S.J. Sheather (2007)*
"All models are wrong but some are useful" – *G.E.P. Box (1978)*

Course Information

Instructor:	*TA:*
Brian Junker, Statistics & Data Science	Ben LeRoy, Statistics & Data Science
http://www.stat.cmu.edu/people/faculty/brian	http://www.stat.cmu.edu/people/students/bpleroy
`brian@stat.cmu.edu`	`bpleroy@stat.cmu.edu`
Office Hours:	*Office Hours:*
Zoom Meeting ID[2]: 923 9360 1531	Meeting ID[3]: 919 9389 2805
Passcode: 491610	Passcode: 664608
9:30–10:30am Mon & Wed, Pittsburgh Time	8–9pm Sunday, Pittsburgh Time
3–4pm Mon & Wed, Pittsburgh Time	Noon–1pm Monday, Pittsburgh Time
(or by appointment).	*(or by appointment).*

Prerequisites

You must be an MSP student to take this class. Beyond that, there are no formal prerequisites for this class. However, *I expect you to be familiar with statistical theory and the statistics of applied linear regression at a junior or senior undergraduate level. You will also be expected to know, or learn quickly, the computational software used for this course. That is, primarily* R. All homework and projects will be submitted online as pdf's on Canvas (`https://canvas.cmu.edu`). There are several ways to prepare pdfs:

- Write your assignment with pen/pencil and paper, and then scan to pdf
- Write your assignment in Microsoft Word and save as pdf
- Write your assignment in LaTeX, and generate pdf output
- Write your assignment using `rmarkdown` or similar tools in `rstudio`, and knit to pdf

`Rmarkdown` is nice for homeworks involving R, but it tends to encourage bad habits when you are writing reports and papers. So for the projects in this course I **strongly recommend** you use LaTeX or MS Word.

Please feel free to contact me if you have any questions or need additional information.

[1] Class web link: https://cmu.zoom.us/j/95460492939?pwd=dFlYbnBsRDdFVXF5MUh4TzlIbWlndz09
[3] BJ Office hours link: https://cmu.zoom.us/j/92393601531?pwd=TE50aDZKRFA5MmFpRnBQQjNUditSUT09
[3] BL Office hours link: https://cmu.zoom.us/j/91993892805?pwd=M2pSbFRiRGJxdFZzK0IwaWFiNThHZz09

Texts and Course Materials

Almost all of the material for this course are available on-line. In particular the main text for the course is available to download as a pdf free of charge to members of the CMU community. In addition, some other reading (especially about writing and communicating in statistics) will be available on-line in Canvas.

Texts

The primary text for this course is

Sheather, S.J. (2009). *A Modern Approach to Regression with R*. New York: Springer Science + Business Media LLC.*

Later in the course we will also take material from either or both of

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R.* New York: Springer Science + Business Media LLC.* See also http://www-bcf.usc.edu/˜gareth/ISL/

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* NY: Cambridge Univ Press.

Other books you may find useful include

Weisberg, S. (2013). *Applied Linear Regression.* John Wiley & Sons.

Weisberg, S. (2013). *Computing Primer for Applied Linear Regression, 4th Edition, Using R*. Available at `http://www.statpower.net/Content/313/R Stuff/alrprimer.pdf`

Berk, R. A. (2016). *Statistical learning from a regression perspective, 2nd Ed.*. New York: Springer Science + Business Media LLC.* (Do not get the 2008 1st edition.)

Course Description and Course Objectives

Linear regression and its generalizations are the basic modeling toolkit—some would say the *whole modeling toolkit*—of applied statisticians and data scientists. Almost every applied statistics or consulting problem that involves "inputs" and "outputs" can be solved, or at least profitably explored, using regression techniques. So it is essential that, as an applied statistician or data scientist, you understand how regression works, and practice using it.

Using regression in practice usually also involves translating a real-world problem into a techncial form that regression can be applied to, performing regression analysis, and translating back into real-world language, a process I sometimes label "ABA^{-1}". The ABA^{-1} process is the first step in a crucial part of the work of statisticians: effective communication. You will also practice communication of statistical results, using a report format that is useful for empirical scientific research.

By the end of this course you should be able to:

- Understand the tension between "valid" models and "wrong but useful" models.
- Carry out the ABA^{-1} process: translate from real world to quantitative terms, analyze quantitatively, and translate back to real world.

[*] You can buy the physical book many places online, e.g. `smile.amazon.com`, but you can also download the pdf for free at `link.springer.com` if you access through the CMU campus or VPN networks. I will also put pdf's on Canvas.

- Understand the machinery of linear regression well enough to use it intelligently.
- Build, fit and critically evaluate linear regression models and some generalizations of them, in a variety of messy, real-world settings.
- Communicate statistical results clearly in writing, from sentence to paragraph to full report, especially using a modified IMRaD format called IDMRaD.

This is a lot of material to cover. I would like to move quickly, but if I start to lose you I will slow down, since a good understanding of a few things is better than a poor understanding of many.

Computing

We'll mostly be working in R with supporting libraries, and as I suggested above, it is good to learn how to use LaTeX to produce documents. You should install this software to run on your own laptop:

- `R`, a statistical analysis and programming environment (best to have the most current version). See `http://cran.r-project.org/`
  - You may also like to install `RStudio`, from `https://www.rstudio.com/`. This provides an integrated development envioronment (IDE) for R, and also provides support for `rmarkdown` and other useful tools. I myself do not use R Studio much, but it is a fine option, and I may try to integrate more of it into the course.
- LaTeX is the academic standard technical typsetting system. Use it, or MS Word, for formal reports, (not `rmarkdown`). LaTeX is distirbuted in most TeX systems. The flavor of TeX that I have always used is called MiKTeX. It is available at `https://miktex.org/`, for both Windows and Mac.
  - There is also a nice online version, called Overleaf (`https://www.overleaf.com/`). Overleaf is a great way to get your feet wet with LaTeX—lots of online help, and a nice user interface—without installing TeX on your own computer.

There is lots of help for R, RStudio and LaTeX on the www. I will post some links on Canvas.

Online Resources and Etiquette

Canvas: Canvas (https://canvas.cmu.edu) has all materials (class notes, handouts, homework assignments and solutions, etc.). It is also where you will

- Take weekly quizzes
- Turn in weekly homework assignments (in the "Gradescope" app within Canvas)
- Submit and review data analysis papers, and
- Ask for help outside of office hours (in the "Piazza" app within Canvas).

You can also find all of your grades for this course, throughout the semester, in Canvas.

Zoom: All classes and office hours will be conducted on Zoom. I plan to record classes and post the recordings on Canvas, so that you can view them at a reasonable hour if you are in a different time zone, and so that you can review a class whereever you are. *I will not be recording office hours, ever.* Zoom links for class and for office hours are on the first page of this syllabus.

During class it would be nice if you have your video on, but I do not require it. You should have your audio off, unless you need to speak. If you have a question, please raise your hand in the "participants" window, or write your question in the "chat" window. I will answer as soon as I see your hand or your question (if your question is urgent, and I am not responding quickly, feel free to turn on your audio and interrupt me verbally). If you are worried about being identified in the class recordings, it is OK to leave your video off and use a pseudonym to identify yourself in Zoom.

Gradescope: You will need to upload weekly homework assignments as pdf files on Canvas using the Canvas app "Gradescope". Your homework will also be graded online, and you can see your grades and the TA's comments, all in Gradescope.

Piazza: If you have questions that cannot be easily asked or answered in class or in office hours, Piazza is a good place to post your questions. You can answer or comment on each others' posts if you wish, or wait for one of us to answer. The TA and I will especially monitor Piazza during our office hours, and I will also check in on Piazza occasionally throughout each week.

Please be kind in your questions, answers and comments: On Piazza, there are no dumb questions, and no dumb mistakes.

Academic Integrity

As members of a top-ranked academic institution, your academic integrity is assumed and expected.

Unless I specifically direct otherwise, your work is expected to be your own. For all work, if you get ideas or words from a website, journal article, book, another person (in or out of this class), etc., cite the source in your writeup, right where you use it. Then put a bibliography or list of sources cited at the end of the writeup. ***If you are not sure what is allowed, or required, please ask me.***

Carnegie Mellon guidelines are listed at `http://www.cmu.edu/academic-integrity/` (click on the "Student" link near the top of the page there); however, I expect each of you to behave well above these lower bounds.

Disability and other Special Needs

Carnegie Mellon makes great efforts to provide physical and programmatic campus access to everyone. Disability Resources ensures that qualified individuals receive reasonable accommodations and that they further receive the rights and protections to equal access programs and services as guaranteed by the Americans With Disabilities Act (ADA) and Section 504 of the Rehabilitation Act of 1973.

If you have a documented disability, please let me know so that we can take whatever steps are needed to accomodate your needs.

Please contact CMU's Disability Resources office (http://www.cmu.edu/hr/eos/disability/) if

- You think you may have a disability and want to document it;
- You have a documented disability that is not being adequately accomodated.

For other issues and special needs, please contact me, your advisor or another trusted mentor, and/or the Office of the Dean of Student Affairs (`https://www.cmu.edu/student-affairs/resources.html`).

Student Work

Your work for this class will consist of:

10-ish Homeworks	20%
Monday Quizzes	10%
2–3 Short Projects	50%
Peer Review	10%
Participation	10%

- *Homework:* I intend to give roughly 8–10 assignments. Homework will provide practice developing and exploring theoretical material, using software to analyze data, and some writing exercises.

  You **are** allowed to work with other students on these problems or refer to other sources if you would like, unless I forbid it on a particular assignment. I also reserve the right to ask you to stop working with a particular group of students, or work with someone else, if I think you are not getting the right things out of the group you are in. ***If you work with others, or use any other sources, please list you collaborators and other sources on your assignment.*** See also the section on Academic Integrity above.

  Prepare each hw as a single pdf and submit it on to Gradescope on Canvas.

- *Monday Quizzes:* These short quizzes (online, on Canvas) are intended to help me gauge your understanding of the reading each week. You may take the quiz at any time within an approximately 24 hour period. There are no makeups, but you may drop your two lowest scores.

- *Short Projects:* With these projects you will gain experience analyzing data and writing clear reports, in IDMRaD (Introduction–Data–Methods–Results–(and)–Discussion) format.

  You are **not allowed** to work with anyone except the instructor or TA on these projects, although you can refer to written sources on the web or in the library (e.g. books, journal articles, websites, blogs, questions asked and answered on quora, stackexchange, etc.) but you are not allowed to pose questions to any individual, group or other entity on line. ***You must list all sources used in a list of references at the end of your paper.*** See also the section on Academic Integrity above.

  Prepare each project report as a single pdf and submit it on Canvas.

- *Peer Review:* You will also be reading each others' project papers and giving feedback to them. You will be graded on the quality of your feedback.

- *Participation:* There will be some in-class discussions, some in-class exercises, and of course there will be office hours. Participate in as much of this as you can. Your first goal is to get me to remember your name. Your second goal is to get me to remember how much you participate in class. If I can't remember your name or I can't remember your participation, you will get a low participation grade.

In all your work, please label all output, plots, variables, etc., appropriately. Always be judicious about including computer output and graphs: show enough that we can clearly see what you are doing, but not so much that we will get lost or bored leafing through your work! A good rule of thumb is to remove any figures, tables, graphs, etc. that you have not written something interesting about.

A Note on Diversity

In this class, I will affirm and promote the inherent worth and dignity of every person, and I expect that every member of the class will do the same. The University is enhanced by the diversity of its members, in gender, sexuality, disability, age, socioeconomic status, ethnicity, race, nationality, religion and culture: each of you can contribute ideas and perspectives that no one else can. I will endeavor to present materials that are respectful of and accessible to all of our backgrounds and perspectives. Please let me know ways to improve the effectiveness of the course for you personally or for other students or student groups.

CMU Resources that may be useful to you include:

- The Center for Diversity and Inclusion: https://www.cmu.edu/student-diversity/
- The Intercultural Communication Center: https://www.cmu.edu/icc/
- The Office of Title IX Initiatives: https://www.cmu.edu/title-ix/

Tentative Schedule of Topics

The timing of topics and projects is approximate below, but this will give you some idea of how the course will progress.

Week	Dates	Tentative Topics	Tentative Sources
Week 1	Aug 31, Sept 2	**Intro, Appl Statistics, R, Writing**	Ch* 1, handouts
Week 2	(no class[1] on 7th), Sep 9	**Simple Linear Regression, Writing**	Ch 2, handouts
Week 3	Sep 14, 16	**Diagnostics & Transformations I**	Ch 3
Week 4	Sep 21, 23	**Multiple Regression** *Project 1 assigned*	Ch 5
Week 5	Sep 28, 30	**Diagnostics II & Variable Selection**	Ch 6
Week 6	Oct 5, 7	**Variable Selection**	Ch 7, handouts
Week 7	Oct 12, 14	**Logistic Regression, GLM's** *Project 1 due*	Ch 8, handouts
Week 8	Oct 19, 21	**Optional Topics:**	Handouts; stuff from G&H[3]
Week 9	Oct 26, 28	**More Optional Topics**	Handouts; stuff from G&H
Week 10	Nov 2, 4	**Weighted Least Squares** *Project 2 assigned*	Ch 9, (Ch 4?)
Week 11	Nov 9, 11	**Mixed Models**	Ch 10; stuff from G&H
Week 12	Nov 16, 18	**More on Mixed Models**	Handouts; stuff from G&H
Week 13	Nov 23 (no class[2] on 25th)	**Catch-up!** *Project 2 due*	N/A
Week 14	Nov 30, Dec 2	**Special Topics**	To be determined...
Week 15	Dec 7, 9	**Special Topics**	To be determined...

[*] All Chapters from Sheather unless otherwise noted.

[1] Labor Day (US Holiday).

[2] Thanksgiving (US Holiday).

[3] G&H: recommended text by Gelman & Hill (2009).

The appendices of G&H contain brief, but *very* useful advice! If you refer to and follow the advice in appendices A and B, your work as an applied statistician or data scientist will be much better!