36-617: Applied Linear Models
Fall 2021

MW 01:25PM--02:45PM BH A53
Zoom Meeting ID: TBA
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	TA TBA
http://www.stat.cmu.edu/people/faculty/brian	http://www.stat.cmu.edu/people/students/[TBA]
brian@stat.cmu.edu	[TBA]@stat.cmu.edu
232E Baker Hall	Room TBA
Office Hours:	*Office Hours:*
11am-Noon, Mon & Wed	TBA
(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.

Some guidance on Python vs R for this class is provided on page 3 below.

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

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

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`

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/

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 should also 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 do not use R Studio much in class, but it is a fine tool, especially for documenting larger data analysis and model fitting projects.
- LaTeX is the academic standard technical typsetting system. Use it, or MS Word, for formal reports, (not `rmarkdown`). LaTeX is distributed 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.

What about Python? More and more students come to the MSP program familiar with Python, either from a computer science course, or from past data science or machine learning projects; and you will learn and use Python in other MSP classes. It is especially useful for manipulating larger data sets, and for "gluing" together numerical, graphing and data manipulation routines from numpy, pandas, plotly, seaborn, etc. (Many of the same capabilities have been built into ggplot and the tidyverse in R.) I do not plan to use Python in class, but I do not mind if you use Python for some parts of your assignments, project work, etc. — *with two important caveats:*

- Most of the data sets in this course will be relatively small, and Python conveys no advantages over R for manipulating smaller datasets.
- I will focus on methods available in R for model fitting, variable selection and diagnostics. You will likely not be able to complete assignments focusing on these topics using Python alone.

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: During the 2020-2021 school year, when all of our classes were remote on Zoom, I recorded my live lectures—especially keeping in mind students in Asia who were not able to travel to Pittsburgh. It turned out that *all* students liked to have the recordings to refer to when reviewing their notes, planning data analyses, etc. I plan to record live classes again this fall qand post the recordings on Canvas, so that you can review a class whenever you like. If we go "remote" at some point due to the pandemic, or if some students are not able to attend class for some reason, we will already have Zoom set up for the class.

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.

Laptops in Class

In this course, it makes sense to bring your laptops to class. The text is a pdf, most class materials will be online, and there will be in-class activities that require your laptop.

This places a responsibility on you, however. As tempting as it is, please do not be distracted by email, online social networks, online shopping, etc. I want your attention focused on the class.

If you get bored and your attention drifts, ask a question in class. If that doesn't do it for you, let's talk in office hours about how we can help keep you focused in class.

Student Work

Your work for this class will consist of:

10-ish Homeworks	20%
Monday Quizzes	10%
2 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 (typically online on Canvas, most Mondays) are intended to help me gauge your understanding of the reading each week. 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.

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 Resources and Other Concerns

If you have a documented disability that is preventing you from doing the work in this class, please let me know so that we can take whatever steps are needed to accommodate your needs. If I am not able to help, or you have other related questions or concerns, please contact your advisor or a trusted mentor, and/or CMU's Disability Resource Office (http://www.cmu.edu/hr/eos/disability/).

For any other issues or special needs, please contact me, your advisor or a trusted mentor, and/or the Office of the Dean of Student Affairs (http://www.studentaffairs.cmu.edu/dean/).

If at some point during the semester you are unable to attend class because of the pandemic, please let me know so that I can make arrangements for you to participate remotely.

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/

A Note on Life-Work Balance

We care deeply about this course material and are excited to teach it, but we care even more so about your well-being. We are always happy to talk with you about life at CMU and Pittsburgh, your future career and education, or just about anything, really, if it supports your well-being here. There are many resources around you—family, friends, advisors, mentors, etc.—and you should take advantage of them. We also encourage you to be aware of professional resources such as Counseling and Psychological Services (CaPS; 412-268-2922 or `http://www.cmu.edu/counseling/`). However, if you or someone you know is in a life-threatening situation, call the police immediately (8-2323 on campus, 911 off campus).

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 30, Sep 1	**Intro, Appl Statistics, Regression Basics**	Ch* 1, 2
Week 2	Sep 6 (no class[1]), Sep 8	**Writing**	Ch 1, handouts
Week 3	Sep 13, 15	**Diagnostics & Transformations I**	Ch 3
Week 4	Sep 20, 22	**Multiple Regression** *Project 1 assigned*	Ch 5
Week 5	Sep 27, 29	**Diagnostics II & Variable Selection**	Ch 6
Week 6	Oct 4, 6	**Variable Selection**	Ch 7, handouts
Week 7	Oct 11, 13	**Logistic Regression** *Project 1 due*	Ch 8, handouts
Week 8	Oct 18, 20	**Generalized Linear Models (GLM's)**	Handouts; stuff from G&H[3]
Week 9	Oct 25, 27	**Causal Reasoning**	Handouts; stuff from G&H
Week 10	Nov 1, 3	**Generalized Least Squares** *Project 2 assigned*	Ch 9, (Ch 4?)
Week 11	Nov 8, 10	**Multilevel and Mixed Effects Models**	Ch 10; stuff from G&H
Week 12	Nov 15, 17	**Residuals, Estimation and Model Selection**	Handouts; stuff from G&H
Week 13	Nov 22 (no class[2] on 24^{th})	**Shrinkage, Examples** *Project 2 due*	Lecture notes; handouts
Week 14	Nov 29, Dec 1	**Multilevel GLM's, & maybe Bayesian Approaches**	Lecture notes; handouts

*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!