Welcome to Regression Analysis!

Prof. Maria Tackett

01.08.20

Click for PDF of slides

Welcome!

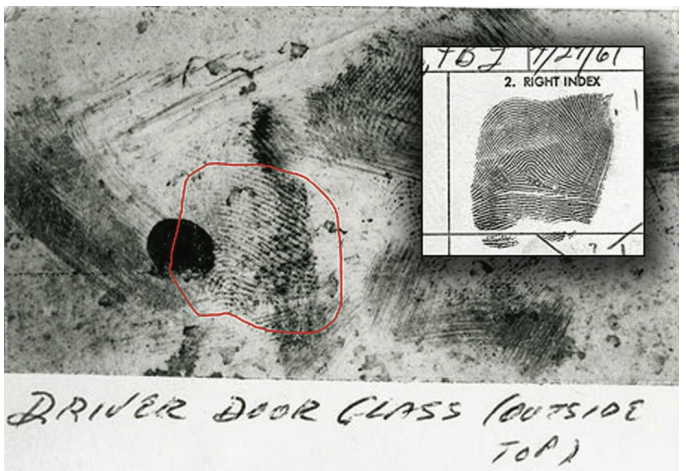
What is regression analysis?

"In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when **the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors')**. More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed."

- Wikipedia

Instructor: Prof. Maria Tackett

Email: maria.tackett@duke.edu
Office: Old Chem 118B



Teaching Assistants

- Youngsoo Baek: Mon 1p - 3p
- Cody Coombs: Tue 1p - 3p
- Sophie Dalldorf: Fri 1p - 3p
- Jonathan Klus: Mon 3p - 5p
- Matty Pahren: Tue 3p - 5p
- Ethan Shen: TBD

Regression in the "real" world

Is it the same as machine learning?

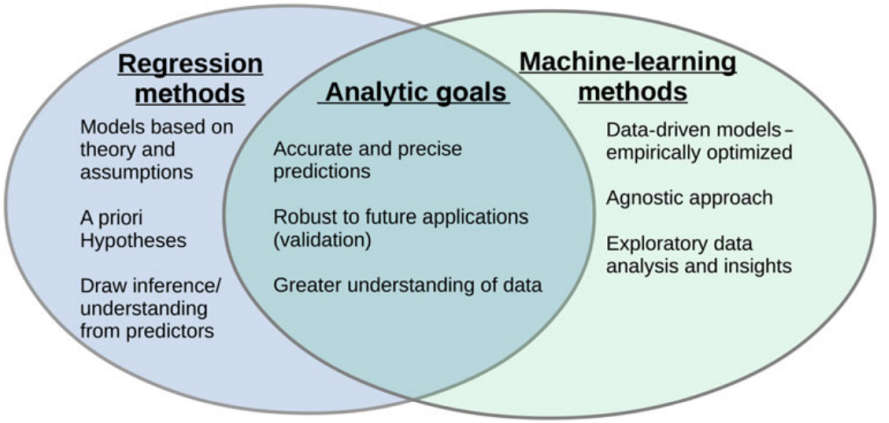
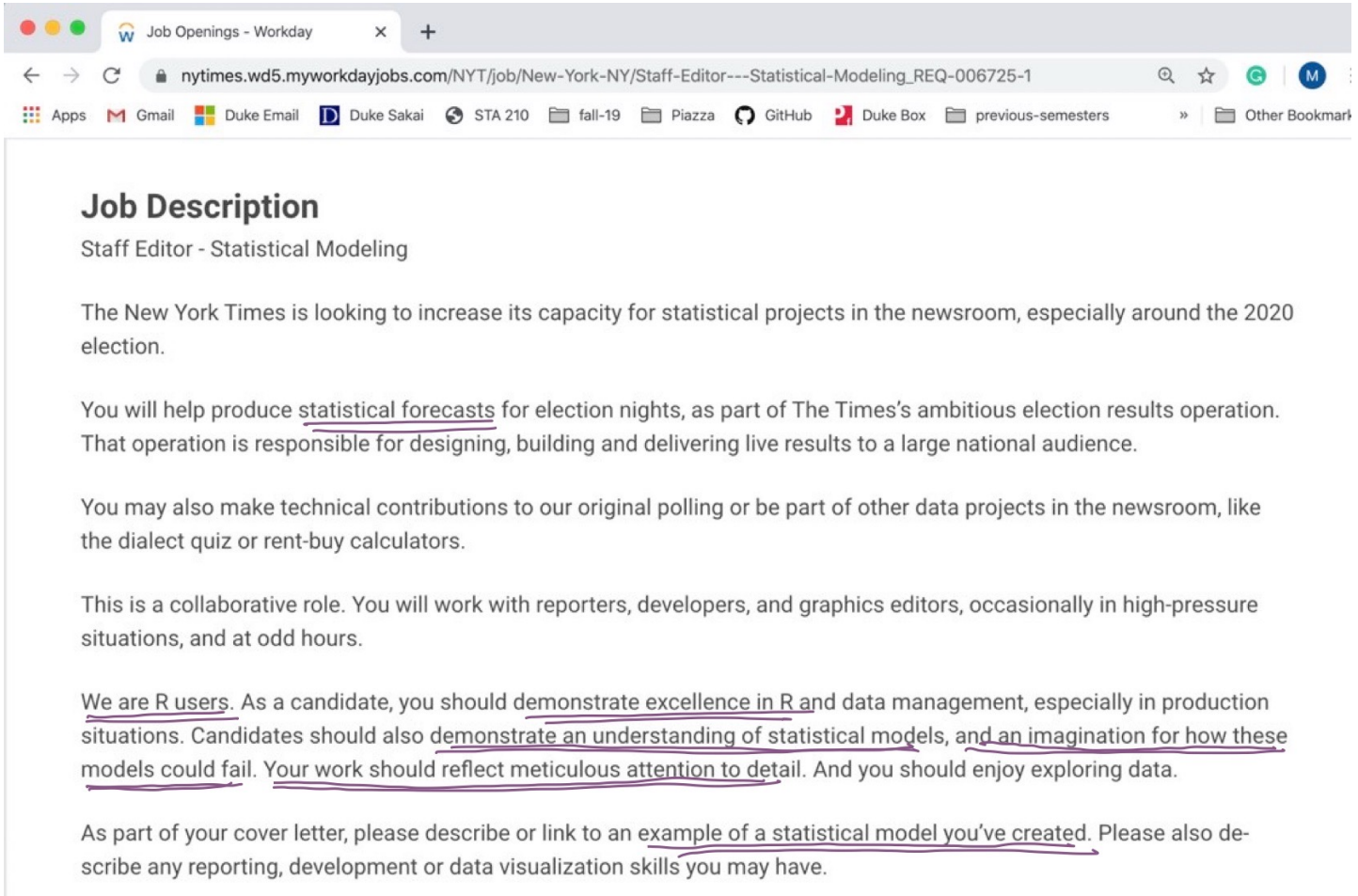


Figure 1 One perspective on the intersection of statistical modelling (blue) and machine-learning (green) goals. The figure highlights that while the processes differ the overarching goals are often the same.

Image source

"Many methods from statistics and machine learning (ML) may, in principle, be used for both prediction and inference. However, **statistical methods have a long-standing focus on inference**...By contrast, ML concentrates on prediction..." -*Statistics vs. machine learning*

Can I get a job that uses regression?



Job Description

Staff Editor - Statistical Modeling

The New York Times is looking to increase its capacity for statistical projects in the newsroom, especially around the 2020 election.

You will help produce statistical forecasts for election nights, as part of The Times's ambitious election results operation. That operation is responsible for designing, building and delivering live results to a large national audience.

You may also make technical contributions to our original polling or be part of other data projects in the newsroom, like the dialect quiz or rent-buy calculators.

This is a collaborative role. You will work with reporters, developers, and graphics editors, occasionally in high-pressure situations, and at odd hours.

We are R users. As a candidate, you should demonstrate excellence in R and data management, especially in production situations. Candidates should also demonstrate an understanding of statistical models, and an imagination for how these models could fail. Your work should reflect meticulous attention to detail. And you should enjoy exploring data.

As part of your cover letter, please describe or link to an example of a statistical model you've created. Please also describe any reporting, development or data visualization skills you may have.

NYT Staff Editor - Statistical Modeling

Requirements

- Familiarity with statistical modeling
- Expertise with R
- Experience with production-level code
- Interest in covering elections and politics
- Familiarity with JavaScript is highly desirable.
- Familiarity with election returns, voter files or polling is desirable.

This position is represented by the NewsGuild of NY

NYT Staff Editor - Statistical Modeling

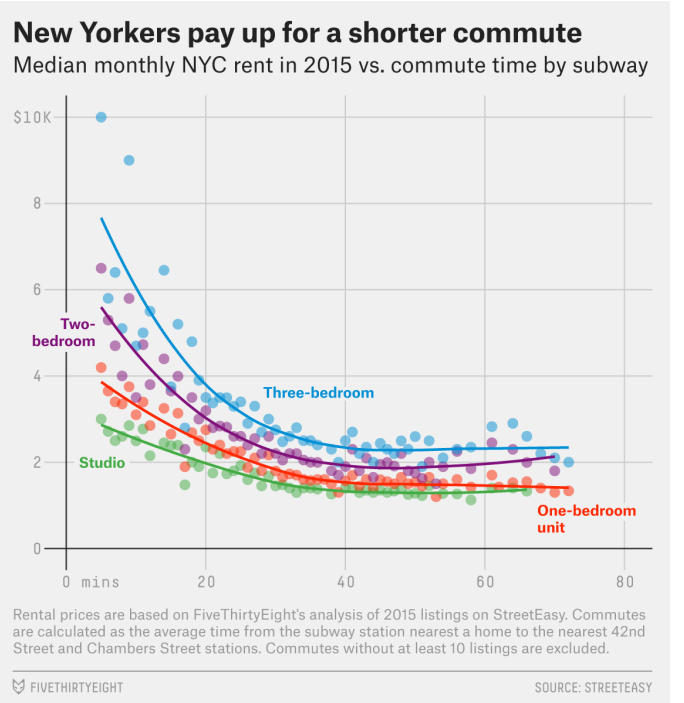
Fingerprint Analysis

*We use **Analysis of Variance (ANOVA) decomposition** to help determine whether the differences in fingerprints are circumstantial or because the prints were produced by different sources.*

Tackett, M., 2018. *Creating Fingerprint Databases and a Bayesian Approach to Quantify Dependencies in Evidence*. PhD dissertation, University of Virginia.

Apartments in New York City

*"We ran a **regression** to find the relationship between the rent of a one-bedroom home and the average of travel time from the station nearest to it to Midtown or downtown. It showed rent increasing by $56 per minute of decrease in average travel time."*



New Yorkers Will Pay $56 A Month To Trim A Minute Off Their Commute

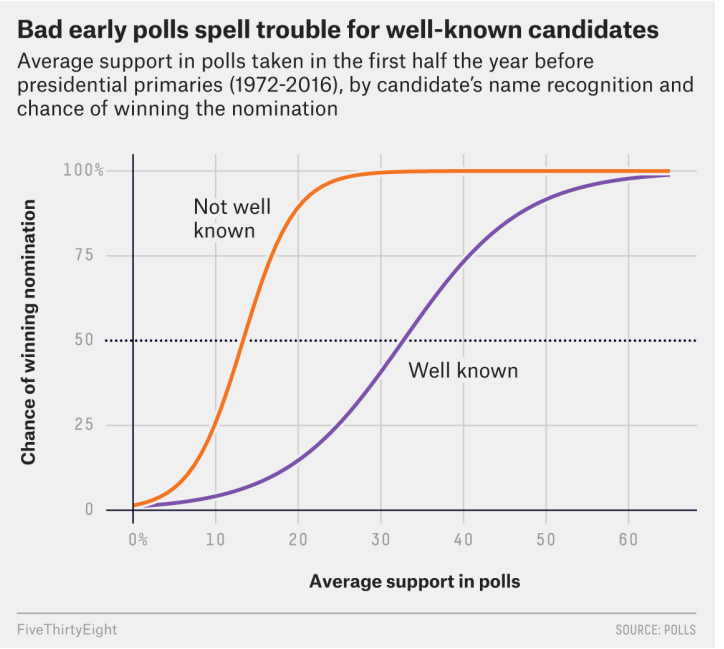
Impact on Educational Achievement

"Our objectives were to ... determine whether there are differences in the impact of lead across the EOG [End of Grade] distribution, and elucidate the impact of cumulative childhood social and environmental stress on educational outcomes. ***Multivariate and quantile regression techniques were employed****....The effects of environmental and social stressors (especially as they stretch out the lower tail of the EOG distribution) demonstrate the particular vulnerabilities of socioeconomically and environmentally disadvantaged children."*

Miranda, M., Dohyeong,K., Reiter, J., Galeano, M., & Maxson, P. (2009). Environmental contributors to the achievement gap. *NeuroToxicology*, 30, 1019-1024.

Analyzing Primary Polls

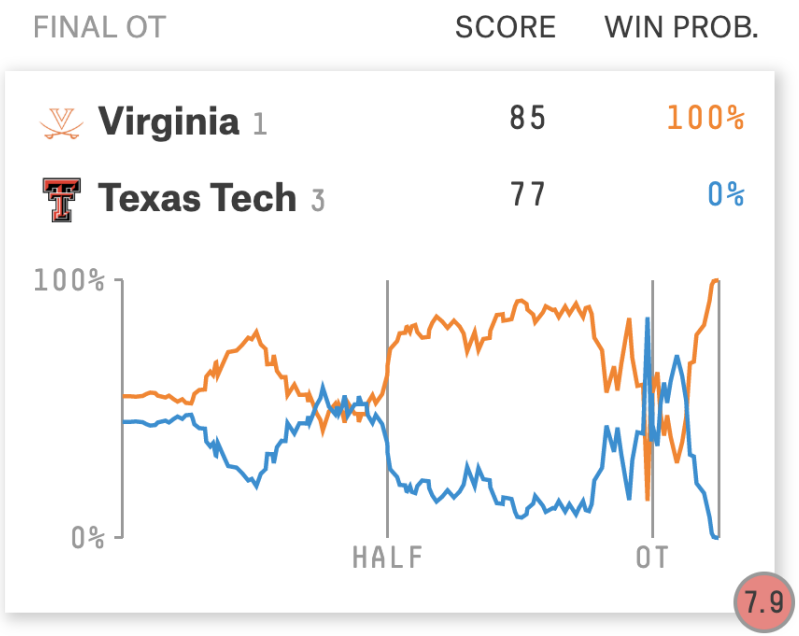
*"...In fact, we can use a **logistic regression** to estimate a high- and low-name-recognition candidate's chance of winning the nomination based on their polling average..."*



"We Analyzed 40 Years Of Primary Polls. Even Early On, They're Fairly Predictive."

March Madness 🏀

*"Live win probabilities are derived using **logistic regression analysis**, which lets us plug the current state of a game into a model to produce the probability that either team will win the game."* -"How Our March Madness Predictions Work"



2019 March Madness Live Predictions

Your Turn!

Movie Data Analysis

Read through the Movie Data Analysis

Answer the questions in the online form. Use **NetId@duke.edu** for your email address.

Discuss your answers with 1 - 2 people around you (don't forget to introduce yourself first!)

Movie Data Analysis

Notice in the graph in Part 2 that `budget` and `gross` are log-transformed. Why are the log-transformed values of the variables displayed rather than the original values (in U.S. dollars)?

Course Info

You will learn...

- How to apply methods for analyzing multivariate datasets, with an emphasis on interpretation
- How to check whether a proposed statistical model is appropriate for the given data
- How to address complex research questions using regression analysis
- How to use R and Git to do statistical analysis in a reproducible way
- The process for conducting data-driven research by applying the methods from this course to a long-term project

Where to find information

- **Course website**: http://bit.ly/sta210-sp20
  - Main hub for all information and course materials
- **Sakai**: https://sakai.duke.edu
  - Gradebook
  - Textbook (in Resources folder)
- **GitHub**: https://github.com/sta210-sp20
  - Work on homework, labs, and final project
- **Gradescope**: https://www.gradescope.com
  - Turn in assignments and receive feedback

Lectures

- Focus on concepts
- Think-Pair-Share (individual ➡️ small group ➡️ full class)
- Bring device to respond to in-class questions
  - Use your Duke email (NetId@duke.edu) to log response
  - Questions provide real-time feedback on understanding of the material
  - Respond to at least 75% of in-class questions over the course of the semester
- Throwable Mic!
  - Use the mic so everyone can hear the class discussion
  - Please wait for the mic before reponding / asking a question
  - You can throw it!

Labs

- Focus on **computing** using RStudio and GitHub
- Most labs done in groups of 3 - 4 students
- Must attend lab and participate to receive credit for group's submission
- If you miss lab for any reason, you may complete the lab for partial credi

Readings

- Handbook of Regression Analysis
  - Free PDF available on Sakai
  - Assigned readings about statistical concepts
  - **NOT** used for coding
- R for Data Science
  - Free online version. Hard copy available for purchase.
  - Some assigned readings and resource for R coding using `tidyverse` syntax.

Grade Calculation

Component	Weight
Homework (*lowest dropped*)	25%
Labs (*lowest dropped*)	15%
Exam 01 - Feb 26	20%
Exam 02 - Apr 15	20%
Final Project	15%
Teamwork & Engagement	5%

- If you have a cumulative numerical average of 90 - 100, you are guaranteed at least an A-, 80 - 89 at least a B-, and 70 - 79 at least a C-. The exact ranges for letter grades will be determined at the end of the semester.
- You are expected to attend lectures and labs. Excessive absences or tardiness can impact your final course grade.

Where to find help

- **Ask during lecture!** There are likely other students with the same question, so by asking you will create a learning opportunity for everyone.
- **Office Hours**: A lot of questions are most effectively answered in-person, so office hours are a valuable resource. Please use them!
- **Piazza**: Outside of class and office hours, any general questions about course content or assignments should be posted on Piazza since there are likely other students with the same questions.
    - If you know the answer to a question posted on Piazza, I encourage you to respond!

Academic Resource Center

The Academic Resource Center (ARC) offers free services to all students during their undergraduate careers at Duke

- Services include Learning Consultations, Peer Tutoring and Study Groups, ADHD/LD Coaching, Outreach Workshops, and more
- Contact ARC@duke.edu or 919-684-5917 to schedule an appointment

Testing Center

- This class will use the Testing Center to provide testing accommodations to students registered with and approved by the SDAO.
- If you need testing accommodations, register with SDAO and make your testing center appointments for Exam 01 and Exam 02 **as soon as possible**.
- For instructions on how to register with SDAO, visit their website at https://access.duke.edu/requests. For instructions on how to make an appointment at the Testing Center, visit their website at https://testingcenter.duke.edu.

Diversity & Inclusion

I strive to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities. To help accomplish this:

- If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. If you prefer to speak with someone outside of the course, your academic dean is an excellent resource.
- I (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to me about it.

Questions?

Please read the syllabus carefully and let me or the TAs know if you have questions.

https://www2.stat.duke.edu/courses/Spring20/sta210.001/syllabus.html

Computing

Reproducibility checklist

What does it mean for a data analysis to be "reproducible"?

Near-term goals:

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear **why** it was done? (e.g., how were parameter settings chosen?)

Long-term goals:

- Can the code be used for other data?
- Can you extend the code to do other things?

Toolkit

Repository (repo) containing all files for an assignment

Uses R programming language

Code, narrative, and output in one place through RMarkdown

Version control with Git commands

Create a GitHub Account

Go to https://github.com/, and create an account (unless you already have one). After you create your account, click here and enter your GitHub username.

Tips for creating a username from Happy Git with R.

- Incorporate your actual name!
- Reuse your username from other contexts if you can, e.g., Twitter or Slack.
- Pick a username you will be comfortable revealing to your future boss.
- Shorter is better than longer.
- Be as unique as possible in as few characters as possible.
- Make it timeless. Don't highlight your current university, employer, or place of residence.

Announcements

- Surveys and consent form
- Reading for Monday
- Office hours start next week
- NO LAB TOMORROW. Labs start on Jan 16
- **New to R or need a refresher?**
  - Attend an Intro to R workshop
  - Offered Jan 13 - 16, 6p - 7:30p, Gross Hall 270 (choose 1 night to attend)