Welcome to Regression Analysis

Dr. Maria Tackett

08.26.19

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

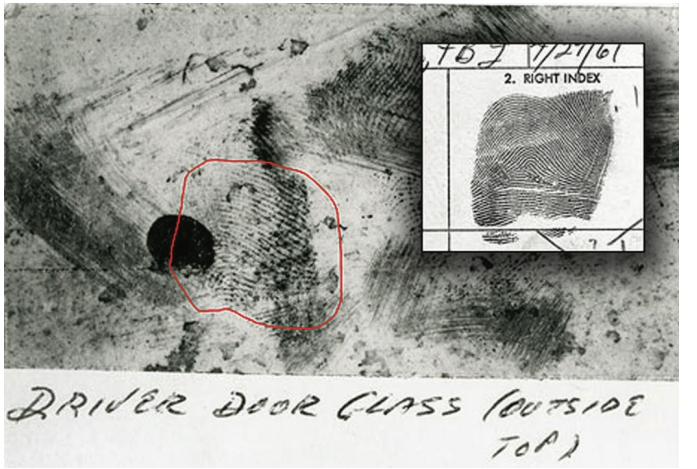
Instructor

Prof. Maria Tackett

maria.tackett@duke.edu

Old Chem 118B

Tue 2:30p - 4p (starting 9/3)



Teaching Assistants

Cody Coombs

- cody.coombs@duke.edu
- Old Chem 203B
- Thu 1p - 3p

Ethan Shen

- ethan.shen@duke.edu
- Old Chem 203B
- Thu 6p - 8p

Matty Pahren

- martha.pahren@duke.edu
- Old Chem 203B
- Tue 10a - 12p

Steven Winter

- steven.winter@duke.edu
- Old Chem 203B
- Wed 12p - 2p

Teaching Assistants

Tong Wu

- shitong.wu@duke.edu
- Old Chem 203B
- Fri 11:30a - 1:30p

Evan Wyse

- evan.wyse@duke.edu
- Old Chem 203B
- Mon 1p - 2p

Where to find information

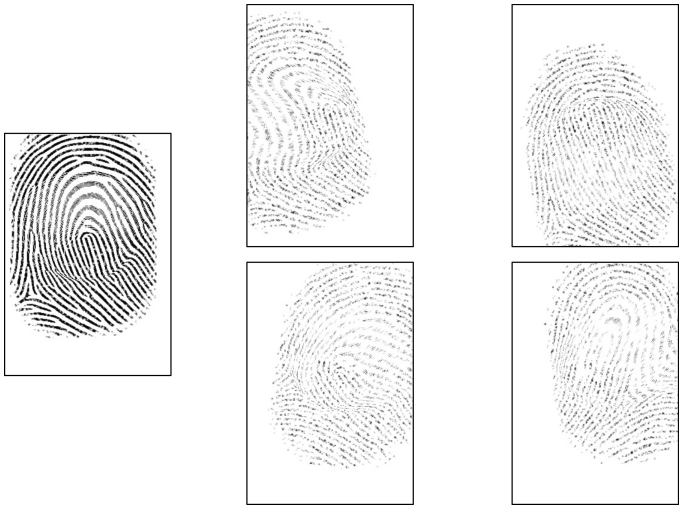
- Course website: http://bit.ly/sta210-fa19
- Sakai (textbook & links): https://sakai.duke.edu
- GitHub (assignments): https://github.com/sta210-fa19

Course Objectives

- Learn and apply methods for analyzing multivariate data sets
- Learn to check whether a proposed statistical model is appropriate for the given data
- Develop proficiency in addressing complex research questions using regression analysis
- Develop proficiency in computing tools used to conduct reproducible statistical analyses, specifically R and Git
- Learn the process of data-based research by applying the methods from this course to a final project

Examples of Regression Analysis

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

*"We can also take these polling averages and estimate the probability of a candidate winning a party's nomination using a **logistic regression**."*

*"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."

FiveThirtyEight March Madness Predictions

*"[Live win probabilities] 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!

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.
- Avoid words laden with special meaning in programming, like `NA`.

Raise your hand if you have any questions.

Join RStudio.cloud

- Go to http://bit.ly/sta210-fa19-rstudio, and log in with your GitHub credentials.
- You should see a project called *Movie Budgets and Revenues*. Click "Copy"; this will create your copy of the project and launch it.

Raise your hand if you have any questions.

Movie Data Analysis

1. In the *Files* pane in the bottom right corner, spot the file called `movies.Rmd`. Open it, and then click on the "Knit" button.

2. Put your name in the author field at the top of the file (in the `yaml` -- we will discuss what this is at a later date). Knit again.

3. Change the genre names in parts 1 and 2 to genres that interest you. The spelling and capitalization must match what's in the data, so you can use the Appendix to see the correct spelling and capitalization. Knit again.

You have made your first data visualization this semester!

Raise your hand if you have any questions.

Discussion

Discuss the following with a partner.

1. Start by introducing yourself! Name, year, major/ academic interest, favorite hobby.

2. Consider the plot in Part 1.

    - Describe how movie revenue has changed over time.
    - Suppose we use revenue as a measure of popularity. How has the popularity of each genre changed over time? In other words, are the genres that were most popular in 1986 still the most popular today?

3. Consider the plot in Part 2.

    - Which genre(s) tend to have the highest budgets?
    - In general, what is the relationship between a movie's budget and its total revenue? Are there any genres that show a different relationship between budget and revenue?

Course Policies

Class Meetings

Lecture

- Focus on concepts of regression analysis
- Interactive lecture that includes examples and hands-on exercises
- Bring fully-charged laptop to every lecture
    - Please let me know as soon as possible if you do not have access to a laptop

Lab

- Focus on computing using R `tidyverse` syntax
- Apply concepts from lecture to case study scenarios
- Work on labs in teams of 3 - 4
- Bring fully-charged laptop to every lab

Textbooks

- Handbook of Regression Analysis
  - Free PDF available through Duke libraries.
  - 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.

Activities & Assessments

- Homework: Individual assignments combining conceptual and computational skills. Have one week to complete. *Lowest score dropped.*
- Labs: Team assignments focusing on computational skills. Start in lab on Thursday and due Monday. *Lowest score will be dropped.*
- Exams: Two in-class exams.
- Final Project: Team project presented during the final exam period, **December 11, 9a - 12p**. You must complete the project and present in class to pass the course.
- Teamwork & Engagement : Score based on periodic peer reviews and completion of 80% of engagement surveys throughout semester (more on this in Lab 01).

Grade Calculation

Component	Weight
Homework	25%
Labs	15%
Exam I	20%
Exam II	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 after Exam 2.
- You are expected to attend lectures and labs. Excessive absences or tardiness can impact your final course grade.

Excused Absences

- Students who miss a class due to a scheduled varsity trip, religious holiday, or short-term illness should fill out the respective form.
    - These excused absences do not excuse you from assigned work.
- If you have a personal or family emergency or chronic health condition that affects your ability to participate in class, please let me and/or your academic dean know.
- Exam dates cannot be changed and no make-up exams will be given.

Late Work & Regrade Requests

- Homework assignments:
  - Late but within 24 hours of deadline: 20% penalty
  - Not accepted if submitted any later
- Late work will not be accepted for the final project
- Regrade requests must be submitted within a week of when the assignment is returned using the link posted on the course website

Academic Honesty

All work for this class should be done in accordance with the Duke Community Standard.

> To uphold the Duke Community Standard:
>
> - I will not lie, cheat, or steal in my academic endeavors;
> - I will conduct myself honorably in all my endeavors; and
> - I will act if the Standard is compromised.

Any violations will automatically result in a grade of 0 on the assignment and will be reported to Office of Student Conduct for further action.

Reusing Code

- Unless explicitly stated otherwise, you may make use of online resources (e.g. StackOverflow) for coding examples on assignments. If you directly use code from an outside source (or use it as inspiration), you must or explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.
- On individual assignments, you may discuss the assignment with one another; however, you may not directly share code or write up with other students.
- On team assignments, you may not directly share code or write up with another team. Unauthorized sharing of the code or write up will be considered a violation for all students involved.

Where to find help

- **If you have a question during lecture or lab, feel free to ask it!** 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.

Academic Resource Center

Sometimes you may need help with the class that is beyond what can be provided by the teaching team. In that instance, I encourage you to visit the 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. Because learning is a process unique to every individual, they work with each student to discover and develop their own academic strategy for success at Duke. Contact the ARC to schedule an appointment. Undergraduates in any year, studying any discipline can benefit! Contact ARC@duke.edu, 919-684-5917, 211 Academic Advising Center Building, East Campus – behind Marketplace.

Technology

- You should bring a laptop to every lecture and lab session. Outlets are limited, so make sure it is fully-charged.
- Ensure the volume on all devices is set to mute.
- Refrain from engaging in activities not related to the class discussion. Browsing the web and social media, excessive messaging, playing games, etc. is not only a distraction for you but is also a distraction for everyone around you.

Accessibility

Please contact the Student Disability Access Office (SDAO) if there is an element of the course that is not accessible to you. There you can engage in a confidential conversation about the process for requesting reasonable accommodations.

Please note that accommodations are not provided retroactively, so please contact them as soon as possible. More information can be found online at access.duke.edu.

Inclusion

In this course, we will strive to create a learning environment that is welcoming to all students and that is in alignment with Duke's Commitment to Diversity and Inclusion. If there is any aspect of the class that is not welcoming or accessible to you, please let me know immediately.

Additionally, if you are experiencing something outside of class that is affecting your performance in the course, please feel free to talk with me and/or your academic dean.

Questions?

Announcements

- Fill out the **Getting To Know You Survey on Sakai** - due 9/2 at 11:59p
- My office hours this week: Wed 1p - 2p or by appointment
  - Regular office hours start next week
- TA office hours start next week
- **New to R or need a refresher?**
  - *Work with Data* primer on RStudio Cloud: https://rstudio.cloud/learn/primers/2
  - "Data Visualization" in *R for Data Science*: https://r4ds.had.co.nz/data-visualisation.html
- **If you are on the waitlist, please see me after class.**