Introduction to predictive analytics

Lecture 1

STA 371G

Course goals

- Use regression and time series analysis to build predictive models
- Utilize simulations to forecast outputs based on uncertain inputs
- Given a new business situation, select an appropriate analysis, carry it out, and effectively communicate the results
- This is a practical course!

About the course staff

- Instructor: **Brian Lukoff, Ph.D.**
  - Office hours: M/W 11 AM-12 PM in CBA 3.440
  - Contact: `brian.lukoff@utexas.edu` or 415-652-8853
- TAs:
  - Office hours: M 11:30 AM-1:30 PM, 2-4 PM, W 12-2 PM, Th 4-6 PM in CBA 4.304
  - Help session: T 6-7 PM in the ModLab

Vasko Lalkov   Nicole Chia   Zameer Vaswani

Who am I?

- **Educator:** Teaching statistics at McCombs since 2014; previously taught at Harvard University and Boston University

- **Entrepreneur:** Currently co-founder and CEO of Perusall; formerly co-founder and CEO of Learning Catalytics (acquired by Pearson)

- **Engineer/statistician:** Software engineering/analytics background

1. **Find someone who...**

2. Course logistics

3. Let's do some statistics, yo

For each box on your bingo card, find someone who matches the description in the box. You must use a different person for each box.

The winner will be crowned the STA 371G Bingo Champion[TM].

1. Find someone who...

2. Course logistics

3. Let's do some statistics, yo

Canvas

- Access at `canvas.utexas.edu`
- This is your home base for the course
- Make sure you can log in and are enrolled in STA 371G in Canvas
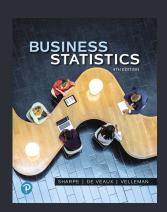
Class participation

- Understanding the concepts really only comes from practice
- We will use Learning Catalytics so you can practice the concepts during class
- No cost to use this
- Graded on participation, not correctness; answer 75% of the questions to get 100% of the credit
- Bring a laptop to every class (check one out from the Media Center if needed)
- A note about devices in class

Homework

- Why homework?
- 10 homework assignments during the semester
- We will use MyStatLab for online homework; purchase and register through Canvas (you'll get a 2-week free trial)

Textbook

- Access the textbook as an eBook through MyStatLab OR buy a hardcover or looseleaf textbook (Sharpe, *Business Statistics*, 4th ed) that includes a MyStatLab access code

- Reading "quizzes" are due by 7 PM; submit 2 questions for each reading about things that you found confusing/interesting

Statistical computing

- We will use R for statistical analysis throughout the course
- This is industrial-strength, state-of-the-art, and free software for statistical computing
- We will access R through RStudio, a graphical interface for R
- Download R and RStudio at `rstudio.com`

Exams

- Two midterm exams and a cumulative final exam
- All are given during class time; you'll take the test on your laptop in this room
- You'll have access to R during every exam
- I will drop the lowest of your three exam scores (unless the final exam is your lowest grade)

Team project

- You will pick a data set (or create one, e.g. through a survey) and apply regression techniques (we'll learn about this!) to build a predictive model
- Six deliverables throughout the semester:
    1. Draft proposal
    2. Final proposal
    3. Review of related work
    4. Exploratory data analysis
    5. Write up results as a paper
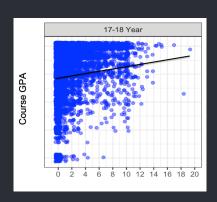   6. Presentation to class (during the finals period, on May 17)

Pretest

- It is critical that you be up-to-speed on the prerequisite material covered in STA 309
- By January 31, complete a pretest online to confirm your proficiency
- If you don't do well, expect to have to review some material on your own to avoid falling behind

Grading

Pretest	**1%**
Learning Catalytics	**5%**
Reading quizzes	**5%**
Homework	**14%**
Team project	**15%**
Exams	**60%**

PLUS (Peer-Led Undergraduate Studying)

- Weekly, student-run study groups
- You can apply to be a facilitator or just participate in any of the sessions
- Students who attend more PLUS sessions tend to get higher grades!

How to get an A in STA 371G

- Work on more problems than are assigned in the homework; can find problems in the book and on MyStatLab
- Consider attending the optional R help sessions in the ModLab
- Consider attending the PLUS sessions
- Get help when you need it
  - My office hours (after every class, until 12 PM)
  - TA office hours (see syllabus for schedule)
  - E-mail me to ask questions or set up a special appointment

1. Find someone who…

2. Course logistics
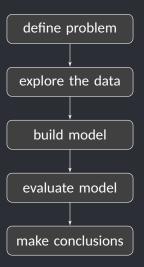
3. Let's do some statistics, yo

Purpose of a model

- **Make a prediction** about one variable based on the others
- **Understand the relationships** between the variables

Data analysis process

Define the problem

What personal characteristics about an instructor do you think are predictive of the scores they receive on student evaluations?
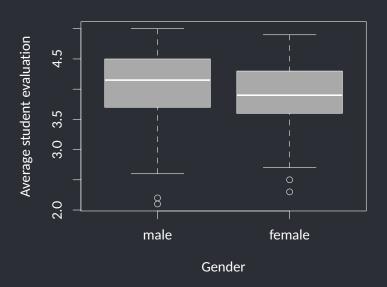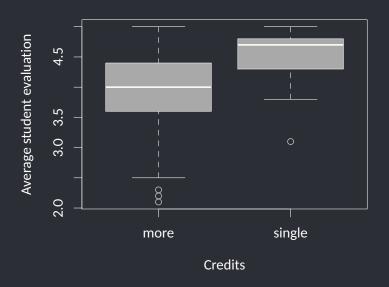
Hamermesh & Parker (2005) data set

- Student evaluations of $N = 463$ instructors at UT Austin, 2000-2002
- For each instructor:
  - **beauty**: average score from a six-student panel)
  - **gender**: male or female
  - **credits**: single- or multi-credit course
  - **age**: age of instructor
  - (and more...)

Explore the data

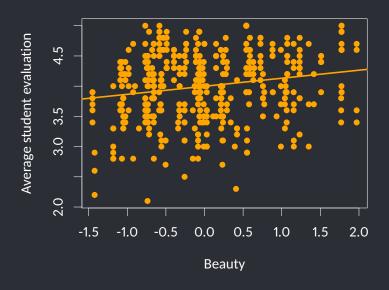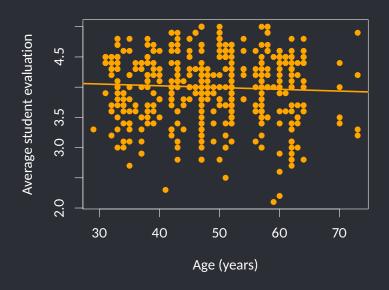# Explore the data

Explore the data

Explore the data

Build the model

A regression model lets us create a model that incorporates all of these relationships to best predict evaluation scores:

$$\widehat{eval} = 4.13 + 0.16 \cdot beauty - 0.2 \cdot female + 0.58 \cdot credits + 0 \cdot age$$

Build the model

A regression model lets us create a model that incorporates all of these relationships to best predict evaluation scores:

$$\widehat{eval} = 4.13 + 0.16 \cdot beauty - 0.2 \cdot female + 0.58 \cdot credits + 0 \cdot age$$

We predict a 40-year-old female, with a beauty score of 2, teaching a multi-credit course would get an evaluation score of

$$\widehat{eval} = 4.13 + 0.16 \cdot 2 - 0.2 \cdot 1 + 0.58 \cdot 0 = 4.18.$$
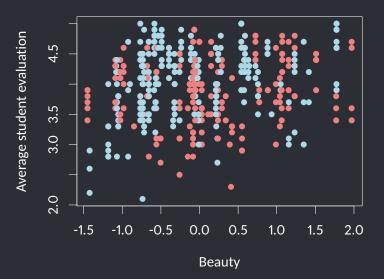
Evaluate the model

How could you evaluate the quality of this model?

Can we do better?

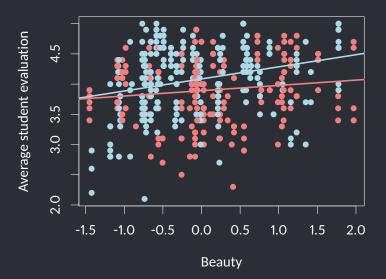Do you see a difference between men (blue) and women (red)?

Can we do better?

Do you see a difference between men (blue) and women (red)?

Five for the weekend

1. Read the syllabus
2. Install R and RStudio on your computer
3. Make sure you can log in to Canvas
4. Bring a laptop to class on Monday (and every day)
5. Take the pretest (by January 31)