San José State University

Math 261A: Regression Theory & Methods

Welcome to the first class

Dr. Guangliang Chen

Agenda

1. Introductions
2. Course overview
3. Syllabus information

Know your professor

Guangliang Chen

Know your professor - life milestones

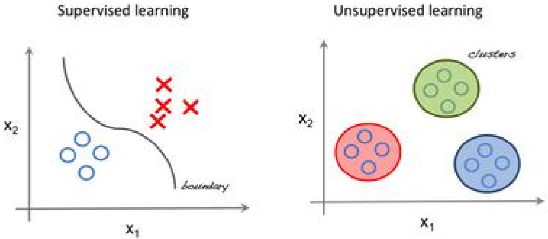
- 198x: Born in China, Anhui Province
- 2003: B.S. in Math, Univ. of Science & Technology of China, Hefei
- 2009: Ph.D. in Applied Math, University of Minnesota
- 2009-2013: Worked at Duke University as visiting faculty
- 2013-2014: Taught at Claremont McKenna College
- 2014-2020: Assistant Professor of Statistics, SJSU
- 08/17/2020-present: Associate Professor of Statistics, SJSU

Know your professor - teaching experience

- **Minnesota**: recitation instructor for calculus and college algebra
- **Duke**: calculus, differential equations, and linear algebra
- **Claremont McKenna**: calculus, and statistics
- **SJSU**:
  - Lower-division: Math 39, 42
  - Upper-division: Math 161a, 163, 164
  - Graduate: Math 250, 251, 261a, 261b, 263, 285, 298, 203

Know your professor - research expertise

I work in different areas of machine learning including classification and clustering, all with applications to image and documents analysis.



I have developed 2 machine learning courses at SJSU - Math 250 and 251.

Know your professor - personal background and interests

- I am married with 3 kids: Mina (8-), Zachary (6) and Noah (4).
- I am a Christian.
- I enjoy
  - reading (history, geography, politics, religion)
  - fishing, playing badminton, and cue sports (billiards)
  - traveling, cooking

Know your classmates

Most people have introduced themselves on Piazza.

I will give you a chance later to meet and talk to each other in small groups (I also have a task for you to do).

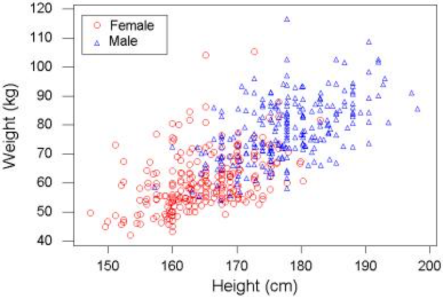
Overview of Math 261a

- **A graduate course on the topic of regression**, covering
  - simple and multiple linear regression,
  - use of categorical variables in regression,
  - model diagnostics,
  - variable transformations, and
  - nonlinear regression techniques.
- **Prerequisites**: Math 39 and 161A (each with a grade of B or better), 163* and 167R* (co-requisites)

An example of linear regression

Consider the following dataset[1] which contains weights and heights of 507 physically active individuals - 247 men and 260 women.

There seems to be roughly a linear relationship between the two variables, weight (y) and height (x). However, there is also uncertainty in y at each value of x.



[1] http://www.amstat.org/publications/jse/datasets/body.dat

Mathematically, we can write down the following equation:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

- x: **predictor**, regressor, or independent variable,
- y: **response**, or dependent variable
- ϵ: **error** accounting for the variability in y
- β_0, β_1: **parameters** (to be determined)

This is called a **simple linear regression** model.

Adding statistical assumptions

Suppose that for each fixed height x, the error follows the same normal distribution

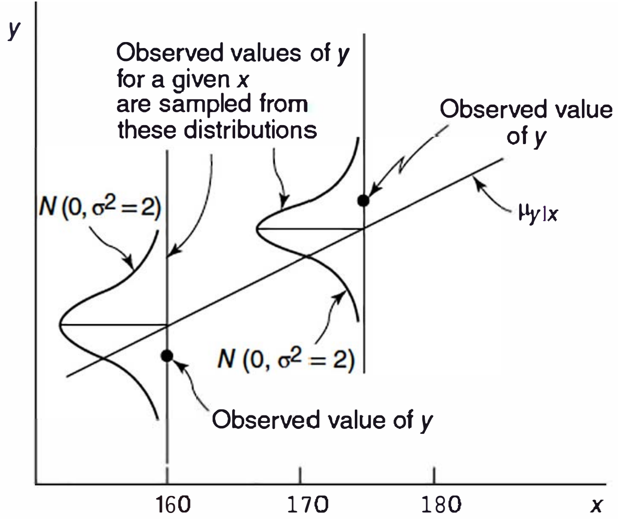
$$\epsilon \sim \mathrm{N}(0, \sigma^2).$$

This implies that

$$y \mid x \;\sim\; \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$$

with

- $\mathrm{E}(y \mid x) = \beta_0 + \beta_1 x$
- $\mathrm{Var}(y \mid x) = \sigma^2$


y
Observed values of y for a given x are sampled from these distributions
Observed value of y
$N(0, \sigma^2 = 2)$
$\mu_{y|x}$
$N(0, \sigma^2 = 2)$
Observed value of y
160
170
180
x

Parameter estimation

Once a regression model is specified, the next step is to choose the values of the unknown parameters (e.g., β_0, β_1 in the simple linear regression model) based on a set of observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

This process is called fitting the model to the data.

There are different ways to find the "optimal" values of the parameters:

- *Method of Least Squares*
- *Maximum Likelihood Estimation*

Model adequacy checking

The major assumptions that we have made for regression analysis are

- The relationship between the response y and the regressor x is linear, at least approximately.
- The errors are iid Gaussian with zero mean and constant variance (at different values of x)

We should always consider the validity of these assumptions to be doubtful and conduct analyses to examine the adequacy of the model we have tentatively built.

Uses of regression

- Data summary
- Description of relationship between the response variable and the explanatory variables
- Prediction

Regression models are only empirical models

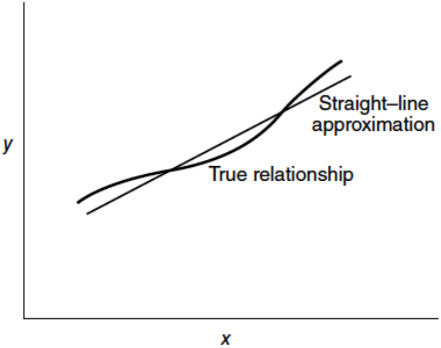


Figure 1.3 Linear regression approximation of a complex relationship.

Never assume the same linear relationship outside of data range

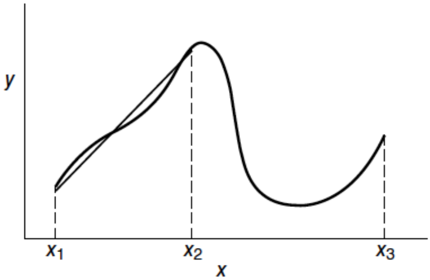


Figure 1.5 The danger of extrapolation in regression.

Multiple linear regression

Sometimes, a response variable may depend linearly on more than one explanatory variable, leading to the task of multiple linear regression.

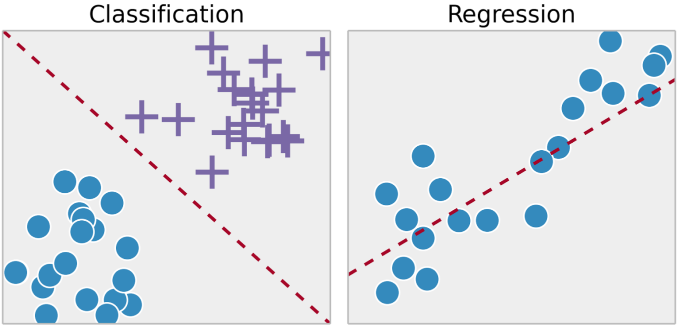
For example, in the weight-height example, we may add age and gender to the regression model:

$$y = \beta_0 + \beta_1 \underbrace{x_1}_{\text{weight}} + \beta_2 \underbrace{x_2}_{\text{age}} + \beta_3 \underbrace{x_3}_{\text{gender}} + \epsilon$$

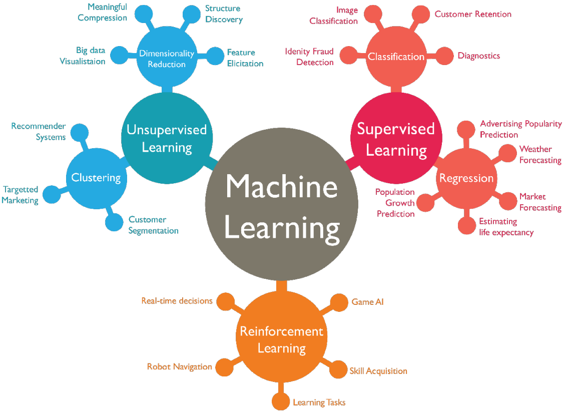
Note that x_3 is a categorical variable (male/ female). Chapter 8 of the book talks about how to handle categorical variables in regression tasks.

Regression vs Classification

In the cases where **the response variable is categorical**, the regression task becomes **classification**, which on its own is a vast, exciting field.



The bigger picture



Syllabus information

We'll use the following activity to explore the syllabus together:

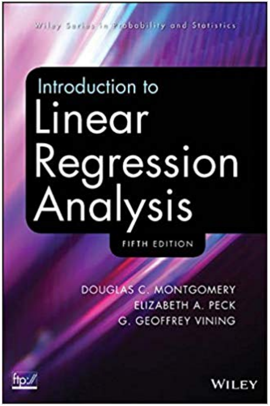
- **Breakout rooms**: You are divided into groups of size 3 at random to meet and talk to each other.
- **Task**: Discuss the course syllabus in your group to thoroughly understand the course organization, requirements, and policies. If any information is missing or unclear, write down your questions.
- **Length**: You have 10 minutes. After that, we will get back together and you will have a chance to ask your questions to me directly.

Required textbook

Introduction to Linear Regression Analysis, 5th ed., by Montgomery, Peck and Vining (2012), publisher: Wiley.

This text is available as an e-book (for free) through the SJSU library.

PDF copies of the required chapters can be found in Canvas.



Technology requirements

- A computer (laptop or desktop) with a camera and microphone
- Chrome browser and Proctorio extension
- R (statistical software) $\longleftarrow$ Math 167R is a co-requisite
- Access to a scanner (physical or cell phone app)
- Calculator

Learning management system

I will use **Canvas** in various ways:

- Post homework assignments and tests
- Record homework and test scores
- Make announcements (e.g. reminders, clarifications, deadline changes)
- Post Zoom recordings and annotated slides

Make sure to check your Canvas settings to receive timely notifications. Also, check if your email address in record is still good.

Zoom classroom etiquette

- Arrive at each Zoom meeting on time
- Have your cameras on (when without privacy concerns)
- Keep yourself muted when you are not speaking
- Avoid inappropriate name, language, or virtual background
- Use "raise hand" or the chat box to ask or answer questions
- Refrain from distracting activities on Zoom

Lecture recording policy

All lectures will be recorded and shared with the whole class afterwards. However, you should still make every effort to attend all classes.

If you would prefer to remain anonymous during these recordings, then please speak with the instructor about possible alternatives.

Students are prohibited from recording class activities, distributing class recordings, or posting class recordings. Materials created by the instructor for the course are copyrighted by the instructor. Students who record, distribute, or post these materials will be referred to the Student Conduct and Ethical Development office.

Course requirements

- **Homework (20%)**: Assigned regularly in Canvas.
- **Midterm 1 (25%)**: September 29, Tuesday.
- **Midterm 2 (25%)**. November 17, Tuesday.
- **Project 1 (10%)**: In this project you will learn something new and present it in class.
- **Project 2 (20%)**: In this project you need to apply what you learned in this course to a large data set.

Homework

Students may collaborate on homework but must independently write their own solutions.

You may write your work on paper or a tablet. Once completed, submit a legible, electronic copy to Canvas (as a single file attachment).

For some assignments, only a subset of the problems may be graded.

You must submit homework on time to receive full credit (late submissions within 24 hours of the due time can still be accepted but will automatically lose 10% of the total grade).

Your lowest homework score will be dropped.

Tests

The course has two midterms, each covering a distinct part of the course.

The midterms will be delivered via Proctorio.[2] Both of them are open book, open notes, but you are not allowed to communicate with other people during the test or use the internet to search for answers.

No make-up exam will be given if you miss a midterm exam unless you have a legitimate excuse such as illness or other personal emergencies and can provide documented evidence.

[2] https://proctorio.com/support

Some reminders about homework and tests

You must show all steps to earn full credit:

- It is your entire work (in terms of *correctness*, *completeness*, and *clarity*) that is graded.
- Correct answers with no or poorly written supporting steps will be given very little credit.

Some homework questions require coding in R, in which case you need to provide the R scripts you used, present your results in an organized, meaningful way, and interpret them carefully.

Please **write legibly** (unrecognizable work will receive no credit).

Project 1 (further learning)

Objective: To supplement the lectures by the instructor while engaging students through group work

Task: Learn a new concept/method about regression along with a classmate and teach it to the rest of the class

Topics: Can come from a skipped textbook section, or outside the textbook

What's required: A 15-minute presentation from each group

How it is graded: It will be graded based on your familiarity with the material and quality of your presentation (clarity, accuracy and completeness)

Project 2 (application)

Objective: To practice the knowledge and skills learned in the course by applying them to a nontrivial data set found from the internet

Task: Along with another classmate (different from your project 1 partner), perform regression analysis including model fitting, variable selection, transformations, assumption checking, diagnostics, and interpretation

What's required: A 10-minute presentation and a 5-page report from each group

How it is graded: It will be graded based on depth, clarity, accuracy and completeness

Grade cutoffs

...will be determined by combining the following **percentages**:

- A+: 97%, A: 93%, A-: 90%
- B+: 86%, B: 80%, B-: 76%
- C+: 73%, C: 68%, C-: 65%
- D: 60%
- F: 59% or less

and **the actual distribution of the class** at the end of the semester.

Your responsibilities in learning

My duty as an instructor is to disseminate knowledge while helping you learn. **The ultimate responsibility of learning is upon the student, not the instructor**. Thus, you should

- Attend all classes
- Participate in classroom discussions
- Read the textbook before and after class
- Take time to think through the concepts
- Do your homework
- ASK whenever you don't understand something!!!

Academic dishonesty

Students who are suspected of cheating during an exam will be referred to the Student Conduct and Ethical Development office and depending on the severity of the conduct, will receive a zero on the assignment or a grade of F in the course.

Special accommodations

If you anticipate needing any special accommodation during the semester (e.g., you have a disability registered with SJSU's Accessible Education Center), please let me know as soon as possible.

Instructor availability

- **Office hours**: TWR 10:30-11:30pm (Zoom ID: 983 3362 3851), and by appointment.
- **Piazza**: `https://piazza.com/class/kdt29tgm77h2wa`
- **Email**: `guangliang.chen@sjsu.edu`. I check my emails frequently, but you should allow a turnaround time of up to 24 hours (on weekdays) or 48 hours (during weekends).

Student feedback

I strive to teach in the best ways to facilitate your learning. To achieve this goal, it is very helpful for me to receive timely feedback from you.

You can choose to

- talk to me in person, or
- send me an email, or,
- submit your feedback anonymously through `http://goo.gl/forms/f0wUD5aZSK`.

Reading assignments

Read Chapter 1 of the textbook about the general concepts of regression.

Next Tuesday I will start teaching

Simple linear regression

from Chapter 2 of the textbook (Sections 2.1-2.8).