STAT 849
Theory and Application of Regression and Analysis of Variance - I

Yinqiu He

Department of Statistics
UW Madison

Email: yinqiu.he@wisc.edu

Instructors

Instructor: Yinqiu He (Pronunciation: In-Cho Her)
Email: yinqiu.he@wisc.edu
Office hours: Wednesday 12:15PM - 1:30PM @7225D MSC

TA: Sijia Fang
Email: sfang44@wisc.edu
Office hours: Friday: 12:15PM - 1:30PM @1276 MSC

(Updates will be posted on Canvas.)

Course description

- This course is an **advanced** graduate study in statistics. It is designed for first or second year statistics PhD students. One of the four core courses to be tested in PhD qualifying exam.
- There are two courses in this sequence; the subsequent one is STAT 850, offered in spring semester.
- Course website: `canvas.wisc.edu`. Lecture notes, homework assignments, and important announcements will be posted there.

Prerequisite

- There are no formal course prerequisites to this class. But we will assume a **solid** background in linear algebra, probability, and statistical theory. Please find the pdf "Mathematics Prerequisites for Success in STAT 849.pdf" on canvas.
- Requires a general ability to do mathematical proofs and hands-on data analysis and programming skills.
- Students who wish to take the course for credit should submit an entrance quiz. A 75-mins countdown timer will start when you download the Quiz0.pdf file.
- Deadline for submission is **Sep 8, 11:59pm** (tomorrow night). Graded by completion. Work independently.

Textbook

There are no required texts. The material that covered does not appear in one single text. The following is a list of useful supplementary reading and references.

1. Julian J. Faraway (2004) *Linear Models with R*.
2. Seber and Lee (2003) *Linear Regression Analysis (2nd ed)*
3. Ronald Christensen. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models.*
4. McCullagh and Nelder (1999) *Generalized Linear Models (2nd ed)*.

Reading instruction will be listed on Canvas "Supplementary Reading" page.

Homework

- Assignments will be posted on canvas and due back in approximately one or two weeks. There will be 5 ± 1 assignments.
- Upload a single PDF on Canvas for the homework assignment. Start each exercise on a new page and make sure they are in the correct order. Typed homework will be given 1 additional bonus point. Use R markdown to present R codes and results.
- Read syllabus requirements and communicate with the TA.

Exams

Two in-class midterms:

- Closed book and closed notes. You may take one (8.5 by 11 inches; both sides) paper as a cheat sheet.
- Midterm 1: **Oct. 17th, M, 11:00AM-12:15PM**.
  Midterm 2: **Nov. 21st, M, 11:00AM-12:15PM**.

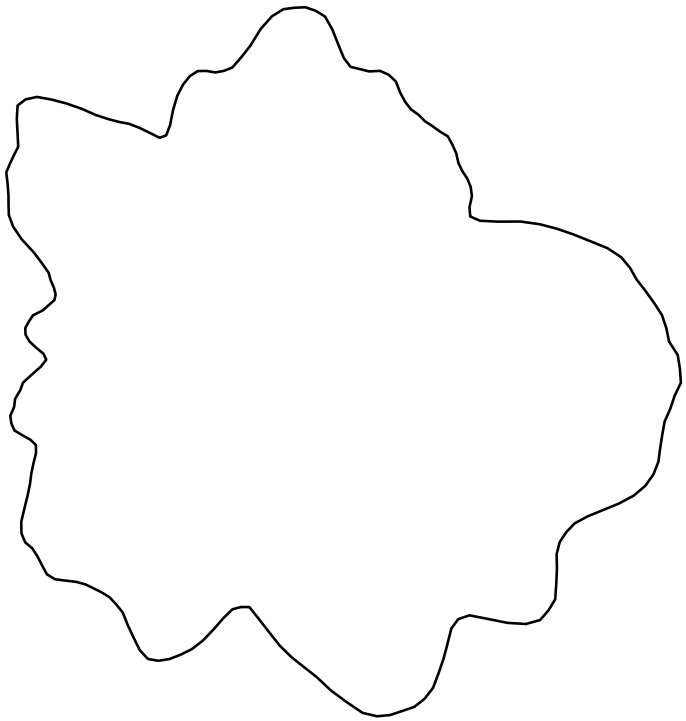
Final is a take-home project:

- You will be provided a dataset with some questions.
- Write a report independently and keep it confidential.
- The deadline for submission is **Dec 21st, M, 12:15PM**.

Grade: The grade will be weighted as: entrance quiz-0 and regular homework (25%), midterm 1 (25%), midterm 2 (25%), and the final (25%).

Email Policy: When sending an e-mail on the course, please include "STAT849" in subject line.

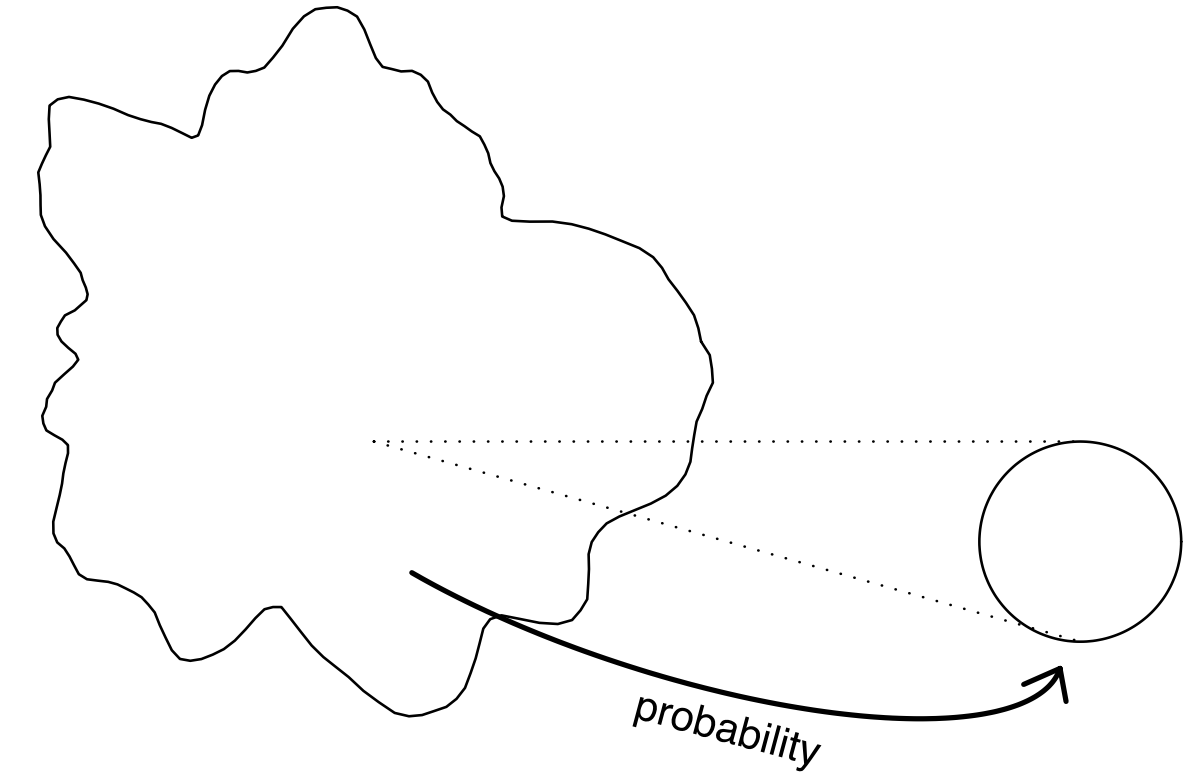
Probability vs. Statistics



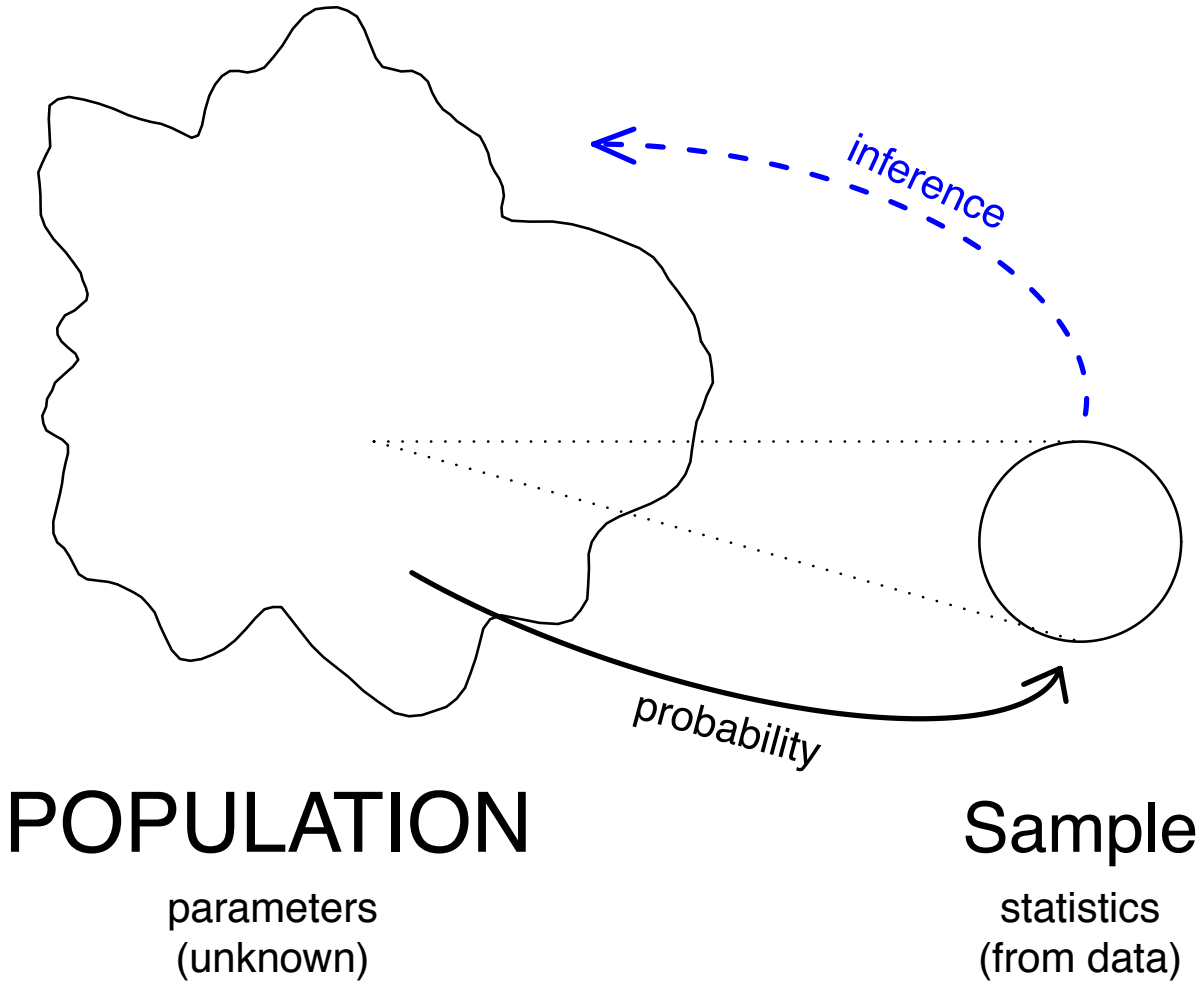
POPULATION

parameters
(unknown)

Probability vs. Statistics



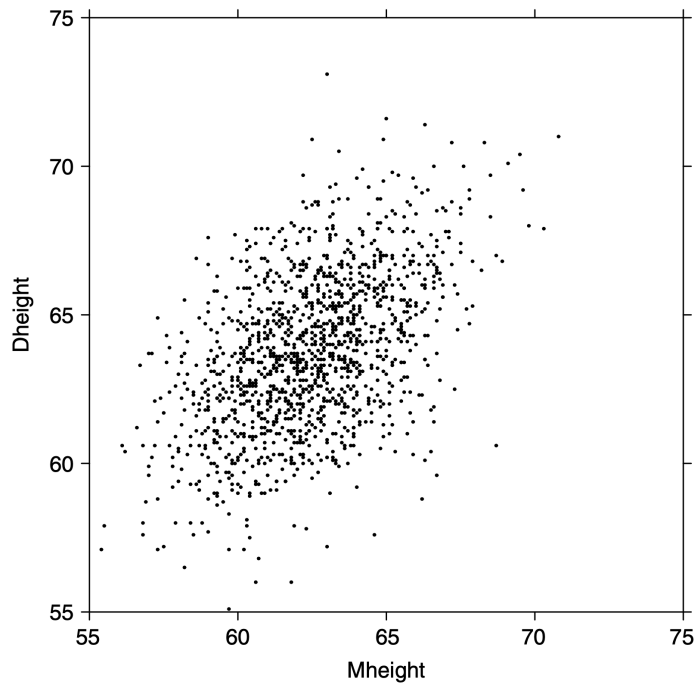
Probability vs. Statistics



Regression Analysis

- Goal: Construct models to explain relationship between variables.
- Karl Pearson, late 19th century,
  studied $n = 1375$ heights of mothers in the United Kingdom under the age of 65 and one of their adult daughters over the age of 18

Figure: Scatterplot of mothers' and daughters' heights in the Pearson's data.



Regression Analysis

Goal: Learn an unknown function f that relates variables $Y \in \mathcal{R}$ and $\boldsymbol{X} \in \mathcal{R}^p$ through $Y \approx f(\boldsymbol{X})$.

Terminology:

- Independent variables (covariates, predictors, regressors, explanatory variables, exogeneous variables):

$$\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathcal{R}^p.$$

- Dependent variables (response, outcome, endogeneous variables):

$$Y \in \mathcal{R}.$$

Remark: The terms "independent" and "dependent" do not imply statistical independence or linear algebraic independence. They refer to the setting of an experiment where the value of $\boldsymbol{X}$ can be manipulated, and we observe the consequent changes in Y.

Regression Analysis

- The regression analysis is empirical (based on a sample of data).
  Collect n pairs of observations $(Y_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$:

$$Y_i \in \mathcal{R}, \quad \boldsymbol{X}_i \in \mathcal{R}^p.$$

- n is the sample size.
- Each pair $(Y_i, \boldsymbol{X}_i)$ tells us what is known about the i-th "observation" ("subject", "case", "analysis unit", "individual").

Why do we want to do regression analysis?

Prediction: predict the value of the response Y given a particular value of covariate $\boldsymbol{X}$.

- What is the price of a 3500ft^2 house in Boston area?
- **supervised learning** in machine learning

Model Inference: inductive learning about the underlying relationship between the response Y and covariate $\boldsymbol{X}$.

- Do taller mothers tend to have taller daughters?
- The goal is to better understand the physical (biological, social, etc.) mechanism underlying the relationship between $\boldsymbol{X}$ and Y.

Examples

- Prediction: An empirical model for the weather conditions 48 hours from now could be based on current and historical weather conditions. Such a model could have a lot of practical value, but it would not necessarily provide a lot of insight into the atmospheric processes that underly changes in the weather.
- Inference: A study of the relationship between childhood lead exposure and subsequent health problems would primarily be of interest for inference, rather than prediction. Such a model could be used to assess whether there is any risk due to lead exposure, and to estimate the overall effects of lead exposure in a large population. The effect of lead exposure on an individual child is probably too small in relation to numerous other risk factors for such a model to be of predictive value at the individual level.

Topics

- Least-squares fitting: estimation and testing;
- Analysis of variation;
- Measurement errors, confounding;
- Regression diagnostics;
- Model selection;
- Prediction, bias and variance trade-off, shrinkage methods;
- Generalized linear models and beyond (if time permits).