Chapter 1: Introduction

Stats 500, Fall 2017
Brian Thelen, University of Michigan
443 West Hall, bjthelen@umich.edu

Statistical Approach to a Scientific Problem

- Collecting data
- Initial (exploratory) data analysis
- **Inferential statistics**

Collecting data

(1) Determining the population

(2) Sampling

- Is this a **random sample** ? An experiment or an observational study?
- Is this a **sample of convenience** ? Is there non-response?
- Is there **missing data** ? How is it handled?

(3) Sometimes can do **designed experiments**

Initial data analysis - Descriptive statistics

- Displaying data graphically
- Summarizing data
- Organizing data

Inferential Statistics

Based on data

- Testing hypotheses
- Reaching conclusions
- Making decisions

via **linear regression analysis or generalized linear model**

Regression Analysis

Build a model to explain the relationship between a single variable Y and other variables $X_1, \cdots, X_p$

Y: **response** variable, output, dependent variable

X: **predictor** variable, input, independent variable

- $p = 1$: regression
- $p > 1$: regression

Goals of Regression Analysis

- **Prediction**
- **Effect of predictor variables**
- Description of data structure
- **Warning:** regression analysis does not establish **causation** (i.e., you cannot tell whether X causes Y or the other way around)

Types of Variables

Qualitative, **categorical** : can't say one is bigger than

Quantitative, **numerical**

- Discrete counts
- Continuous measures

Examples

Population: STATS 500 students

- Categorical: ethnicity
- Binary (two values only): gender
- Discrete: # of credits, # of house mates
- Continuous: age, height

What We Will Cover

- Y is a continuous variable: linear regression
- Y is a binary variable: logistic regression
- Y is a discrete count: Poisson regression
- X: continuous, discrete or categorical

Pima Data Example

- Data collected on 768 adult female Pima Indians
- Variables: number of times pregnant, plasma glucose concentration, diastolic blood pressure, skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, age, and a test whether the patient showed signs of diabetes

Pima Data Example - ctd

Many possibilities

- Y: diabetes; X_1, X_2: diastolic, BMI
- Y: BMI; X_1, X_2: diastolic, test
- Y: test; X_1, X_2: diastolic, BMI
- Y: number of times pregnant; X_1, X_2: age, BMI

Emphasis of the Course

- **Practice** of linear regression models
- Goal: **what methods are available & when** they should be applied
- Many **examples** , less mathematical theory
- More **intuition** , less derivation of formulas
- Will still learn **mathematical foundations** behind practical tools