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.
- **Typo: It should be Sep 8, 11:59PM. See Canvas for more accurate info.**
- Deadline for submission is **Sep 6, 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**.

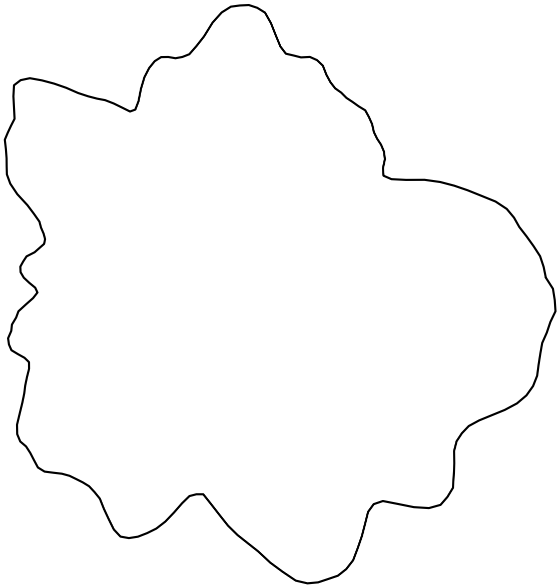
Final is a take-home project:

- 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.

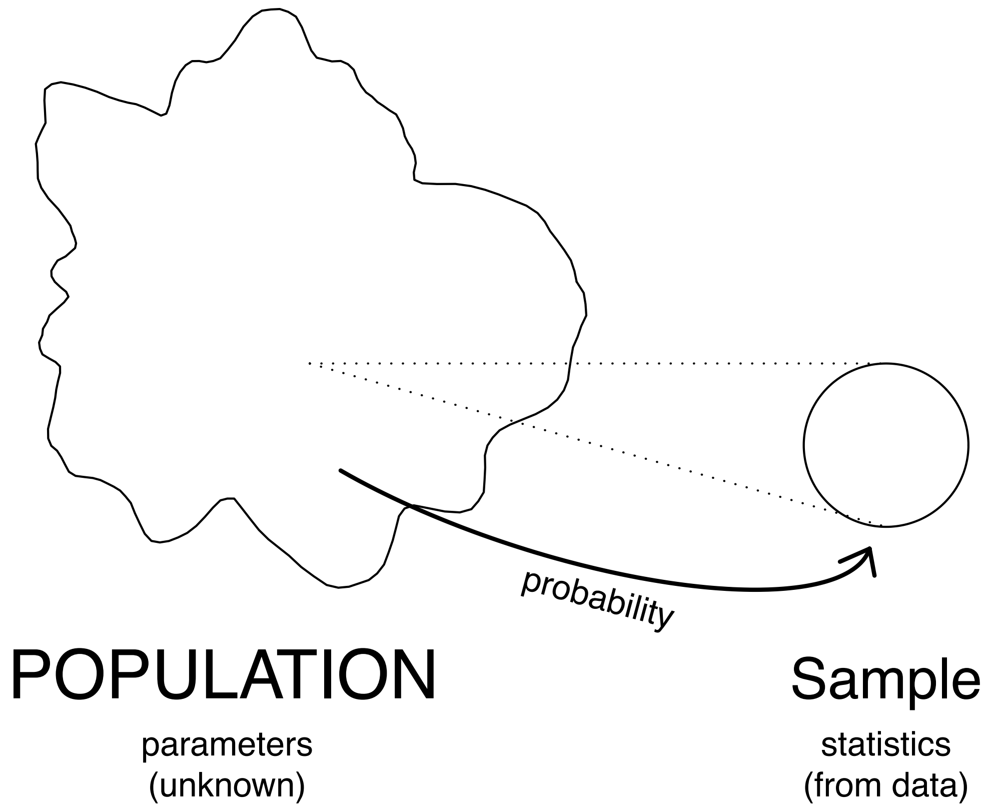
Probability vs. Statistics



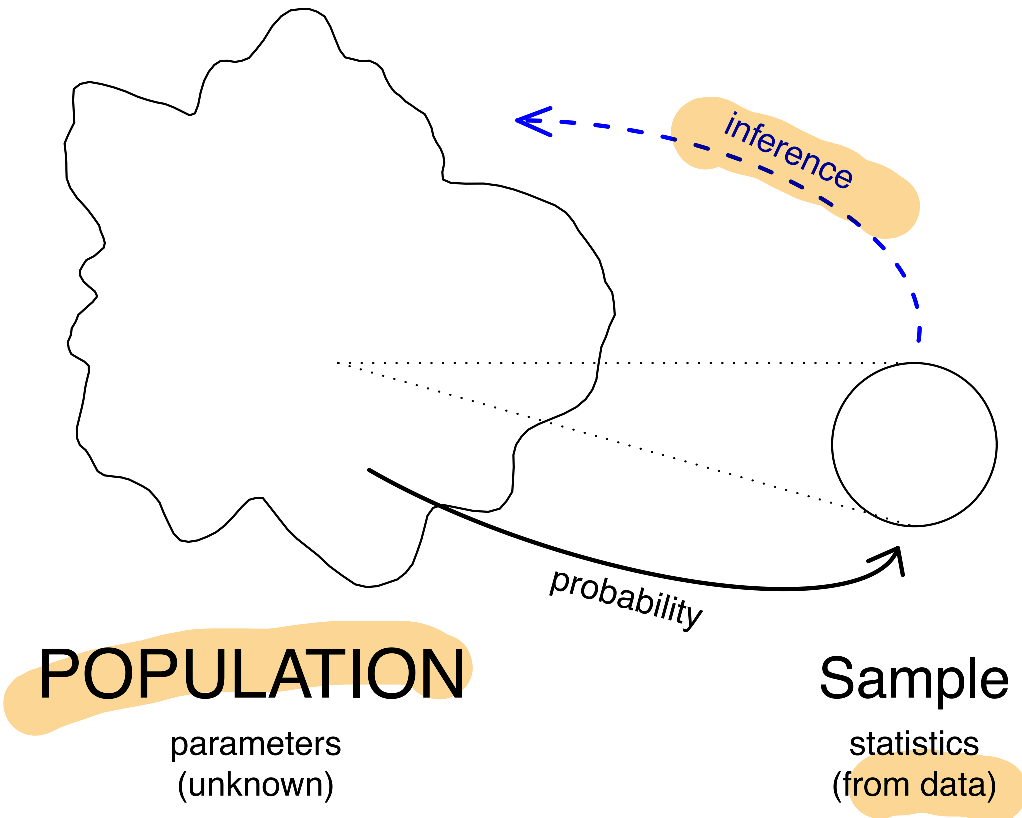
POPULATION

parameters
(unknown)

Probability vs. Statistics



Probability vs. Statistics

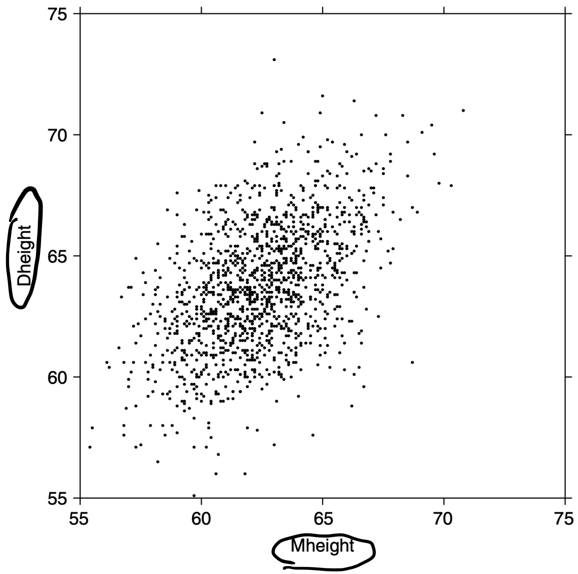


Regression Analysis

- 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).

Regression function / model

① The most common way of relating Y & X is through conditional mean

$$E(Y \mid X_1 = x_1, \cdots, X_p = x_p) = f(x_1, \cdots, x_p)$$

or $E(Y \mid \mathbf{X} = \mathbf{x}) = f(\mathbf{x})$ $\qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$

$\mathbf{X}$ ↓ vector

② The regression can be defined as a conditional quatile. Such as median

$$\text{Median}(Y \mid \mathbf{X} = \mathbf{x}) = f(\mathbf{x})$$

↓

other quantile α $\quad Q_\alpha(\cdot)$ $\quad$ α-level quantile

⇒ [Quantile regression]

Focus: conditional mean

$E(Y|X)$ can viewed in 2 ways

(1) A deterministic function of a realization $\mathbf{x}$ of random vector $\mathbf{X}$

$$E(Y \mid \mathbf{X} = \mathbf{x}) = \int y \times f(y \mid \mathbf{X} = \mathbf{x})\, dy$$

(2) A scalar random variable. A realization of $E(Y|X)$ by sampling X from its marginal distribution. Plugging realization of X into deterministic function in (1).

(Regression Aanalysis: (1) more common than (2))

Linear Model

special model

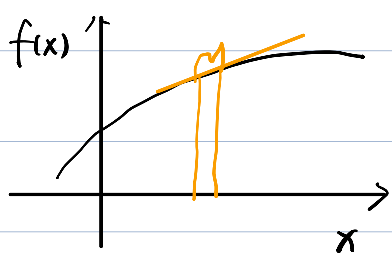
$$f(\mathbf{x}) = f(x_1 \cdots x_p) = \underline{\beta_0} + \beta_1 \underline{x_1} + \cdots + \beta_p x_p$$

linear in the parameters $(\beta_0, \beta_1 \cdots \beta_p)$

(NOT because a linear function of $\mathbf{x}$)

Linearity restriction: not as restrictive as one might think

① Many function can be approximately linear over a sufficiently small region



② Model may be made linear with transformations

Ex 1. Theory of gravitation $\Rightarrow F = \alpha / d^{\beta}$

F: force of gravity between two objects

α: constant related to masses of 2 objects

d: distance between objects

$$\log F = \log \alpha - \beta \log d \qquad (\beta, \text{ constant})$$

Generally, covariate X_i in the linear model can be functions of other variables

Ex 2. $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} z \\ z^2 \\ \vdots \\ z^p \end{pmatrix} \Rightarrow$ polynomial model

"linear" in parameters. not variables

Ex 3. "categorical" model

Dummy variable: takes 0 or 1 to indicate the presence or absence of certain categorical effect

Model mean hourly wage Y of married/non

$\mu_1 = E(U_1)$: mean of hourly wage of married

$\mu_2 = E(U_2)$: mean of hourly wage of non-married

Combine into one linear model, through dummy X

$$\begin{cases} X = 0 & \text{if } Y \text{ an observaation from married} \\ X = 1 & \text{otherwise} \end{cases}$$

$$E(Y \mid X = x) = \mu_1 + (\mu_2 - \mu_1) x$$

$$= \underbrace{\beta_0} + \underbrace{\beta_1 x}$$

Estimation / fitting of linear model

To estimate f, equivalently $\boldsymbol{\beta} = (\beta_0 \cdots \beta_p)^T$ in linear models

One approach least squares fitting / estimation (LS)

Minimize the loss function of $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_j - \beta_0 \right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\beta}^T \mathbf{x}_i \right)^2 \qquad \text{(squared error)}$$

Understand? Solve?

Simple Linear Regression (SLR)

[SLR 1] Model: intercept + $p=1$ covariate

$\Downarrow$

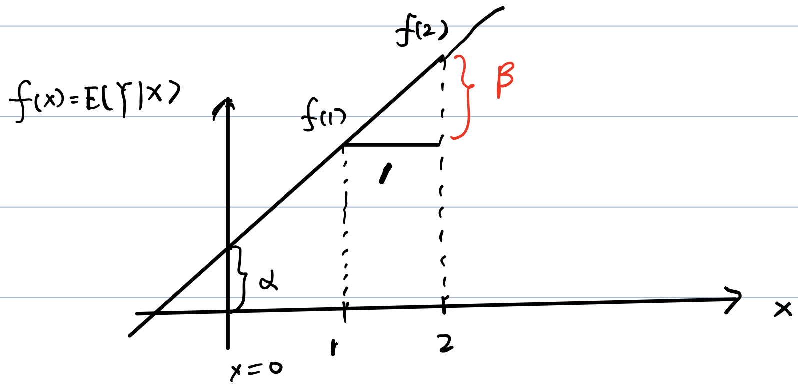
(can be viewed as a covariate whose value $=1$)

$$E(Y \mid X) = \alpha + \beta X$$

α: Intercept mean of Y if $x=0$

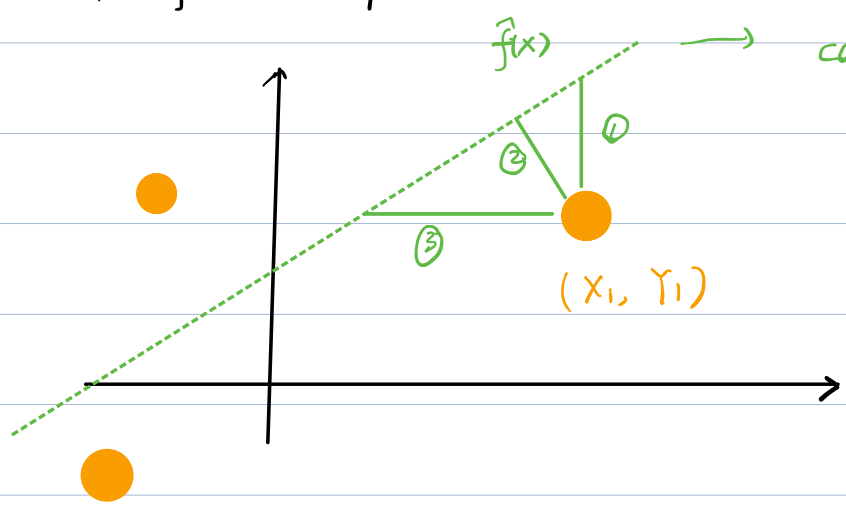
β: Slope change of mean of Y per unit change in X

α & β: Regression coefficients



[SLR 2] Estimation / fitting based on data

To fit in practice, observe data pairs (X_i, Y_i) $i=1 \cdots n$



Interprotations can be different

① Vertical $\quad Y_i - \hat{f}(X_i)$

Given x, the distance between observed Y & fitted values of Y

⇒ ordinay least square

② Perpendicular ⇒ Orthogonal regression (Deming regression)

(errors - in - variable), PCA

③ Horizontal $\quad X_i - \hat{f}^{-1}(Y_i)$

Ordinary LS is the most common

[SLR 3] Ordinary LS estimates in SLR

Loss function: $L(\alpha, \beta) = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$

Minimize $(\alpha, \beta) \in \mathbb{R}^2$

$L(x) \Rightarrow$ stationary, gradient $\Rightarrow$ (Quiz 0, 7)

Differentiate

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial \alpha} \\ \frac{\partial L}{\partial \beta} \end{pmatrix} = \begin{pmatrix} -2 \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i) \\ -2 \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i) X_i \end{pmatrix} \circledast$$

(Quiz 0, 6)

By setting $\circledast = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

$$\Leftrightarrow \begin{cases} n\hat{\alpha} + \hat{\beta} \times (n\bar{X}) - n\bar{Y} = 0 \\ n\bar{X} \times \hat{\alpha} + \sum_{i=1}^{n} X_i^2 \cdot \hat{\beta} - n \sum_{i=1}^{n} Y_i X_i = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \\ \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \hat{\beta} = \sum_{i=1}^{n} Y_i X_i - n\bar{X}\bar{Y} \end{cases}$$

Solution:

$$\begin{cases} \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \\ \hat{\beta} = \dfrac{\sum_{i=1}^{n} Y_i X_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \end{cases}$$

("hat" ^ : LS optimal values of α & β)