Statistical Modeling I

Introduction

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Introduction to Regression Analysis

Statistical Analysis

Data Analysis

Problem Formulation

1. Understand the physical background.
2. Understand the objective.
3. Learn what the *client* wants.
4. Set the problem in statistical terms.

Data Collection

- Observational vs. Experimental Studies.
- Is there a missing response?
- Are there missing values?

Exploratory Data Analysis

Before fitting any model, you should first

- compute the summary statistics for each variable under consideration.
- draw boxplots, histograms, density plots, etc. for each variable under consideration.
- draw scatter plots, interactive graphics, etc. for pairs or multiple variables.
- look for outliers, typing errors, skewed distributions (are the prior distributions as expected?)

Main Course Objective

Regression Analysis

It is a "tool" used to examine the relationship between

- a **Dependent Variable** or **Response** Y, and
- one (or more) **Independent Variables** or **Regressors** or **Predictors** $X_1, X_2, \ldots, X_p$.

Regression Analysis Objectives

- describe the relationship between the mean of the response and the predictors.
- predict the response using a function of the regressors.
- control.

Types of Regression

1. Simple Regression: $p = 1$
2. Multiple Regression: $p > 1$
3. Multivariate multiple regression: More than one response variable (not covered in this class)
4. Linear
5. Polynomial
6. Logistic
7. Lasso
8. Ridge
9. ....

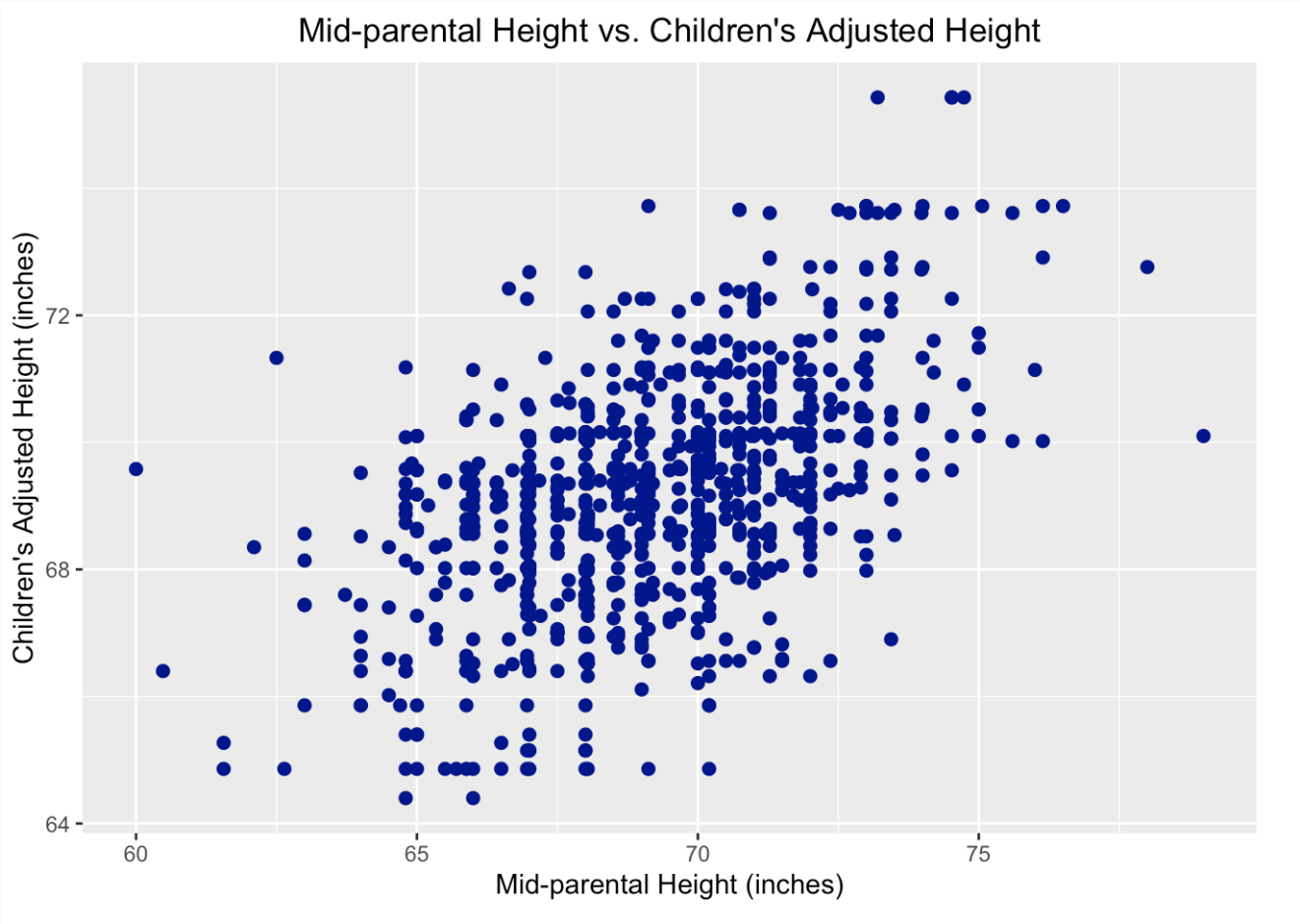
Francis Galton

- Francis Galton (1822–1911) was an English statistician, sociologist, psychologist, anthopologist among other things.
- He is responsible for many concepts and innovations in statistics, such as correlation, standard deviation, quartile, percentile, bivariate normal distribution and regression among other.
- ‘Regression Towards Mediocrity in Hereditary Stature.‘ published in ‘The Journal of the Anthropological Institute of Great Britain and Ireland‘ in 1886.
- He coined the term *regression to mediocrity* in 1875. This is where the term regression comes from.

Francis Galton's Regression to Mediocrity

- Scatter plot of child's height against a combined parents height:

$$MP = \frac{(\text{father's height} + 1.08\ \text{mother's height})}{2}$$



Francis Galton's Regression to Mediocrity

- Based on the plot, is the height of a child from tall (height above average) parents, also above the average?
  - The *solid* line on the plot is the regression line, while the *dotted* line is the line that corresponds to the correlation between the two heights.
  - The child's height is not going to be above average unless the correlation (dotted line) is close to 1.
- F. Galton: *Regression to mediocrity*, or *Regression to the mean*:

$$\underbrace{\frac{y-\bar{y}}{SD_y}}_{\text{“}\mathcal{Y}\text{”}} = r \cdot \underbrace{\frac{x-\bar{x}}{SD_x}}_{\text{“}\mathcal{X}\text{”}} \longrightarrow^{\textit{symbolically}} \text{“}\mathcal{Y} = r \cdot \mathcal{X}\text{”}$$

where r is the correlation between x and y.