Chapter 1: Introduction

Chong Zhang
Email: chong.zhang@uwaterloo.ca
Office: M3 4114

Fall, 2014

Chapter 1

Introduction

What is regression?

- Regression: one of the most important topics in modern applied statistics
- Models the functional relationship between y and $x_1, x_2, \ldots, x_p$
  - y: a response/dependent variable
  - $x_1, x_2, \ldots, x_p$: predictors/independent variables/explanatory variables
- A typical regression model is:

$$y = f(x_1, x_2, \ldots, x_p) + \epsilon$$

  - ϵ: random error term

Applications

- Applications: social science, business, engineering, etc.

Application	y	x's
Finance	Stock Price	Unemployment Rate Consumer Price Index Money Supply
Marketing	Sales	Advertising Expenditures
Manufacturing	Hardness	Temperature

Regression and Causality

- The existence of a statistical relation between y and x does not imply that y depends causally on x.

  "Correlation does not imply causation."

- For example,
  - readings of the thermometer (x) and temperatures (y)
  - aptitude test scores (x) and performance of an employee (y)

  Although a strong statistical relationship exists, the causal condition actually acts in the opposite direction, from y to x.

Why Regression?

Regression models can be used to:

- Identify important predictors
- Estimate the response for given values of predictors
- Predict future values of response

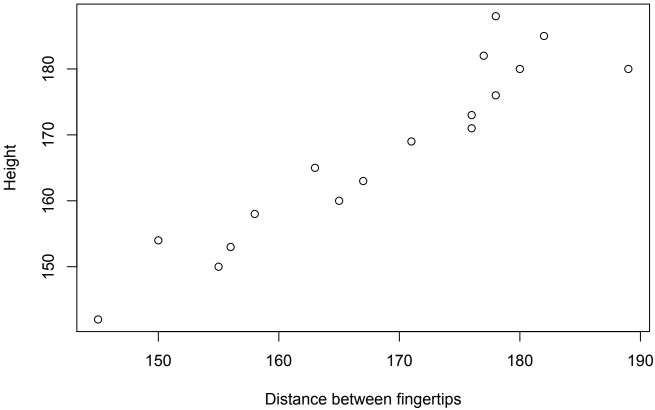
An Example

- Relationship between the height (y) and the distance between fingertips (DF, x) for all students at U Waterloo.
- We take a simple random sample with $n = 18$ (sample size):

DF (cm)	156	176	167	155	180	178	145	177	189
Height (cm)	153	171	163	150	180	188	142	182	180
DF (cm)	165	176	178	182	158	163	171	150	188
Height (cm)	160	173	176	185	158	165	169	154	186

- **Question:** Why random sample?

Scatter Plot



Simple Linear Regression

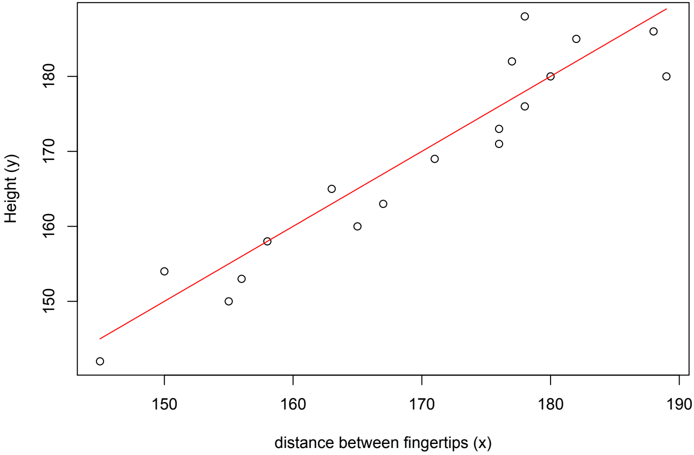
- STAT 231: a simplest form of regression
  - only one predictor x
  - $f(x)$ is a linear function

$$y = f(x) + \epsilon = \beta_0 + \beta_1 x + \epsilon$$

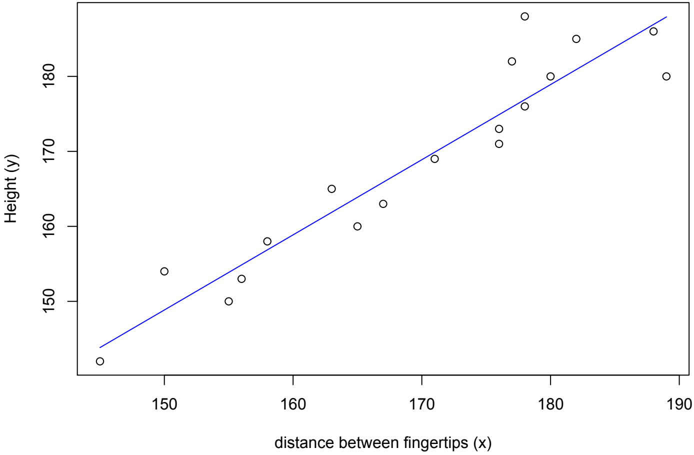
Note: we refer to this model as linear in the parameters β's ($\frac{\partial f}{\partial \beta_i}$do not depend on the parameters).

Are the following models linear?

1. $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
2. $f(x) = \beta_0 + \beta_1 e^{\beta_2 x}$



Red line represents the true underlying relationship between x and y:
$f(x) = \beta_0 + \beta_1 x$



Blue line is the fitted regression line: $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{f}$: "f–hat"; ^ stands for fitted/estimated value

Outline of STAT 331

- Ch2. Review of Simple Linear Regression
- Ch3. Random Vectors and Matrix Algebra
- Ch4. Multiple Linear Regression
- Ch5. Model Evaluation-Residual Analysis
- Ch6. Variance-Stabilizing Transformations
- Ch7. Model Evaluation-Outliers and Influential Cases
- Ch8. Model Building and Selection
- Ch9. Binary Outcome: Logistic Regression

Course Objectives

By the conclusion of this course, students should have achieved the following objectives:

- Understand how to use a broad range of regressions techniques and their limitations,
- Be able to understand and use the software R for regression problems,
- Be able to undertake statistical inference in a regression context and be able to interpret the results in the context of the problem,
- Understand the theoretical foundations of regression and be able to derive the fundamental mathematical results,
- Be able to give a critical review of an applied regression analysis, interpreting the results and be able to describe its strengths and weakness.

R

- In this course, R will be the computer language used for implementation
- In exams, you will be expected to interpret R output
- Some useful information about R on the department webpages: `math.uwaterloo.ca/sas/research/resources/essential-software-statistics`
- Questions about a specific function in R? Google/wiki the answer first!

Review of Normal (Gaussian) Distribution

- Commonly used distribution
- Bell shaped
- Symmetric with respect to the central value
- and so on...
- $X \sim N(\mu, \sigma^2)$:
  - $E(X) = \mu$
  - $Var(X) = \sigma^2$