Regression Analysis
What and Why ?

Buddhananda Banerjee

Department of Mathematics
Centre for Excellence in Artificial Intelligence
Indian Institute of Technology Kharagpur

bbanerjee@maths.iitkgp.ac.in

Example: Hight-Weight chart

What is Considered the Right Weight for My Height?

**The table below has been updated to show both Metric and Imperial measurements i.e. Inches/Centimeters - Pounds/Kilograms.*

Adults Weight to Height Ratio Chart

Height	Female	Male
4' 6" (137 cm)	63/77 lb (28.5/34.9 kg)	63/77 lb (28.5/34.9 kg)
4' 7" (140 cm)	68/83 lb (30.8/37.6 kg)	68/84 lb (30.8/38.1 kg)
4' 8" (142 cm)	72/88 lb (32.6/39.9 kg)	74/90 lb (33.5/40.8 kg)
4' 9" (145 cm)	77/94 lb (34.9/42.6 kg)	79/97 lb (35.8/43.9 kg)
4' 10" (147 cm)	81/99 lb (36.4/44.9 kg)	85/103 lb (38.5/46.7 kg)
4' 11" (150 cm)	86/105 lb (39/47.6 kg)	90/110 lb (40.8/49.9 kg)
5' 0" (152 cm)	90/110 lb (40.8/49.9 kg)	95/117 lb (43.1/53 kg)
5' 1" (155 cm)	95/116 lb (43.1/52.6 kg)	101/123 lb (45.8/55.8 kg)
5' 2" (157 cm)	99/121 lb (44.9/54.9 kg)	106/130 lb (48.1/58.9 kg)
5' 3" (160 cm)	104/127 lb (47.2/57.6 kg)	112/136 lb (50.8/61.6 kg)
5' 4" (163 cm)	108/132 lb (49/59.9 kg)	117/143 lb (53/64.8 kg)

Example: Hight-Weight chart

Adult Male and Female Height to Weight Ratio Chart [1]

Author: Disabled World : Contact: www.disabled-world.com

Published: 2017-11-30 : (Rev. 2020-03-05)

[1] Ref: https://www.disabled-world.com/calculators-charts/height-weight.php

Weight-Hight regression

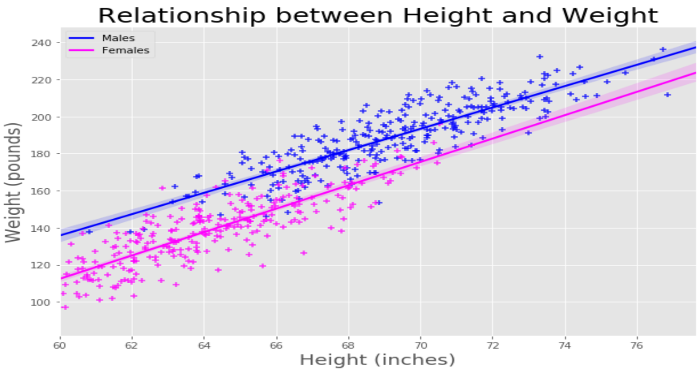


Figure: Weight vs Hight

Example: Obesity

- Worldwide, at least 2.8 million people die each year as a result of being overweight or obese, and an estimated 35.8 million (2.3%) of global DALYs are caused by overweight or obesity. [2]

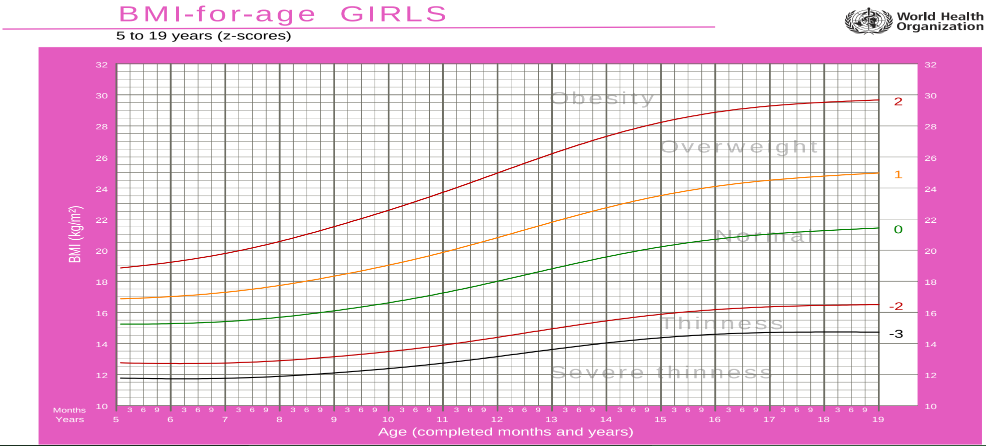
- What are obesity and overweight ?
  Overweight and obesity are defined as abnormal or excessive fat accumulation that may impair health.

- For adults, WHO defines overweight and obesity as follows:
  - overweight is a BMI greater than or equal to 25; and
  - obesity is a BMI greater than or equal to 30.

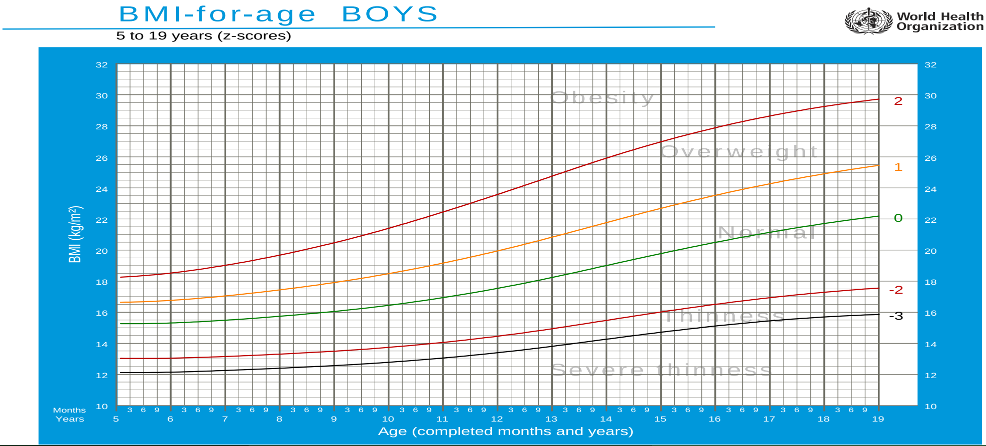
- Body mass index (BMI) is a simple index of weight-for-height that is commonly used to classify overweight and obesity in adults. It is defined as a person’s weight in kilograms divided by the square of his height in meters (kg/m^2).

[2] Ref: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Example: Obesity chart for girls (5-19yr)



Example: Obesity chart for boys (5-19yrs)



What is the value of ‘g’ ?

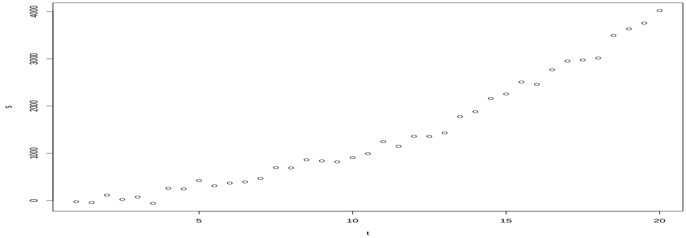


Figure: Free fall

$$S = ut + \frac{1}{2}gt^2$$

Two treatment comparison

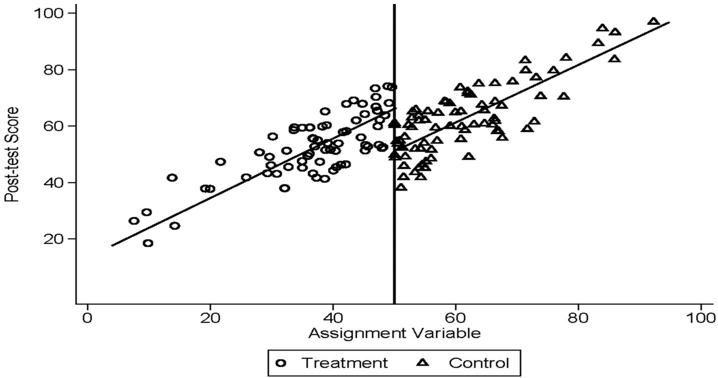


Figure: Linear Treatment effect model

Why regression?

- Regression is a very natural attempt to answer many queries that come in human mind and scientistic work.
- The information we gather about a natural phenomena or a controlled experiment are often incomplete.
- Regression is one of the ways to make these information complete based on the available data.
- In other words, it an attempt to access beyond than that has been already observed.

What is regression?

Definition

Let $(Y, \mathbf{X})$ be a random vector. The conditional expectation of Y given $\mathbf{X} = \mathbf{x}$, is known as the regression of Y on $\mathbf{X}$. It can be denoted as

$$\hat{y} = g(\mathbf{x}, \beta) = E(Y|\mathbf{X} = \mathbf{x})$$

- $g(\mathbf{x}, \beta)$ can be a line, curve, plane, surface etc. or may be unknown
- $\mathbf{x}$ can be stochastic or non-stochastic
- Y is always stochastic or a random viable

Linear and non-linear regression

Consider a data set $D = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^k, y_i \in \mathbb{R}, \ \forall i = 1, 2, \cdots, n\}$ where x_is are non-stochastic but y_i are stochastic and realized values of random variable Y_is respectively.

Definition

If the relation, $g(\mathbf{x}, \beta)$,between the **response variable** y and the **regressor variable** $\mathbf{x}$ is linear in parameter (β) then it is called a **linear regression.**

e.g. Linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$y = \beta_0 + \beta_1 e^x + \epsilon$$

e.g. Non-Linear regression:

$$y = \frac{1}{\beta_0 + \beta_1 x} + \epsilon$$
$$y = \beta_0 \cos(\beta_1 + \beta_2 x) + \epsilon$$

where, ϵ is random error.

Some varieties of regression

- Linear Model (LM):
  - Simple linear regression
  - Multiple linear regression
  - Polynomial regression
- Generalized linear model (GLM)
  - Logit-modle
  - Probit-model
  - Poisson-regression
- Isotonic regression
- Spline regression etc.

Simple linear regression

- Consider a data set $D = \{(x_i, y_i) | x_i \in \mathbb{R}, y_i \in \mathbb{R}, \ \forall i = 1, 2, \cdots, n\}$
- x_is are non stochastic
- y_is are stochastic and realized values of random variable Y_is

Problem statement

We are interested to have a prediction line

$$\hat{y} = g(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x$$

which will approximate well the y values if the x values are known.

Simple linear regression : Example

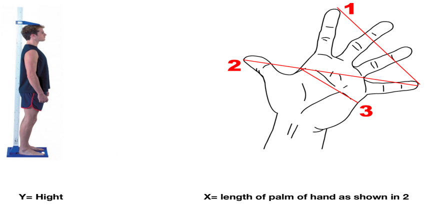


Figure: Palm length vs Hight

Can we have a prediction line $\hat{y} = \beta_0 + \beta_1 x$ which will approximate well the hight of a person if his/her palm length(2) is known ?

Least square estimate

- What do we mean by “approximate well"?

ANS: Minimum distance between the predicted ($\hat{y}_i s$) and the true ($y_i s$) values of Y.

- What will the notion of distance ?

ANS: There could be many. But we will consider either absolute or square/ Euclidean distance.

- Given the data how can we obtain the values of β_0 and β_1?

ANS: We will consider such values of of β_0 andβ_1 that will minimize the square/ Euclidean distance between the predicted ($\hat{y}_i s$) and the true ($y_i s$) values of Y.

This is known as the least squared method of estimation

Least square estimate

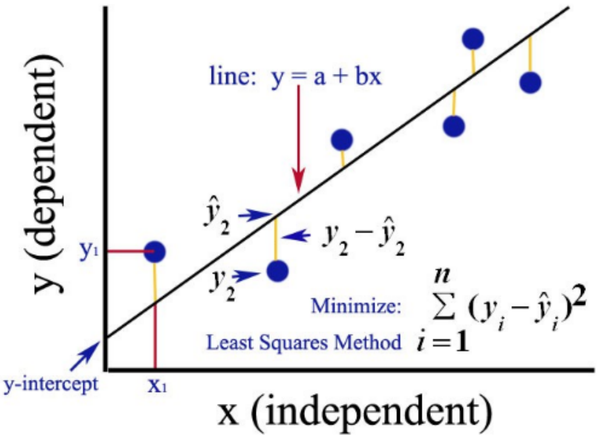


Figure: Least square

Least Squared condition

The least squared condition to estimate the model parameters is to minimize

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2. \tag{1}$$

with respect to β_0 andβ_1.

If $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes $S(\beta_0, \beta_1)$ then their values can be obtained by solving the **normal equations**

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \quad \Longrightarrow \quad n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i = \sum_i y_i \tag{2}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \quad \Longrightarrow \quad \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i \tag{3}$$

Prediction

Estimated parameters

Defining $S_{xy} = \sum_i (y_i - \bar{y})(x_i - \bar{x})$ we have the solutions as

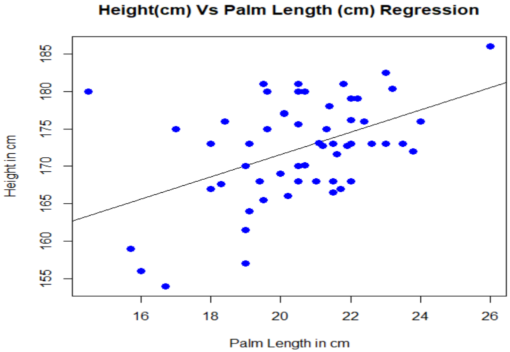
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least squared prediction line

For any x such as old x_is or some x_{new} the prediction line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Prediction



$\hat{\beta}_0 = \mathbf{141.8916}$ $\hat{\beta}_1 = \mathbf{1.4833}$

What more?

- Can we have a prediction interval for $\hat{y}$?
- Can we test for $H_0 : \beta_0 = 140$ vs $H_1 : \beta_0 > 140$?
- Can we test for $H_0 : \beta_1 = 1.5$ vs $H_1 : \beta_1 \neq 1.5$?
- What will be the distribution of estimated error ?
- If we incorporate more regressor variables then how significantly the error can be reduced ?

We are not yet ready to answer these important questions!!!!
Results from Linear Algebra and Multivariate Analysis can help us.


THANK
YOU