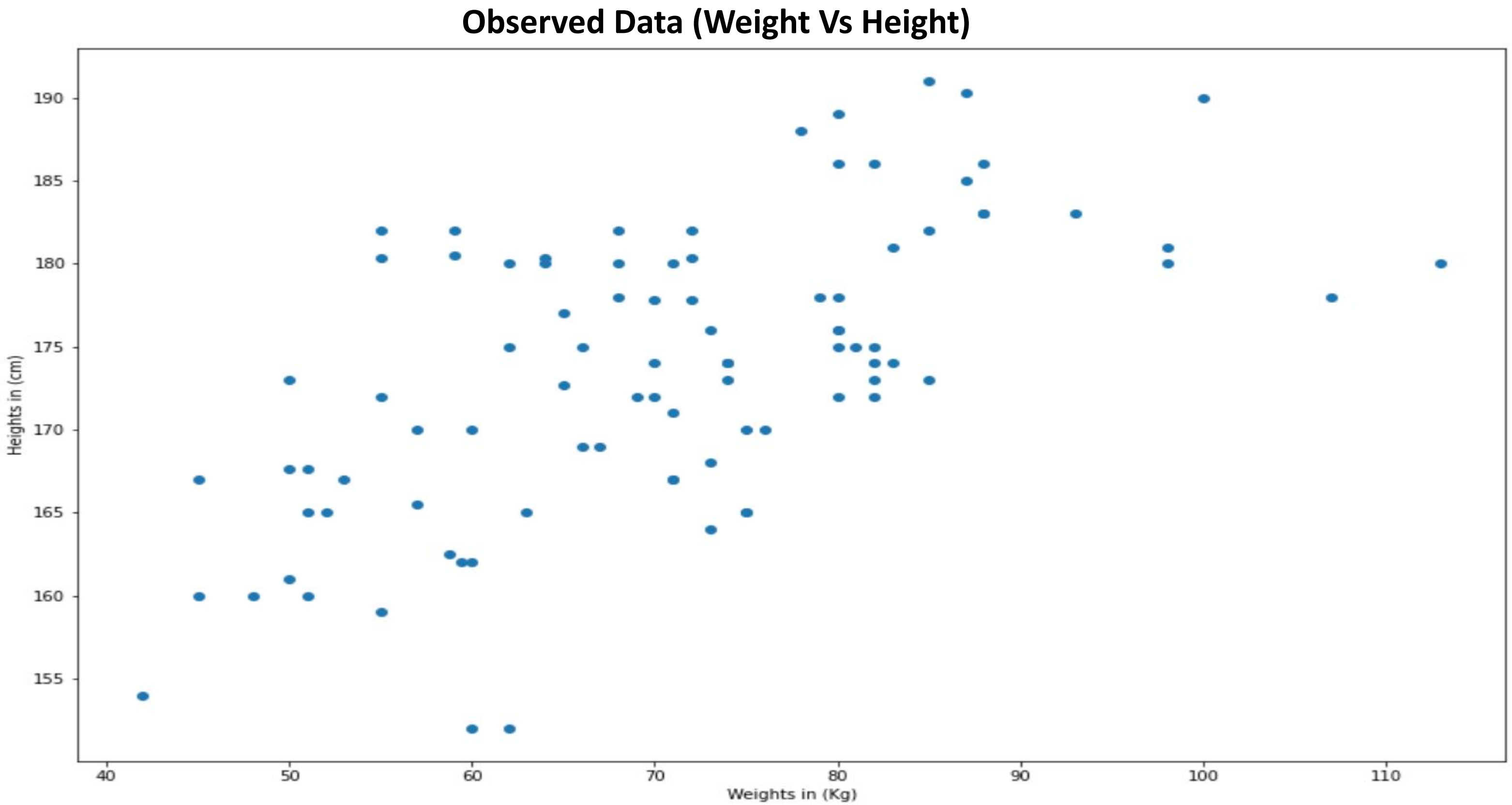
Linear Models for Regression

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

The Problem



Observed Data (Weight Vs Height)
Heights in (cm)
Weights in (Kg)
190
185
180
175
170
165
160
155
40
50
60
70
80
90
100
110

Motivation

- Height and weight are random.

- Can we accurately predict what will be someone's height given her weight?
  - Difficult to estimate from "a priori" models
  - But, we have lots of data from which to build a model

- Assumptions:
  - Linear correspondence (relationship) exists between height and weight.
    - Usually true
  - Observation noise follows a normal distribution.
    - Also usually true!

Formal Problem Settings

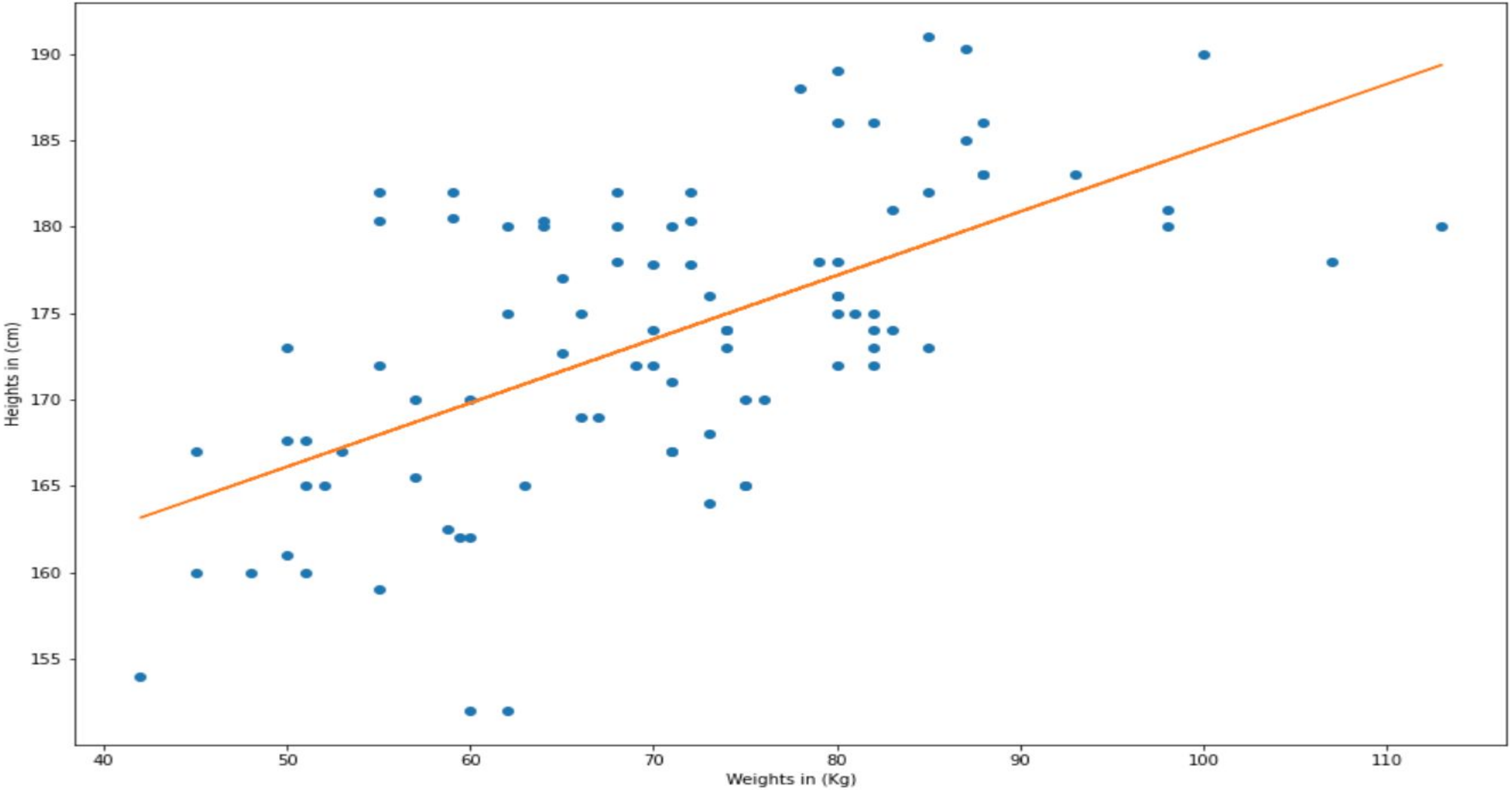
- **Input: D** = $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
- $\mathbf{x_i} \in \mathbf{X}, \mathbf{y_i} \in \mathbf{Y}$ where X, Y ε R [real number]

- Hypothesis $y_i = w^T x_i + \varepsilon$
  - Linear Correspondence -> $w^T x_i$
  - Normal Distribution of noise -> $\varepsilon \sim N(0, \sigma^2)$

- $y_i \sim N(w^T x_i, \sigma^2)$

$$P(y_i|\vec{x}_i; w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{w^T x_i - yi}{2\sigma}\right)^2}$$

A simple model



How to get W?

- **Maximum Likelihood Estimation (MLE)** gives us the solution which maximises the likelihood.
  - Find w that maximizes the probability of the data D
    - $argmax\ P(D|w)$

- **Maximum A Posterior (MAP)** gives us the solution which maximises the posterior probability.
  - Find w that is most likely given the data D.
    - $P(w|D) = P(D|w) * P(w)/P(D)$
    - Assumes the availability of the prior $P(w) \sim N(0, \sigma_0^2)$
      - E.g. in case of transfer learning as initial weights w_0

How to get W: Maximum Likelihood Estimation (MLE)

- MLE gives us the solution which maximises the Likelihood

$$\underset{w}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i|\vec{x}_i; w) = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} log P(y_i|\vec{x}_i; w)$$

- Both will provide the same maxima. Now putting the value of the normal distribution probability:

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(w^T x_i - y_i)^2$$

- The log term is independent of w, *hence we can get rid of it*

Finding model parameters [Optimization]

- Simplifying it further:

$$= \underset{w}{\operatorname{argmax}} -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$$

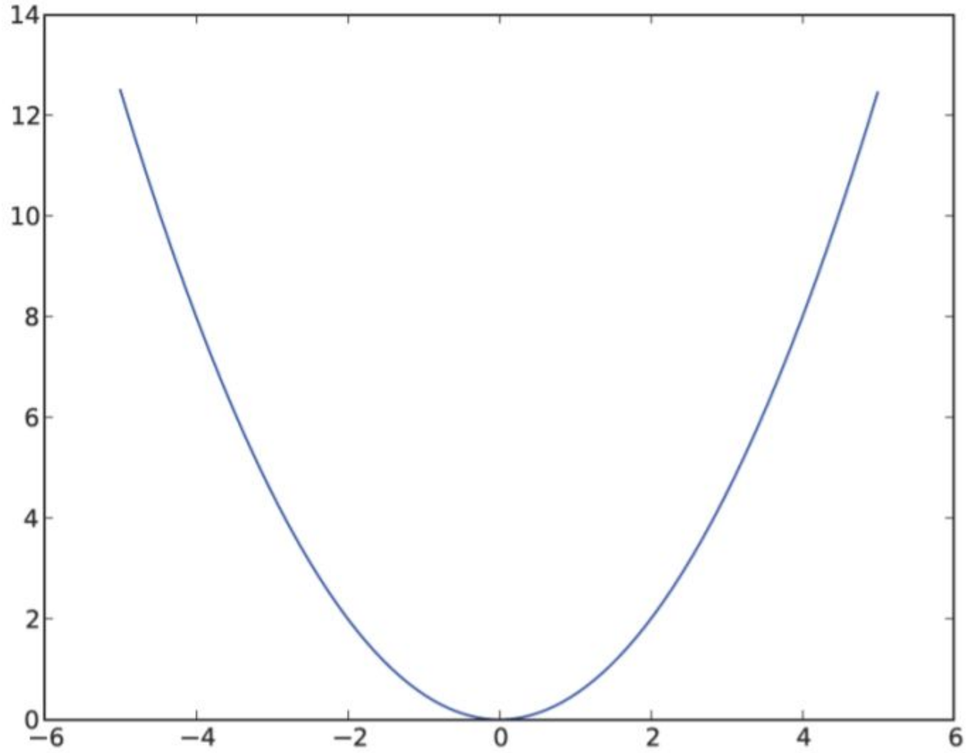
- $1/2\sigma^2$ is again a constant. Further, negative of maximization is same as minimization, hence:

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n}(w^T x_i - y_i)^2$$

- This is the L_2 loss. To add interpretability to the above, we need to take an average:

$$J(w) = \underset{w}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(w^T x_i - y_i)^2$$

Loss functions: Squared Loss (L_2 Loss)



$$l(\hat{y}, y_i) = (\hat{y} - y_i)^2 = (w^T x_i - y_i)^2$$

Analytical Solution

- In the multivariate case:

$$W = (X'X)^{-1}X'Y$$

- Exercise 1: Find the the solution for unidimensional case:

Hint:

$$J(w) = \sum_{i=1}^{n} (w_0 + w_1 x_i - y_i)^2$$

$$\frac{\partial}{\partial w_0} = ?$$

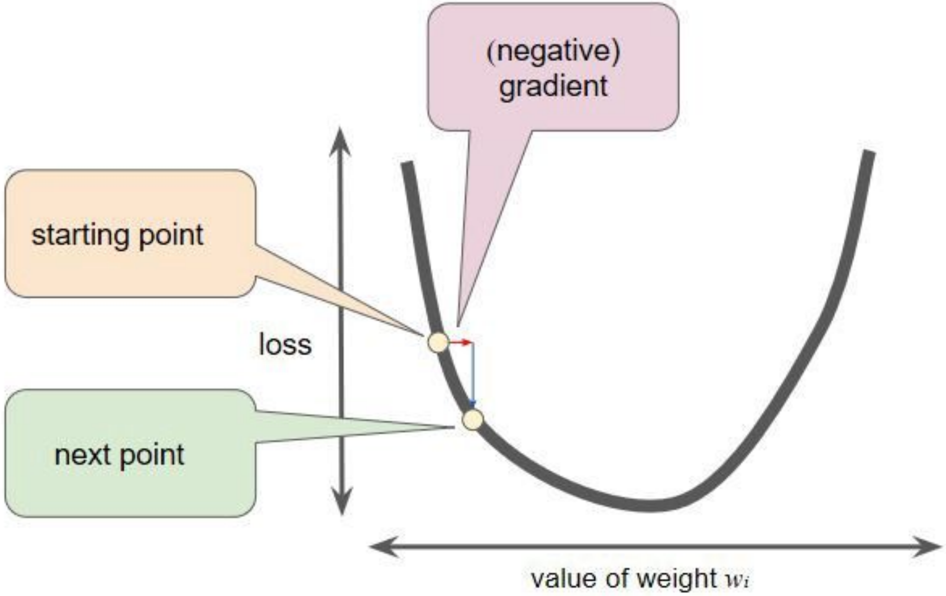
$$\frac{\partial}{\partial w_1} = ?$$

- For a multivariate case, the computational complexity is very high!
  - Hence, an iterative solution such as gradient descent is preferred.

Gradient descent

$$w_i = w_i - \eta \frac{\partial J(w)}{\partial w_i}$$

- Repeat until "convergence"



Gradient descent: Demo

- https://lukaszkujawa.github.io/gradient-descent.html

How to get W: Maximum A Posterior Estimation (MAP)

- Study the Maximum A Posteriori (MAP) solution for Linear Regression