Linear Regression

Anubha Gupta, PhD.
Professor
SBILab, Dept. of ECE,
IIIT-Delhi, India
Contact: anubha@iiitd.ac.in; Lab: http://sbilab.iiitd.edu.in

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Linear Regression

Motivation

House price prediction

- You visit a recently constructed township in your city and wish to guess the price of a house in the new township. Can you do that?
- Answer: Yes, using some past statistics and your ML-skills
- You know that the price of a house is directly proportional to:
  - the size of the house
  - the distance from the market

  and inversely proportional to
  - the reported crime cases in the area

- These are some intuitive (qualitative) assumptions based on past experiences
- To validate these relationships, we need a data-driven quantitative approach
  - A regression ML model does exactly that
  - It will gain experience from the data
  - It helps the ML model learn these relationships
  - Further, it will predict the house price based on these relationships.

Learning Objectives

In this module, we will learn the concepts of

- Simple linear regression
- Multiple linear regression

- Solving Regression problem
  - Using data driven approach
  - Using Statistical approach

Definition

- A mathematical modeling to learn the relationship between one or more dependent variables (output) and one or more independent variables (features)

- Uses:
  - **Predictive Modeling**: To anticipate future outcomes
  - **Relationship Analysis**: To analyze relationships between variables

Simple Linear Regression

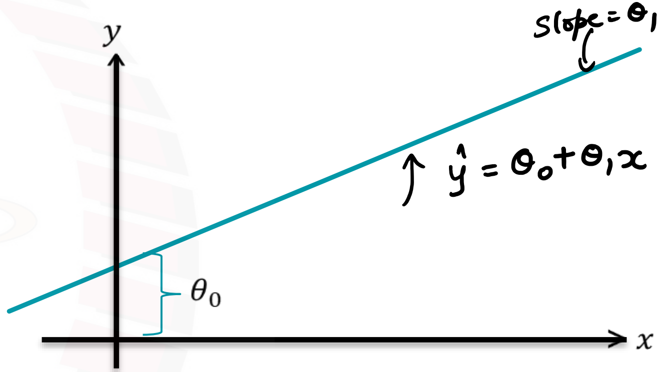
- Hypothesis: The output variable is a linear function of input variables

parameters

$$y = \theta_0 + \theta_1 x + \varepsilon$$

error

observation feature



where,

- y : dependent variable or response variable or observations
- x : independent variable or explanatory variable or covariates or features
- ε : measurement error or model error or the residual

- θ_0: Bias parameter/Intercept
- θ_1: Weight parameter/ Slope of the line

} Parameters (to be learned)

- Simple linear regression:

  One independent variable and one output variable

Simple Linear Regression

- Matrix Notation: For a pair of feature and the response (output) variable

$$
\begin{aligned}
y &= \underbrace{[1 \quad x]}\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix} + \varepsilon \\
&= \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} + \varepsilon \\
&= \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} + \varepsilon \\
&= \hat{y} + \varepsilon
\end{aligned}
$$

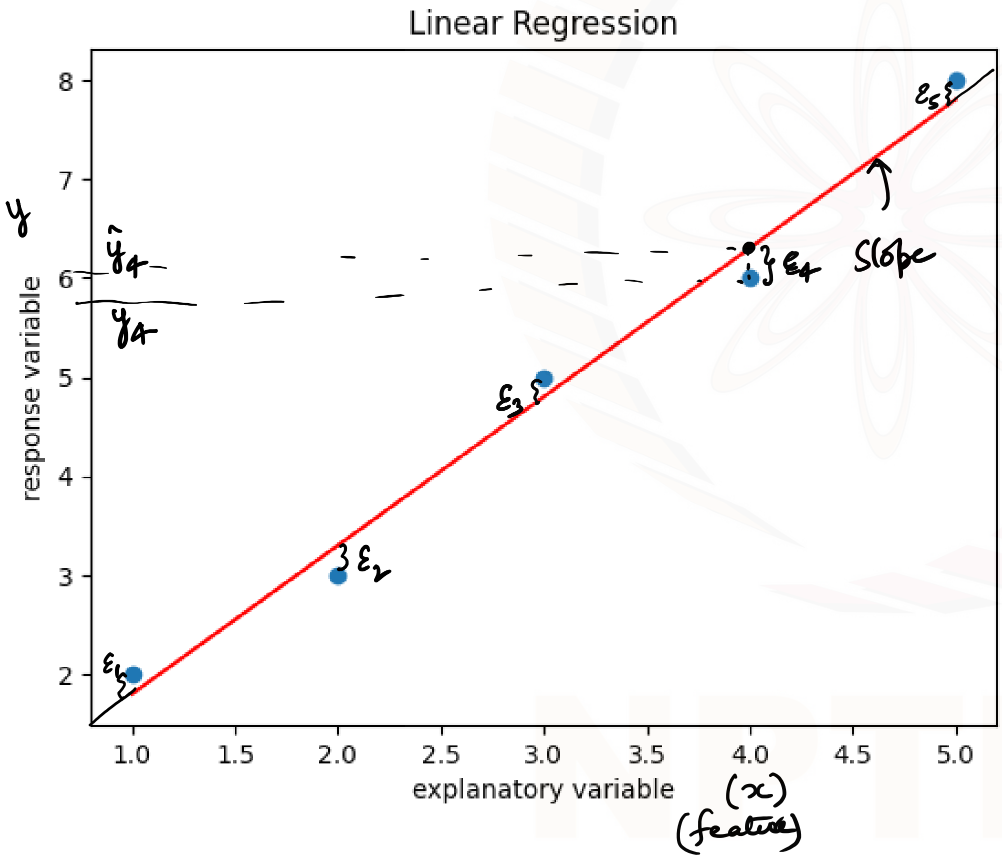
$[1 \quad x] \equiv \underline{x}^{T}$

$\begin{bmatrix}1\\ x\end{bmatrix} \equiv \underline{x}$

$\hat{y} = \theta_0 + \theta_1 x$

- Error $\varepsilon = y - \hat{y}$

Example



Explanatory variable (x)	Response variable (y)
1 x_1	2 y_1
2 x_2	3 y_2
3 x_3	5 y_3
4 x_4	6 y_3
5 x_5	8 y_4

$$y-\hat{y}=\varepsilon$$

$$\hat{y}_1=\theta_0+\theta_1 x_1$$

$$\hat{y}_2=\theta_0+\theta_1 x_2$$

Multiple Linear Regression

- A linear regression technique that uses several (more than two) explanatory variables to predict the outcome of a response variable.

$n \equiv$ no. of observation

- **System of linear equations**

$$\begin{cases} y_1 = \theta_0 + \theta_1 x_{1,1} + \theta_2 x_{1,2} + \cdots + \theta_{p-1} x_{1,p-1} + \varepsilon_1 \\ y_2 = \theta_0 + \theta_1 x_{2,1} + \theta_2 x_{2,2} + \cdots + \theta_{p-1} x_{2,p-1} + \varepsilon_2 \\ \vdots \\ y_n = \theta_0 + \theta_1 x_{n,1} + \theta_2 x_{n,2} + \cdots + \theta_{p-1} x_{n,p-1} + \varepsilon_n \end{cases}$$

$y_1, \underline{x}_1$

$y_2, \underline{x}_2$

- **Matrix Notation**

$$[\mathbf{y}]_{n\text{x}1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,(p-1)} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,(p-1)} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,(p-1)} \end{bmatrix}_{n\times p} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{p-1} \end{bmatrix}_{p\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Multiple Linear Regression

- In general, for a pair of feature vector and the response (output) variable

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_{p-1} x_{p-1} + \varepsilon$$

$$y = \begin{bmatrix} 1 & x_1 & x_2 & x_3 & \dots & x_{p-1} \end{bmatrix}_{1\times p} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{p-1} \end{bmatrix}_{p\times 1} + \varepsilon$$

$\langle \underline{x}, \underline{\theta} \rangle$

$$= \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} + \varepsilon \quad = \langle \underline{\theta}, \underline{x} \rangle$$
$$= \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} + \varepsilon$$
$$= \hat{y} + \varepsilon$$

$$\boxed{y = \underline{x}^T \underline{\theta} + \varepsilon}$$

$\mathcal{D}$: $y_1, \underline{x}_1^{\theta} = [\;]$

$y_2, \underline{x}_2 = [\;]$

- Error $\varepsilon = y - \hat{y}$

Linear Regression

- Train data: Seen known observations on which the model is required to be trained/built

$$\mathcal{D}_{train} = \begin{Bmatrix} (\mathbf{x}_1, y_1) \\ (\mathbf{x}_2, y_2) \\ \vdots \\ (\mathbf{x}_n, y_n) \end{Bmatrix}$$

- Test data [Hold-out set]: Unseen, but known observations on which the trained model is required to be tested

Ground Truth ≡ GT

$$\mathcal{D}_{test} = \begin{Bmatrix} (\mathbf{x}^t{}_1, y^t{}_1) \\ (\mathbf{x}^t{}_2, y^t{}_2) \\ \vdots \\ (\mathbf{x}^t{}_q, y^t{}_q) \end{Bmatrix}$$

- Generally, $n \gg q$ if the dataset is small

Seen/unseen: by the model
Known/unknown: by the model developer

Linear Regression

- Model is trained (fitted) on $\mathcal{D}_{train}$
- Predictions are taken on each $\mathbf{x}^{t}{}_{k}$ test sample of $\mathcal{D}_{test}$, producing $\widehat{y_k}$, for k=1,…, q, where q are the number of test set samples. performance metric
- $\widehat{y_k}$'s are compared with the true labels y_k's of $\mathcal{D}_{test}$ using an appropriate error function

Assumptions: Linear Regression

1. Parameter θ_i's are constant and the regression equation is linear in parameters (θ_i's)
2. Errors ε_i are Gaussian i.i.d. random variables

$$\varepsilon_i \approx \mathcal{N}(0, \sigma^2)$$

$$\boxed{y_1 = \theta_0 + \theta_1 x_1 + \epsilon_1}$$
$$y_2 = \theta_0 + \theta_1 x_2 + \epsilon_2$$

ϵ_1 s.t. ϵ_2
$\epsilon_1 \sim N(0, \sigma^2)$
$\epsilon_2 \sim N(0, \sigma^2)$

- **Homoscedasticity**: when the variance of random variables is same, we say that they exhibit homoscedasticity. In linear regression, it is exhibited by the errors/residuals (ε_i's).
- **Independence**: Observations are independent of each other
- **Normality**: For any fixed feature vector $\mathbf{x}$, y is normally distributed

3. The independent variables (features) should be uncorrelated to the residual error (ε_i's).

4. There is no multicollinearity in the independent variables, i.e., the features are not highly correlated with each other. If so, dimensionality reduction via PCA or feature selection helps get better results.

x_1
x_2

$x_1 = \alpha x_2$

In general, the number of features < number of training samples

Approaches for Error Minimization

1. Data-driven Approach

- Least Squares Solution
- Objective: $\min e = \min \sum_{i=1}^{n} \varepsilon_i^{\,2}$

where, ε_i: Error = $y_i - \hat{y}_i$

$$e = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = (\mathbf{y} - \hat{\mathbf{y}})^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{y}})$$

$$= \sum_{i=1}^{n} \varepsilon_i^{\,2} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

- This is the T**otal Squared Error**.

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

$\theta_0 + \theta_1 x_1$ ✓

$$\underline{\varepsilon} = \begin{bmatrix} \underline{y} \end{bmatrix} - \begin{bmatrix} \hat{\underline{y}} \end{bmatrix}$$

$$\underline{\varepsilon}^{T} \underline{\varepsilon} = \langle \underline{\varepsilon}, \underline{\varepsilon} \rangle$$

2. Statistical Approach ✓

- Maximum Likelihood Estimation (MLE)
- Objective: $\min \mathbb{E}[e]$

  where $\mathbb{E}$: Mean or the expectation operator