Contents

Regression

```
a statisical tool for the investigation of relationships between variables.
```

- finding the relationship between dependent and independent variables

Regression Analysis

```
a form of predictive modeling which investigates the relationship etween a
dependent (target) and independent variable(s). independent variables are also
known as predictors.
```

- indicates the strength of impact of multiple independent variables on a dependent variables

Simple Linear Regression

- measures how strong the relationship is between 2 variables
- uses a straight line (of best fit) to model the relationship

Nonlinear regression models

- representing the relationship with a curved line
  Example:

$$Y = a + bx + cx^2$$

Examples of Linear Regression

- predicting house prices
  - given some dataset of bedroom numbers, square footage, location, etc.

- stock price prediction
  - using historical stock data, predict the future price of a stock
- employee performance prediction
  - given data on employee characteristics, predict their future job performance or productivity

Linear Regression Model

$$Y = a + bX$$

- X = independent variablle
- Y = dependent variable

where...

- a - constant which equals the value of Y when X=0
- X - independent variable that is predicting Y
- Y - value of dependent variable that is being predicted
- b - slope of regression line u.e. how much y changes for each one unit change in X

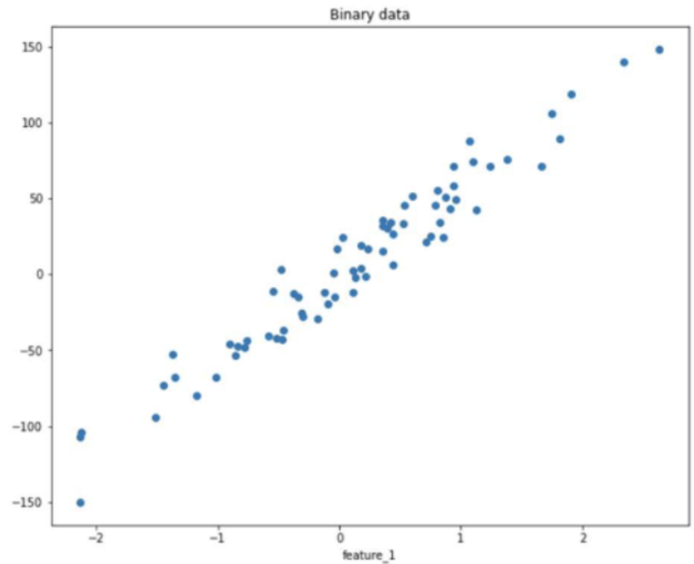
Line of Best Fit

```
a straight line that best represents the data on a scatter plot
```

- best linear approximation for some set of data

How to obtain best fit line (Value of a and b)?

- Ordinary Least Square
  - Find a model whose line fits best with respect to the given data



https://nandeshwar.in/100-days-of-deep-learning/what-is-linear-regression-with-derivation/

How to obtain best fit value

```
we define some function to represent the error between our prediction and the
actual data. We then minimize this error function. Its common to use the ordinary
least square method.
```

- Ordinary Least Square
    - Find parameters which minimize the sum of the square of the distance (d) of the actual point from the model fit line $\hat{y}_i = \beta_0 + \beta_1 x_i$

$$\boxed{E = \sum_{i=0}^{n} (y_{(actual)} - y_{(predicted)})^2}$$

Minimize the sum of the square of error distance

$$E = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

$$E = \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial E}{\partial \beta_0} = \sum_{i=1}^{n} -2\left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

$$\frac{\partial E}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i\left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

Finding the value of Beta_0

- Ordinary Least Square
  - Find β_0

$$\frac{\partial E}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad \sum_{i=1}^{N} y_i = N\bar{y}$$

$$n\beta_0 = n\bar{y} - n\beta_1\bar{x}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

- note that this value depends on some average values from the data set

Finding Beta_1

- Ordinary Least Square
  - Find β_1

$$\frac{\partial E}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i \left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \beta_0 x_i - \beta_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x}) x_i - \beta_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \xrightarrow{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.} \beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- in the bottom part, we meaningfully represent this information to be useful for our model
- note that the sum((x1-x_hat)(y-y_hat)) is the covariance between x and y
- we then normalize this value to x (dividing by sum of squares of x)
- putting it together, we divide the covariance of x and y by the covariance of just x

R-Squared (goodness of fit)

```
a numebr that indicates how well data fits a statistical model
```

- defined between [0,1]
- higher value indicates a better fit

equation: 1 - ((sum of squares between y actual and y predicted)/(sum of squares between actual y values and their mean))