Linear Regression

Prerequisite:

Statistics- mean, median, mode, variance and standard deviation, correlation, and covariance.

Exploratory Data Analysis- Data distribution, Scatter plot, correlation matrix, Heat map.

Objectives:

- Understand, what is Linear Regression and motivation behind linear regression
- What is the best fit line and residual of regression
- Least Square method to find Best Fit line of Regression.
- Gradient Descent method to find Best Fit line of Regression

Linear Regression

Linear regression is a way to identify a relationship between two or more variables and use these relationships to predict values for one variable for given value(s) of other variable(s). Linear regression assume the relationship between variables can be modeled through linear equation or an equation of line. The variable, which is used in prediction is termed as independent/explanatory/regressor where the predicted variable is termed as dependent/target/response variable. Linear regression assumes that independent variables are **related linearly** to response variable.

$$y = c + mx$$

In machine learning and regression literature above equation is used in the form:

$$y = w_0 + w_1x$$

Where w_0 is intercept on y-axis, w_1 is slope of line, x is an explanatory variable and y is the response variable.

Motivational Examples

1. Let we have given with sales data of house prices. For each house we have complete information of it plot size area and the price at which the house was sold. Can we use this information to predict the price of a house for a given plot size area? The problem can be

modeled as linear regression with plot_ size(x) is explanatory variable and house price(y) as response variable.

$$HousePrice(y) = w_0 + w_1 PlotSize$$

2. Consider a scenario where we have given a medical data about some patients. The data contains the information of the blood pressure for a patient along with his/her age. Can we use this information to predict the blood pressure level of patient for given age. This problem is modeled as regression problem with age as an explanatory variable and blood pressure is the response variable.

$$BloodPressure(y) = w_0 + w_1 Age$$

3. Next consider a problem where we need to predict the price of a used car. The sale price of a used car depends on many attributes, some of them may be mileage (km/litre), model (Maruti, Hundai, Honda, Toyota, Tata), segment (Small, Medium, Luxury. In this scenario the sale price is response or target variable depends on mileage, model and segment(explanatory variables). This problem is model as linear regression problem but belongs to **multiple linear regression** as there are more than one explanatory variables are involved in the prediction of target variable.

$$SalePrice(y) = w_0 + w_1 Mileage + w_2 Model + w_3 Segment$$

In real scenarios, we rarely have one explanatory variable, so we use multiple linear regression rather than simple linear regression. However, here we take an example of simple linear regression to understand the fundamentals of regression.

Example: Consider a toy example where we are interested to find the effect of studying hours per day over grades in examination and predict the grads of a student for given study hours. We have sample data about six students for their grades and total study hours per day.

From the given data we get an idea that study hours per day and grades have positive relationship. So one can say that if a student spends more hours studying per day he is likely to get good grades in his/her examination.

The scatter plot of given data is shown in Figure-1. Scatter plot is a useful tool to judge the strength of relationship between two variables. If scatter plot does not conclude any relationship then fitting a linear model to data is probably not useful. A valuable measure to quantify the relationship between two variables is the **correlation coefficient,** The correlation coefficient has range values between -1 to 1 to indicate the strength of relationship. -1(minus one) indicates the strong negative relation where an increase in one variable results in a decrease of other variable. 1(plus one) indicates a strong positive relationship with increase in one variable results in increase in other variable too. 0 (zero) shows no correlation between the two variables.

From the scatter plot shown in Figure-1, we get some intuition that there is a positive effect of studying hours per day over grades in exam.

Table-1	
Study Hours per day	**Grades**
6	7
5	5
4	2
7	9
8	6
2	3

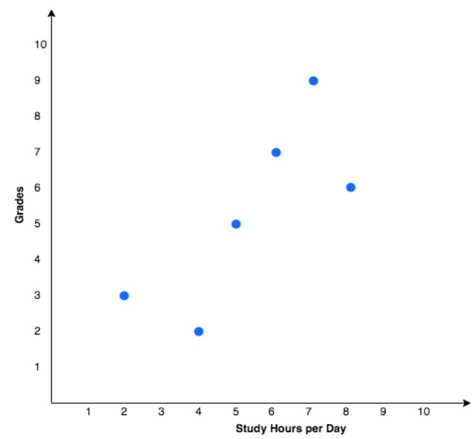

Figure-1: Scatter Graph

To fit, given data we can draw multiple lines out of them one will be our best fit line. Let the equation of the linear model is given by:

$$y(Grades) = w_0 + w_1\, X(Study\ Hours\ per\ day)$$

Now we need to define the criteria for best fit line.

Any line we might come up with has some fixed intercept w_0 and a slop w_1.This line may include some data points on it but cannot cover all of them. In our example we have given with six data points let us label these points by (x1,y1), (x2, y2), (x3, y3), (x4,y4), (x5, y5) and (x6, y6), with values (6, 7), (5, 5), (4, 2), (7, 9), (8, 7) and (2, 3). For any given point X_i the prediction of $yhat_i$ is given by:

$$yhat_i = w_0 + w_1 x_i$$

Unless the line passes through (x_i, y_i) the value of $yhat_i$ differs from the observed value of y_i. The difference between the two values is denotes as **error** or **residual** of regression.

$$e_i = y_i - yhat_i$$

The **best line** is the line which minimizes the **sum of the squared error:**

$$e_i^2 = (y_1 - yhat_1)^2 + (y_2 - yhat_2)^2 + \ldots\ldots\ldots + (y_n - yhat_n)^2$$

Following graphs illustrate the process to find the best line of regression.

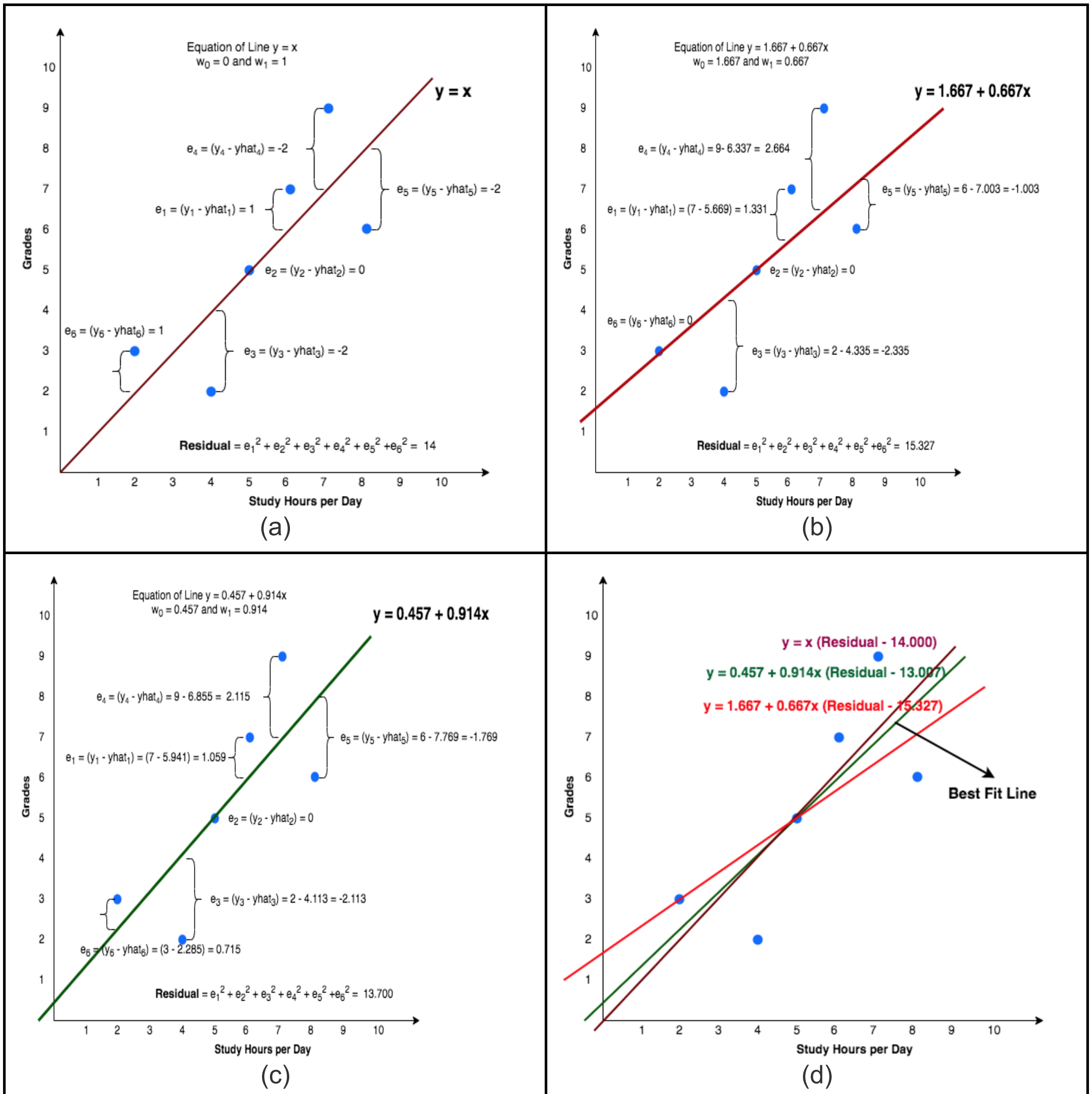

Figure-2 (a) Fit a line of equation y =x. (b) Fit a line of equation y = 1.667 + 0.667x (c) Fit a line of equation 0.457 + 0.914x (d) Combine all three lines and choose the Best Fit Lines with minimum Residual.

Methods to Find Best Fit Line - We can use two different methods to find best fit line of regression.

1. Principle of Least Squares.
2. Gradient Descent.

Least Square- Let the equation of regression line of y on x is:

$$y = w_0 + w_1 x$$

According to the least square principle the equations to estimate the values of w_0 and w_1 are:

$$\sum_{i=1}^{n} y_i = n w_0 + w_1 \sum_{i=1}^{n} x_i \ldots\ldots\ldots(1)$$

$$\sum_{i=1}^{n} X_i y_i = w_0 \sum_{i=1}^{n} x_i + w_1 \sum_{i=1}^{n} x_i^2 \ldots\ldots\ldots.(2)$$

Dividing equation (1) by n we get,

$$\frac{1}{n}\sum_{i=1}^{n} y_i = w_0 + \frac{w_1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y_i} = w_0 + w_1 \bar{x_i} \ldots\ldots\ldots.(3)$$

Thus we can say the line of regression will always passes through the points $(\bar{x}, \bar{y})$

Now we need to estimate the values for w_0 and w_1,

We know,

$$cov(x, y) = \frac{1}{n}\sum_{i} x_i y_i - \bar{x}\bar{y} \Rightarrow \frac{1}{n}\sum_{i} x_i y_i \Rightarrow cov(x, y) + \bar{x}\bar{y} \ldots\ldots\ldots..(4)$$

Also,

$$var(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = var(x) + \bar{x}^2 \ldots\ldots\ldots..(5)$$

Dividing equation (2) by n and using equation (4) and (5),

$$cov(x, y) + \bar{x}\bar{y} = w_0 \bar{x} + w_1 (var(x) + \bar{x}^2) \ldots\ldots..(6)$$

By solving equation (3) and (6) we get,

$$w_1 = \frac{cov(x, y)}{var(x)} \ldots\ldots\ldots(7)$$

and

$$w_0 = \bar{y} - w_1\bar{x}$$

The straight line defined by $y = w_0 + w_1 x$ satisfies the residual (least squares) condition error = E{(y -(w_0 + w_1x))^2} is minimum for variations in a and b, is called the line of regression of y on x. Let us try these equations to estimate best fit line on our data given in Table-1.

To estimate w_0 and $w_1$we need to find covariance between x and y cov(x, y), variance of x var(x)and mean of x and y variables (x, and y) .For given data we get,

$$\bar{x} = \frac{6+5+4+7+8+2}{6} = 5.333$$

$$\bar{y} = \frac{7+5+2+9+6+3}{6} = 5.333$$

$$cov(x, y) = 3.5555$$

$$var(x) = 3.8889$$

when we substitute these values in equation (7) and (8) we get,

$$w_0 = 0.4571 \; and \; w_1 = 0.9143$$

which are exactly the same as shown in Figure-2(c) for the line **y = 0.457 + 0.914x**, which gives the minimum residual among all the lines.

Performance metric for least square regression- Performance metrics are the way to quantify and compare the efficiency of any machine learning model. Least square regression uses R^2(R-squared) and R_{adj}^2 (Adjusted R-Square) metrics to measure the performance of regression model. R_{adj}^2 (Adjusted R-Square) is used with multiple linear regression. Both of these metrics denotes the power of explain ability of selected independent variable(s) to the variation of response variable. The equations of R^2(R-squared) and R_{adj}^2(Adjusted R-Square) are given by:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The numerator term gives the average of squares of residuals and denominator shows the variance in y(response) value. A small value for R^2 or higher mean residual error denote poor model.

$$R^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1}$$

Where n is the total number of observations in data and k is the number of explanatory variables. R_{adj}^2 (Adjusted R-Square) is slight improvement over R^2(R-squared) by adding an additional term to it. The problem with R^2 (R-squared) is that, R^2(R-squared) increases with increase in number of terms in the model irrespective of whether the added terms are significantly contribute in prediction or not . On the contrary, the value of R_{adj}^2 (Adjusted R-Square) is only affected by if useful terms are added to the model. The relation between R^2(R-squared) and R_{adj}^2 (Adjusted R-Square) is:

$$R_{adj}^2 \leq R^2$$

Gradient Descent- Let the equation of regression line of y on x is:

$$y = w_0 + w_1 X$$

This straight line tries to approximate the relationship between x and y for given set of data. By varying the values of w0and w1we can find the best fit line. With the above discussion we know that the best fit line is one which minimizes the total error in prediction. Gradient descent method defines a cost function of parameter w0and w1and uses a systematic approach to optimize the values of parameters to get minimum cost function. Let us dive into mathematics of algorithm.

Let the model is defined as:

$$y = w_0 + w_1 x$$

Now define the cost function of gradient descent as Mean Squared Error of prediction:

$$cost\ (w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2$$

The cost function includes two parameters w0and w1, which control the value of cost function. As we know the derivatives give us the rate of change in one variable with respect to others, so we can use partial derivatives to find the impact of individual parameter over the cost function.

The principle of gradient descent is that we always make progress in the direction where the partial derivatives of w_0 and w_1 are steepest. If the derivatives of parameters are become zero or very less, point the situation of either maxima or minima on the surface of cost function. The process of gradient descent is started with the random initialization of w_0 and w_1. Every iteration of gradient descent improves in the direction of optimal values for w_0 and w_1 parameters which will have minimum cost function value. Following figure illustrate the process of optimization.

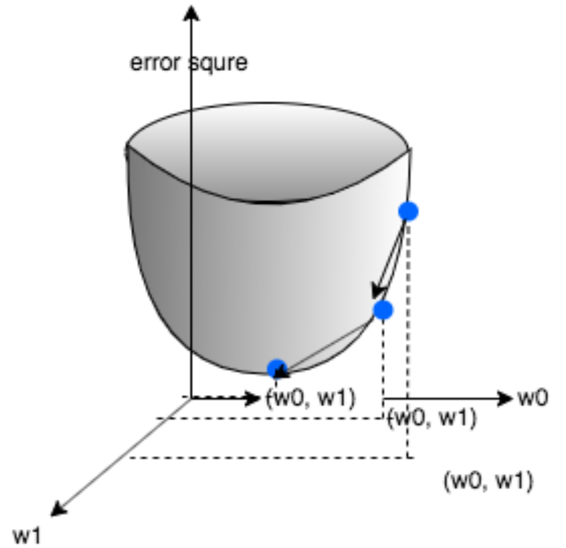

Figure 3: Gradient Descent Iteration

Gradient descent works in following steps:

1. Random initialization of parameters.
2. Calculate the partial derivatives of the cost function with respect to each parameter(gradients).
3. Update the parameters in the opposite direction of gradients.
4. Repeat step 3 and 4 till maximum iteration reached or minimum cost function value achieved.

Partial derivatives:

We have,

$$error = cost\ (w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2$$

Partial derivative w. r. t. w_0 and w_1:

$$\frac{\partial cost(w_0, w_1)}{\partial w_0} = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)(-2) = \frac{-2}{n}\sum_{i=1}^{n} error_i$$

$$\frac{\partial cost(w_0, w_1)}{\partial w_1} = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)(-2x_i) = \frac{-2}{n}\sum_{i=1}^{n} error_i\ (x_i)$$

Parameter updates:

$$w_0 = w_0 - lrate \frac{\partial cost(w_0, w_1)}{\partial w_0}$$

$$w_1 = w_1 - lrate \frac{\partial cost(w_0, w_1)}{\partial w_1}$$

lrate is the learning rate which controls the step size of parameter update.

Let's run it on our example:

X:	6	5	4	7	8	2
y:	7	5	2	9	6	3

Let's initialize both coefficient w_0 and w_1 with 0.0,

$$w_0 = 0.0$$

$$w_1 = 0.0$$

Iteration #1:

$$yhat_i = 0.0 + 0.0\, x_i$$

Calculate gradients:

$$\frac{\partial cost(w_0, w_1)}{\partial w_0} = \frac{-2}{6}(7 + 5 + 2 + 9 + 6 + 3) = -10.6667$$

$$\frac{\partial cost(w_0, w_1)}{\partial w_1} = \frac{-2}{6}(7 * 6 + 5 * 5 + 2 * 4 + 9 * 7 + 6 * 8 + 3 * 2) = -64$$

Update parameters: (*lrate* = 0.01)

$$w_0 = w_0 - lrate \frac{\partial cost(w_0, w_1)}{\partial w_0} = 0.0 - 0.01\,(-10.6667) = 0.1066$$

$$w_1 = w_1 - lrate \frac{\partial cost(w_0, w_1)}{\partial w_1} = 0.0 - 0.01\,(-64) = 0.64$$

Iteration #2:

$$yhat_i = 0.1067 + 0.64\, x_i$$

Calculate gradients:

$$\frac{\partial cost(w_0, w_1)}{\partial w_0} = \frac{-2}{6}(3.0533 + 1.6933 - 0.6667 + 4.4133 + 0.7733 + 1.6133) = -3.6266$$

$$\frac{\partial cost(w_0, w_1)}{\partial w_1} = \frac{-2}{6}(3.0533 * 6 + 1.6933 * 5 - 0.6667 * 4 + 4.4133 * 7 + 0.7733 * 8 + 1.6133 * 2)$$

$$= -21.475$$

Update parameters: (*lrate* = 0.01)

$$w_0 = w_0 - lrate \frac{\partial cost(w_0, w_1)}{\partial w_0} = 0.1067 - 0.01\ (-3.6266) = 0.14296$$

$$w_1 = w_1 - lrate \frac{\partial cost(w_0, w_1)}{\partial w_1} = 0.64 - 0.01\ (-21.475) = 0.8547$$

Similarly the number of iteration for gradient descent are performed till the minimum value of cost function of error is achieved or some finite iterations are reached.

Multiple Linear Regression: Till we have discussed the case linear regression with just one explanatory variable. But in real scenarios the target variable depends on multiple explanatory variable which need to be cater during the development of linear regression model. The model is expressed as:

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 \ldots \ldots \ldots \ldots .. + w_nx_n$$

Where x_1, x_2, x_3.......x_n are explanatory variables and y is target variable.

Evaluation of Linear regression model- Evaluation helps to judge the performance of any machine learning model that would provide best results to our test data. Fundamentally three types of evaluation metrics are used to evaluate linear regression model.

- R2 measure (discussed with least square method)
- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)

Mean Absolute Error(MAE)- Mean Absolute Error is the average of the difference between actual and predicted value of target variable.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - yhat_i|$$

Root Mean Square Error(RMSE)- defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i\ )^2}$$

Pros and cons of Linear Regression:

Pros- Linear regression models are very simple and easy to implement. These models are said to be most interpretable.

Cons- Linear regression models are largely affected by the presence of outlier in training data. These models assume linear relationship between target and explanatory variables which is sometimes is not true.

****