greatlearning
Learning for Life

Linear Regression

Introduction to machine learning

greatlearning
Learning for Life

Linear Regression Models -

a. The term "regression" generally refers to predicting a real number. However, it can also be used for classification (predicting a category or class.)

b. The term "linear" in the name “linear regression” refers to the fact that the method models data with linear combination of the explanatory variables.

c. A linear combination is an expression where one or more variables are scaled by a constant factor and added together.

d. In the case of linear regression with a single explanatory variable, the linear combination used in linear regression can be expressed as:

$$\text{response} = \text{intercept} + \text{constant} * \text{explanatory}$$

e. In its most basic form fits a straight line to the response variable. The model is designed to fit a line that minimizes the squared differences (also called errors or residuals.).

Introduction to machine learning

Linear Regression Models -

greatlearning
Learning for Life

a. Before we generate a model, we need to understand the degree of relationship between the attributes Y and X

b. Mathematically correlation between two variables indicates how closely their relationship follows a straight line. By default we use Pearson's correlation which ranges between -1 and +1.

c. Correlation of extreme possible values of -1 and +1 indicate a perfectly linear relationship between X and Y whereas a correlation of 0 indicates absence of linear relationship
    I. When r value is small, one needs to test whether it is statistically significant or not to believe that there is correlation or not

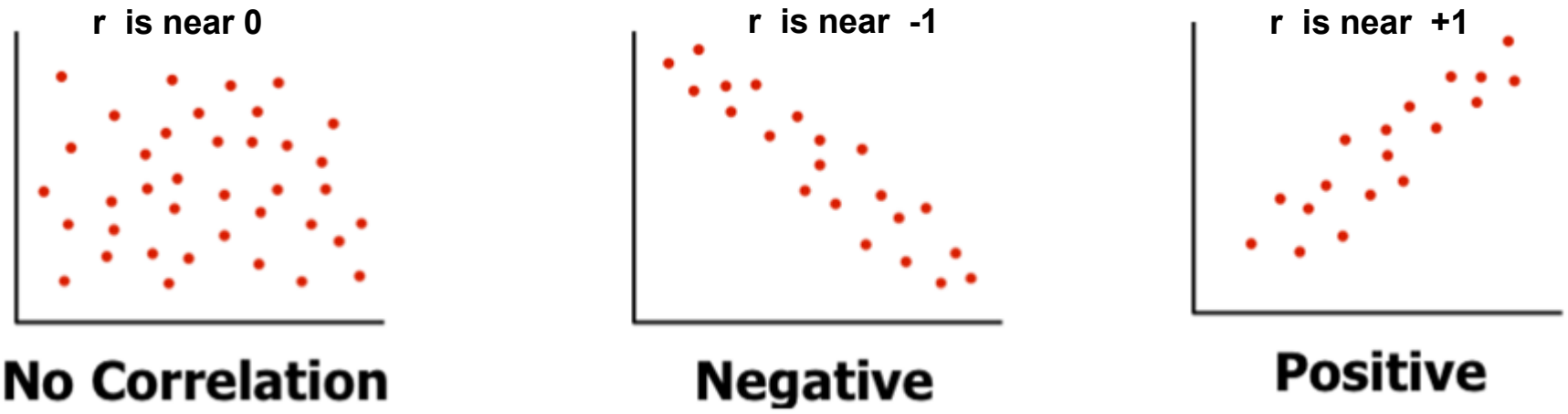
Introduction to machine learning

Linear Regression Models -

greatlearning
Learning for Life

d. Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



e. Generating linear model for cases where r is near 0, makes no sense. The model will not be reliable. For a given value of X, there can be many values of Y! Nonlinear models may be better in such cases

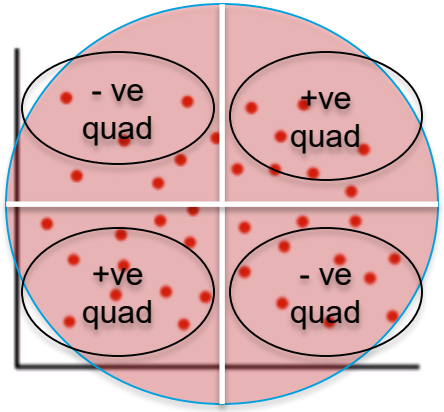
Introduction to machine learning

Linear Regression Models (Recap) -

greatlearning
Learning for Life

f. Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = 0$$

http://www.socscistatistics.com/tests/pearson/Default2.aspx

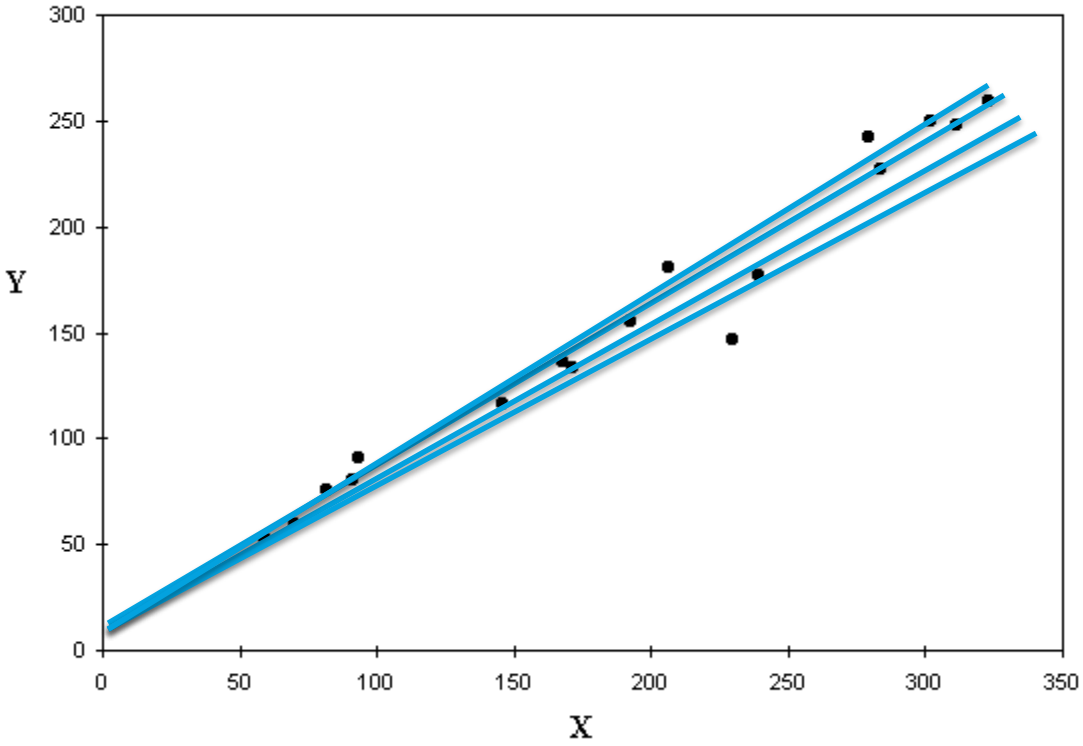
Introduction to machine learning

Linear Regression Models -

greatlearning
Learning for Life

g. Given Y = f(x) and the scatter plot shows apparent correlation between X and Y Let's fit a line into the scatter which shall be our model

h. But there are infinite number of lines that can be fit in the scatter. Which one should we consider as the model?



i. This and many other algorithms use gradient descent or variants of gradient descent method for finding the best model

j. Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors

Introduction to machine learning

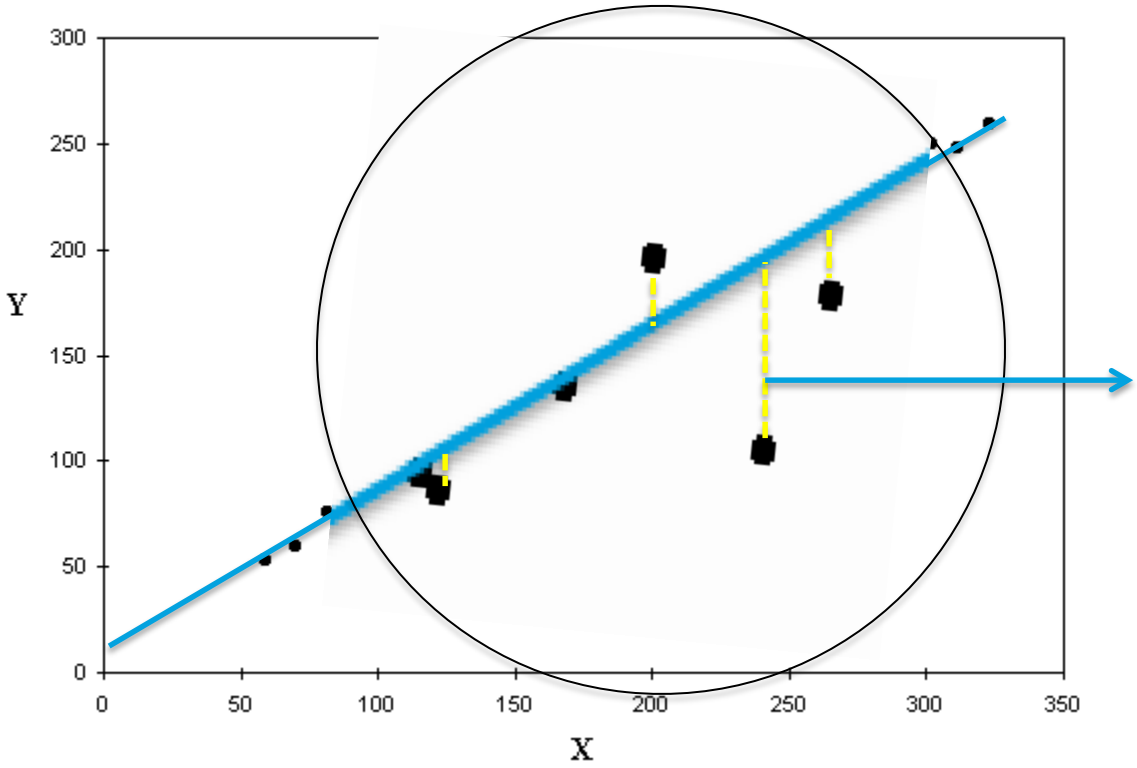
greatlearning
Learning for Life

Linear Regression Models (Recap) -

k. Whichever line we consider as the model, it will not pass through all the points.

l. The distance between a point and the line (drop a line vertically (shown in yellow)) is the error in prediction

m. That line which gives least sum of squared errors is considered as the best line



Error = (T – (mx + C)

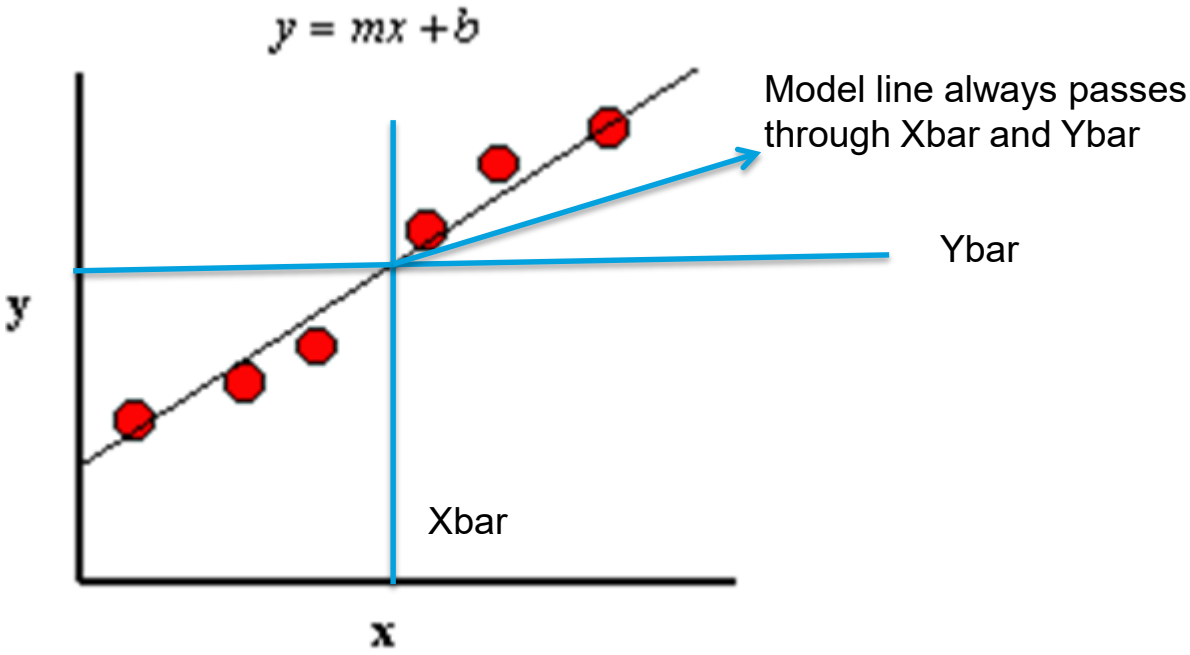
Sum of all errors can cancel out and give 0

We square all the errors and sum it up. That line which gives us least sum of squared errors is the best fit

Introduction to machine learning

Linear Regression Models -

n. Coefficient of determinant – determines the fitness of a linear model. The closer the points get to the line, the R^2 (coeff of determinant) tends to 1, the better the model is



Introduction to machine learning

Linear Regression Models -

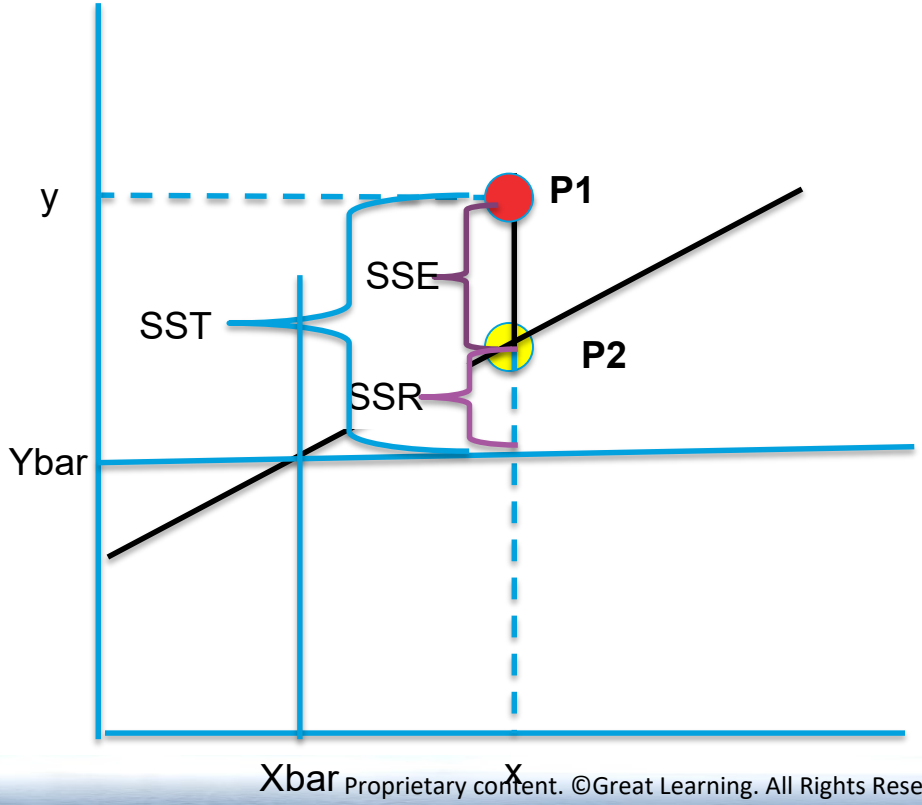
greatlearning
Learning for Life

o. Coefficient of determinant (Contd…)

I. There are a variety of errors for all those points that don't fall exactly on the line.

II. It is important to understand these errors to judge the goodness of fit of the model i.e. How representative the model is likely to be in general

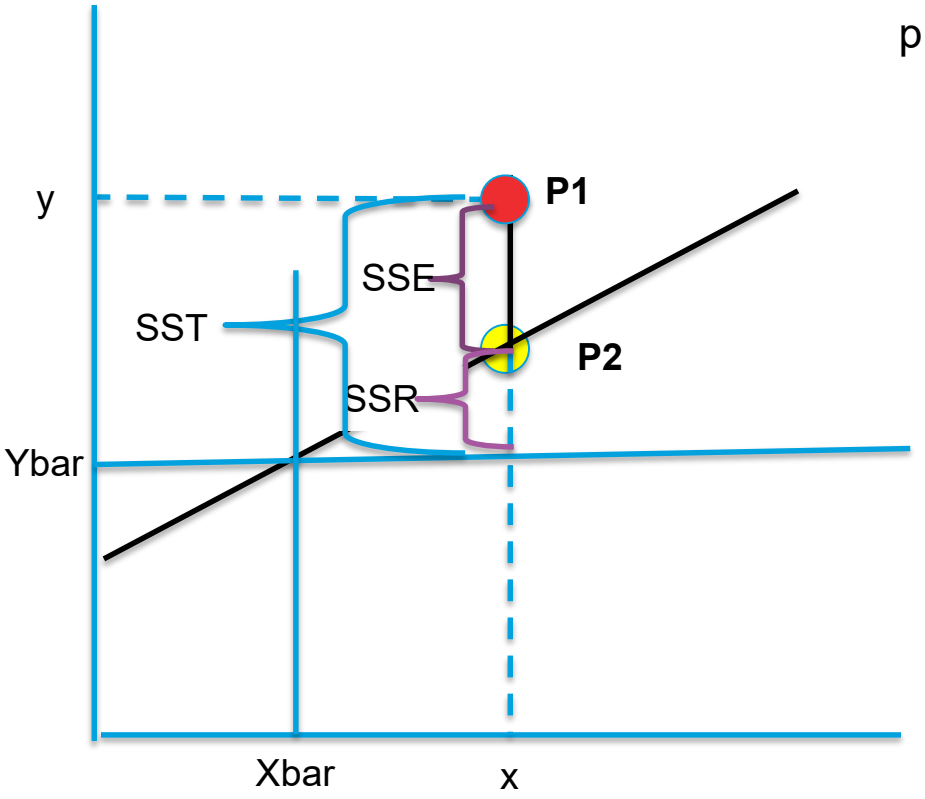
III. Let us look at point P1 which is one of the given data points and associated errors due to the model



1. P1 – Original y data point for given x
2. P2 - Estimated y value for given x
3. Ybar – Average of all Y values in data set
4. SST – Sum of Square error Total (SST) Variance of P1 from Ybar (Y – Ybar)^2
5. SSR - Regression error (p2 – ybar)^2 (portion SST captured by regression model)
6. SSE - Residual error (p1 – p2)^2

Introduction to machine learning

Linear Regression Models -



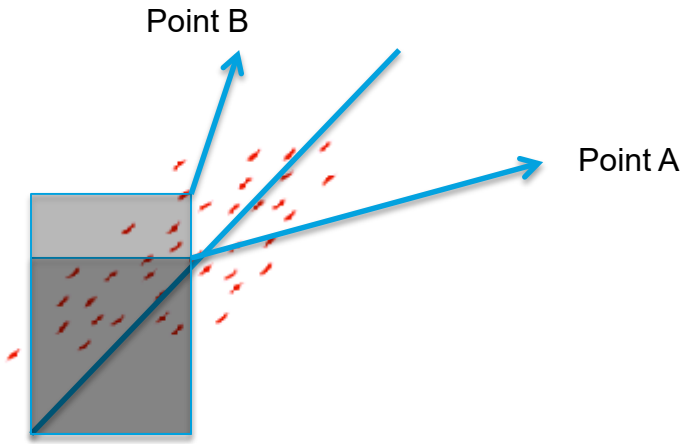
p. Coefficient of determinant (Contd…)

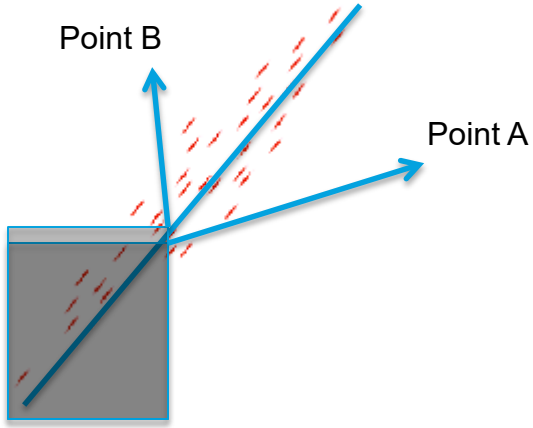
1. That model is the most fit where every data point lies on the line. i.e. SSE = 0 for all data points
2. Hence SSR should be equal to SST i.e. SSR/SST should be 1.
3. Poor fit will mean large SSE. SSR/SST will be close to 0
4. SSR / SST is called as r^2 (r square) or coefficient of determination
5. r^2 is always between 0 and 1 and is a measure of utility of the regression model

Linear Regression Models -

greatlearning
Learning for Life

q. Coefficient of determinant (Contd…) -





In case of point “A”, the line explains the variance of the point

Whereas point “B” the is a small area (light grey) which the line does not represent.

%age of total variance that is represented by the line is coeff of determinant

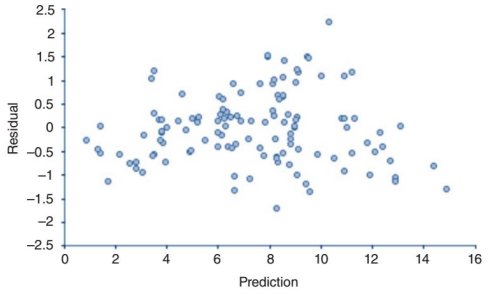
greatlearning
Learning for Life

Linear Regression Assumptions

Linear regression model is based on a set of assumptions. If the underlying dataset does not meet these assumptions, then data may have to be transformed or linear model may not be good fit

1. Assumption of linearity. assumes a linear relation between the dependent / target variable and the independent / predictor variables.

2. Assumption of normality of the error distribution.
    a. The errors should be normally distributed across the model.
    b. This assumption can be tested using a frequency histogram, skew and kurtosis of a normal plot. If the distribution does not approximate normal distribution, data transformation may be necessary
    c. A scatter plot between the actual values and the predicted values should show the data distributed equally across the model.
    d. Another way of doing this is to plot residual values against the predicted values. We should not see any trends



greatlearning
Learning for Life

Linear Regression Assumptions

3. Assumption of homoscedasticity of errors. The variation of the error or residuals across each of the independent variable should remain constant. There should be no trend visible in plots of errors against predicted values, independent variables

4. Assumption of independence of errors. There should be no trend in the residuals based on the order in which the observations were collected. A scatter plot of the errors against an order in which the data was collected should show not trend. Durbin Watson test can also be employed… Ref. https://www.investopedia.com/terms/d/durbin-watson-statistic.asp

Introduction to machine learning

greatlearning
Learning for Life

Linear Regression Model -

Advantages –

1. Simple to implement and easier to interpret the outputs coefficients

Disadvantages -

1. Assumes a linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them
2. Outliers can have huge effects on the regression
3. Linear regression assume independence between attributes
4. Linear regression looks at a relationship between the mean of the dependent variable and the independent variables.
5. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables
6. Boundaries are linear

Introduction to machine learning

greatlearning
Learning for Life

Linear Regression Model -

Lab- 1- Estimating mileage based on features of a second hand car

Description – Sample data is available at
https://archive.ics.uci.edu/ml/datasets/Auto+MPG

The dataset has 9 attributes listed below that define the quality

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

Sol : mpg-linear regression.ipynb

greatlearning
Learning for Life

ThankYou