**Simple Progression Towards Simple Linear Regression**

Introduction:

The goal of the blogpost is to get the beginners started with basics of the linear regression concepts and quickly help them to build their first linear regression model. We will mainly focus on the modeling side of it. The data cleaning and preprocessing parts would be covered in detail in an upcoming post.

Linear Regression are one of the most fundamental and widely used Machine Learning Algorithm. Linear regression is usually among the first few topics which people pick while learning predictive modeling. Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).The dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear relationship can either be positive or negative. Positive relationship between two variable basically means that that increase in one variable should also increase the value in other value by some constant value. Negative relationship between two variable means that that increase in one variable should decrease the value in other value by some constant value.

Mathematical Explanation:

A simple linear regression has one independent variable. Mathematically, the line representing a simple linear regression is expressed through a basic equation:

|  |
| --- |
| Y = mX + b + e |

Here ,

m is the slope

X is the predictor variable

b is the intercept/bias term

Y is the predicted target variable

e is the error term

Enough of theory now let’s dive into the implementation of both; a simple linear regression and a multivariate linear regression.

We will use implementation provided by the python machine learning framework known as scikit-learn.

**Problem Statement:**

Predict the price of a car given its compression ratio.

**Data details**

These data sets are originally from the <http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>.

The label is the identifier of the column you are trying to predict. The identified features are used to predict the label.

**vendor\_id**: The ID of the taxi vendor is a feature.

**rate\_code**: The rate type of the taxi trip is a feature.

**passenger\_count**: The number of passengers on the trip is a feature.

**trip\_time\_in\_secs**: The amount of time the trip took. You won't know how long the trip takes until after it is

**Completed**: You exclude this column from the model.

**trip\_distance**: The distance of the trip is a feature.

**payment\_type**: The payment method (cash or credit card) is a feature.

**fare\_amount**: The total taxi fare paid is the label.

**Tools used**:

Pandas , Numpy , Matplotlib , scikit-learn

**Python Implementation with code:**

**1. Import necessary libraries**

Import the necessary modules from specific libraries.

|  |
| --- |
|  |

**2. Load the data set**

Use pandas module to read the taxi data from the file system. Check few records of the dataset.

|  |
| --- |
|  |

**3. Select the predictor feature for Simple Regression, select the target variable**

|  |
| --- |
|  |

**4. Check 5-num summary of selected predictor feature**

|  |
| --- |
|  |

**5. Train test split :**

|  |
| --- |
|  |

Dome reshaping of the variable for visualization

|  |
| --- |
|  |

**6. Perform visual inspection between predictor and target variable**

|  |
| --- |
|  |

**7. Training / model fitting**

Fit the model to selected supervised data

|  |
| --- |
|  |

**8. Study the Model parameters**

|  |
| --- |
|  |

**9. Accuracy report with test data:**

Visualize the goodness of the fit with the predictions being visualized by a line

|  |
| --- |
|  |

**10. Prediction:**

**11. What are its advantages?**

**12. What are its disadvantages?**