Name: Gokhan Koyunlu

1. Linear Regression

Is it supervised /unsupervised ?

Yes, Linear regression is a supervised learning algorithm because it **uses true labels for training**.

1. What does the algorithm do?

Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example

1. In which situations will it be most useful

Regression analysis can handle many things. For example, you can use regression analysis to do the following:

1. Model multiple independent variables
2. Include continuous and categorical variables
3. Use polynomial terms to model curvature
4. Assess interaction terms to determine whether the effect of one independent variable depends on the value of another variable
5. Can you find any examples of where this algorithm has been used ? ( optional )

The dataset that we are using has below columns –

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Hyundai(Thousands of dollars) | Maruti (Thousands of dollars) | Mahindra (Thousands of dollars) | Sales(Thousands of dollars) |
| 1 | 230.1 | 37.8 | 69.2 | 22.1 |
| 2 | 44.5 | 39.3 | 45.1 | 10.4 |
| 3 | 17.2 | 45.9 | 69.3 | 9.3 |
| 4 | 151.5 | 41.3 | 58.5 | 18.5 |
| 5 | 180.8 | 10.8 | 58.4 | 12.9 |

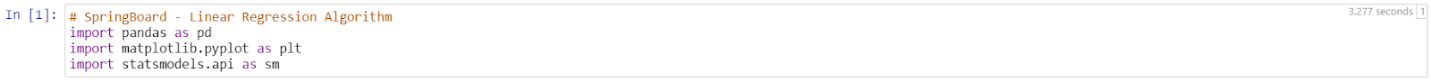
In our data Frame, the first column is an index column and the first row is the column header. Our dataset has 200 rows. Let’s describe the dataset:

1. Hyundai – This column indicates the money spent on advertising the Hyundai cars in the given market.
2. Maruti – Similarly, money spent on advertising by Maruti car.
3. Mahindra – Similarly, money spent on advertising by Mahindra car.
4. Sales – This column indicates the sales of cars in the given market. (Value of the sales in thousands)

Implementation of linear regression:

For our implementation, we are using [Jupyter](https://jupyter.org/try) Notebook and executing our algorithm in python v3.0.

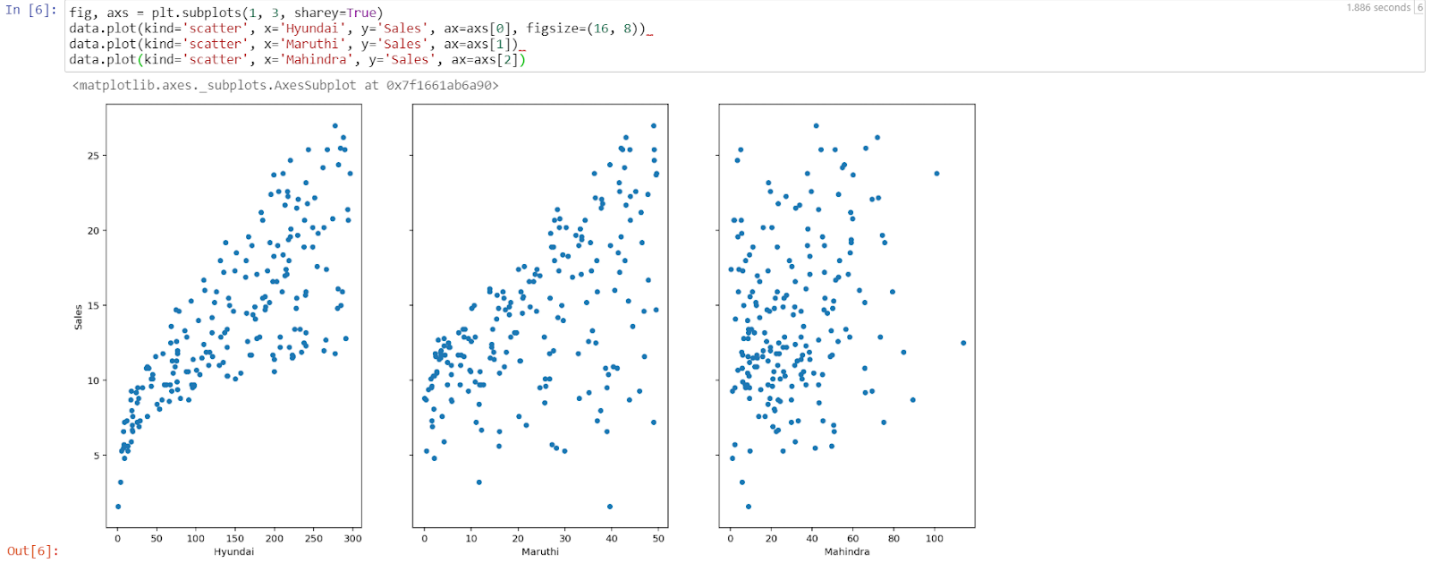
1. Let us start by loading the necessary libraries. Firstly, we are importing Pandas which is the most popular python library for data exploration, manipulation and analysis. Please follow this document [to install Pandas](https://pandas.pydata.org/pandas-docs/stable/getting_started/install.html). We are importing Mathplotlib for multiplatform data visualization.  For linear regression, we need to use Statsmodels to estimate the model coefficients for the advertising data.



2. Next steps we are going to load the dataset, read the data into a data frame and display the head (top 5 rows). Also, we can see the total number of rows. There are 200 observations in the given dataset.

what-is-linear-regression-4


3. Let’s visualize the relationship between the features and the sales response using scatterplots.



Now by looking into the above scatterplots we can easily say that Hyundai advertising and sales have a strong relationship. However, Maruti’s and Mahindra’s datapoints look scattered all over the graph that implies that they have a weak relationship between advertisement and sales.

4. Now let’s Estimate the model coefficients for Linear Regression by using single feature to predict quantitative response. It takes the following form:

what-is-linear-regression-6

Where y will be the response

X will be the feature

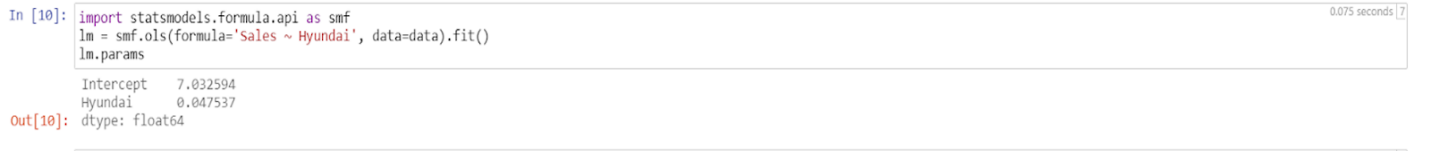
a is the intercept

b is the coefficient for x, a & b are called Model coefficient.

To calculate coefficients, we will use the least square criterion, which means we will find a line that will decrease the sum of squared errors.

 In this step we will load statsmodels to estimate the model coefficients for the advertising data. Statsmodels allows users to fit statistical models by importing OLS. As shown below we are going to fit the model using statsmodels OLS.

Here we are finding the model coefficient between Sales column and Hyundai column. lm.params will print the coefficient.



5. From step 4- we got the value of A and B. we will use the model to predict the future sales of Hyundai cars. Let’s say in the new market Hyundai is spending 50 thousand dollars in advertising. That means the new value of X will be 50. Now using Y = A + BX to predict the new value.

Graphical user interface, text, application, email, Teams

Description automatically generated

6. Now let’s plot the least square line by creating a data frame with the minimum and maximum

values of Hyundai and predict for x value and store that value in preds variable.

Graphical user interface, text, application, email, Teams

Description automatically generated

7. Let’s plot the observed data graph and the least square line using preds value and new x value.

A picture containing shape

Description automatically generated

In linear regression, the observation blue dots are assumed to be the result of random deviation from an underlying relationship (red line) between a dependent variable (y) and an independent variable (x). Here the goal is to decrease the distance between the red line and the blue dots. If all the blue dots are on red lines that means Root means square error will low and better.

However, Linear Regression is a very vast algorithm and it will be difficult to cover all of it. You can improve the model in various ways could be by detecting collinearity and by transforming predictors to fit nonlinear relationships. This article is to get you started with simple linear regression. Let’s quickly see the advantage and disadvantage of linear regression algorithm:

1. Linear regression provides a powerful statistical method to find the relationship between variables. It hardly needs further tuning. However, it’s only limited to linear relationships.
2. Linear regression produces the best predictive accuracy for linear relationship whereas its little sensitive to outliers and only looks at the mean of the dependent variable.

Springboard’s 1:1 mentoring-led, project-based [data science](https://www.springboard.com/courses/data-science-career-track/), [data analytics](https://in.springboard.com/courses/data-analytics-career-program-online/https:/in.springboard.com/courses/data-analytics-career-program-online/?utm_source=Blog&utm_medium=Social&utm_campaign=K-Means-Clustering-Web-Blog), [AI/Machine Learning](https://www.springboard.com/courses/ai-machine-learning-career-track/) career tracks are industry-focused online learning programs that come with a job guarantee, designed to prepare you for a meaningful and successful career in future technologies.