Linear Regression

Here we go!

Introduction

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values

Types of Linear Regression Models

01 Simple Linear Regression

Single independent variable (x) is used to estimate a dependent variable(y).

<u>Example</u>: Predicting housing price using house area only.

02 Multiple Linear Regression

Multiple independent variables (x1,x2,x3 …) are used to predict a dependent variable (y).

<u>Example</u>: Predicting housing price using area, facilities and architecture.

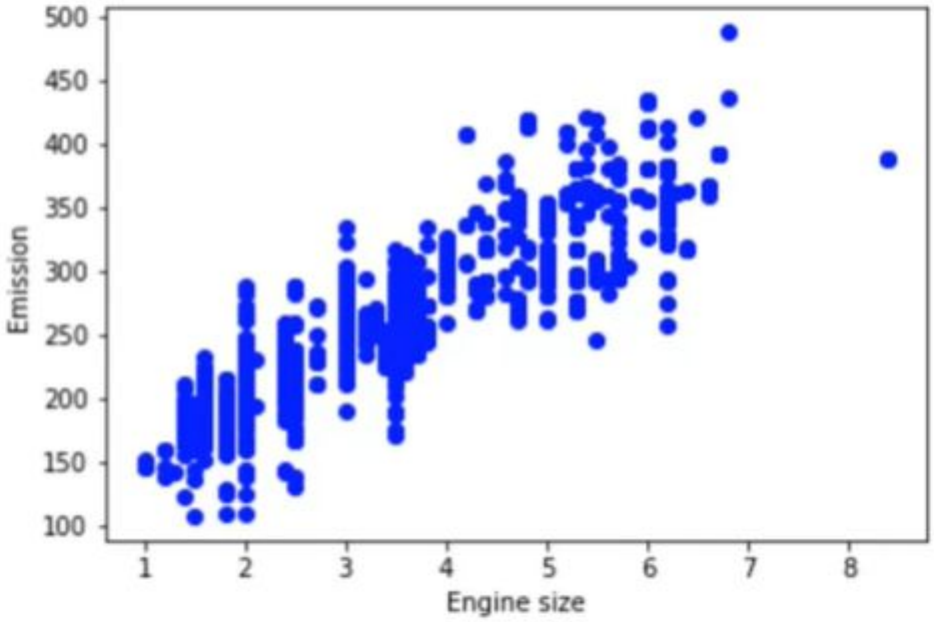
Working

This is our dataset. First, Let's plot our variables in a scatter plot for this dataset considering only engine size as independent variable(x) (Single Linear Regression) and CO2 emissions as dependent variable (y).

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

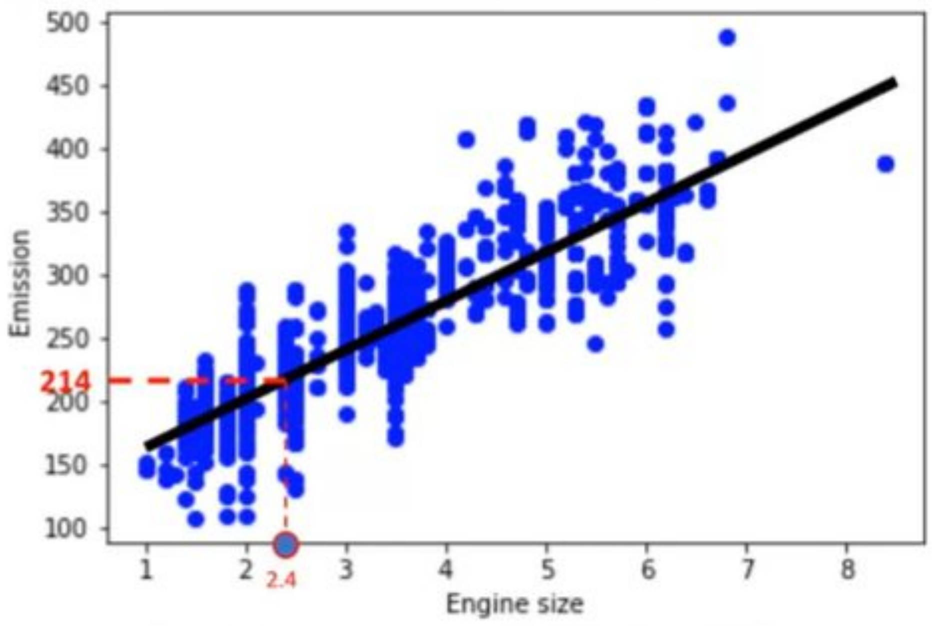
Working

From the scatter plot we can clearly see that change in one variable clearly causes change in another variable. Also, It indicates that these variables are linearly related.



Working

Using Linear Regression, we can model this plot and generate a regression model to fit a straight line in the plot. Then, with the help of model, we can predict the value of dependent variable(y) for entirely new independent variable(x).



Working

Yes, we can now predict the independent variable now, but what is that fitting line?

Generally, the fitting line is a polynomial function. In this case, (linear) it is a polynomial function of degree 1.

Working

The fit line is given as:

$$\hat{y} = \theta_0 + \theta_1 x_1$$

where;

y = response (prediction value)
x = predictor value
θ_1 *and* θ_2 *are 2 parameters we need to adjust.*

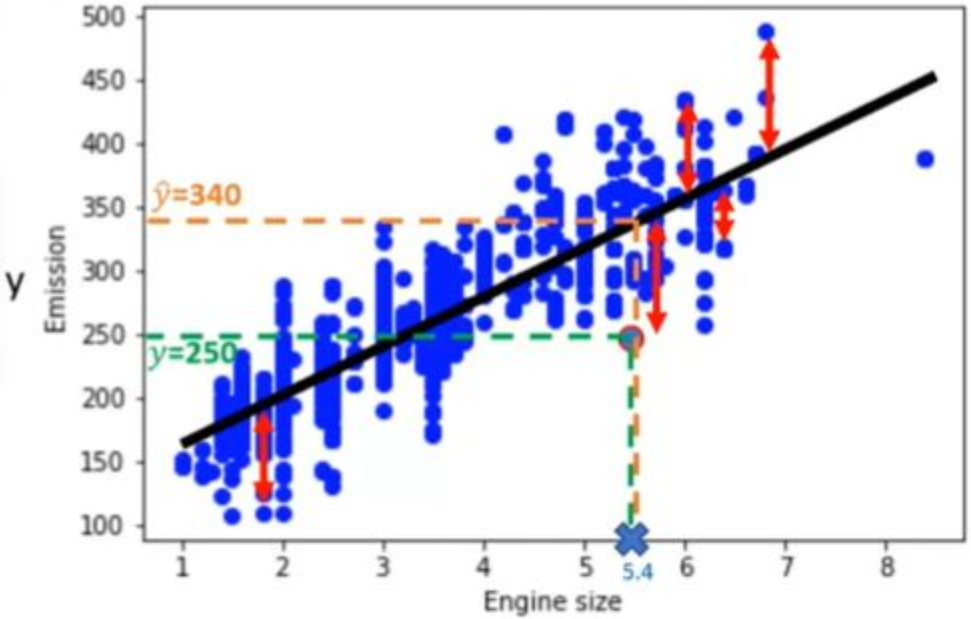
θ_1 *= Slope/gradient of the fit. &* θ_2 *= intercept*

<u>Note</u> :We need to find the best values for θ_1 *and* θ_2 *to make the best estimate.*

Working

So, How can we find the best fit?

Let's say that our fit is the line in the figure. Observing the real value (at 5.4) and the prediction value, we can say there is large difference. We got an error which can also the distance between the original point and predicted point.



Working

The mean of all the errors shows how line fits with the dataset. Mathematically, it can be shown by the equation of Mean Squared Error (MSE) which is given as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Our objective is to minimize the MSE i.e. Mean Squared Error. In order to minimize it, we need to find the best value for parameters θ_1 and θ_2.

Working

As we have our dataset, we can calculate parameter values in following manner:

We don’t need to remember this, libraries will do this for us.

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$$\bar{x} = (2.0 + 2.4 + 1.5 + \ldots)/9 = 3.03$$

$$\bar{y} = (196 + 221 + 136 + \ldots)/9 = 226.22$$

$$\theta_1 = \frac{(2.0 - 3.03)(196 - 226.22) + (2.4 - 3.03)(221 - 226.22) + \ldots}{(2.0 - 3.03)^2 + (2.4 - 3.03)^2 + \ldots}$$

$$\theta_1 = 39$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_0 = 226.22 - 39 * 3.03$$

$$\theta_0 = 125.74$$

$$\hat{y} = 125.74 + 39 x_1$$

Working

After finding the fighting equation, we can predict other values of y for various cases of x.

Pros of Linear Regression:

- Easy
- No Tuning Required
- Fast

Model Evaluation

The goal of regression model is to accurately predict an unknown case. To be sure that the model is performing efficiently, we need to evaluate it. Generally, there are two types of approach for model evaluation:

1. Train and Test on the same dataset
   -Train with Entire Dataset and Test all of them

2. Train/Test Split
   - Train with (70-80%) of dataset and test with remaining.

Evaluation Metrics

Evaluation metrics are used to measure the performance of a model.

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Test

y

Actual values

$$Error = \frac{(232 - 234) + (255 - 256) + \ \dots}{4}$$

$$Error = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

- MAE
- MSE
- RMSE
- ...

$\hat{y}$

	Prediction
6	234
7	256
8	267
9	210

Predicted values

Evaluation Metrics

There are various evaluation metrics such as:

- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error) ....etc.

But First, What is an Error?
In the context of regression, error of the model is the difference between the data points and the trend line generated by the algorithm.

Evaluation Metrics

Mean Absolute Error(MAE) is the mean of absolute value of errors.

Mean Squared Error(MSE) is the mean of the squared form of the errors.

Root Mean Squared Error (RMSE) is just the root of MSE. It is most popular of the metrics because RMSE is interpretable in the same units as response vector.

Relative Absolute Error (RAE) is the residual sum of errors. i.e It takes the total absolute error and normalizes it.

Relative Squared Error (RSE) is very similar to RAE and is used by Data Science Community to calculate R^2 which is used to calculate accuracy. [$R^2 = 1 - RSE$]

Evaluation Metrics

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

Thank You