Linear Regression

In data analytics we come across the term "Regression" very frequently. Before we continue to focus topic i.e. "Linear Regression" lets first know what we mean by Regression. Regression is a statistical way to establish a relationship between a dependent variable and a set of independent variable(s). e.g., if we say that

Age = 5 + Height * 10 + Weight * 13

Here we are establishing a relationship between Height & Weight of a person with his/ Her Age. This is a very basic example of Regression.

Simple Linear Regression

Introduction

Least Square "Linear Regression" is a statistical method to regress the data with dependent variable having continuous values whereas independent variables can have either continuous or categorical values. In other words "Linear Regression" is a method to predict dependent variable (Y) based on values of independent variables (X). It can be used for the cases where we want to predict some continuous quantity. E.g., Predicting traffic in a retail store, predicting a user's dwell time or number of pages visited on Dezyre.com etc.

Prerequisites

To start with Linear Regression, you must be aware of a few basic concepts of statistics. i.e.,

- Correlation (r) – Explains the relationship between two variables, possible values -1 to +1
- Variance (σ^2)– Measure of spread in your data
- Standard Deviation (σ) – Measure of spread in your data (Square root of Variance)
- Normal distribution
- Residual (error term) – {Actual value – Predicted value}

Assumptions of Linear Regression

Not a single size fits or all, the same is true for Linear Regression as well. In order to fit a linear regression line data should satisfy few basic but important assumptions. If your data doesn't follow the assumptions, your results may be wrong as well as misleading.

i. **Linearity & Additive**: There should be a linear relationship between dependent and independent variables and the impact of change in independent variable values should have additive impact on dependent variable.

ii. **Normality of error distribution:** Distribution of differences between Actual & Predicted values (Residuals) should be normally distributed.

iii. **Homoscedasticity:** Variance of errors should be constant versus,

a. Time

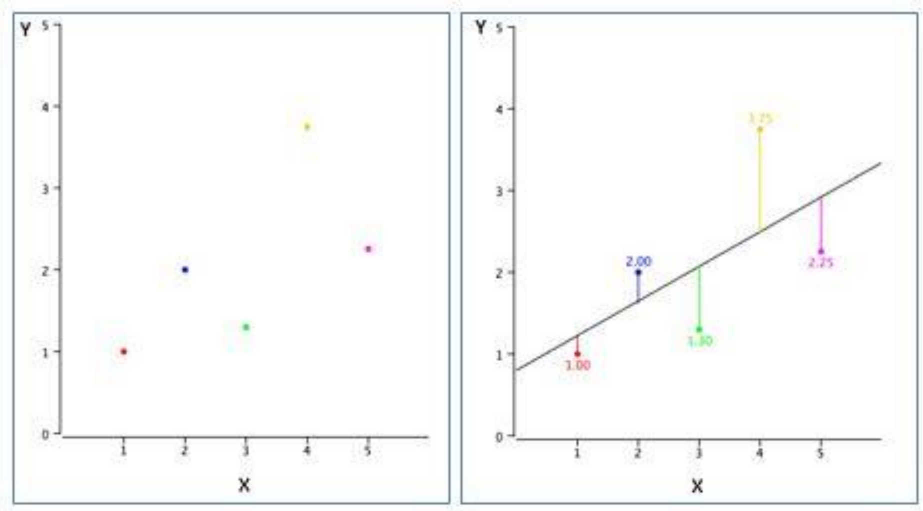
b. The predictions

c. Independent variable values

iv. **Statistical independence of errors:** The error terms (residuals) should not have any correlation among themselves. E.g., In case of time series data there shouldn't be any correlation between consecutive error terms

Linear Regression Line

While doing linear regression our objective is to fit a line through the distribution which is nearest to most of the points. Hence reducing the distance (error term) of data points from the fitted line.



For example, in above figure (left) dots represent various data points and line (right) represents an approximate line which can explain the relationship between 'x' & 'y' axes. Through, linear regression we try to find out such a line. For example, if we have one dependent variable 'Y' and one independent variable 'X' – relationship between 'X' & 'Y' can be represented in a form of following equation:

$Y = B_0 + B_1X$

Where,

- Y = Dependent Variable
- X = Independent Variable
- B_0 = Constant term a.k.a Intercept
- B_1 = Coefficient of relationship between 'X' & 'Y'

Few properties of linear regression line

- Regression line always passes through mean of independent variable (x) as well as mean of dependent variable (y)
- Regression line minimizes the sum of "Square of Residuals". That's why the method of Linear Regression is known as "Ordinary Least Square (OLS)"

Food for thought: Why to reduce "Square of errors" and not just the errors?

- B_1 explains the change in Y with a change in X by one unit. In other words, if we increase the value of 'X' by one unit then what will be the change in value of Y

  Food for thought: Will correlation coefficient between 'X' and 'Y' be same as B_1?

Finding a Linear Regression Line

Using a statistical tool e.g., Excel, R, SAS etc. you will directly find constants ($B_0$and B_1) as a result of linear regression function. But conceptually as discussed it works on OLS concept and tries to reduce the square of errors, using the very concept software packages calculate these constants.

For example, let say we want to predict 'y' from 'x' given in following table and let's assume that our regression equation will look like "y=B0+B1*x"

x	y	Predicted 'y'
1	2	B_0+B_1*1
2	1	B_0+B_1*2
3	3	B_0+B_1*3
4	6	B_0+B_1*4
5	9	B_0+B_1*5
6	11	B_0+B_1*6
7	13	B_0+B_1*7
8	15	B_0+B_1*8

9	17	B_0+B_1*9
10	20	B_0+B_1*10

Where,

Table 1:

Std. Dev. of x	3.02765
Std. Dev. of y	6.617317
Mean of x	5.5
Mean of y	9.7
Correlation between x & y	.989938

If we differentiate the Residual Sum of Square (RSS) wrt. B_0 & B_1 and equate the results to zero, we get the following equations as a result:

*B_1 = Correlation * (Std. Dev. of y/ Std. Dev. of x)*

*B_0 = Mean(Y) – B_1 * Mean(X)*

Putting values from table 1 into the above equations,

$B_1 = 2.64$

$B_0 = -2.2$

Hence, the least regression equation will become –

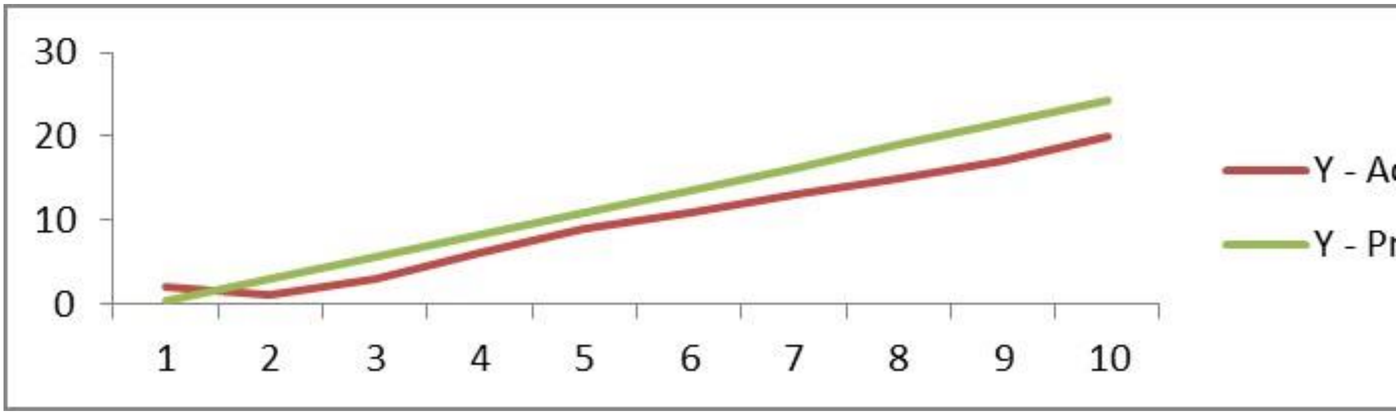
Y = -2.2 + 2.64*x

Let see, how our predictions are looking like using this equation

x	**Y - Actual**	**Y - Predicted**
1	2	0.44

2	1	3.08
3	3	5.72
4	6	8.36
5	9	11
6	11	13.64
7	13	16.28
8	15	18.92
9	17	21.56
10	20	24.2

Given only 10 data points to fit a line our predictions are not pretty accurate but if we see the correlation between ‘Y-Actual’ & ‘Y – Predicted’ it will turn out to be very high; hence both the series are moving together and here is the graph for visualizing our prediction values:



Model Performance

Once you build the model, the next logical question comes in mind is to know whether your model is good enough to predict in future or the relationship which you built between dependent and independent variables is good enough or not.

For this purpose there are various metrics which we look into-

R – Square (R^2)

Formula for calculating R^2 is given by:

$$R^2 = \frac{TSS - RSS}{TSS}$$

- **Total Sum of Squares (TSS) :** TSS is a measure of total variance in the response/ dependent variable Y and can be thought of as the amount of variability inherent in the response before the regression is performed.
- **Residual Sum of Squares (RSS) :** RSS measures the amount of variability that is left unexplained after performing the regression.
- (TSS – RSS) measures the amount of variability in the response that is explained (or removed) by performing the regression

Where N is the number of observations used to fit the model, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

- R^2 ranges from 0 to 1.
- R^2 of 0 means that the dependent variable cannot be predicted from the independent variable
- R^2 of 1 means the dependent variable can be predicted without error from the independent variable
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.20 means that 20 percent of the variance in Y is predictable from X; an R^2 of 0.40 means that 40 percent is predictable; and so on.

ii. **Root Mean Square Error (RMSE)**

RMSE tells the measure of dispersion of predicted values from actual values. The formula for calculating RMSE is

$$R^2 = \{ (1 / N) * \Sigma [(x_i - \text{mean}(x)) * (y_i - \text{mean}(y))] / (\sigma_x * \sigma_y) \}^2$$

N : Total number of observations

Though RMSE is a good measure for errors but the issue with it is that it is susceptible to the range of your dependent variable. If your dependent variable has thin range, your RMSE will be low and if dependent variable has wide range RMSE will be high. Hence, RMSE is a good metric to compare between different iterations of a model.

iii. **Mean Absolute Percentage Error (MAPE)**

To overcome the limitations of RMSE, analyst prefer MAPE over RMSE which gives error in terms of percentages and hence comparable across models. Formula for calculating MAPE can be written as:

$$\text{RMSE} = \sqrt{\frac{\Sigma(Y_{Actual} - Y_{Predicted})^2}{N}}$$

N : Total number of observations

Multiple Linear Regression

Till now we were discussing about the scenario where we have only one independent variable. If we have more than one independent variable the procedure for fitting a best fit line is known as "Multiple Linear Regression"

How is it different

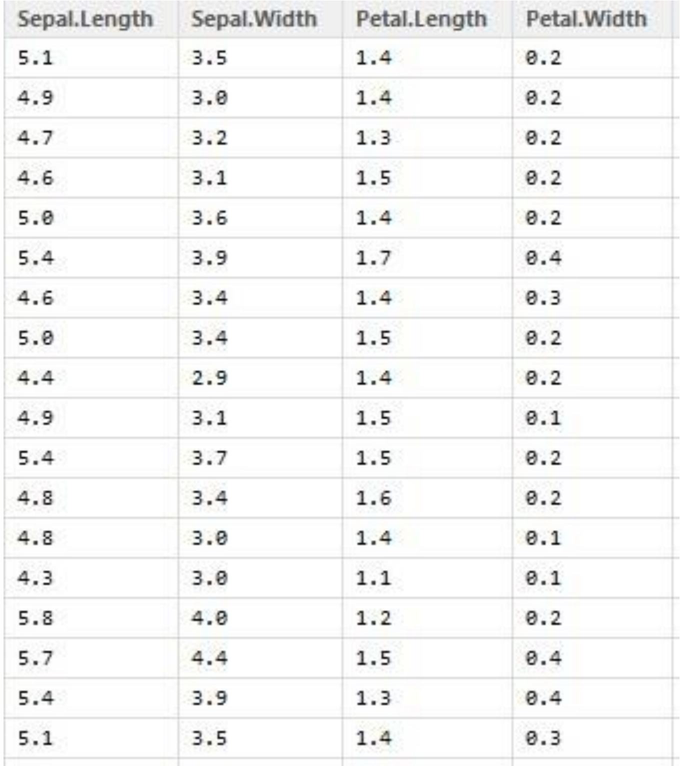
Fundamentally there is no difference between 'Simple' & 'Multiple' linear regression. Both works on OLS principle and procedure to get the best line is also similar. In the case of later, regression equation will take a shape like:

$Y=B_0+B_1X_1+B_2X_2+B_3X_3$.....

Where,

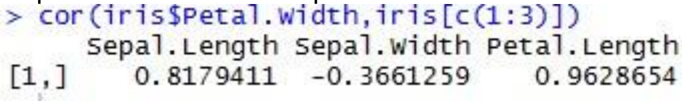
- B_i : Different coefficients
- X_i : Various independent variables

Let's take a case where we have to predict 'Petal.Width' from the given data set (We are using 'iris' dataset which comes along with R). We will be using R as a platform to run the regression on our dataset (If you are not familiar with R, you can use Microsoft Excel as well for your learning purpose). Firstly, let see how data looks like-



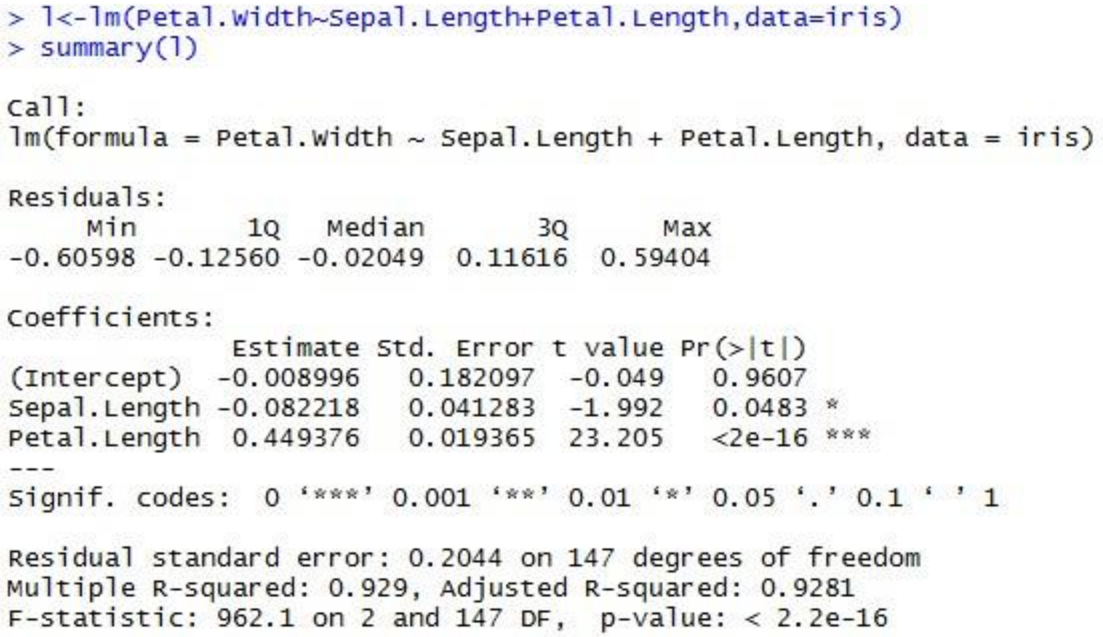
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
5.0	3.4	1.5	0.2
4.4	2.9	1.4	0.2
4.9	3.1	1.5	0.1
5.4	3.7	1.5	0.2
4.8	3.4	1.6	0.2
4.8	3.0	1.4	0.1
4.3	3.0	1.1	0.1
5.8	4.0	1.2	0.2
5.7	4.4	1.5	0.4
5.4	3.9	1.3	0.4
5.1	3.5	1.4	0.3

We can see that there are 4 variables in our data out of which **Petal.Width** is a dependent variable while rests of them are predictors or independent variables. Now, let's calculate correlation between our dependent & independent variables:



```
> cor(iris$Petal.Width,iris[c(1:3)])
     Sepal.Length Sepal.Width Petal.Length
[1,]    0.8179411  -0.3661259    0.9628654
```

From the correlation matrix we can see that not all the variables are strongly correlated to *Petal.Width*, hence we will only include significant variables to build our model i.e. '*Sepal.Length*' & '*Petal.Length*'. **Let's run our first model:**



```
> l<-lm(Petal.Width~Sepal.Length+Petal.Length,data=iris)
> summary(l)

Call:
lm(formula = Petal.Width ~ Sepal.Length + Petal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.60598 -0.12560 -0.02049  0.11616  0.59404 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.008996   0.182097  -0.049   0.9607    
Sepal.Length -0.082218   0.041283  -1.992   0.0483 *  
Petal.Length  0.449376   0.019365  23.205   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2044 on 147 degrees of freedom
Multiple R-squared: 0.929, Adjusted R-squared: 0.9281 
F-statistic: 962.1 on 2 and 147 DF,  p-value: < 2.2e-16
```

Here, Intercept estimate is same as B_0 in previous examples, while coefficient values written next to the variable names are nothing but our beta coefficients (B_1, B_2, B_3 ....Etc.). Hence we can write our linear regression equation as:

Petal.Width = -0.008996 – 0.082218*Sepal.Length + 0.449376*Petal.Length

When we run a linear regression, there is an underlying assumption that there is some relationship between dependent and independent variable. To validate this assumption, linear regression module validates the hypothesis that "**Beta coefficient B_i for an independent variable X_i is 0**". The P-Value which we are seeing in the last column is nothing but the probability of this hypothesis being true. Generally if P-Value is less than or equal to 0.05 we consider this hypothesis to be false and establish a relationship between dependent and independent variable

Multi-collinearity

Multi-collinearity tells us the strength of relationship between independent variables. If there is Multi-Collinearity in our data, our beta coefficients may be misleading. VIF (Variance Inflation Factor) is used to identify the Multi-collinearity. If VIF value is greater than 4 we exclude that variable from our model building exercise

Iterative Models

Model building is not one step process, one need to run multiple iterations in order to reach a final model. Take care of P-Value and VIF for variable selection and R-Square & MAPE for model selection.