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).

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 ($\sigma 2$)– 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

- Dependent variable is continuous
- Linear relationship between Dependent Variable and Independent Variable.
- No Multicollinearity (no relationship between Independent variables)
- Residuals should follow Normal Distribution.
- Residuals should have constant variance: Homoscedasticity
- Residuals should be independently distributed/no autocorrelation

To check relationship between dependent and independent variable:

1. Perform Bivariate Analysis
2. Calculate Variance Inflation factor: value which is closer to 1 and till maximum 4

To find whether residuals are normally distributed or not:

1. Perform Histogram/ Boxplot
2. Perform Kolmogorov Smirnov K’s test

To check Homoscedasticity:

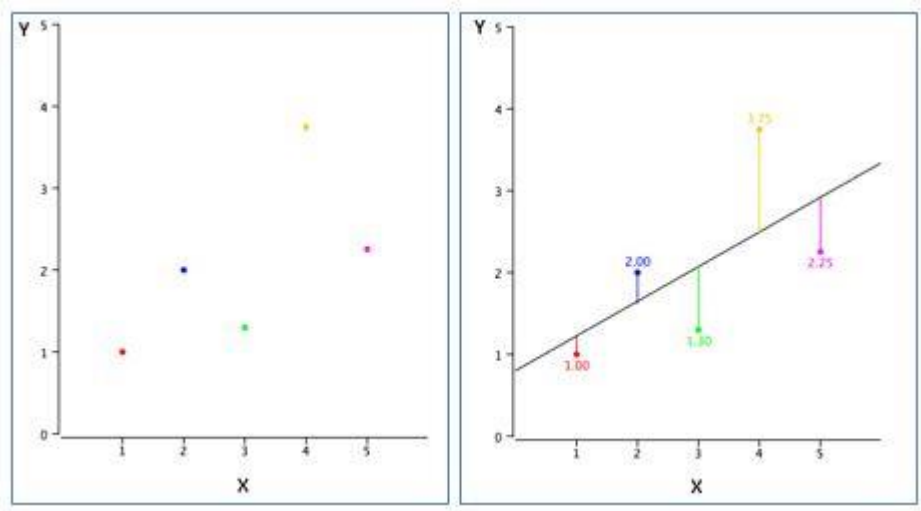
1. Plot Residuals Vs. Predicted values, and there should be no pattern.

This file is meant for personal use by jainharshal1997@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

2. Perform Non Constant Variance Test.

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:

$$Y = \beta_0 + \beta_1 X$$

Where,
Y = Dependent Variable
X = Independent Variable
β_0 = Constant term a.k.a Intercept
β_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)

This file is meant for personal use by jainharshal1997@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

- Regression line minimizes the sum of "Square of Residuals". That's why the method of Linear Regression is known as "Ordinary Least Square (OLS)".
- β_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.

Finding a Linear Regression Line

Using a statistical tool e.g., Excel, R, SAS etc. you will directly find constants (β_0 and β_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

This file is meant for personal use by jainharshal1997@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

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.

This file is meant for personal use by jainharshal1997@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

i. 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, σx is the standard deviation of x, and σy is the standard deviation of y.

- R2 ranges from 0 to 1.
- R2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- R2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.20 means that 20 percent of the variance in Y is predictable from X; an R2 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

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

N: Total number of observations

This file is meant for personal use by jainharshal1997@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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:

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

This file is meant for personal use by jainharshal1997@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.