**Linear Regression**

Regression Objectives:

1. If predictor variable(s) can predict the dependent variable
2. Variables which can be significant predictors of the outcome variable

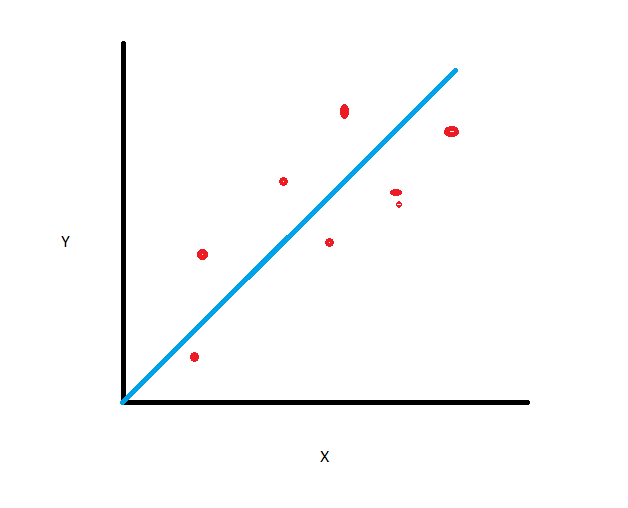
X: Input or predictor variable

Y: Output variable we want to predict

Y1=a1+b1X+e

The objective is to minimize the difference between Y & Y1, by finding optimum values for a1 & b1. This will give us the best fit line in the graph that can be used to predict Y Values based X

To estimate the optimum value of a1&b1, we use OLS (Ordinary Least Squares) by minimizing the sum of squared difference between Y & Y1



**Assumptions:**

* Linearity: Relationship between X & Y is linear
* Homoscedasticity: When noise is same across all values of independent variable (X). Variance around regressor line is same for all values of a predictor.
* Independence of errors
* Normality: Independent variables should be normally distributed for better results
* No multi collinearity between different features
* Both variables should be quantitative

**R-squared**: Statistical measure to show how the regression line is the best fit line for all the real data points: = variance explained by model/ total variance. It represents the proportion of variance for a dependent variable that is explained by an independent variable. Example, if R-squared value = 0.70, then 70% of the observed variation can be explained by mode.

Hence, larger the R-squared, better the regression model fits the observations. It can never be 100%, as that would mean all data points fall on the regressor line.

However, R-squared value always does not give you correct picture (except when assumptions of linear regression are fulfilled). It needs to be considered in combination with other residual plots, statistics, eg: Adjusted R-square & Predicted R-square.

Also, a low R-values model is always bad & high R-value model is always good is also not true. There are many factors affecting such behavior. Independent variables selection plays major role in such scenarios. Correlation between independent variables & impact of

**Exceptions**: For high R-squared value, when taking physical process into consideration & have very accurate measurements already. For low R-squared value, if you are working on human behavior, it would generally have low r-square as there is lot of inexplicable variations.

**Reasons when R-Square can shoot high:**

* R-square is a biased estimator: There is a high chance for the R-squared value to be higher or lower than expected value. One would need to adjust the R-squared value till the time it becomes unbiased estimate of population value, this brings in the concept of adjusted R-square, To determine the correct adjustment/shrinkage, one need many samples per term need less shrinkage.
* Overfitting of model: It’s a condition when model works on random data error instead of correlation between variables. R-squared value is very high in such scenarios. Predicted R-squared can help in detecting overfitting.
* Independent & dependent variable should not be derived from each other: Basically, they should not mean the same thing directly/indirectly. Ex: Units & revenue. Because, obviously when the unit will be more, revenue will be more.
* Time series data: If the dependent & independent variables both are dependent on time, then the correlation will increase & R-squared value will inflate.
* Independent variables selection: For larger datasets, one should have subject knowledge to check if independent variable is significant or not for the model. Even if the variable has no significance with the dependent variable, you can get relationship because of larger dataset and hence R-squared value will get inflated.

**Errors are squared:**

Because prediction can be either above or below the true value, giving negative or positive difference, respectively.

Squaring, penalizes large difference and hence minimizing the squared errors, guaranteeing a better model

**Cost function (J) of linear regression** is the Root Mean Squared Error(RMSE) between predicted value & actual value.

**Types of Regression:**

Simple Linear Regression: 1 dependent & 1 independent variable

Multiple Linear Regression: 1 dependent & 2+ independent variables

Logistic Regression: 1 **binary** dependent & 2+ independent variables

Ordinal Regression: 1 **ordinal** dependent & 1+ independent variable

Multinomial Regression: 1 **nominal** dependent & 1+ independent variable

**Terms:**

Independent Variable (Predictors or Regressors): Variables which act as inputs for dependent variables.

Dependent Variable (Target): Output variables that get derived from the model based on the independent variables.

Residual (Error): Difference between predicted value & actual value of dependent variable.

Correlation: Strength between two variables relationship

Adjusted R-square: It corrects the R-squared value till the time it becomes an unbiased estimate of the population value.

Deterministic Component: For eq: Y1=a1+b1X+e, deterministic component is a1+b1X

Stochastic Component: e in the above eq is stochastic component

**Scikit versus Stats Model Libraries**

|  |  |
| --- | --- |
| **Scikit** | **Stats Model** |
| More widely used | - |
| Conventionally, its for Machine Learning (esp, prediction) | Conventionally, its for advanced statistics |
| Machine-Learning, Data Science | Econometrics, Generalized Linear Models, Time Series Analysis |
| Provides easy algorithms that requires data to be organized in right way. Then one can run classification, regression or clustering method | Works best for linear regression models & other complex linear models. |
| Package is very well documented | New package, hence documentation isn’t great |
| It follows simple approach(fit->transform->predict). Its more “Pythonic” | Syntax similar to R language. |
| Gives R^2, MSE etc. | Gives other diagnostic values like: p-value, t-tests, standard errors of each parameters |
| Adds intercept on its own for best fit | Doesn’t add on its own, need to add it |