LINEAR REGRESSION –CHAPTER 3

Linear regression is a Supervised learning technique which is-

1. Simple and good
2. Assumes ***linear Dependencies*** between Xi(inputs) and target variable(Y).
3. True Regression functions are never **Linear**.(Actual relation b/w Xi and Y )
4. Linear Regression generates a model which is mostly incorrect but does a good **Approximation** of the true Function.
5. Actually Simple Models and algorithms are good .
6. Not always complicated , flexible and fancy methods are good.(good only under some circumstances)

Questions that a regression Model might ask ?

1. Is there a relationship between the Target(Response-Y) and Predictors Variables(Xi).
2. How strongly the variables of the Data set are related and dependent on each other.-Using Correlation coefficient ‘r’-which is nothing but a measure of the strength of relationship b/w between X and Y variables.
3. How changes in X changes Y in what magnitude-i.e explained by the slope’m’
4. Is the relationship linear?
5. Is there a predictor which is releted to Response variable ?

Etc

Regression Equation

Y = a + bx + e , e =errors & noise in data = a equation of a straight line.

a , b are regression coefficients or parameters

a = y-intercept , b = slope of the line ,multiplicative change in Y with 1 unit change in X var.

**e = Residuals** = ( Yactual - Ycalculated ) , difference b/w the actual Y value of the training data and the Y values which the model calculated .

NOTE- The Residuals should always be Normally Distributed , ie Bell shaped curve, highest frequency at 0

Error = RSS-Residual Sum of Squares = e1^2 + e2^2 + e3^2 …… ei^2 = Least squared Errors =DEVIANCE

RSS = sum( (Yiactual – Yical)^2)

Our main aim is to have a regression line which Minimize the **RSS value , i.e least squared Error** value.(Least number of mismatches)

Slope is given by formula = r \* SD(Y)/SD(X) , r =cor coeff , SD = standard Deviation

Also the Standard Errors of the Regression coefficients a ,b should be least and significant

If slope = 0 , it means X and Y have no relation amongst them and are independent.

Slope is directly dependent on the ‘r’ value i.e correlation coefficient which explains the Linear relationship b/w X and Y , how strong it is etc.

‘r’ value lies b/w -1 < r < 1 , r = 0 (no relation) , r > 0 (positive correlation) , r < 0 (inverse relation)

Confidence Interval – It is the interval which says that how confident we are regarding a population unknown true parameter will lie within that interval.

Eg 95 % CI for slope b- says that there is approximately 95% chance that the interval will contain the true value of slope(b) .

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter

**NOTE-Confidence intervals are a "frequentist" concept: the interval, and not the true parameter, is considered random. A 95% confidence interval is a random interval that contains the true parameter 95% of the time .**

**HYPOTHESIS TESTING- Any statement or conclusion about a population parameter**

1. **H0-NULL HYPOTHESIS – X and Y variables have no relation , ie slope b = 0**
2. **HA-Alternate Hypothesis – X and Y are Related , b != 0**

***To test Null hypothesis we use t-test.***

**t-statistics should be large enough as possible to reject Null hypothesis and accept that A relation b/w X and Y exists . t-value should be greater than 2 for 5% significance level.**

**and the corresponding p-value should always be < level of significance(mostly 5% or 1%) to reject Null hypothesis.**

**p-value and t-statistic values can be used to find out that which input features are related to Y(target) or not.**

**If these values are not significant for a particular Input(Xi) then we can remove that feature from the Model .**

**RSE(Residual Standard Error) = SQRT(RSS)**

**R^2 = Goodness of fit measure , also tells us about the amount of variance explained in Y due to X. = 1 – TSS/RSS , TSS = true error from mean(y) values or = ‘r’^2 (r=correlation coeff)**

**MULTPIPLE REGRESSION – Multiple Predictors (X –inputs)- will fit a Hyperplane**

**Y = a + b1X1 + b2X2 + b3X3 . .. . . .**

**a,b1,b2,b3 are the regression coefficients**

**NOTE-The ideal scenario is when the Predictors(INPUTS) are Independent of each other i.e they have no relation with each other and not correlated.**

**Correlation amongst Predictors(Input variables) actually causes lots of problems.**

**In order to make a Causal statement for the effect of Xi on output variable(Y) , it is important that we keep other Inputs(X) fixed at that point.**

**Which means to study the Causal statements it is better to have a Univariate Regression Model with only one Predictor(X) than multiple to study the effects of a input precisely on The output (response-Y) value.**

Adding lots of extra predictors to the model can just as easily muddy the interpretation of β^1 as it can clarify it.Because most of the times the Inputs (Predictors) would be correlated with each other which would affect each other’s relationship with Y in multiple Regression.

**F-Statistic** – It actually tells us about the effect of Predictors on the Outcome(Y variable)

F = (TSS – RSS/ p )/ (RSS / n-p-1) , n = data points in sample , p = no of parameters

DECIDING WHAT ARE THE IMPORTANT VARIABLES (PREDICTORS) & MODEL SELECTION TECHNIQUES

1. **Forward Selection Technique –**

**-Start with a Null model with no variables**

**-Start adding variables(X) to the model one by one and access the Model’s Perfomance via least Squared Errors.**

**-Finally Select the model with least Error and most accuracy and which best fits the training data.**

1. **Backward Selection –**

**-Start a Model by including all variables(Xi) in it.**

**-Remove the variable with largest p-value or least t-statistic (which implies H0 is accepted that X and Y are not related)**

**-Create multiple models with such pruning of variables which are not significant and effective**

**-Stop when a stopping rule is reached, eg all predictors (Xi) in the model have a significant p-value and large t-statistic which means they all are related to the Y variable.**

1. **AIC, BIC ,R^2 values and Cross Validation for The best Regression Model Selection with least Squared Errors and highest Accuracy and also tells us the Goodness of fit , i.e is how well the model fits the data.**

**R^2 the higher the better – the amount of variance explained**

**HOW TO DEAL WITH NOMINAL/CATEGORICAL INPUT FEATURES**

**What happened when we have factor variables – ie ‘qualitative’, which takes a finite set of values. E-g :Yes/No , Married/Not Married, 0-10 , 0/1 etc, gender –Male/Female**

**We convert such categorical values to 0 or 1 .**

**-If we have got K levels in a categorical variable we will make K-1 Dummy variables each having 0 or 1 values**

**The level with no dummy variable is called *Baseline* Level.**