Problem Statement - Answer the following questions to the best of your knowledge including the concepts taught to you in the level.

1. On what basis we choose data scaling method (Normalization/Standardization)?

2. If the VIF is 2 then what is value of correlation coefficient (r^2)

3. How do you interpret chi-square result?

4. Why do we choose boxplot method than other for outlier detection and removal?

5. How do we choose best method to impute missing value for a data?

1. On what basis we choose data scaling method (Normalization/Standardization)?

Solution:

Feature Scaling methods are used to to keep all the features in a dataset within a range so that none of the variable dominate in predictions of the model.

Following Techniques are used for feature Scaling:

1. Normalization:

It focuses on rescaling the values of features in the range from 0 to1 without distorting differences between them. Normalization uses the minimum and maximum values of the feature to accomplish rescaling in range of 0 to 1 . Thus using normalization we can bring dataset to a common scale.

Xnormalized  = (Xi - Xmin) / (Xmaximum - Xminimum)

1. Standardization:

Standardizing the features around the center and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bais.

Xstnadardizes = (XI - XMEAN) / XSD

We should consider following points while selecting scaling method:

1. The data distribution :

Data distribution is very important factor to be considered in selecting scaling method.

If the data is normally distributed (Bell Curve) i.e if data lies equally to both the sides of mean at given multiples of SD then data is normally distributed.

If data is normally distributed then we should go for Standardization and not Normalization. Since Standardization will give us location of every data-point in multiple of SD from mean.

1. Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks.
2. Standardization assumes that data is normally distributed so it must be used when the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression, and linear discriminant analysis.

2. If the VIF is 2 then what is value of correlation coefficient (r^2)

Answer:

Pearson’s correlation coefficient (R) is an indicator which shows linear relationship between two variables.

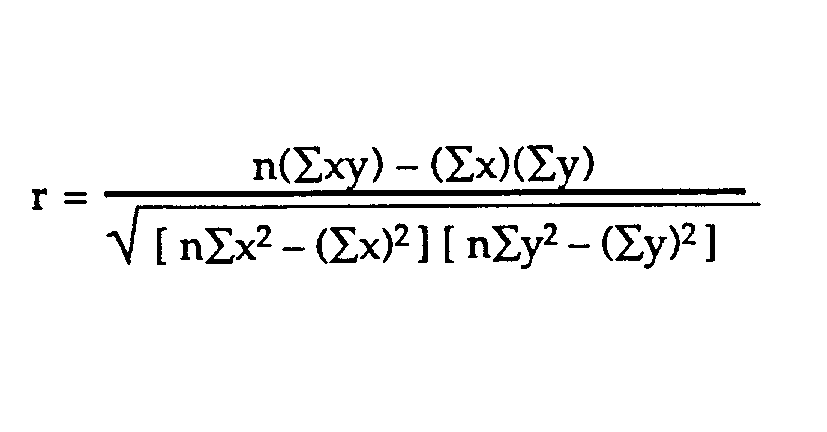
Range of R is -1 to 1

-1 denotes extreme negative correlation

0 denotes no correlation

1 denotes extreme positive correlation

It is calculated as follows:



VIF (variance Inflation Factor ):

Variance Inflation Factor (VIF) is a common simple stat used to quantify multicollinearity in regressions. It is calculated for each covariate in a regression, with higher values meaning that the covariate is more colinear with the other covariates. It [technically](https://en.wikipedia.org/wiki/Variance_inflation_factor)measures “how much the variance of an estimated regression coefficient is increased because of collinearity.” The equation is:

VIF = 1/1-R2

If ,

VIF = 2

Then,

2 = 1/1-R2

2(1-R2) = 1

2 – 2R2 = 1

2R2 = 1

**R2  = 1/2**

3. How do you interpret chi-square result?

**Answer:**