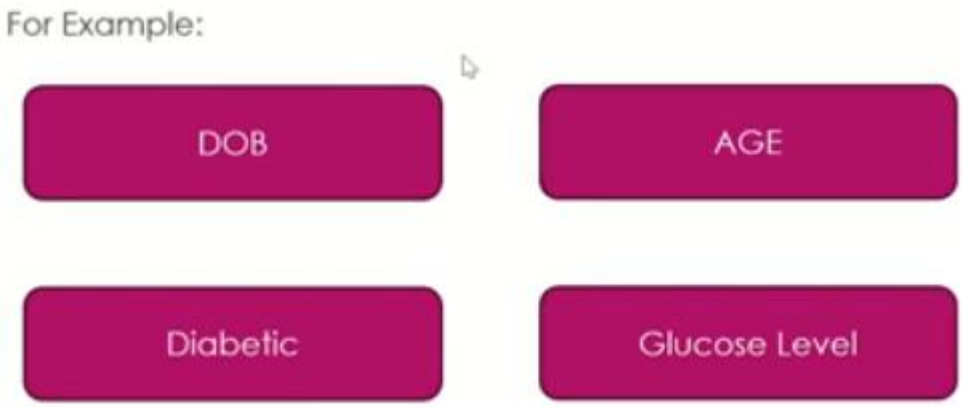
Multicollinearity

Multicollinarity is a statistical phenomenon where two or more independent variable in a linear regression model are strongly correlated. It means that the variables have an almost perfect or exact relationship between them.

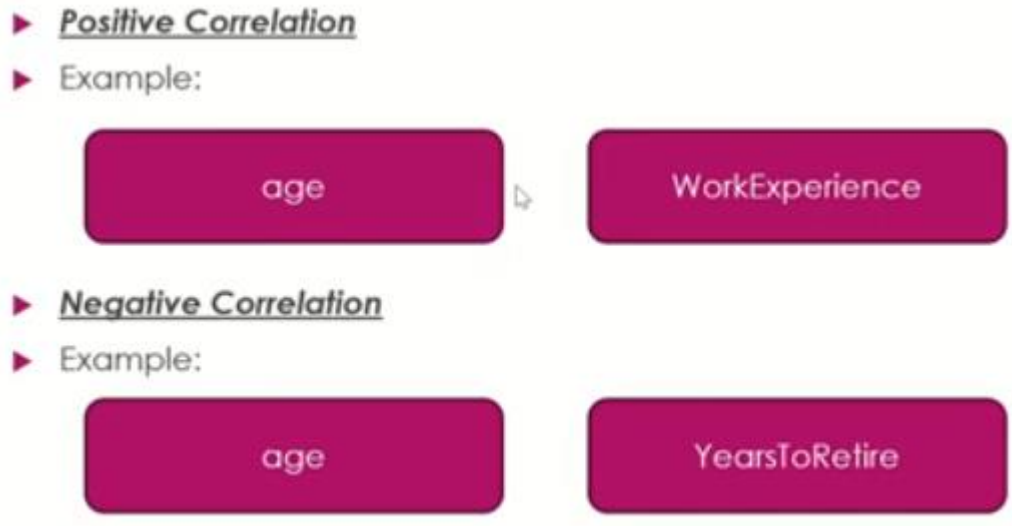


- Predict salary of a person. Salary is dependent variable. DOB, Age is feature.
- As DOB has age will understand. As age data has will get DOB data. here two feature don't required because both are correlated. so we will remove any one data.
- Same like Diabetic and glucose level, both are co related. We will remove diabetic's data.

Types of Multicollinearity

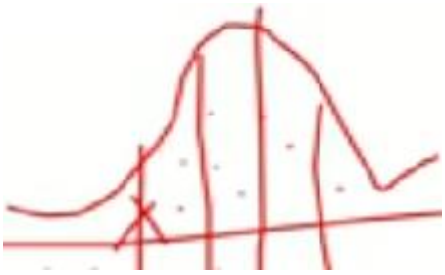
- **Positive Correlation**
- **Negative Correlation**

Predict salary, two feature is age and work experience. when age increase work experience increase vice versa.



Why Multicollinarity is a problem?

- y=mx+c
- y=B0+B1X1 + B2X2 + … + Bn Xn
- Sales = 0.5+ 0.8Advertiemnt + 0.45Schemas + 0.1TVAdvertisement
- Result in huge swings based an independent variable with in a modal and reduce the strength of the coefficient's used with in a model
- Causes the coefficient estimates of the model (and even the sgns of the coefficients) to fluctuate significantly based on which other predictor variable are included in the model.
- Reduces teprecisn of the coeffcen estimatesm which makes the p-values unreliable.



Step to Detect Multicollinearity

- Scatter plot and Correlation Matrix
- Correlltion heat map
- VIF – Vriation inflation factor

Step to Avoid multicolinarity

- Set VIFvalue and remove variables abouve the value
- Use Regularization techniques like Rdege, Lasso and Elastic Net
- Using Haeatmap / Correlation matrix detect the highly collinear variable and remove them manually
- Using feature engineering combine te correlated variable

Homoscedasticity and Heteroscedasticity

In statistical models, homoscedasticity refers to the condition where the variance (spread) of the error terms (residuals) is constant across all

levels of the independent variable(s), while heteroscedasticity refers to the condition where this variance is not constant.

When to use Homoscedasticity:

- **Assumptions of Ordinary Least Squares (OLS) Regression:**

  Homoscedasticity is a key assumption of OLS regression, along with normality and independence of errors. When this assumption is met, OLS regression provides unbiased, efficient, and consistent estimates of the regression coefficients.

- **Valid Inference:**

  When the error term has constant variance, the standard errors of the regression coefficients are accurate, and statistical tests like t-tests and F-tests are valid for inferring the significance of the independent variables.

When to use Heteroscedasticity:

- **Identifying Variance Issues:**

  Heteroscedasticity indicates a violation of the homoscedasticity assumption and can lead to inaccurate standard errors and biased inferences.

Step to detect Multicollinearity