What exactly happens in Multicollinearity ?

When multicollinearity is present in a model, it can lead to several issues, including:

1. Difficulty in identifying the most important predictors: Due to high correlation between independent variables, it becomes challenging to determine which variable has the most significant impact on the dependent variable.

2. Inflated standard errors: Multicollinearity can lead to larger standard errors for the regression coefficient, which decrease the statistical power and can make it challenging to determine the true relationship between the Independent and dependent variables.

3. Unstable and unreliable estimates: The regression coefficient becomes sensitive to small changes in the data, making it difficult to interpret the results accurately.

Perfect Multicollinearity

Perfect multicollinearity occurs when one independent variable in a multiple regression model is an exact linear combination of one or more other independent variables. In other words, there is an exact linear relationship between the independent variable, making it impossible to uniquely estimate the individual effects of each variable on the dependent variable.

$$\boxed{percent = 10 \times cgpa + error}$$

Corr linear

$$\boxed{x_1 = a_1 x_2 + a_0 + error}$$

example

add deciplinemarks

Cgpa	percent	Lpa
8.5	83	-
9.12	95	-

Perfect collinearity

$$\boxed{x_1 = a_1 x_2 + a_0}$$

example

Cgpa	percent	Lpa
8.5	8.5	-
9.12	91.2	-

$$\boxed{percut = 10 \times Cgpa}$$

Example

Cgpa	percent	lpa
8	80	3
6	60	4

$$\boxed{lpa = \beta_0 + \beta_1 Cgpa + \beta_2 percent + error}$$

OLS →

$$\boxed{\beta = (X^T X)^{-1} X^T Y}$$

→ design Matrix

$$\underbrace{\begin{bmatrix} 1 & 8 & 80 \\ 1 & 6 & 60 \end{bmatrix}}_{\text{input data}}$$

3x2 2x3

$$\underset{X^T}{\begin{bmatrix} 1 & 1 \\ 8 & 6 \\ 80 & 60 \end{bmatrix}} \begin{bmatrix} 1 & 8 & 80 \\ 1 & 6 & 60 \end{bmatrix} = \underset{X^T X}{\begin{bmatrix} 2 & 14 & 140 \\ 14 & 84 & 8400 \\ 140 & 8400 & 840000 \end{bmatrix}}$$

* before finding Inverse check Determinate (det)

$$\begin{bmatrix} 2 & 14 & 140 \\ 14 & 84 & 8400 \\ 140 & 8400 & 840000 \end{bmatrix}$$

Det = $2(0) - 14(0) + 140(0) = 0$

if det is zero then it is singular matric and we can't find inverse. and also $|\beta|$ not find.

Proofe

if our data has multicollinearity

→ unstable coefficient } Problem

→ High S.E

Formula

$$\boxed{\beta = (X^T X)^{-1} X^T Y}$$

$$\mathrm{Var}(\beta) \text{ or } SE(\beta)$$

Underroot of var(β)

$\beta_0 \quad \beta_1 \quad \beta_2$

$SE(\beta_0) \quad SE(\beta_1) \quad SE(\beta_2)$

$$\boxed{SE(\beta) = \sqrt{\mathrm{diag}(\sigma^2 \underline{(X^T X)}^{-1}}}$$

$$\boxed{\beta = (X^T X)^{-1} X^T Y}$$

→ perfect multicoli → $\det(X^T X) = 0$

not inverse

→ Strong multicoli

$\det(X^T X) \to$ very small

$$\text{inverse} = \frac{1}{\det} \begin{bmatrix} - & - & - \\ - & - & - \\ - & - & - \end{bmatrix} \to \text{high}$$

very small

inflate

How to remove multicollinearity

1. **Collect more data:** In some cases, multicollinearity might be a result of a limitted sample size. Collecting more data, if possible, can help reduce multicollinearity and improve the stability of the model.

2. **Remove one of the highly corellated variable:** If two or more are highly corelated, consider removing one of them from the model. This step can help eliminate redundancy in the model and reduce multicollinearity. Choose the variable to remove based on domain knolwedge, variable importance, or the one with the highest VIF.

3. **Combine Corelated Variable:** If correlated Independent variable represent similar information, consider combining them into a single variable. This combinatio can be done by average, summing or any other mathematical operations, depending on the content and the nature of the variables.

4. Use partial least square regression (PLS): PLS is a technique that combines features of both principal component analysis and multiple regression. It Identifies linear combinations of the predictor variables (called latent variable) that have the highest covariance with the response variable, reducing multicollinearity while retaining most of the predictive power.

* First find PCA → then →LR