Multi collinearity

Multi collinearity is a statistical phenomenon that occur when two or more independent variable in a multiple ~~linear~~ regression model are highly correlated. In other words, these variables exhibit a strong linear relationship, making it difficult to isolate the individual effects of each variable on the dependent variable.

$\boxed{\text{corr}}$ → linear regression

eg:- 1

cgpa	iq	lpa
8	80	8

iq ↑ → cgpa ↑

2

iq	backlogs	lpa
↑↑	↓↓	

corr.

3

cgpa	dob	lpa

not corr

$$\overset{??}{lpa} = \beta_0 + \beta_1 \overset{1}{cgpa} + \beta_2 \overset{c}{iq}$$

How its work?

1. let iq is constant
2. and if cgpa is increase from 1.
3. Then how much lpa increase.

But in Multicollinearity how its work?

(1) let assume iq is constant

(2) and cgpa increase from 1.

(3) If cgpa and iq is multicollinearity then iq is also increase from 1 (no constant)

(4) then this formula is not working

* **When is Multicollinearity bad?**

1. **Inference:**

* Inference focus on understanding the relationship betn variable in a model.

eg:- cgpa | iq | lpa

* How much cgpa and iq perform role to give output data (lpa)

let assume

cgpa	iq	lpa
perform 75%.	perform 25%.	to give output

eg:- (in game)

User activity	ban / unban

User activity → bullet shots, Running, guns

if user use auto bulletshot Hack ⇒ Ban

perform 75%.

if user not using Hack ⇒ Unban

Prediction

* Prediction focuses on using a model to make accurate forecasts or estimate for new, unseen data.

eg:- learn from old or train data set to predict new unseen data.

** Multicollinearity doesn't affect the model when you are building a predictive model. But if you are using for inference (find the relationship betn input and output) then multi-collinearity is bad.

x_1	x_2	y

Explaining

In Prediction → $\boxed{x_1 = a_0 + a_1 x_2 + \lambda}$ → Reduced Error

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + t$$

$$y = \beta_0 + \beta_1(a_0 + a_1 x_2 + \lambda) + t$$

$$y = \beta_0 + \beta_1 a_0 + \beta_1 a_1 x_2 + \beta_1\lambda + t$$

$$y = \underbrace{(\beta_0 + \beta_1 a_0)}_{\text{constant}} + \beta_1 a_1 \underbrace{x_2}_{x_2} + \underbrace{(\beta_1\lambda + t)}_{\text{some error}}$$

$\boxed{x_1 \text{ not in new form}}$

* which means automatically convert into x_2 in multicollinearity in prediction

Simple Linear Regression

direct formula

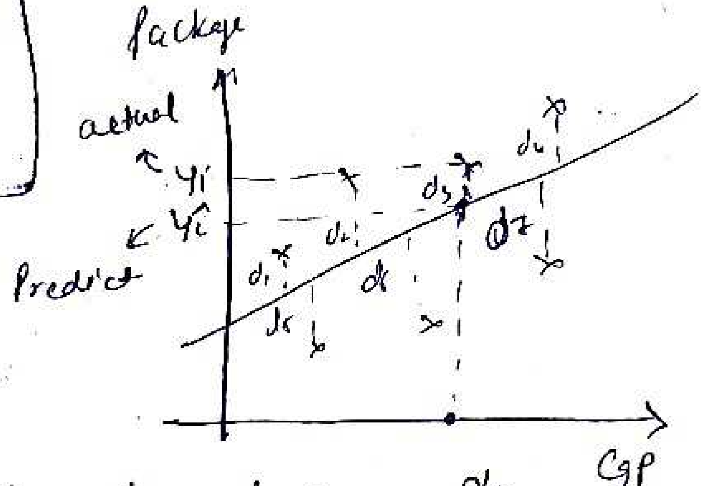
↓

Closed form — Non-closed form

OLS — Gradient Descent

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b = \bar{y} - m\bar{x}$$



$$E = d_1 + d_2 + d_3 + \cdots d_n$$

$$E = d_1^2 + d_2^2 + d_3^2 + \cdots d_n^2$$

$$\boxed{E = \sum_{i=1}^{n} d_i^2}$$

$$d_i = (y_i - \hat{y_i})$$

$$\boxed{E = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2} \rightarrow \text{Avg Error}$$

$$\boxed{E = \sum_{i=1}^{n} (y_i - \hat{y_i})^2} \rightarrow \text{Total Error}$$

$$\boxed{\hat{y}_i = mx_i + b}$$

$$\boxed{E(m,b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2}$$

Types of Multicollinearity

1. **Structural Multicollinearity** :- Multicollinearity create by data scientist.

eg:- ① Apply One Hot encoding

category

City
Delhi
Mumbai
Kolkata

apply One hot encoding

Delhi	Mumbai	Kolkata
1	0	0
0	1	0
0	0	1

⇒ { Perfect Multicollinearity }

↳ β ⇒ we cannot find β because Perfect Multicollinearity

k = 1 - Delhi - Mumbai

k = 1 - 1 - 0

$\boxed{k = 0}$

↳ Kolkata

② If we apply polynomail regression

② **Data driven multicollinearity** :- Data-driven multi-collinearity occur when the independent variables in the dataset are highly correlated due to the specific data ~~due~~ being analysed. In this case, the high correlation between the variables is not a result of the way the variables are defined or the model is constructed but rather due to the observed data patterns.

Flat data

Size	no. of washroom

↳ data-driven multicollinearity

How to detect Multicollinearity

- Corr
- VIF — Variance inflation factor
- Condition Number → more than 30 = multicollinearity

① Correlation

↳ using $\boxed{\text{cov()}}$ function

② Variance Inflation factor

eg:- We have dataset

TV	NP	radio	sales

x_1 x_2 Independent   dependent $\rightarrow y$

assume

y (dependent)

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

If linear relation have then linear regression bta dega.

TV	NP	radio	sales

x_1 x_2

Independent

dependent

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

TV	NP	radio	sales

x_2

x_1

y   Independent

dependent

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

$\boxed{\text{3 input}}$

↓

3 LR (Linear Regression)

↳ R2 score

$$VIF = \frac{1}{1-R^2}$$

if VIF value (greater than 5 or 10, depending on the context) then multicollinearity exist.

if VIF value is near 1 then multicollinearity doesn't exist.

Code

```
from statsmodels.stats.outliers_influence
                    import variance_inflation_factor

vif = []

for i in range(3)
        vif.append(variance_inflation_factor(df.iloc
                                          [:, 1:4], i))
```

```
pd.DataFrame({'vif': vif}, index = df.columns[1:4]).T
```

How to remove Multicollinearity

(1) collect more data :- In some cases, multicollinearity might be a result of a limited sample size. Collecting more data, if possible can help reduce multicollinearity and improve the stability of the model.

(2) Remove one of highly correlated variable.

(3) Combine correlated variable

(4) use Partial least squares regression

↳ $\boxed{PCA} \rightarrow \boxed{LR}$:

$\boxed{x_1 \quad x_2} \xrightarrow[PCA]{} \boxed{x'_1 \quad x'_2}$