COLLINEARITY

→ When one regressor is highly correlated with another regressor. (Refer to 1st example)

OR (Refer to 2nd example)

→ When one regressor is highly correlated with a linear combination of other regressor.

Example →

$$\begin{array}{ccc} C_1 & C_2 & C_3 \\ \end{array}$$
$$\left[\begin{array}{ccc} 1 & 2 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array}\right] \rightarrow C_2 = 2C_1$$

In this if we take 2 times of C_1, then it is C_2. So it collinearity

$$\begin{array}{ccc} C_1 & C_2 & C_3 \\ \end{array}$$
$$\left[\begin{array}{ccc} 1 & 0 & 2 \\ 1 & 0 & 2 \\ 0 & 1 & 4 \\ 0 & 1 & 4 \end{array}\right] \rightarrow C_3 = 2C_1 + 4C_2$$

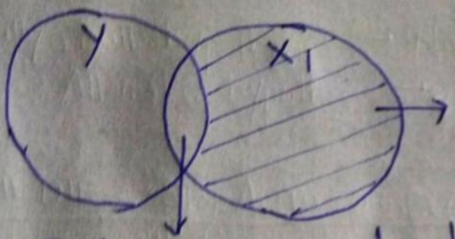
2 times C_1 addition to 4 times C_2 is C_3. So it is collinearity

Why is Collinearity a problem?

→ Each regressor is trying to "tell a story" about the dependent variable. The p-values, size etc reflect how well the story overlaps.

Suppose we want to find out what cause road accident. Suppose we investiage a person.

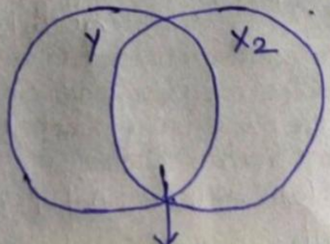
Y (dependent variable) → Road accident (Yes), X1 (Dependent variable) → Reason of a person who met with an accident.



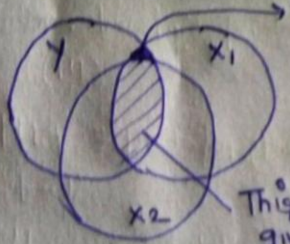
Reasons which may not lead to the accident. ~~So~~ May be bluff reasons/ reason not related.

True/Exact reason why it met with an accident. So p value will have that value which is only affected / reasons which may lead to accident.

· X2 → Another person.

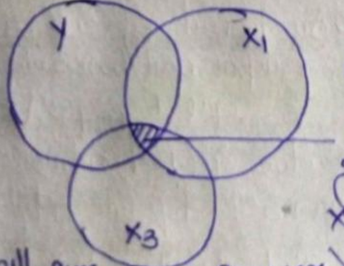


Here we find more reasons which can lead to accidents. Here we will have more P value.



Only this much new information is given compare to X2 in regards to accident.

This is the common info given in regards to accident.



Only this much common information given by X1 and X2 towards Y.

→ Perfect collinearity - No unique solution (all regressor will give same answer)

High collinearity - "Woobly" estimates. They will have high variances.

TECHNIQUE TO CHECK COLLINEARITY →

Variation Inflation Factor (VIF) identifies correlation between independent variables and the strength of that correlation.

zedstatistics

MULTI COLLINEARITY → Correlations , Variance Inflation factors (VIFs)

Intution → Lawyer's salary = $\beta_0 + \beta_1$ (Years Experience) + β_2 (Age) + ε_i

- Multicollinearity occurs when the X variables are themselves related
  More experience you have as a lawyer, more is the age of lawyer.
- Regression model try to find single / individual effects.
- Individual effects should be considered.

β_1 → Marginal effect on salary of 1 additional year experience, holding other variable constant.

β_2 → Marginal effect on salary of 1 additional year of age, holding other variable constant.

But β_1 and β_2 are interrelated. Other is increasing, if one is increased (cannot be constant). This is called multi collinearity.

→ Multicollinearity generally occurs when there are high correlations between two or more predictor variables. In other words, one predictor variable can be used to predict the other. This create redundant information, skewing the result in a regression model.

→ An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables. If the correlation coefficient is exactly -1 or +1, then it is perfect multicollinearity, and one of the variable should be removed from the model.

CHECK MULTI COLLINEARITY - i) Correlation
ii) VIF

MAIN CAUSES OF MULTICOLLINEARITY →

i) Include two (identical or almost identical variables) - For example weight in pounds, weights in kg OR amount in ₹ or amount in $.

ii) Include a variable in regression that is actually combination of two other variables - For example, total investment income = income from stocks and bonds + income from saving interest

iii) Dummay variable may be incorrectly used.

WHY WE CARE ABOUT MULTICOLLINEARITY -

Lawyer's salary = $\beta_0 + \beta_1$ (Years of Experience) + β_2 (Ag) + ε_i

Dependent variable = Salary	Coefficient	Standard Error	t stat	P-value
Intercept	19074.53	51499.7	0.3704	0.7221
Experience	3886.147	2093.61	1.8562	0.1508
Age	2023.351	1928.49	1.0492	0.329

Coefficient means if we change one unit in independent var, what change is in dependent variable. If there is a increase of 1 year in Age, then salary is increased by 2023.

Multicollinearity affects SE, t stat and Pvalue. For eg. p value < 0.05 then only we select the variable, but both experience & age pvalue > 0.05, so we reject statistically but this does not make sense as Salary is dependent on experience. Issue is Multicollinearity