# WHAT,WHY & HOW MULTICOLLINEARITY
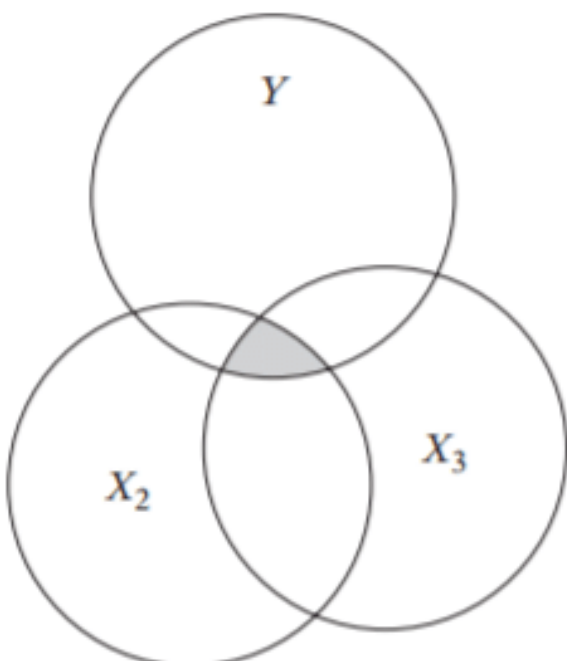
What is multicollinearity?

- **Multicollinearity** occurs when independent variables in a regression model are correlated.
- This correlation is a problem because independent variables should be independent.
- If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.
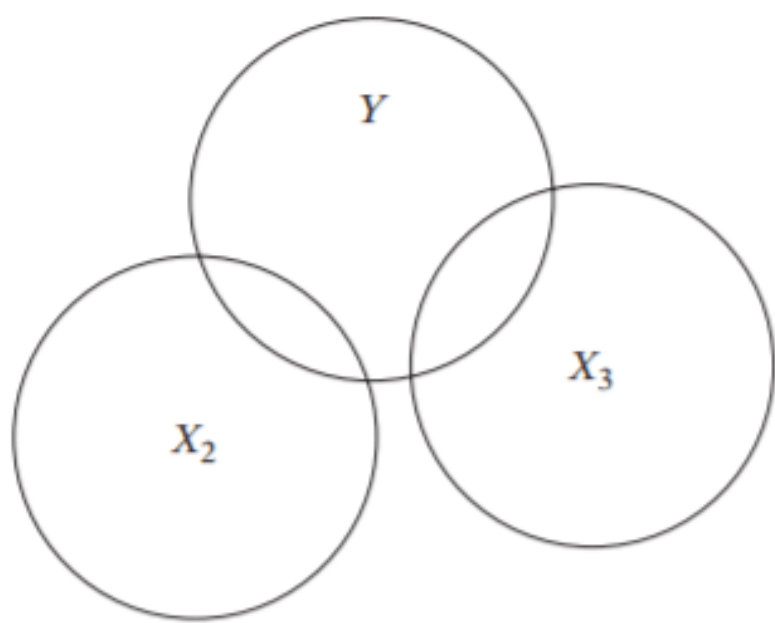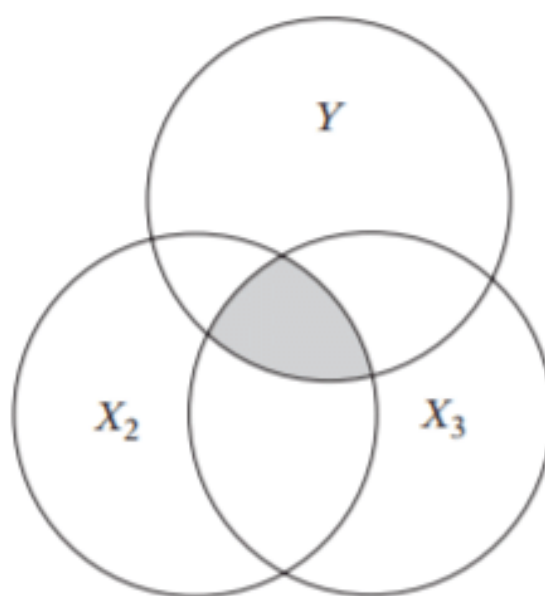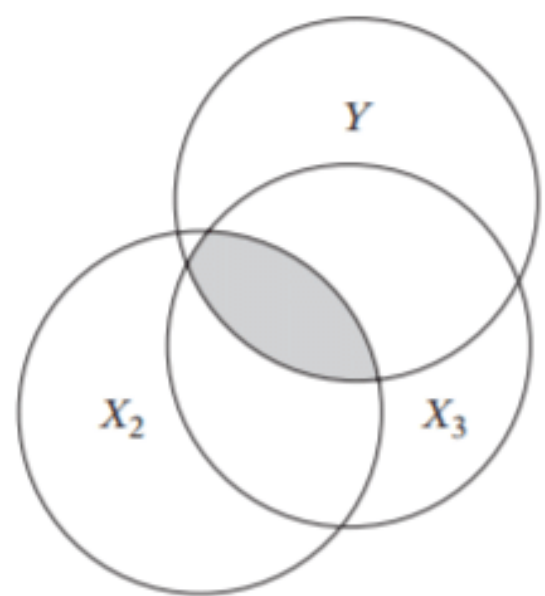

(a) No collinearity
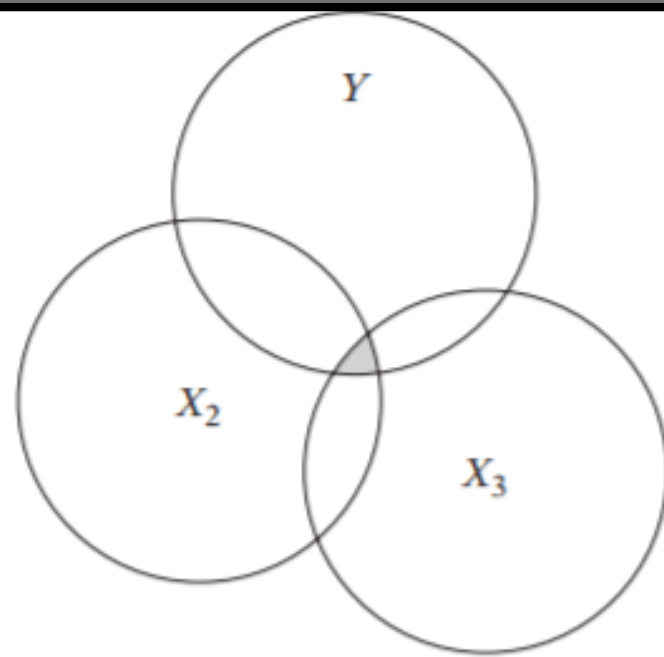
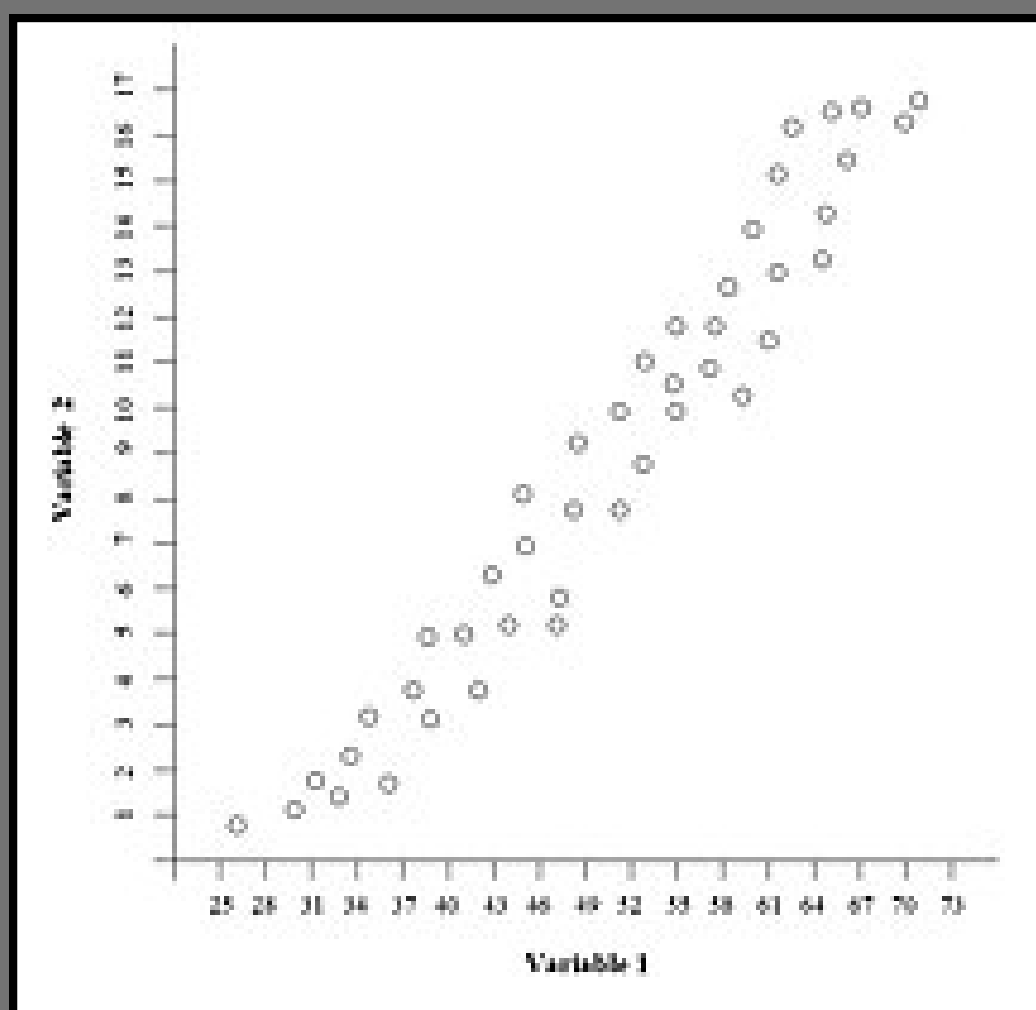(b) Low collinearity

(c) Moderate collinearity

(d) High collinearity

(e) Very high collinearity

Why is Multicollinearity a Potential Problem?

- A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable.

- The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you hold all of the other independent variables constant.

Two basic kind of multicollinearity

1. Structural multicollinearity:

- This type occurs when we create a model term using other terms.
- In other words, it's a by product of the model that we specify rather than being present in the data itself.
- For example, if you square term X to model curvature, clearly there is a correlation between X and X2.

2. Data multicollinearity:

- This type of multicollinearity is present in the data itself rather than being an artefact of our model.
- Observational experiments are more likely to exhibit this kind of multicollinearity.

How to Deal with Multicollinearity?

- Remove some of the highly correlated independent variables.
- Linearly combine the independent variables, such as adding them together.
- Partial least squares regression uses principal component analysis to create a set of uncorrelated components to include in the model.
- LASSO and Ridge regression are advanced forms of regression analysis that can handle multicollinearity.

What are the ways of handling multicollinearity?

1. **Increase sample size** to strengthen the statistical power.
2. **Remove highly correlated predictors by** checking the **Variance Inflation Factor (VIF).**
3. **Combine** correlated **variables into a single predictor through** Principal Component Analysis **(PCA) or factor analysis.**