# Multicollinearity

## What is Multicollinearity?

Multicollinearity occurs when **two or more independent variables (features)** in a regression model are **highly correlated**, making it difficult to isolate their individual effects on the dependent variable (target).

- In case of Multicollinearity, often there is **linear relationship** between independent features.
- Pearson correlation coefficient → **0.9 or 0.8**
- **Perfect Multicollinearity**: One variable is an exact linear combination of others
    - e.g., $X_1 = 2X_2 + 3X_3$
- **High Multicollinearity**: Variables are strongly but not perfectly correlated
    - e.g., $X_1 \approx 0.9X_2$

## Why is Multicollinearity a Problem?

- **Unstable Coefficients**: Small changes in data can drastically alter coefficient estimates.
- **Inflated Standard Errors**: Reduces statistical power (larger p-values, harder to detect significance).
- **Misleading Interpretations**: Coefficients may have unexpected signs or magnitudes.
- **Redundancy**: Wastes computational resources on correlated features.
- Unstable and unreliable estimates: The regression coefficients become sensitive to small changes in the data, making it difficult to interpret the results accurately.

In this equation:

$y = ß_0 + ß_1X_1 + ß_2X_2$

- When variables are not related:
    - If we keep $X_2$ constant & change $X_1$, $y$ changes wrt $X_1$
- But when there's collinearity:
    - $X_2$ changes with $X_1$
    - So, we won't be able to interpret the value of $ß_2$ or $ß_1$
- Therefore, it becomes difficult to calculate the relationship between $y$ and $X_1$

# Inference vs Prediction

## Inference 🔍

- **Goal**: **Understand** the relationships between variables and the underlying data structure.
- **Focus**: Draw conclusions about the **population** or **process** that generated the data.
- **Methods**: Hypothesis testing, confidence intervals, significance of variables.
- **Interpretability**: *Very important* because you want to understand the **"why"** behind the data.
- **Examples**: **Linear regression**, **logistic regression**, **ANOVA**.

## Prediction 📊

- **Goal**: Make **accurate forecasts** for new, unseen data.
- **Focus**: Use the model to **generalize** and predict outcomes based on observed patterns.
- **Methods**: Minimize error metrics like **mean squared error**.
- **Interpretability**: *Less important* since the main goal is **accuracy**.

- **Examples**: **Decision trees**, **support vector machines**, **neural networks**, **random forests**.

### Key Difference ⚖️:

- **Inference** helps **understand** data and relationships, while **Prediction** focuses on making **accurate predictions** for new data.

> 💡
>
> **Multicollinearity does not affect the model when it’s predictive model.**
>
> **It affects the model when used for inference (To find out relationship between Input & Output.)**

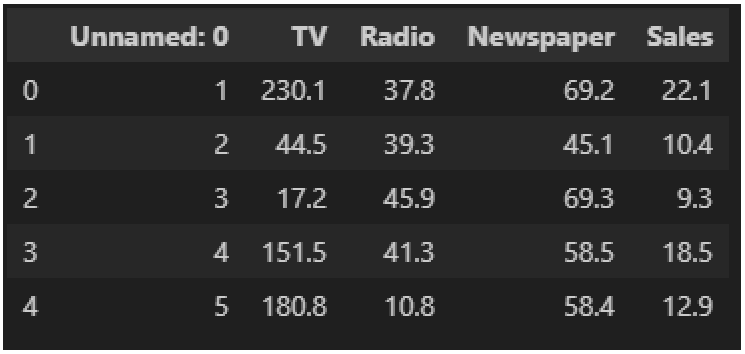
# How to Detect Multicollinearity

## 1. Correlation Matrix

- A table showing pairwise correlations among all predictor variables.
- **Purpose**: Identify pairwise linear relationships between predictors.
- **Method**:
  - Compute the correlation matrix for all independent variables.
  - Look for absolute correlation values **> 0.7–0.8**.

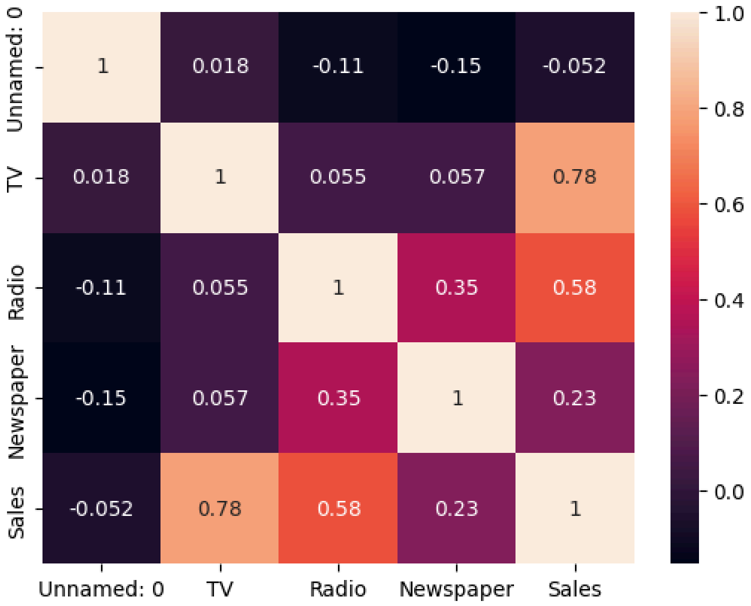
```
import pandas as pd
import seaborn as sns

df = pd.read_csv('https://raw.githubusercontent.com/justmarkham/scikit-learn-videos/master/data/Advertising.csv')
```

```
df.head()
```



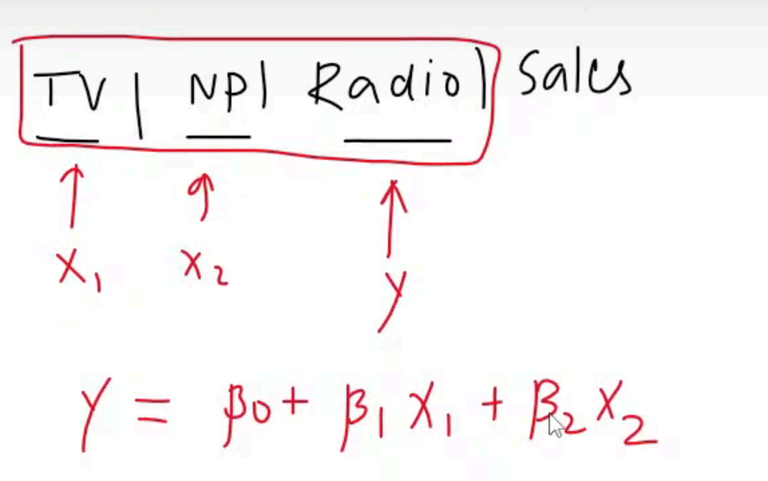
| | Unnamed: 0 | TV | Radio | Newspaper | Sales |
|---|---|---|---|---|---|
| 0 | 1 | 230.1 | 37.8 | 69.2 | 22.1 |
| 1 | 2 | 44.5 | 39.3 | 45.1 | 10.4 |
| 2 | 3 | 17.2 | 45.9 | 69.3 | 9.3 |
| 3 | 4 | 151.5 | 41.3 | 58.5 | 18.5 |
| 4 | 5 | 180.8 | 10.8 | 58.4 | 12.9 |

```
sns.heatmap(df.corr(),annot=True)
```



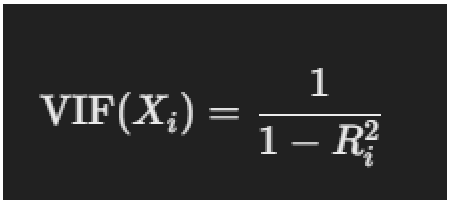
## 2. Variance Inflation Factor (VIF)

- If you have 3 input columns, you make 1 column as Output column & calculate linear regression & calculate the **R2 Score**
- Then you do this 1 by 1 with other 2 columns as well



- From R2 score, you calculate the VIF score.

- **Purpose**: Quantify how much the variance of a coefficient is inflated due to multicollinearity.
- **Formula**:



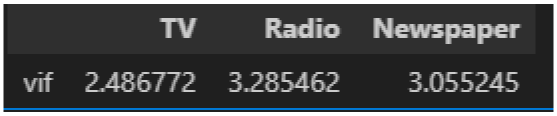
$$\mathrm{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

where $R_i^2$ is the coefficient of determination when $X_i$ is regressed on all other predictors.

- **Threshold**:
  - VIF > 5–10: **Moderate to severe** multicollinearity.
  - A VIF of 1 means **no correlation**.

```
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
```



|  | TV | Radio | Newspaper |
|---|---|---|---|
| vif | 2.486772 | 3.285462 | 3.055245 |

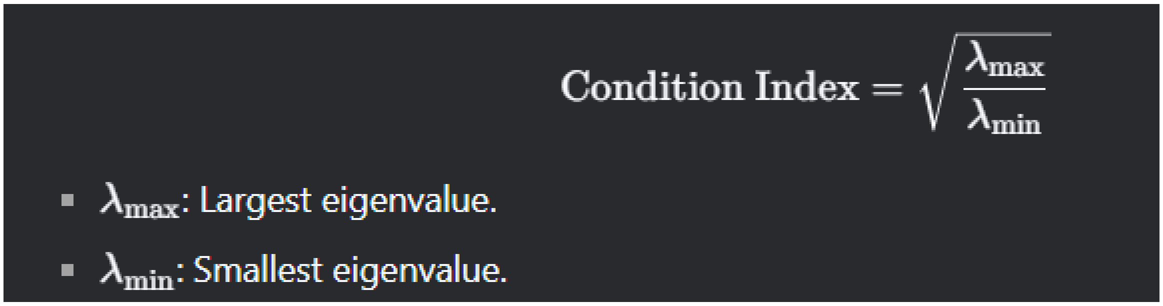
- By making TV as output column, the VIF is 2.48... and so on

## 3. Condition Number

- The condition number is a metric derived from the eigenvalues of the predictor matrix.
- **It indicates how sensitive the regression coefficients are to small changes in the data.**

**Method**:

- Compute eigenvalues of the **correlation matrix**.
- Calculate the **condition index**:



$$\text{Condition Index} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

- $\lambda_{\max}$: Largest eigenvalue.
- $\lambda_{\min}$: Smallest eigenvalue.

## How to detect multicollinearity:

- If the **condition number** is large (typically greater than **30**), multicollinearity may be present.

**Example:**

If the condition number of the matrix of predictors is 150, then small changes in the data may lead to large changes in the regression coefficients.

```
import numpy as np
from numpy.linalg import cond

# Assuming 'X' is your independent variable dataset
condition_number = cond(X)
print("Condition Number: ", condition_number)
```

## How to remove multicollinearity

- Collect more data
- Remove one of the highly correlated variables
- Combine correlated variables
- Use partial least squares regression (PLS)