CHAPTER 4: MULTICOLLINEARITY

Multicollinearity

The problem of multicollinearity exists when two or more regression variables are strongly correlated or linearly dependent.

Suppose we wish to fit the model $Y = X\beta + \epsilon$

LSE : $\hat{\beta} = (X'X)^{-1}(X'Y)$

If $(X'X)$ is singular then we can't perform the inverse. This happens when at least one column of X is LD (linearly dependent) on the other columns.

Effect of Multicollinearity / problems due to Multicollinearity:-

Consider MLR model with two regressors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon \quad , i = 1(1)n$$

$$\underset{\sim}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ & \vdots & \\ 1 & X_{n1} & X_{n2} \end{bmatrix}$$

① Strong multicollinearity between regressors result in large variance and covariance of regression coefficients.:-

Centering & Scaling Regression Data:- Original data: X_{i1}, X_{i2}, Y_i

We write $x_{i1} = \dfrac{X_{i1} - \bar{X}_1}{\sqrt{S_{11}}}$, $x_{i2} = \dfrac{X_{i2} - \bar{X}_2}{\sqrt{S_{22}}}$; $y_i = \dfrac{Y_i - \bar{Y}}{\sqrt{S_{YY}}}$;

where $\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{i1}$, $S_{11} = \sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2$, $\bar{Y} = \frac{1}{n}\sum_{1}^{n} Y_i$

$\bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{i2}$, $S_{22} = \sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2$, $S_{YY} = \sum_{1}^{n}(Y_i - \bar{Y})^2$

ORIGINAL DATA:

X_1 X_2 Y

Mean (original data) $\bar{X}_1$ $\bar{X}_2$ $\bar{Y}$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$

(for original data)

TRANSFORMED DATA:

x_1 x_2 y

$\bar{x}_1 = 0$ $\bar{x}_2 = 0$ $\bar{y} = 0$ [Mean of transformed data using centering & scaling]

$\sum_{i=1}^{n} x_{i1}^2 = 1$ $\sum_{i=1}^{n} x_{i2}^2 = 1$ $\sum_{i=1}^{n} y_i^2 = 1$ [variance of transformed data using ...]

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

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 = 0$ [for transformed data]

The model, assuming that X_1, X_2 and Y are centered and scaled, is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad [\text{intercept} = \beta_0 = 0 \text{ always}]$$

[Example of Multicollinearity:

Compute $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}(X'Y)$

→ $X'X$ is singular since X_1 is linearly dependent on X_2 and X_3 and check that $|X'X| = 0$.

X_1	X_2	X_3	Y_i
1	-2	4	81
2	-7	11	88
4	3	5	94
7	1	13	95
8	-1	17	123

[Note that: $X_1 = \dfrac{X_2}{2} + \dfrac{X_3}{2}$]]

For the centered and scaled data, we have the following X matrix for the transformed data:

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \\ x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} \frac{X_{11}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{12}-\bar{X}_2}{\sqrt{S_{22}}} \\ \frac{X_{21}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{22}-\bar{X}_2}{\sqrt{S_{22}}} \\ \vdots & \\ \frac{X_{n1}-\bar{X}_1}{\sqrt{S_{11}}} & \frac{X_{n2}-\bar{X}_2}{\sqrt{S_{22}}} \end{pmatrix}, \quad Y = \begin{pmatrix} \frac{Y_1-\bar{Y}}{\sqrt{S_{YY}}} \\ \frac{Y_2-\bar{Y}}{\sqrt{S_{YY}}} \\ \vdots \\ \frac{Y_n-\bar{Y}}{\sqrt{S_{YY}}} \end{pmatrix}$$

[since $\beta_0 = 0$]

Normal equation:- [since $\sum x_{i1}^2 = \sum x_{i2}^2 = 1$]

$$(X'X)\hat{\beta} = X'Y \Rightarrow \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix};$$

where r_{12} is sample correlation between x_1 and x_2.
r_{1y} is sample correlation between x_1 & y.
r_{2y} is sample correlation between x_2 & y.

$$r_{1y} = \frac{\sum_1^n (X_{i1}-\bar{X}_1)(Y_i-\bar{Y})}{\sqrt{S_{11}S_{yy}}} \quad \& \quad r_{2y} = \frac{\sum (X_{i2}-\bar{X}_2)(Y_i-\bar{Y})}{\sqrt{S_{22}S_{yy}}}$$

and $$r_{12} = \frac{\sum (X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2)}{\sqrt{S_{11}S_{22}}};$$ $X'X = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}$ is called the correlation matrix.

Inverse of $(X'X)$ is

$$(X'X)^{-1} = \begin{pmatrix} \frac{1}{1-r_{12}^2} & -\frac{r_{12}}{1-r_{12}^2} \\ \frac{-r_{21}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix} \quad \text{—— (A)}$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$
$$= \begin{pmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{21}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix}\begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix}$$

The estimates are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1-r_{12}^2} \quad \& \quad \hat{\beta}_2 = \frac{r_{2y} - r_{21}r_{1y}}{1-r_{12}^2}$$

Thus, $\hat{\beta}_1$ and $\hat{\beta}_2$ (LSEs on the transformed data) are estimated using (in terms of) the sample correlation coefficients.

$$V(\hat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj}$$ [See in Chapter 2]

$$V(\hat{\beta}_1) = \sigma^2 (X'X)^{-1}_{11} = \frac{\sigma^2}{1-r_{12}^2} \to \infty \text{ as } |r_{12}| \to 1$$ [From (A)]

[strong multicollinearity between x_1 & x_2]

$$V(\hat{\beta}_2) = \sigma^2 (X'X)^{-1}_{22} = \frac{\sigma^2}{1-r_{12}^2} \to \infty \text{ as } |r_{12}| \to 1$$ [From (A)]

If there is strong multicollinearity between x_1 and x_2, then the correlation coefficient (r_{12}) will be large, i.e., $|r_{12}| \to 1$.

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = \sigma^2 (X'X)^{-1}_{12} = -\frac{\sigma^2 r_{12}}{1-r_{12}^2} \to \pm\infty$$

depending on whether $r_{12} \to +1$ or $r_{12} \to -1$.

For More than two regressors (MLR):-

It can be shown that the diagonal elements of $(X'X)^{-1}$ matrix are $\frac{1}{1-R_j^2}$ $\forall\ j = 1(1)K-1$;

where R_j^2 is the coefficient of multiple determination from the regression of x_j on the remaining $(K-2)$ regressors, i.e,

[Here $x_j = f(x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_{K-1})$.] $R_j^2 = \frac{SS\,reg}{SST}$ measures the proportion of variability in the response variable that is explained by regressor variables

$$V(\hat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj} = \frac{\sigma^2}{1-R_j^2} \to \infty \text{ as } R_j^2 \to 1$$

If there is strong multicollinearity between x_j and any subset of other $(K-2)$ reg., then R_j^2 will be close to unity, i.e., $R_j^2 = 1$

(2) Multicollinearity tends to produce LSE $\hat{\beta}$ that are too far from the true parameter β:-

Squared Distance $= L^2 = \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta_i)^2$

Expected squared distance $= E(L^2) = \sum_i E(\hat{\beta}_i - \beta_i)^2$; $E(\hat{\beta}_i) = \beta_i$ since LSEs are UEs.

$$= \sum_i E(\hat{\beta}_i - E(\hat{\beta}_i))^2$$

$$= \sum_i V(\hat{\beta}_i) = \sum_i \sigma^2 (X'X)^{-1}_{ii}$$

$$\therefore E(L^2) = \sum_i \frac{\sigma^2}{1-R_i^2} = \sigma^2 \sum \frac{1}{1-R_j^2}$$

When there is multicollinearity, $\frac{1}{1-R_i^2}$ value will be large for at least one i.

i.e., $\frac{1}{1-R_i^2} \to \infty$ as $R_i^2 \to 1$.

(3) Model coefficient with '-'ve sign when '+'ve sign is expected.

(4) High significance in a GLOBAL F-test but in which none of the regressors are significant in partial F-test.
[Example of Multiple Linear Regression in chapter 2].

(5) Different model selection procedures yield different models.

TECHNIQUES FOR DETECTING MULTICOLLINEARITY:

- Examination of Correlation Matrix $(X'X)$:

- A simple measure of multicollinearity is the inspection of off-diagonal elements r_{ij} in $X'X$;
- Suppose $|r_{ij}| > 0.9$ indicates the multicollinearity problem.
- Examining the correlation matrix $(X'X)$ is helpful in detecting linear dependence between pairs of regressors.
- However, examining the correlation matrix $(X'X)$ is not helpful in detecting multicollinearity problem arising from linear dependence between more than two regressors.

Example: Unstandardized regressors and response variable from Webster, Gunst and Mason (1974):

X_1	X_2	X_3	X_4	X_5	X_6	Y
8	1	1	1	0.541	-0.099	10.006
8	1	1	0	0.130	0.070	9.737
8	1	1	0	2.116	0.115	15.087
0	0	9	1	-2.397	0.252	8.422
0	0	9	1	-0.046	0.017	8.625
0	0	9	1	0.365	1.504	16.289
2	7	0	1	1.996	-0.865	5.958
2	7	0	1	0.228	-0.055	9.313
2	7	0	1	1.380	0.502	12.960
0	0	0	10	-0.798	-0.399	5.541
0	0	0	10	0.257	0.101	8.756
0	0	0	10	0.440	0.432	10.937

- This data has the problem of MULTICOLLINEARITY.
- However, $X'X$ examination is unable to find it (since it is for more than two regressors)

$$X'X = \begin{bmatrix} 1 & 0.052 & -0.843 & -0.498 & 0.417 & -0.192 \\ & 1 & -0.432 & -0.371 & 0.485 & -0.317 \\ & & & & & 0.498 \\ & & 1 & -0.355 & -0.505 & -0.087 \\ & & & 1 & -0.215 & -0.123 \\ & & & & 1 & 1 \end{bmatrix}$$

Since $|r_{ij}| > 0.9$ indicates multicollinearity problem. Here, none of the pairwise correlations are suspiciously large, and we have no indication of near linear dependence. Thus, $X'X$ fails to detect multicollinearity.

- Eigen System Analysis on $X'X$:-

– Multicollinearity can also be detected from the eigenvalues of the correlation matrix $X'X$.

– For a $(k-1)$ regression in the model, there will be $(k-1)$ eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$.

– If there are one or more linear dependences in the data, then one or more eigen values will be small.

– Define the condition number of $(X'X)$ as

$$K = \frac{\lambda_{max}}{\lambda_{min}} = \text{condition number (CN)}.$$

– As a general rule,

$K < 100$ indicates no serious problem with multicollinearity.
$100 \leq K \leq 1000$ indicates moderate to strong multicollinearity.
$K > 1000$ indicates severe problem with multicollinearity.

– The condition indices of the $(X'X)$ matrix are

$$CI = K_j = \frac{\lambda_{max}}{\lambda_j}, \quad j = 1(1)k-1.$$

Advantages:
- Detecting multicollinearity.
- Measuring no. of linear dependences.
- Identify the nature of linear dependence

– Clearly the largest condition index is the condition number.

– The number of $K_j > 1000$ is a useful measure of the number of near linear dependence in $X'X$.

– The correlation matrix $(X'X)$ may be decomposed as

$$X'X = TDT'$$

where, $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{k-1})$ and

$$T^{k-1 \times k-1} = (t_1, t_2, \ldots, t_{k-1});\ \text{where}$$

- Eignsystem analysis can be used for identifying the nature of linear dependence in the data.

$$t_i = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{k-1} \end{pmatrix}$$ is the eigen vector associated with eigen value λ_i.

If the eigenvalue λ_i is close to zero, the elements of the associated eigenvector t_i describe the nature of linear dependence, i.e., $a_1x_1 + a_2x_2 + \cdots + a_{k-1}x_{k-1} = 0$

$$\Leftrightarrow \sum_{i=1}^{k-1} a_i x_i = 0 \quad \text{and} \quad t_i = (a_1\ a_2 \cdots a_{k-1})'.$$ (*)

[P.T.O (*)]

Example of Webster Data:

The eigen values are: $\lambda_1 = 2.4288$, $\lambda_2 = 1.5462$, $\lambda_3 = 0.9221$, $\lambda_4 = 0.794$, $\lambda_5 = 0.3079$, $\lambda_6 = 0.0011$.

There is only one small eigenvalue.

$\therefore$ K (Condition number) $= \frac{2.4288}{0.0011} = 2188.11 > 1000$

which indicates severe problem with multicollinearity.

Compute K_j's for $j = 1(1)6$.

Check $K_6 = \frac{2.4288}{0.0011} > 1000$ implies there is only one near linear dependence in the data. (examination of $X'X$ fails here)

- Variance Inflation Factor:- (VIF):-

Variance of the i^{th} regression coefficient

$$V(\hat{\beta}_i) = \sigma^2 (X'X)^{-1}_{ii} = \frac{\sigma^2}{1-R_i^2}$$

- R_i^2 is the coefficient of multiple determination when x_i is regressed on the remaining regressors.
- If x_i is nearly orthogonal to the remaining regressors, then R_i^2 is small and $\frac{1}{1-R_i^2}$ is close to unity. (meaning x_i is independent of the remaining regressors)
- If x_i is nearly linearly dependent on some subset of the remaining regressors, R_i^2 is near unity & $\frac{1}{1-R_i^2}$ is very large.
- $V(\hat{\beta}_i)$ can be viewed as the factor by which the $V(\hat{\beta}_j)$ is increased due to linear dependence among the regressors.
- The VIF associated with regressor x_i is defined by

$$VIF_i = \frac{1}{1-R_i^2}$$

large value of VIF_i indicates possible multicollinearity associated with x_i.

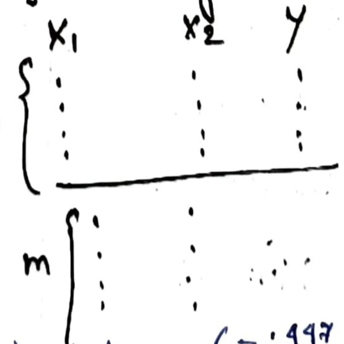
- In general, $VIF_i \geq 5$ indicates possible multicollinearity problem.
  $VIF_i \geq 10$ indicates almost certainly multicollinearity problem.

Dealing with Multicollinearity:-

→ However, VIF can't identify the nature of multicollinearity.

- Collect Additional data:- Collecting additional data has been suggested as the best method of dealing with multicollinearity

Additional data should be collected in manner to break up the multicollinearity in the existing data



✱ Webster Data:

The smallest eigenvalue is $\lambda_6 = 0.0011$ and

$$t_6 = \begin{pmatrix} -.447 \\ -.421 \\ -.541 \\ -.573 \\ -.006 \\ -.002 \end{pmatrix}$$

$$\sum a_i x_i = 0 \Rightarrow -.447x_1 - .421x_2 - .541x_3 - .573x_4 - .006x_5 - .002x_6 = 0.$$

$$\Rightarrow x_1 = -0.941x_2 - 1.21x_3 - 1.28x_4.$$

very small values thus ignoring x_5 and x_6

This linear dependence is associated with λ_6.

- **Remove regressors from the model:**

– If two regressors are linearly dependent then it means that either of them contain redundant information. Thus, we can pick one regressor to keep in the model and discard the other one.

– If x_1, x_2, and x_3 are linearly dependent, then eliminating one regressor may be helpful to reduce the effect of multicollinearity.

– However, elimination of regressor(s) from the data may damage the predictive power of the model.

- **Collapse Variables:** Combine two or more variables that are linearly dependent into single composite variables.

- **Ridge Regression:** To be discussed in the next chapter.

TIPS FOR MODELLERS:

Commonly used Multicollinearity detection tools in Analytics:

1. High R^2 but few significant t ratios.
2. High pairwise correlation among regressors.
3. Examination of partial correlations.
4. Eigen value method and calculation of CI and CN. (Higher CI ⇒ Multicollinearity)
5. VIF (most commonly used).

Commonly used techniques to deal with Multicollinearity in Business analytics:

1. Dropping variables.
2. Using Principal components
3. Ridge Regression

} common

4. Collection of additional data
5. Collapsing variables.