

Multicollinearity:

It is a statistical problem where multiple terms (regressors) of a model are heavily correlated which reduces the accuracy of identifying the effect of individual terms. [linear dependency exists]

That means we cannot identify the influence of individual terms so its effects cannot be interpreted meaningfully. Model becomes unstable + Redundant ~~when~~ Terms are present.

Diagnosis:

[Usually if multicollinearity exists two or more terms are very similar to each other (correlated), so, that is what we check for here] [They are linearly dependent]

original model: $\hat{y} = b_1 x_1 + b_2 x_2 + \dots + b_k x_k + a$
Try to predict $\hat{x}_1$ using the others!

$$\hat{x}_1 = b_2 x_2 + \dots + b_k x_k + a \quad [\text{if success, we don't need } x_1]$$

x_2 :

$$\hat{x}_2 = b_1 x_1 + \dots + b_k x_k + a$$

x_k :

$$\hat{x}_k = b_1 x_1 + b_2 x_2 + \dots + a$$

Tolerance:

$$T = 1 - R^2 \quad \text{coefficient of determination}$$

MLC could exist if

$$T < 0.1$$

[VIF] Variance inflation factor:

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

MLC could exist if

$$VIF > 10$$

$$1 - 0.2 = 0.8 - 1 = 0.24$$

Sources:

- The data collection: Sampling over a limited range of values taken by regressors in the pop
- Constraints on the model: $x_1 = \text{Income}$, $x_2 = \text{House size}$, to predict electricity bill, there is a phy constraint in the pop
- model Specification: Adding polynomial terms when range of x is small [Ex: x^2]
- An Overdetermined model: when a model has more data points than unknown parameters.

Consequences:

1. large variance & covariance of OLS estimators:

$$\text{model} \Rightarrow y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$X = X\beta + \epsilon$$

$$\text{Var-covariance matrix} \Rightarrow \sigma^2 (X'X)^{-1}$$

↑
var of error term

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_{1i}^2 (1 - r_{12}^2)}$$

$r_{12} = \text{correlation b/w } x_1 \text{ \& } x_2$

$(1 - r_{12}^2) = \text{portion of var in } x_1 \text{ that is not exp by } x_2$

$$\frac{\sigma^2}{\sum_{i=1}^n x_{1i}^2 (1 - r_{12}^2)} = \frac{\sigma^2}{\sum x_i^2} \times \left[\frac{1}{(1 - r_{12}^2)} \right] \text{ VIF}$$

shows how the var of the estimator is inflated by the presence of multicollinearity.

Detection of Multicollinearity:

1. High R^2 but few / no significant t -ratios.
[It's a classic symptom of mcl]

If R^2 is high, > 0.8 , the F test will reject H_0 :
 $\beta_1 = \beta_2 = \dots = \beta_k = 0$, but the individual t test will
show that none / few slope coefficients are statistically
correct

2. High pair-wise correlations among regressors. [D.R.]

3. Using VIF:

The larger the value, the more collinear it is.
Suspected to be > 0.8 is mid, > 10 is highly
collinear.

4. Eigen system analysis of $X'X$:

The eigenvalues of $X'X$ say $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_k$
can be used to measure the extent of multicollinearity
in the data.

Remedial measures:

- Do Nothing: It's an inherent prob with the data which
may not be subject to change
- A priori Information: $\hat{y}_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$

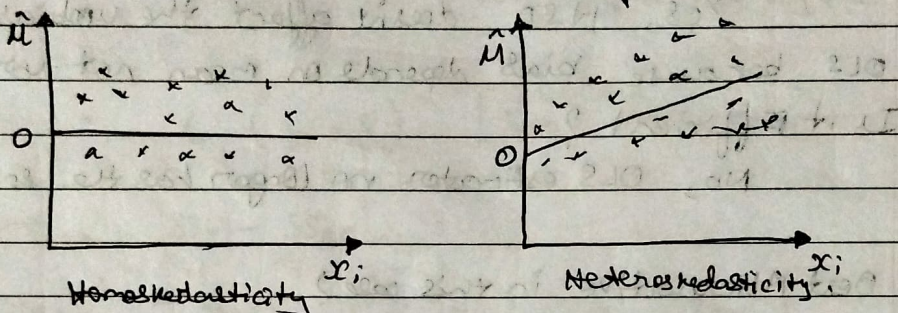
If $\beta_3 = 0.1 \times \beta_2 \Rightarrow \beta_1 + \beta_2 x_{2i} + 0.1 \beta_2 x_{3i} + \epsilon_i$
One we obtain $\tilde{\beta}_2 \Rightarrow \beta_1 + (x_{2i} + 0.1 x_{3i}) \beta_2 + \epsilon_i$
we can get $\beta_3 \Rightarrow \beta_1 + \beta_2 x_{1i} + \epsilon_i$ where $x_{1i} = x_{2i} + 0.1 x_{3i}$

- Drop one of the collinear parameter.
- Transformation of variables
- Collect better data
- Use methods like factor analysis / component
analysis / ridge reg

SMC [Unit 2 - 2nd half]

Heteroskedasticity:

In a regression model $y = \beta_0 + \beta_1 x_1 + \epsilon$, but cross-sectional data occurs when the variance of the error term ϵ changes as the value of X changes.



$\hat{\mu} = \text{variance of } \epsilon$

Causes:

- Presence of outliers
- Variations in scale or range of dependent variable for different values of independent variables
- Omitted var's / incorrect model space
- Nature of the data
- Measurement errors / Non-linearity.

Consequences:

Consider the 2 var model: $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$

By applying OLS estimation:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

And variance:

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= \frac{SS_{RES}}{n-1} \cdot \frac{1}{\sum x_i^2} \\ \text{[For Homoskedasticity]} \quad &\Rightarrow \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

[For Heteroskedasticity]

$$\text{Var}(\hat{\beta}_2) = \frac{\sum (x_i - \bar{x})^2 \leftarrow \text{var } x \text{ the differences each from } \bar{x}}{[\sum (x_i - \bar{x})^2]^2}$$

Is $\hat{\beta}_2$ still BLUE?

Yes, HSD doesn't affect the unbiasedness of OLS because bias depends on mean not var.

Is it Efficient?

No, OLS estimator no longer has the smallest var.

Best Alternative in this case?

WLS: Assign weights to each observation based on its var.

GLS (Generalized Least Square): It is capable of producing BLUE by taking into consideration the info of change in variance/spread. The procedure of transforming the original var in such a way that the transformed var satisfy the assumption of classical model & then apply OLS to them is the method of GLS (Adjusted account for heteroskedasticity.)

$$S_{xx} = \sum (x_i - \bar{x})^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$SS_{RES} = \sum (y_i - \hat{y})^2 \quad [\text{var not explained by the reg model}]$$

$$SS_{REG} = \sum (\hat{y}_i - \bar{y})^2 \quad [\text{var explained by the reg model}]$$

$$SST = \sum (y_i - \bar{y})^2 \quad [\text{Total Variance}]$$

$$SS_{RES} = SST - SS_{REG}$$

$$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} \quad \left[\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \right]$$

Example:

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
5	10	-1	0.6	1	-0.6
7	9	1	-0.4	1	-0.4
6	11	0	1.6	0	0
4	7	-2	-2.4	4	4.8
8	10	2	0.6	4	1.2

$$\bar{x} = 6 \quad \bar{y} = 9.4$$

① Perform LR on y : $\hat{y} = 6.4 + 0.5x$

$$S_{xx} = \sum (x_i - \bar{x})^2 = 1 + 1 + 0 + 4 + 4 = 10$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = -0.6 - 0.4 + 0 + 4.8 + 1.2 = 5$$

$$SST = 9.2$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5}{10} = 0.5$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 9.4 - (0.5)(6) = 9.4 - 3 = 6.4$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 6.4 + 0.5x$$

Predicted values:

$$\hat{y}_i = 8.9, 9.9, 9.4, 8.4, 10.4$$

Find Residuals & Squares:

$$[y_i - \hat{y}_i] \hat{e}_i = 1.1, -0.9, 1.6, -1.4, -0.4$$

$$\hat{e}_i^2 = 1.21, 0.81, 2.56, 1.96, 0.16$$

② Calculate $\hat{\sigma}^2 = \frac{SS_{RES}}{n} \quad (SS_{RES} = \sum \hat{e}_i^2)$

$$= 6.7/5 = 1.34$$

Calculate P_i :

$$P_i = \frac{\hat{e}_i^2}{\hat{\sigma}^2}$$

$\hat{e}_i^2$	P_i	Δ_i	$\hat{P}_i$
1.21	0.9	5	1.176
0.81	0.6	7	0.714
2.56	1.91	6	0.938
1.96	1.46	4	1.414
0.16	0.12	8	0.452

③ Regress P_i on Δ_i

$$\bar{P}_i = 0.834$$

$$P_i = \frac{2.366 - 0.238 \Delta_i}{0.2178} = 0.274$$

④ $\theta = \frac{1}{2} SS_{REG} \quad (SS_{REG} = SS_T - SS_{RES})$

~~$$SS_T = \sum (P_i - \bar{P}_i)^2 = 0.96632$$~~

~~$$SS_{RES} = \sum (P_i - \hat{P}_i)^2 = 0.21556$$~~

~~$$\sum (\hat{y}_i - \bar{y})^2 = 0.62052$$~~

~~$$SS_{REG} = 0.96632 - 0.21556 = 0.75076$$~~

~~$$\theta = \frac{1}{2} \cdot SS_{REG} = 0.37538$$~~

~~$$0.62052$$~~

$$SS_{REG} = \hat{\beta}_1 \cdot S_{xp} \quad \left[\hat{\beta}_1 = \frac{S_{xp}}{S_{xx}} \Rightarrow S_{xp} = \hat{\beta}_1 \cdot S_{xx} \right]$$

$$\Rightarrow \hat{\beta}_1 \cdot \hat{\beta}_1 \cdot S_{xx}$$

$$\Rightarrow -0.238x - 0.238 \times 10 \quad ??$$

$$SS_{REG} = 0.56644 \quad ??$$

$$\theta = \frac{SS_{REG}}{2} = \frac{0.56644}{2} = 0.28322$$

$$\chi^2_{0.05, 1} = 3.841$$

(0.05, n-1)

$$I \nmid \theta > \chi^2_{\alpha, n-1}$$

Reject H_0

n = no. of points

$$0.28 < 3.841$$

$\therefore$ NO HSD ??

Procedure 2:

① Same as Prev Procedure

② Regress $\hat{e}_i^2$ on x_i

$$e_i = 3.896 - 0.426x_i$$

③ calculate R^2 for $\hat{e}_i^2$

$$R^2 = 1 - \frac{SS_{RES}}{SST} = \frac{1.968}{3.563} = 1 - 0.552 = 0.448$$

e_i^2	x_i	$\hat{e}_i^2$
1.21	5	1.766
0.81	7	0.91
2.56	6	1.34
1.96	4	2.192
0.16	8	0.488
$\hat{e}_i^2 = 1.34$		

$$SS_{RES} = 1.968$$

$$SST = 3.563$$

Calculate: $n \times R^2$??

Reject H_0 if $n \times R^2 > \chi^2_{\alpha, n-1}$

$$5 \times 0.448 = 2.25$$

$$2.25 < 3.841$$

$\therefore$ Accept H_0 ; NO HSD ??

Autocorrelation:

It refers to the degree of correlation b/w the values of the same variables across different observations in the data.

The concept of autocorrelation most often discussed in the context of timeseries data.

Occurs in Timeseries, Crosssectional & Panel Data.

In the context of ~~the~~ classical regression model, it assumes that autocorrelation doesn't exist in the disturbances. If we relax this assumption, that is, the error term ε_t at time t is correlated with the error terms $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-k}$.

The correlation b/w ε_t & ε_{t-k} is an autocorrelation of order k . The correlation b/w ε_t & ε_{t-1} is the first order autocorrelation & denoted by ρ_1 .

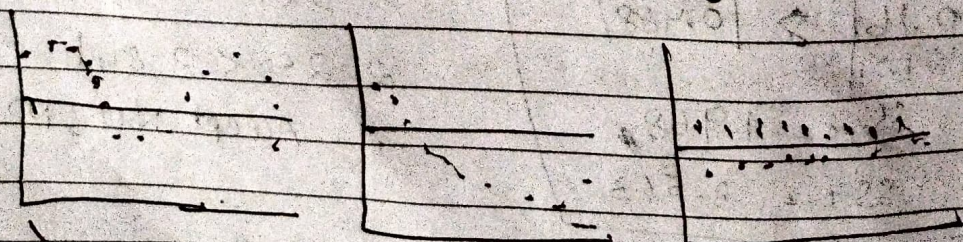
Detection:

We can obtain a rough idea of autocorrelation in the disturbances by plotting the values of the regression residuals on a 2D plot in the following way -

the points will be $(\hat{\varepsilon}_1, \hat{\varepsilon}_2), (\hat{\varepsilon}_2, \hat{\varepsilon}_3), \dots, (\hat{\varepsilon}_{n-1}, \hat{\varepsilon}_n)$

If most of the points lie in the 1st & 3rd quadrants, its +ve autocorrelation.

If most lie in 2nd & 4th quadrant its -ve



Some autocorrelation

No autocorrelation

Date _____
Page _____

Sources/Reasons:

Inertia: The series has a memory of its past values & tends to continue its current trend.

Specification Bias: When adopted relationship is different from the original relationship.

Ex: If a linear relationship is assumed b/w X & Y , while the true relation is quadratic, (Y depends on x^2 not x) in such cases the influence of x^2 is accommodated into ϵ & hence ϵ and x^2 become inter-correlated.

Cob-web Phenomena: Manipulation of data, data transformations, non-stationarity are other reasons for autocorrelation.

Consequences: Although model remains linear,

Unbiased & asymptotically normally distributed, it is no longer the least-variance estimator [not BLUE]

TEST:

Durbin-Watson Test

- Applicable for small sample sizes
- Applicable only for first order autoregressive scheme only.

$H_0: \rho = 0$; ϵ_i are not autocorrelated

If $\hat{d} = 2 \Rightarrow$ Accept H_0

$$\hat{d} = 0 \quad [0 < \hat{d} < 2] \Rightarrow \text{+ve autocorrelation}$$

$d = 4$ $[a < \hat{a} < 4]$ $\Rightarrow$ -ve autocorrelation

$n = \text{Instances}$; $K = \text{no. of predictors including the intercept}$

If $\hat{d} < d_L$ Reject H_0 indicating true auto...
 $\hat{d} > (y - d_U)$ " " " " " " " " " " " "
 $d_L \leq \hat{d} < d_U$ || $(y - d_U) < \hat{d} < (y - d_L) \Rightarrow$ Inconclusive

- Exact distribution of this test is unknown
- Chance for inconclusiveness is a Big Con
- Cannot be used for highest order correlation

X	Y	e_i	$e_i - e_{i-1}$	e_i^2	$(e_i - e_{i-1})^2$
5	10	1.1	-	1.21	-
7	9	-0.4	-2.0	0.16	4.0
6	11	1.6	2.5	2.56	6.25
4	7	-1.4	-3.0	1.96	9.0
8	10	-0.4	1	0.16	1.0
				<u>6.7</u>	<u>20.25</u>

$$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} = \frac{20.25}{6.17}$$

27 3.02