Application of statistical Methods to measure economic relationship.

Economic Data were Classified into 5 – But 3 will be discussed

① Cross sectional data – at one time period
② Time series data
③ Panel data.

} differentiable btw the 3

Model is Linear in Parameter

obs Assumptions
- Linearity / Multicollinearity
- Normality
- Independence of obs
- Homoscedasticity (some variance across groups)
- spherical error / unexplained

Ts.: Monthly sale data – Yearly, Monthly, Quarterly, Weekly } basis.

Panel data: Combines X-ic of cross sectional data and Time series – More than one variables More than one-one time period.

Fixed Model
Pooled model
Random model
} Carry test on the data to know which model to be used. depending on the results of that test

$$y_i = \beta_0 + \beta_1 X_{i1} + e_i \quad i = 1, 2, 3, \ldots, n$$

$$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK} \quad i = 1, \ldots n$$

Matrix form, $y = X\beta + e$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad X = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1K} \\ X_{21} & X_{22} & \ldots & X_{2K} \\ \vdots & & & \\ X_{n1} & X_{n2} & \ldots & X_{nK} \end{bmatrix}_{(n\times K)} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}_{K\times 1} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK}$$

But $\beta_i = \frac{\partial y_i}{\partial X_i}$ $\beta_1 = \frac{\partial y_i}{\partial X_2}$ $\beta_2 = \frac{\partial y_i}{\partial X_3}$ $\beta_3 = \frac{\partial y_i}{\partial X_3}$

Rank of matrix X — obs from ...

Number of linearly independent columns. X_i doesn't affect other X_is. If Model is not fully ranked, then Inverse of X cannot be defined.

Full rank matrix — obs are not independent of one another.

No column is linearly depend on one another.

Homoscedacity of error variances are constant

X_1	X_2	...	X_n	
x_1	-	-	-	$E(e_i)=0$
-	-	-	-	$y = X\beta + e$
-	-	-	-	

Rank = no of non zero columns in the matrix

Full Rank =>

$V(e_i) = \sigma_i^2$ Variance

$= E[(e - E(e))(e - E(e))']$

$= E(ee' - eE(e') - E(e)e' + E(e)E(e'))$

$= E(ee')$

$\Rightarrow$

$$V(e_i) = \begin{vmatrix} e_1 & e_2 & \cdots & e_n \\ e_1 & E(e_1e_1) = \sigma_1^2 & E(e_1e_2) = \sigma_{12} & \cdots & E(e_1e_n) \\ e_2 & E(e_2e_1) & E(e_2e_2) = \sigma_2^2 & \cdots & E(e_2e_n) \\ \vdots & \vdots & \vdots & & \vdots \\ e_n & E(e_ne_1) & E(e_ne_2) & \cdots & E(e_ne_n) = \sigma_n^2 \end{vmatrix}$$

$i \neq j$

Giving a scalar

$E(e_1e_1) = Var(e_1) = \sigma_1^2$

Generally $E(e_ie_i) = E(e_i^2) = \sigma_i^2$

- Covariance b/w errors are zero — OLS/MLR assumptions.

$E(e_i e_j) = Cov(e_i e_j) = 0$

$E(e_i - E(e_i))(e_j - E(e_j))$.

$E(e_i e_j) = 0 \quad \forall i \neq j$

If $i = j$ then we get the variances of e_i

$$V(e_i) = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & & & & \sigma_n^2 \end{pmatrix} = \sigma^2 I_n$$

Normality assumptions. — Assumption 5

Since $Y = X\beta + e$

$E(y) = X\beta$

$V(y) = V(X\beta + e)$

$= V(X\beta) + V(e)$

$V(y) = 0 + V(e)$

$= \sigma_e^2$

Then sampling distribution of y is

$y \sim N(X\beta, \sigma_e^2)$

Least Square Estimation

$$SSe = (y - X\beta)(y - X\beta)'$$

$$= (y - Xb)(y - Xb)'$$

$$= y'y - b'X'y - y'Xb + b'X'Xb$$

$1 \times n \;\; n \times 1$ $\quad$ $1 \times n \;\; n \times n \;\; n \times 1$

1×1 scalar $\quad$ 1×1 scalar

$$= y'y - 2b'X'y + b'X'Xb$$

differentiate partially wrt b and equate to zero

Closing for

Bismillahi Rahman Raheem

Define Econometrics

- Econometrics is the field of economics in which statistical methods are developed and applied to estimate the economic relationships, testing economic theories and evaluating and implementing government and business policy. I.e policies implemented by private industry, government and supernational organizations are evaluated.

Define Econometric Models. Vs Economic Model

Econometric Models

- Econometric Models consists of 3 things.

I. A set of equations describing the behaviour of variables involved. The Equations are derived from the econometric model and are of 2 parts ① Observed variables ② disturbances

II Statement about errors in the in the observed values of the variables.

III specification of the probability distribution of the disturbances or error.

- Econometric Models specifies statistical relationship b/w the various economic quantities pertaining to a particular economic phenomenon.

eg Monthly Expenditure Spending by customers is linearly dependent on their income in the previous month.

$$C_t = a + b Y_{t-1} + \varepsilon_t$$

- C_t → Customer's spending in a month
- a → constant or intercept
- b → slope
- Y_{t-1} → customer's income during previous month
- ε_t → error term showing the extent to which model cannot fully explain the consumptions.

Objectives:

1. To estimate a and b and fit the model
2. To predict for future observations.

Econometric Models consists of a set equations designed to provide a quantitative explanation of the behaviour of the economic variables.

Types of Econometric Models.

- Linear Regression
- Generalised Linear Model — a generalization of ordinary linear regression that allows for response to have an error distribution other than the normal distribution
- ARIMA — it analyses future values on past values. It is form of regression analysis that gauges strength of one dependent variable relative to other changing variables.
- Probit — The dependent variable can only take 2 values {married/unmarried}. It focuses binary response.
- Logit — also Logistic regression model used to model probability of certain class or event existing such as win/lose, alive/dead etc
- Tobit — A regression model in which observed range of dependent variable is censored - used to estimate linear relationship btw variables where there is either left or right censoring in dependent variable.
- Vector Auto Regression (VAR) — VAR model is a multivariate multivariable time series model that relates current observations of a variable with past observations of itself and past observations of other variables in the system.

Economic Models

- Economic model is a set of assumptions that describes behaviour of an economy or phenomenon under study

- Economic model Consists of mathematical equations that describes various relationships.

Utility maximization is made use of here., Individual make choices to maximise their well being which gives possible framework for creating economic model.

- Utility maximization lead to a set of demand equations

In demand equations, quantity demanded depends on price of goods (1), price of substitute and complementary goods (2), the consumer's income (3) & Individual tastes (4) that affects taste.

$$y \sim X_1 + X_2 + X_3 + X_4$$

y = GDP y = food expenditure

X_1 = Income, X_2 = education

X_3 = family size X_4 = age etc.

Another definition of Econometrics

- Econometrics include the study of methods for (1) selecting the models, estimating (2) the model and making inference on the model

Hint Econometrics — select model, estimate model, make inference

Structure of economic data

3 structures

1. Cross sectional data
2. Time series data
3. Panel data OR longitudinal data

} CTP (Cross section data, Time series data, panel data or longitudinal data)

⊕ Others are clustered data AND spatial data

Cross sectional Data:

- Consists of sample of Individuals, households, firms, cities, states, countries or a variety of other units taken at a given point in time
- Cross sectional data have one observation per individual
- It can be assumed that the CSD (data) have been obtained by random sampling underlying the population.

Example:

CSD on 526 random sampled persons in 1976 with variables

obs no	Wage	Education level	expenditure	sex	Marital status
1	34	11	2	1	1
2	3.24	12	22	0	0
⋮	⋮	⋮	⋮	⋮	⋮
526	3.50	14	5	1	0

Time series Data:

- Set of observations on a variable or several variables which are collected over time.

Example: ① GDP ② CPI ③ Money supply ④ Stock prices ⑤ Automobile sales figures etc.

- In time series data, we have lags. Past events can influence future events
- With lags and influence in the data; it is more difficult to analyse a time series data because observations are strongly related to recent histories.

② data frequency at which data is collected:

The frequencies may be Weekly Monthly Quarterly ~~Yearly~~ Daily etc

- Daily ⇒ Stock prices
- Weekly ⇒ Money supply (US)
- Monthly ⇒ Inflation, Unemployment rates
- Quarterly (every 3 months) ⇒ GDP
- Yearly ⇒ Infant Mortality Rate.

Example: Time series on Minimum Wage effect in a certain country

Panel data OR Longitudinal data.

- Panel data combines both cross sectional data and time series data
- Data are collected on time series basis: for each cross sectional member in the data set.
- Panel data consists of set of individuals, persons household, Corporation etc which are measured repeatedly over time

Example:

① Wage, Education and Employment data for a set of individuals followed over a 10-year period

② Investment & financial information about the same set of firms (eg Banks) over a 5-year period

Assumptions of panel data modelling

✓ Individuals are mutually independent of one another BUT a given Individuals observations are mutually dependent

Example (3)

3-year period panel data set on Crime and related statistics for 120 cities in a country

Crime variables may be Murder, Rape, Theft etc.

Basic Multiple Regression Model

- Regression Models are used to study relationship b/w variables
- This model can be econometric model which seeks to explain a dependent DV variable y in terms of some IV, X
- OR seeking to study how DV varies with changes in IV

$$y = f(x_1, x_2, \ldots x_k, \beta) + \varepsilon \quad \text{---} (1)$$

y = Regressand | DV | response var | explained variable.

$X = x_1, x_2, x_3 x_4 \ldots x_k$ = Regressors | IV | explanatory vars

β = parameter vector

ε = random error or disturbance

- We will use OLS which states that model is linear in parameters.

f in general may be ① Linear | Non linear in variables
② Linear or Non linear in parameters
③ Parametric or Non parametric.

eqn ① can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon \quad \text{---} (2)$$

here ε is unobserved factors

If model is linear in parameters we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

βs and ε are unobservable which are generated by ① Deterministic component

β_i's are marginal effects of x_i on y with other factors held constant

Aim : To estimate Unknown parameters
To test hypothesis about parameters
To predict values on y outside sample

Assumptions of classical Linear Regression Model (5 assumptions)

1. Linearity
2. Full rank of X
3. Errors have zero Mean
4. Homoscedasticity & Spherical errors
5. Normality of errors

(1) Linearity

Model is linear in parameters. This allows us to write the model in matrix form

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i$$

Linearity in parameters helps us to write the model in Matrix form.

$$y = X\beta + \varepsilon$$

$$\underset{n \times 1}{y} = \underset{n \times K}{X} \quad \underset{K \times 1}{\beta} + \underset{n \times 1}{\varepsilon}$$

(rows × cols)

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = x_i \beta + \varepsilon_i$$

(2) Full Rank

- There are no Exact linear dependencies among the columns of X
- If there is (linear dependencies) then one or more regressor is redundant
- Rank = no of linearly independent columns in the matrix X
  Since X is $n \times K$ matrix then $\text{rank}(X) = K \quad \forall n > K$

i.e. $\text{rank}(A) \leq \min(\#\text{rows}, \#\text{columns})$ which indicates that $n > K$

③ Errors have zero mean

It is assumed in the population that $E(\varepsilon_i) = 0 \quad \forall i = 1, 2, 3, \ldots n$

ε is $n \times 1$ matrix

$$E(\varepsilon) = E\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = 0$$

④ Homoscedasticity. {constancy of Variance of ε_i}

- In the population, the errors are generated by a process whose Variance, σ_ε^2 is constant. and that ε_i is Uncorrelated with each other.

$$Var(\varepsilon_i) = \sigma_\varepsilon^2 \qquad \forall i = 1, 2, 3, \ldots n$$

- Constancy of variance = assumption of homoscedasticity.
- ε_i are uncorrelated.

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \qquad \forall i \neq j \quad \{\text{No autocorrelation}\}$$

- Covariance Matrix for random vector ε can be generated

$$V(x) = E(x-\mu)^2$$

$$Var(\varepsilon) = E\{\varepsilon - E(\varepsilon)\}^2 = E\{(\varepsilon - E(\varepsilon))(\varepsilon - E(\varepsilon))'\}$$

$$= E\{\varepsilon\varepsilon' - \varepsilon E(\varepsilon)' - E(\varepsilon)\varepsilon' + E(\varepsilon)E(\varepsilon)'\}$$

$$= E\{\varepsilon\varepsilon'\} - E(\varepsilon E(\varepsilon)') - E(E(\varepsilon)\varepsilon') + E(E(\varepsilon)E(\varepsilon)')$$

but $E(\varepsilon) = 0 \quad E(\varepsilon') = 0$, then

$$Var(\varepsilon) = E(\varepsilon\varepsilon') - E(0) - E(0) + E(0)$$

$$= E(\varepsilon\varepsilon') - 0 - 0 - 0$$

$$Var(\varepsilon) = E(\varepsilon\varepsilon')$$

Now $E(\varepsilon\varepsilon') = Var(\varepsilon) =$

i \ j	ε_1	ε_2	$\cdots$	ε_n
ε_1	$\varepsilon_1\varepsilon_1$	$\varepsilon_1\varepsilon_2$	$\cdots$	$\varepsilon_1\varepsilon_n$
ε_2	$\varepsilon_2\varepsilon_1$	$\varepsilon_2\varepsilon_2$	$\cdots$	$\varepsilon_2\varepsilon_n$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
ε_n	$\varepsilon_n\varepsilon_1$	$\varepsilon_n\varepsilon_2$		$\varepsilon_n\varepsilon_n$

where $E(e_i) = E(e_i^2) = E(e_i - E(e_i))^2$

But $E(e_i) = 0$

$E(e_i e_i) = E(e_i - 0)^2 = E(e_i^2) = \text{Var}(e_i) = \sigma^2$

Likewise

$E(e_i e_j) = E\{(e_i - E(e_i))(e_j - E(e_j))\}$ But $E(e_i) = E(e_j) = 0$

$= E(e_i - 0)(e_j - 0)$

$= E\{(e_i) e_j\}$

$= \text{Cov}(e_i, e_j) = 0 \quad \forall i \neq j$ { No auto correlation between errors

$$\text{Var}(e) = \begin{Bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & & 0 \\ 0 & 0 & \cdots & & \\ \vdots & & & & \\ 0 & 0 & \cdots & & \sigma_n^2 \end{Bmatrix} = \sigma^2 I_n$$

I_n = Identity Matrix

Var(e) is a scalar matrix

(5) Normality of errors

Errors e_i are normally distributed as $e_i \sim N(0, \sigma^2 I_n)$ (mean / Variance)

Since $y = X\beta + e$

$E(y) = E(X\beta) + E(e)$

$E(y) = X\beta + E(e)$

$E(e) = 0$

$E(y) = X\beta + 0$

$E(y) = X\beta$

Variance

$V(y) = V(X\beta + e)$

$V(y) = V(X\beta) + V(e)$

$V(y) = 0 + V(e)$

$V(y) = \sigma_e^2$

$\therefore$ y has sampling distn

$y \sim N(X\beta, \sigma^2 I_n)$

Least squares Estimation of parameters β_s

From $y_i = X\beta + \varepsilon$ $\quad e_i \sim N(0, \sigma^2 I_n)$

We use Least squares strategy because we have $(k+1)$ parameters.

Our Unobserved error $e_i = y_i - x_i\beta$

equivalently for estimate $\hat{e}_i = y_i - x_i b$

$$y_i = X\hat{\beta} + \varepsilon$$

$$\hat{y}_i = Xb \text{ or } X\beta$$

then $y_i = \hat{y}_i + \varepsilon$

$$\varepsilon = y_i - \hat{y}_i$$

To use least square, i.e we minimize e_i

$$\hat{e}_i = y_i - \hat{y}_i$$

$$\hat{e}_i^2 = (y - \hat{y}_i)^2$$

$$= (y - \hat{y}_i)(y - \hat{y}_i)$$

Since $\hat{y} = X\beta$

$$\underset{n\times 1}{Y} = \underset{n\times k}{X}\ \underset{k\times 1}{\beta} + \underset{n\times 1}{e}$$

$e = Y - \hat{Y}$ But $\hat{Y} = X\beta$

$e_i = Y - X\beta$

$e_i' = (Y - X\beta)'$

$e'e = (Y - X\beta)(Y - X\beta)$

$= Y'(Y - X\beta) - X'\beta'(Y - X\beta)$

$= Y'Y - Y'X\beta - X'\beta'Y + X'\beta'X\beta$

$= Y'Y - Y'X\beta - X'\beta'Y + X'X\beta'\beta$

$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$

diff wrt β

$$\frac{\partial e'e}{\partial \beta} = \frac{\partial}{\partial \beta}\left(Y'Y - 2\beta'X'Y + \beta'X'X\beta\right)$$

$= 0 - 2X'Y + 2X'X\beta$

$= -2X'Y + 2X'X\beta$

To minimise, $\frac{\partial e'e}{\partial \beta} = 0$

$0 = -2X'Y + 2X'X\beta$

$-2X'Y = 2X'X\beta$

$X'Y = X'X\beta$

$\beta = \frac{X'Y}{X'X}$

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

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

Variance of error $s_e^2 = \frac{S_{yy} - bS_{xy}}{n-2}$ or $\frac{S_{yy}}{n-2}$

$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$ $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$ $S_{yy} = SST$

$b = \frac{S_{xy}}{S_{xx}}$ $s^2 = \frac{SSE}{n-2} = \frac{SST - SSR}{n-2}$

$s_e^2 = \frac{Y'Y - \hat{\beta}' X'Y}{n-r-1}$ r = no of variables X_1 X_2

$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$ But $\bar{y} = \frac{\Sigma y}{n}$ $\therefore n\bar{y} = \Sigma y$ then $\frac{(\Sigma y)^2}{n} = \frac{(n\bar{y})^2}{n} = \frac{n^2\bar{y}^2}{n} = n\bar{y}^2$

We deduct $n\bar{y}^2$ from $\hat{\beta}' X'Y$ i.e $\hat{\beta} X'Y - n\bar{y}^2$ when cal s_e^2

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

$V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ where $\sigma^2 \equiv s_e^2$ calculated

Testing parameter β_i

$t_{cal} = \frac{\hat{\beta}_i - 0}{Se(\hat{\beta}_i)}$ with $H_0 : \hat{\beta}_i = 0$ Reject H_0 if $t_{cal} > t_{tab}$

$H_1 : \hat{\beta}_i \neq 0$ $t_{tab} = t_{1-\frac{\alpha}{2}, n-r-1}$

CI $\Rightarrow \hat{\beta}_i \pm t_{1-\frac{\alpha}{2}, n-r-1} Se(\hat{\beta}_i)$

$R^2 = 1 - \frac{SSE}{SST}$ have r = no of pred

$R_a^2 = 1 - \frac{(n-1)}{(n-r)}\left(\frac{SSE}{SST}\right)$

$R_a^2 = 1 - \left(\frac{n-1}{n-r-1}\right)\left(\frac{SSE}{SST}\right)$

ANOVA Table

SV	df	SS
Reg on X_1	1	$\hat{\beta}_1 S_{1y}$
Reg on X_2	1	$\hat{\beta}_2 S_{2y}$
Error	$n-3$ no of variables	SSE
Total	$n-1$	$SST = S_{yy} = Y'Y$

$S_{1y} = \Sigma X_1 Y - \frac{\Sigma X_1 \Sigma Y}{n}$

$S_2 = \Sigma X_2 Y - \frac{\Sigma X_2 \Sigma Y}{n}$

$SSR = \hat{\beta}_1 S_{1Y} + \hat{\beta}_2 S_{2Y}$

have r = no of all variables

SV	df	SS	MS	F
Reg	$p-1$			
Error	$n-p$			
Total	$n-1$			

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad \text{—①}$$

from ① xply through by Σ

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

$$\Sigma Y = n\beta_0 + \beta_1 \Sigma X_1 + \beta_2 \Sigma X_2 \quad \text{—②}$$

from ① xply through by ΣX_1

$$\Sigma X_1 Y = \beta_0 \Sigma X_1 + \beta_1 \Sigma X_1^2 + \beta_2 \Sigma X_1 X_2 \quad \text{—③}$$

from ① xply through by ΣX_2

$$\Sigma X_2 Y = \beta_0 \Sigma X_2 + \beta_1 \Sigma X_1 X_2 + \beta_2 \Sigma X_2^2 \quad \text{—④}$$

Combine ② ③ ④

$$\Sigma Y = n\beta_0 + \beta_1 \Sigma X_1 + \beta_2 \Sigma X_2$$

$$\Sigma X_1 Y = \beta_0 \Sigma X_1 + \beta_1 \Sigma X_1^2 + \beta_2 \Sigma X_1 X_2$$

$$\Sigma X_2 Y = \beta_0 \Sigma X_2 + \beta_1 \Sigma X_1 X_2 + \beta_2 \Sigma X_2^2$$

$$\begin{pmatrix} \Sigma Y \\ \Sigma X_1 Y \\ \Sigma X_2 Y \end{pmatrix} = \begin{pmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2^2 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X^2 \end{pmatrix}^{-1} \begin{pmatrix} \Sigma Y \\ \Sigma X_1 Y \\ \Sigma X_2 Y \end{pmatrix}$$

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

$$\hat{Y} = X\beta$$

$$Y = X\beta + e$$

1. Multicollinearity
2. Heteroscedasticity
3. Autocorrelation

MULTICOLLINEARITY

- Presence of perfect linear or nearly perfect linear relationship among some or all the predictor variables of the regression model
- The parameters becomes indeterminate and OLS breaks down if the predictor variables are perfectly linearly correlated.
- There is no problem of multicollinearity if regressor are orthogonal i.e correlation coefficients (r_{xixj}) for the x's is zero. Then each parameter β_i can be estimated by SLR of Y versus X_i
- Multicollinearity can be present, but the issue is its magnitude or severity
- Multicollinearity is common/inherent in most economic relationships

Consequences of Multicollinearity

- OLS is still BLUE {Best Linear Unbiased estimate} But have large variances & covariances; making precision estimation difficult
  - If there is perfect correlation;
    1. estimates of coefficients are indeterminate
    2. std error of estimates are infinitely large
- If wider confidence interval, t ratio of one or more coefficients tends to be statistically insignificant.
- R^2 is high, even if t-ratio of one or more coeff is statistically insignificant. {There must be agreement between t-ratio and R^2

- OLS estimators and their standard errors can be very sensitive to small changes in the data
  If small changes in the data can change the values of OLS estimators and their standard errors

Detection of Multicollinearity (5)

- VIF
- High R^2 but few significant t-ratios
- High pair-wise correlation among regressors
- Through partial correlations
- Auxiliary regressions

F procedure $F_j = \frac{R_j^2/(p-1)}{(1-R_j^2)/(n-p)} \sim F(p-1, n-p)$

Rule: If $F_j > F^{\alpha}_{(p-1)(n-p)}$, we reject H_0

n = no of samples
p = no of regressors/predictors
R_j^2 = coeff of determination from regression of X_j on remaining predictors

H_0: X_j is not collinear
H_1: X_j is collinear

OR Use Klein Rule of thumb

H_0: No multicollinearity
H_1: Multicollinearity exists

Rule: If $R_j^2 > R^2$ (overall from reg of Y on all X_s), reject H_0

Auxiliary regression is the regression of the predictors/IV/regressors.
We regress X_j on remaining X_s and Compute R_j^2
Each of these regressions is known as Auxiliary regression.

Partial correlation: measuring the association btw 2 variables while controlling the effect of one or more additional variables
Considering X_1 X_2 X_3
Partial correlation of X_1 and X_2 is done by keeping X_3 fixed
X_2 and X_3 is " " X_1 fixed
X_1 and X_3 " " " X_2 fixed,

VIF (Variance Inflation Factor)

- To measure how variables are associated with each other
- VIF reflects how the variance of an estimator is inflated by presence of Multicollinearity
- VIF is used to check/detect Multicollinearity b/w variables

For X_1 and X_2,

$$VIF = \frac{1}{1 - r^2_{X_1,X_2}}$$

- High correlation between X_1 and X_2 suggest high VIF

R^2_j = coeff of determination of reg of X_j on remaining regressors in mo

If R^2_j increases towards Unity (i.e 1) collinearity of X_j with other regressors increases & the VIF also increases and can even be infinite

- If VIF > 10, X_j is highly collinear

The higher the value of VIF, the more collinear the X_j

eg If $R^2_j > 0.90$

Example

Regressors predictors $X_6 = 6$

$n = 70$. R^2 values for the Auxillary Regression on X_6 are given And also given is R^2 overall [for Reg of Y on all X_s]

Test for multicollinearity Using ① F test procedure

② Klein's Rule.

X_j	R_j^2	$R^2_{overall}$	F_j	F_{tab}	F Conclusion	Klein Conclusion
X_1	0.982	0.623	698.311	2.362	> *	> *
X_2	0.803	0.623	52.1746	2.362	> *	> *
X_3	0.651	0.623	23.876	2.362	> *	> *
X_4	0.314	0.623	5.8589	2.362	> *	< * *
X_5	0.213	0.623	3.4643	2.362	> *	< * *
X_6	0.124		1.8119	2.362	< * *	< * *

Hypothesis: F procedure

H_0: X_j is not collinear

H_1: X_j is collinear

Klein Rule

H_0: No multicollinearity

H_1: Multicollinearity exists

If $R_j^2 > R^2$ overall, reject H_0

R^2 overall $= 0.623$

Test statistic:

$$F_j = \frac{R_j^2 / k-1}{(1-R_j^2)/(n-k)} \quad j = 1, 2, 3 \ldots 6$$

$$F_1 = \frac{R^2/(k-1)}{(1-R_1^2)(n-k)} = \frac{(0.982)(6-1)}{(1-0.982)/(70-6)} = \frac{(0.982)/5}{0.018/64} =$$

$$F_1 = \frac{0.1964}{0.00028125} = 698.311.$$

$$F_2 = \frac{0.803/5}{(1-0.803)/64} = \frac{0.1606}{0.003078} = 52.1746$$

$$F_3 = \frac{0.651/5}{(1-0.651)/64} = \frac{0.1302}{0.005453} = 23.8762$$

$$F_4 = \frac{0.314/5}{(1-0.314)/64} = \frac{0.0628}{0.01072} = 5.8589$$

$$F_5 = \frac{0.243/5}{(1-0.243)/(64)} = \frac{0.0426}{0.012247} = 3.4643$$

$$F_6 = \frac{(0.124)/5}{(1-0.124)/64} = \frac{0.0248}{0.01369} = 1.8119$$

DR: If $F_j > F^{\alpha}_{(k-1)(n-k)}$ (df reg, df error)

$F^{\alpha}_{(6-1)(70-6)} = F^{0.05}_{(5,64)} = 2.362$

Alternatively, as $F_{5,64}$ is not readily available in F table,

$F_{5,64} = x$ $F_{560} = 2.37$ $F_{5120} = 2.29$

then $\frac{2.37-2.29}{60-120} = \frac{2.37-x}{60-64}$

$\frac{0.08}{60} = \frac{2.37-x}{4}$

$0.001333 = \frac{2.37-x}{4}$

$0.005333 = 2.73-x$

$x = 2.37 - 0.005333$

$x = 2.365$ {As our F tab}

Comments:

F-procedure :- X_1 X_2 X_3 X_4 and X_5 are all collinear as we reject H_0

Only X_6 is not collinear as we fail to reject H_0

Klein Rule: X_1 X_2 and X_3 are collinear WHILE X_4 X_5 and X_6 are not collinear as

collinear → R_1^2 R_2^2 R_3^2 > R^2 overall and however

Not collinear → R_4^2 R_5^2 R_6^2 < R^2 overall.

Solutions to Multicollinearity

① Increase the sample size:
gather more observation. It is possible that in another sample of greater size multicollinearity may not be serious as seen in 1st sample taken. Since covariances are inversely proportional to sample size, multicollinearity can be reduced.
This is true when ① multicollinearity characterizes the original sample and not the population
② if multicollinearity is due to error of measurement

② – Transformation of Variables
First difference & Ratio transformation can be used.

③ – Combine cross sectional data and time series data

④ – Use Factor Analysis & Principal component Analysis

⑤ – Introduce additional equations to express the relationship between multicollinear X_s. This can reduce multicollinearity
Hint: We now have simultaneous equation system instead of one equation.

Given $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + U_t$ ——— (1)

First differencing i

$$Y_{t-1} = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{2t-2} + U_{t-1} \quad \text{——— (2)}$$

(1) minus (2)

$$Y_t - Y_{t-1} = \beta_0 - \beta_0 + \beta_1 X_{1t} - \beta_1 X_{1t-1} + \beta_2 X_{2t} - \beta_2 X_{2t-2} + U_t - U_{t-1}$$

$$Y_t - Y_{t-1} = \beta_1 (X_{1t} - X_{1t-1}) + \beta_2 (X_{2t} - X_{2t-2}) + U_t - U_{t-1}$$

But let $V_t = U_t - U_{t-1}$, then

$$Y_t - Y_{t-1} = \beta_1 (X_{1t} - X_{1t-1}) + \beta_2 (X_{2t} - X_{2t-2}) + V_t \quad \text{——— *}$$

eqn * is reg model of the differences not that of the original variables.

NB: X_1 X_2 have high correlation doesn't mean their differences will have high correlation.

Ratio Transformation

From $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + U_t$ ——— (1)

Divide (1) by one of the Xs say X_2 i.e divide by X_{2t}

$$\frac{Y_t}{X_{2t}} = \frac{\beta_0}{X_{2t}} + \frac{\beta_1 X_{1t}}{X_{2t}} + \frac{\beta_2 X_{2t}}{X_{2t}} + \frac{U_t}{X_{2t}}$$

$$\frac{Y_t}{X_{2t}} = \frac{\beta_0}{X_{2t}} + \beta_1 \frac{X_{1t}}{X_{2t}} + \beta_2 + \frac{U_t}{X_{2t}}$$

NB: Transformation of error term may not satisfy some of OLS assumptions again.

OLS assumptions
- Linearity
- Normality of ε_i
- Homoscedasticity of error
- Full rank
- error term zero mean $E(\varepsilon_i) = 0$

Notables on Multicollinearity

I If sole purpose of Analysis is prediction, then multicollinearity is not a serious problem.
Reason: The higher the R^2, the better the prediction.

II If purpose of Regression Analysis is reliable estimation of parameters, Multicollinearity should not be ignored.
Reason: Estimators would have large standard errors and this suggests multicollinearity.

VIF
Farrar
Klein

Test of Multicollinearity.

Farrar and Glauber set of Tests.
- 3 tests

I – Chi square – detects existence & severity of multicollinearity.
II – F-test – locating variables that are multicollinear
III – t-test – determine variables responsible for appearance of multicollinear variables

1 Chisquare

- Test whether the regressors are pairwise orthogonal or not

orthogonality means $r_{x_i x_j} = 0 \quad \forall i \neq j$

- We use standardized determinant, D

D is of order $(K+1)$ by $(K+1)$

$$D = \begin{vmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_K} \\ r_{x_1x_2} & 1 & r_{x_2x_3} \cdots r_{x_2x_K} \\ \vdots & & & \\ r_{x_Kx_1} & r_{x_Kx_2} & \cdots & 1 \end{vmatrix}$$

$r_{x_1x_1} = r_{x_2x_2} = r_{x_3x_3} = \cdots = r_{x_Kx_K} = 1$

$r_{x_1x_2} = r_{x_2x_1}$

$r_{x_1x_K} = r_{x_Kx_1}$

at $K=3$ regressor

$$D = \begin{vmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_1x_2} & 1 & r_{x_2x_3} \\ r_{x_1x_3} & r_{x_2x_3} & 1 \end{vmatrix}$$

std det

$r_{x_1x_2} = r_{x_2x_1}$

$r_{x_1x_3} = r_{x_3x_1}$

$r_{x_2x_3} = r_{x_3x_2}$

for $K=2$ regressor, But X_s are perfectly linearly correlated.

then $D = \begin{vmatrix} 1 & r_{x_1x_2} \\ r_{x_1x_2} & 1 \end{vmatrix}$ i.e $r_{x_1x_2} = 1$

then

$$D = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} \Rightarrow D = 1 \times 1 - 1 \times 1 = 1 - 1 = 0$$

for $K=2$ regressors, But X_s are Orthogonal i.e $r_{x_ix_j} = 0$

$$D = \begin{vmatrix} 1 & r_{x_1x_2} \\ r_{x_1x_2} & 1 \end{vmatrix}$$

i.e $r_{x_1x_2} = r_{x_2x_1} = 0$

$$D = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \Rightarrow D = 1 \times 1 - 0 \times 0 = 1 - 0 = 1$$

Standard determinant

Hypothesis

H_0: X's are orthogonal { No multicollinearity }

H_1: Xs are not orthogonal { Multicollinearity exists }

Test statistics χ^2

$$\chi^2_{cal} = -\left\{n-1-\frac{1}{6}(2K+5)\right\}\log_e D$$

$$\chi^2_{tab} = \chi^2_{\frac{1}{2}K(K-1),\alpha}$$

K = no of Regres
OR
no of predict

DR

If $\chi^2_{cal} > \chi^2_{tab}$ reject H_0

Notables on D {Standardized determinant}

- If D lies btw 0 and 1, there is some degree of multicollinearity
- If D is close to 0 (zero) — then stronger the multicollinear
- If D is close to 1 — the weaker the multicollinearity.

Example:

A cross-sectional sample of 10 families expenditure on clothing (Y), disposable income (X_1), family size (X_2) and index for clothing (X_3) revealed the fg

$n = 10$

$K = 3$ regressors

$Y = 20.3 + 2.1X_1 + 1.5X_2 - 0.2X_3$ For Chi Square

$r_{X_1X_2} = 0.82 \quad r_{X_1X_3} = 0.70 \quad r_{X_2X_3} = 0.43$

$R^2_{X_1 \cdot X_2X_3} = 0.69 \quad R^2_{X_2 \cdot X_1X_3} = 0.32 \quad R^2_{X_3 \cdot X_1X_2} = 0.59$ Fisher

$r_{X_1X_2 \cdot X_3} = 0.71 \quad r_{X_1X_3 \cdot X_2} = 0.91 \quad r_{X_2X_3 \cdot X_1} = 0.36$ t-test

Q Perform Farrar and Glauber set of tests for multicollinearity at $\alpha = 0.05$

Solution

Chi-square

Hypothesis:

H_0: X_s are orthogonal { No multicollinearity }

H_1: X_s are not orthogonal { Multicollinearity exists }

Test Statistic

$$\chi^2_{cal} = -\left\{n - 1 - \frac{1}{6}(2k+5)\right\} \log_e D \quad \ln D$$

DR If $\chi^2_{cal} > \chi^2_{tab}$, reject H_0

$$\chi^2_{tab} = \chi^2_{\frac{1}{2}k(k-1),\,\alpha}$$

$k = 3$ regressors / predictors

Then $D = \begin{vmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_1x_2} & 1 & r_{x_2x_3} \\ r_{x_1x_3} & r_{x_2x_3} & 1 \end{vmatrix}$

$$D = \begin{vmatrix} 1 & 0.82 & 0.70 \\ 0.82 & 1 & 0.43 \\ 0.70 & 0.43 & 1 \end{vmatrix}$$

$$D = 1\begin{vmatrix} 1 & 0.43 \\ 0.43 & 1 \end{vmatrix} - 0.82\begin{vmatrix} 0.82 & 0.43 \\ 0.70 & 1 \end{vmatrix} + 0.70\begin{vmatrix} 0.82 & 1 \\ 0.70 & 0.43 \end{vmatrix}$$

$D = 1\{1 - 0.1849\} - 0.82\{0.82 - 0.301\} + 0.70\{0.3526 - 0.7$

$D = 1(0.8151) - 0.82(0.519) + 0.70(-0.3474)$

$D = 0.8151 - 0.42558 - 0.24318$

$D = 0.14634$

$D \simeq 0.1463$

$n = 10$
$k = 3$

$$\chi^2_{cal} = -\left\{n - 1 - \frac{1}{6}(2k+5)\right\}\log_e D$$

$$= -\left\{n - 1 - \frac{1}{6}(2k+5)\right\}\ln D$$

$$= -\left\{10 - 1 - \frac{1}{6}(2(3)+5)\right\}\ln 0.1463$$

$$= -\left\{9 - \frac{1}{6}(6+5)\right\}\ln 0.1463$$

$$\chi^2_{cal} = -\left\{9 - \frac{11}{6}\right\} \ln 0.1463$$

$$= -\left\{\frac{54-11}{6}\right\} \ln 0.1463$$

$$= -\left\{\frac{43}{6}\right\} \ln 0.1463$$

$$= -[7.1667] \ln 0.1463$$

$$= -7.1667 \times -1.9229$$

$$= 13.77511407$$

$$\simeq 13.7751$$

$$\chi^2_{tab} = \chi^2_{\frac{1}{2}K(K-1),\alpha}$$

$$= \chi^2_{\frac{1}{2}3(3-1),0.05}$$

$$= \chi^2_{\frac{1}{2}3(2),0.05} \quad \chi^2_{\frac{6}{2},0.05}$$

$$= \chi^2_{3,0.05}$$ {Check 3 under 0.05 straight up}

$$\chi^2_{tab} = 7.815$$

DR: If $\chi^2_{cal} > \chi^2_{tab}$, reject H_0

$13.7751 > 7.815$, We reject H_0 at 0.05 SL

Conclusion: there is multicollinearity i.e X_s are not orthogonal. Multicollinearity exists.

Test II : F test

— We use R^2 ~~Test~~ from auxiliary regressions. To locate variables which are collinear

Hypothesis

$H_0 : R^2_{X_j \cdot X_1 X_2 \ldots X_K} = 0$ { X_j is not multicollinear }

$H_1 : R^2_{X_j \cdot X_1 X_2 \ldots X_K} \neq 0$ { X_j is multicollinear }

Test statistics :-

$$F_J = \frac{R^2_{X_j \cdot X_1 X_2 \ldots X_K} / (K-1)}{(1 - R^2_{X_j \cdot X_1 X_2 \ldots X_K}) / (n-K)} \sim F^{\alpha}_{(K-1)\,n-K}$$

$$F_{tab} = F^{\alpha}_{(K-1)(n-K)}$$

DR :

If $F_J > F^{\alpha}_{tab}$, reject H_0

Example :

$n = 10$ $K = 3$ regressors. Locate variable that are collinear

Given $R^2_{X_1 \cdot X_2 X_3} = 0.69$ $R^2_{X_1 \cdot X_2 X_3} = 0.32$ $R^2_{X_1 \cdot X_2 X_3} = 0.5$

$H_0 : R^2_{X_j \cdot X_1 X_2 \ldots X_K} = 0$ { X_j is not multicollinear }

$H_1 : R^2_{X_j \cdot X_1 X_2 \ldots X_K} \neq 0$ { X_j is multicollinear }

Test statistic

$$F = \frac{R^2_{X_j \cdot X_1 X_2 \ldots X_k} / (k-1)}{\left(1 - R^2_{X_j \cdot X_1 X_2 \ldots X_k}\right) / (n-k)}$$

$n = 10 \quad k = 3$

$\Rightarrow R^2_{X_1 \cdot X_2 X_3} = 0.69$

$$F_{X_1} cal = \frac{0.69/(3-1)}{(1-0.69)/(10-3)} = \frac{(0.69)/2}{(0.31)/7} = \frac{0.345}{0.044} = 7.840 \simeq 7.84$$

$R^2_{X_2 \cdot X_1 X_3} = 0.32$

$$F_{X_3} cal = \frac{0.32/2}{(1-0.32)/7} = \frac{0.16}{0.68/7} = \frac{0.16}{0.0971} = 1.647785788 \simeq 1.65$$

$R^2_{X_3 \cdot X_1 X_2} = 0.36$

$$F_{X_3} cal = \frac{0.36/2}{(1-0.36)/7} = \frac{0.36/2}{0.64/7} = \frac{0.18}{0.0143} = 1.968279239 \simeq 1.97$$

$R^2_{X_3 \cdot X_1 X_2} = 0.59$

$$F_{X_3} cal = \frac{0.59/2}{1-0.59/7} = \frac{0.59/2}{0.41/7} = \frac{0.295}{0.05857} = 5.036708212 \simeq 5.03$$

$$F_{tab} = F^{\alpha}_{(k-1)(n-k)} = F^{0.05}_{(3-1)(10-3)} = F^{0.05}_{2(7)} = 4.74$$

$F_{tab} = F_{2,7,0.05} = 4.74$

DR: If $F_{cal} > F_{tab}$, reject H_0

- $7.84 > 4.74$, we reject H_0
- $1.65 < 4.74$, we fail to reject H_0
- $5.03 > 4.074$, we reject H_0

Conclusion:

X_1 is multicollinear

X_2 is not multicollinear

X_3 is multicollinear

Test III : t-test

- We detect Variable Variables that causes the multicollin
- We use partial correlation coefficients and test for their statistical significance using t-test:

Hypothesis

H_0 : $r_{x_i x_j \cdot x_1 x_2 \cdots x_k} = 0$ {X_i and X_j are not responsible for multicollinearity}

H_1 : $r_{x_i x_j \cdot x_1 x_2 \cdots x_k} \neq 0$ {X_i and X_j are responsible for multicol

Test statistic

$$t_{cal} = \frac{\left(r_{x_i x_j \cdot x_1 x_2 \cdots x_k}\right)\sqrt{n-k}}{\sqrt{1 - r^2_{x_i x_j \cdot x_1 x_2 \cdots x_k}}} \sim t_{(n-k)\frac{\alpha}{2}}$$

DR: If $|t_{cal}| > t_{tab}$ reject H_0 at α %.

$t_{tab} \equiv t_{n-k, \frac{\alpha}{2}}$

Example

$n = 10$ $k = 3$ detect the variable.

Given

$r_{x_1x_2 \cdot x_3} = 0.71$ $r_{x_1x_3 \cdot x_2} = 0.91$ $r_{x_2x_3 \cdot x_1} = 0.36$

Hypothesis:

H_0: $r_{x_ix_j \cdot x_1x_2 \cdots x_k} = 0$ {X_i AND X_j are not responsible for the multicollinearity}

H_1: $r_{x_ix_j \cdot x_1x_2 \cdots x_k} \neq 0$ {X_i AND X_j are responsible for the multicollinearity}

Test statistic

$$t_{cal} = \frac{r_{x_ix_j \cdot x_1x_2 \cdots x_k} \sqrt{n-k}}{\sqrt{1 - r^2_{x_ix_j \cdot x_1x_2 \cdots x_k}}} \sim t_{(n-k), \frac{\alpha}{2}}$$

$$t_{cal} = \frac{r\sqrt{n-k}}{\sqrt{1-r^2}}$$

$$t_{tab} = t_{(n-k)\frac{\alpha}{2}} = t_{(10-3), \frac{0.05}{2}} = t_{7, 0.025} = 2.365$$

$t_{tab} = 2.365$

$$t_{cal_1} = T_1 = \frac{r_{x_1 x_2 \cdot x_3}\sqrt{n-k}}{\sqrt{1 - r^2_{x_1 x_3 \cdot x_2}}}$$

$$= \frac{(0.71)\sqrt{10-3}}{\sqrt{1-(0.71)^2}} = \frac{(0.71)(\sqrt{7})}{\sqrt{1-0.5041}} = \frac{1.8785}{\sqrt{0.4959}}$$

$$= \frac{1.8785}{0.7042} = 2.667566032$$

$$\simeq 2.67$$

$$t_{cal_2} = T_2 = \frac{r_{x_1 x_3 \cdot x_2}\sqrt{n-k}}{\sqrt{1 - r^2_{x_1 x_3 \cdot x_2}}}$$

$$= \frac{0.91\sqrt{10-3}}{\sqrt{1-(0.91)^2}} = \frac{0.91(\sqrt{7})}{\sqrt{1-0.8281}} = \frac{0.91 \times 2.645}{\sqrt{0.1719}}$$

$$= \frac{2.407675}{0.4146082458} = 5.807115528$$

$$\simeq 5.81$$

$$t_{cal_3} = T_3 = \frac{r_{x_2 x_3 \cdot x_1}\sqrt{n-k}}{\sqrt{1 - r^2_{x_2 x_3 \cdot x_1}}}$$

$$= \frac{0.36\sqrt{7}}{\sqrt{1-(0.36)^2}} = \frac{0.36(2.6458)}{\sqrt{1-0.1296}}$$

$$= \frac{0.952488}{\sqrt{0.18704}} = \frac{0.952488}{0.9329523032} = 1.020934655$$

$$\simeq 1.02$$

$t_{tab} = t_{(n-1c)\alpha/2} = t_{7,10.025} = 2.365$

DR: If $t_{cal} > t_{tab}$, reject Ho

- 2.67 > 2.365, we reject Ho, ∴
- 5.81 > 2.365, we reject Ho
- 1.02 < 2.365, we fail to reject Ho

Conclusion:

X_1 AND X_2 are Interconvelated. X_1 and X_2 are responsible for multicollinearity

X_1 and X_3 are Interconvelated, X_1 and X_3 are responsible for multicollinearity

X_2 and X_3 are not interconvelated, X_2 AND X_3 are not responsible for Multicollinearity

$t_{tab} = t_{(n-k)\frac{\alpha}{2}} = t_{7,0.025} = 2.365$

DR: If $t_{cal} > t_{tab}$, reject H_0

- $2.67 > 2.365$, we reject H_0,
- $5.81 > 2.365$, we reject H_0
- $1.02 < 2.365$, we fail to reject H_0

Conclusion:

X_1 AND X_2 are Interconvelated. X_1 and X_2 are responsible for multicollinearity

X_1 and X_3 are Inteconvelated, X_1 and X_3 are responsible for multicollinearity

X_2 and X_3 are not interconvelated, X_2 AND X_3 are not responsible for Multicollinearity

Auto correlation refers to the degree of correlation between the values of same variables across different observations in the data

Auto correlation

$\begin{cases} E(u_i u_j) = 0 \\ Cov(u_i u_j \mid x_i x_j) = 0 \end{cases}$

- Assumption of auto correlation:
- This is about error terms.
- error terms are independent

i.e error term relating to any observation is not influenced by the error term relating to any other observation.

i.e one error term of observation doesn't influence error term of other observation

Successive values of error terms are independent

The assumption of $E(u_i u_j) = 0 \quad \forall i \neq j$ & $Cov(u_i, u_j \mid X_i X_j) = 0 \quad \forall i \neq j$ must hold

If it doesn't hold, we say we have auto-correlation or serial correlation.

Examples

① Effect of labour strike on output over fuel price hike should affect output over fuel price hike should not affect output in the following/Next month. {i.e successive values of u_i are ind

② Cross sectional data: Effect of an increase in one family's income on its consumption is not expected to affect consumption expenditure of another family.

Serial correlation is also known as auto correlation. It occurs when error of one time period a is correlated with error for a subsequent time period b.

i.e $E(u_i u_j) \neq 0$ as well as $Cov(u_i u_j | x_i x_j) \neq 0$

- Autocorrelation is the correlation between members of series of observations ordered in time (time series data) or ordered in space (cross-sectional data)

- Auto correlation is the relationship between successive values of same variable. BUT NOT relationship between two or more variables.

- Auto correlation occurs well in time series. It can be found in time series data.
- Auto correlation occurs least in cross sectional data except when sample is random.

- Auto correlation ranges btw -1 and $+1$

Sources of autocorrelation: (4)

1. (1) omitted predictor/explanatory variables
2. (2) Mis-specification of Model (3) Mis specification of true random error term
3. (4) Interpolations in observations
   (estimation of a value within known value in a sequence of values)

5 Consequences of Autocorrelation.

1. OLS estimates are not BLUE
2. OLS estimates remain Unbiased
3. OLS estimates have larger Variances.
4. Underestimated Variance of error term.
5. Prediction based on OLS estimates may not be effi

Solutions to Autocorrelation.

⟹ solution depends on whether autocorrelation is from misspecification or it is a case of pure autocorrelati

If it is a case of ~~omit~~ omitted explanatory Variab the solution/remedy is to Include such Variables. Otherwise we resort to ~~trans~~ transformation of original ~~model~~ model leading to method of Generalised Least square (GLS)

In GLS, we use knowledge of ρ usually not know It is Usually obtained through iterative - proce

Autoregression of order p

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + v_t$$

where $E(v_t) = 0 \quad E(v^2) = \sigma_v^2 \quad E(v_i v_j) = 0 \quad \forall i \neq j$

If $p = 1$, then we have first order autoregressive AR(1)

$$u_t = \rho u_{t-1} + v_t$$

AR(1) is mostly used for autocorrelation problem

A(1) is first order Autoregressive

Test for Autocorrelation.

- Von Neumann Test
- Durbin-Watson test ✓
- Rw test
- Breusch-Godfrey (BG) test

Detection of Autocorrelation

- Plot residuals e_t vs e_{t-1} residuals
- plot residuals e_t vs time (t)

Durbin Watson Test

- Suitable for first order auto regressive AR(

$$U_t = \rho U_{t-1} + V_t$$

$H_0: \rho = 0$ { U's (errors) are not auto correlated }

$H_1: \rho \neq 0$ { U's (errors) are auto correlated }

$$dw = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

$$dw = \frac{\sum_{t=2}^{n} (e_t^2 - 2e_t e_{t-1} + e_{t-1}^2)}{\sum_{t=1}^{n} e_t^2}$$

opening brackets $\sum_{t=2}^{n}$

$$dw = \frac{\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1} + \sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2}$$

$$dw = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

But recall that

$\sum_{t=1}^{n} e_t^2$, $\sum_{t=2}^{n} e_{t-1}^2$ and $\sum_{t=2}^{n} e_t e_{t-1}$ are

approximately same, then ✱ becomes.

[illegible]

$$dw = \frac{\sum e_t^2 + \sum e_{t-1}^2 - 2\sum e_t e_{t-1}}{\sum e_t^2}$$

$$dw = \frac{2\sum e_t^2 - 2\sum e_t e_{t-1}}{\sum e_t^2}$$

$$dw = \frac{2\left(\sum e_t^2 - \sum e_t e_{t-1}\right)}{\sum e_t^2}$$

$$dw = 2\left[\frac{\sum e_t^2}{\sum e_t^2} - \frac{\sum e_t e_{t-1}}{\sum e_t^2}\right]$$

$$dw = 2\left(1 - \frac{\sum e_t e_{t-1}}{\sum e_t^2}\right)$$

$$dw = 2(1 - \rho)$$

Where $\rho = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$

$d_w = 2(1-\rho)$

Zero correlation, $d = 2$

+ve correlation; $0 < \rho < 1$ $0 < d < 2$

-ve correlation, $-1 < \rho < 0$ $2 < d < 4$

i.e d lies btw 0 and 4

i.e $0 \leq d \leq 4$

DW assumptions (6)

- Regression Model Includes Intercept term
- Regressors are truly exogenous
- u_t are from first order AR(1) autoregressive
- u_t is normally distributed
- Reg Model doesn't have lagged values of dependent variable (Y)
- No missing obs in the data

Limitations / problems of dw

(1) If there are lagged ~~variables~~ values of dependent variable, we cannot use dw

(2) There is inconclusive region.
Inconclusiveness nature of results may arise

(3) Can only be used for first order auto-regressive AR(1)
i.e If ~~the~~ cannot be used for higher order auto correlation.