REGRESSION ANALYSIS AND TIME SERIES

SUBJECT CODE: 18KPISELSI

UNIT-I

Simple Linear Regression Model :

the functional relation of a dependent variable with one or more independent variable is Called regression equation. If the relationship between the independent variable x and dependent variable y. it is known as simple linear regression,

The Straight line equation is

$$Y = \beta_0 + \beta_1 x$$

Where,

β_0 is intercept.

β_1 is slope.

x is independent variable

Y is dependent variable.

Let the difference between the observed value of Y and the straight line $\beta_0 + \beta_1 x$ is an error, it is known and random error and is denoted by E,

Thus the simple linear regression model is written as

$$Y = \beta_0 + \beta_1 x + E$$

the errors are assumed to have mean zero and Variance σ^2. the parameters β_0 and β_2 are regression Co-efficient.

Estimation of $\hat{\beta_0}$ and $\hat{\beta_1}$ using Linear Square Method:

Suppose that we have n pairs data Say (x_1, y_1), $(x_2, y_2) \ldots (x_n, y_n)$. The method of least Square is used to estimate β_0 and β_1., so that the sum of the Squares of the difference between |t| observation y_i and the straight line minimum thus the least Square estimation is

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_i)^2$$

the LS estimates β_0 and β_1 Say $\hat{\beta_0}$ and $\hat{\beta_1}$ must Satisfy differentiating,

(i.e) $$\frac{\partial S}{\partial \beta_0}\Big|\hat{\beta_0}, \hat{\beta_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0 \rightarrow (1)$$

and

$$\frac{\partial S}{\partial \beta_1}\Big|\hat{\beta_0}, \hat{\beta_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0 \rightarrow (2)$$

from (1)

$$\sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0$$

$$\Rightarrow \Sigma y_i = n\hat{\beta_0} + \hat{\beta_1} \Sigma x_i \rightarrow (3)$$

From (2) multiply by xi

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i = 0$$

$$\Rightarrow \Sigma x_i y_i = \hat{\beta}_0 \,\Sigma x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \rightarrow (4)$$

Equ (3) & (4) are called least Square normal equations.

To estimate the value of $\hat{\beta}_0$ and $\hat{\beta}_1$

Consider,

$$\Sigma y_i = n\hat{\beta}_0 + \hat{\beta}_1 \,\Sigma x_i \rightarrow (4)$$

Multiply Σx_i on both sides.

$$\Sigma x_i \,\Sigma y_i = n\hat{\beta}_0 \,\Sigma x_i + \hat{\beta}_1 (\Sigma x_i)^2 \rightarrow (5)$$

Multiply Equ (5) by n we get.

$$n\,\Sigma x_i y_i = n\hat{\beta}_0 \,\Sigma x_i + n\hat{\beta}_1 (\Sigma x_i)^2 \rightarrow (6)$$

$\textcircled{6} - \textcircled{5} \Rightarrow$

$$n\,\Sigma x_i y_i - \Sigma x_i \,\Sigma y_i = n\hat{\beta}_1 \left[\Sigma x_i^2 - 1/n\,(\Sigma x_i)^2\right]$$

$$\cancel{n}\left[\Sigma x_i y_i - 1/n\,\Sigma x_i y_i\right] = \cancel{n}\,\hat{\beta}_1 \left[\Sigma x_i^2 - 1/n\,(\Sigma x_i)^2\right]$$

$$\therefore \hat{\beta}_1 = \frac{\Sigma x_i y_i - 1/n\,\Sigma x_i y_i}{\Sigma x_i^2 - 1/n\,(\Sigma x_i)^2}$$

÷ equation (3) by n, we get,

$$\frac{\Sigma y_i}{n} = \hat{\beta}_0 + \hat{\beta}_1 \frac{\Sigma x_i}{n}$$

$\therefore \bar{y} - \frac{\Sigma y_i}{n}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\therefore \bar{x} = \frac{\Sigma x_i}{n}$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

$$= \Sigma (x_i - \bar{x})^2$$

$$S_{xy} = \Sigma x_i y_i - \frac{\Sigma x_i \, \Sigma y_i}{n}$$

$$= \Sigma y_i (x_i - \bar{x})$$

$$\therefore \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} - \sum_{i=1}^{n} c_i y_i$$

where,

$$c_i = \left(\frac{x_i - \bar{x}}{S_{xx}}\right) \quad (i = 1, 2 \cdots n).$$

Statistical Assumptions in linear model :

Let $E_i = \mu_i - \bar{\mu}$ where μ_i is the disturbance term.

i) Normality :

The disturbance term are normally distributed

$$\mu_i \sim N(\sigma, \sigma^2)$$

ii) Zero mean :

The mean of μ_i is Zero (ie) $E(\mu_i) = 0$

$$\forall i = 1, 2 \ldots n$$

iii) Homo cedasiticity :

Every disturbance variance has the same variance, whose value is unknown.

(ie) $E(u_i^2) = \sigma_u^2$, $\forall\ i = 1, 2 \cdots n$.

iv) Non Auto Regression :

The various disturbance terms are uncorrelated.

(ie) $E(\mu_i \mu_j) = 0$ for $i \neq j$ $i, j = 1, 2, \ldots n$.

v) Non - Stochastic :

If $x - x_i$ is a non - Stochastic variable Such that for each sample of size (n).

Properties of least Square estimators for Simple linear regression model.

i) the least Square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of the model parameters β_0 and β_1.

Proof :

$$E(\hat{\beta}_1) = E\left[\sum_{i=1}^{n} c_i y_i\right]$$

$$= \sum_{i=1}^{n} c_i \, \Sigma(y_i)$$

$$= \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 x_i\right).$$

$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

But We know that,

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})}{S_{xx}} = \sum_{i=1}^{n} c_i = 0 \quad \text{and}$$

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})\, x_i}{S_{xx}} = \frac{\Sigma x_i^2 - \bar{x}\, \Sigma x_i}{S_{xx}}$$

$$= \Sigma x_i^2 - \frac{\Sigma x_i \, \Sigma x_i}{n}$$

$$= 1$$

$\therefore E(\hat{\beta_1}) = 0 + \beta_1 v_1$

$E(\hat{\beta_1}) = \beta_1$

$E(\hat{\beta_0}) = E[\bar{Y} - \hat{\beta_1}\bar{x}]$

$= E(\bar{Y}) - \hat{\beta_1}\ E(\bar{x})$

$= \bar{Y} - \beta_1 \bar{x}$

$= \beta_0$

ii) the least Square estimators are linear estimators.

We know that,

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1} \cdot \bar{x} \text{ and}$$

$$\hat{\beta_1} = \frac{\Sigma y_i x_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

are linear combinations of the observations y_i,

Thus $\hat{\beta_0}$ and $\hat{\beta_1}$ are linear estimators.

iii) Least Square estimators are Best estimators.

$$V(\hat{\beta_1}) = V\left(\sum_{i=1}^{n} c_i y_i\right)$$

$$= \sum_{i=1}^{n} c_i^2\, V(y_i)$$

because the observations y_i are uncorrelated and So the variance of the sum is just the sum of the variance of each term in the sum is

$C/V(y_i)$ and

We have assumed $V(y_i) = \sigma^2$

$$V(\hat{\beta_1}) = \sigma^2 \Sigma c_i^2$$

$$= \frac{\sigma^2 \Sigma (x_i - \bar{x})^2}{S_{xx}^2}$$

$$= \frac{\sigma^2}{S_{xx}}, \quad S_{xx} = \Sigma (x_i - \bar{x})^2$$

The variance of $\hat{\beta_0}$ is,

$$V(\hat{\beta_0}) = V(\bar{y} - \hat{\beta_1}\bar{x})$$

$$= V(\bar{y}) + \bar{x}^2 V(\hat{\beta_1}) - 2\bar{x} \operatorname{cov}(\bar{y}, \hat{\beta_1})$$

We know that,

$$V(\bar{y}) = \sigma^2/n \quad \text{and} \quad \operatorname{cov}(\bar{y}, \hat{\beta_1}) = 0$$

$$\therefore V(\hat{\beta_0}) = V(\bar{y}) + \bar{x}^2 V(\hat{\beta_1})$$

$$= \sigma^2 \left(1/n + \frac{\bar{x}^2}{S_{xx}} \right)$$

Thus the least square estimators are unbiased linear and have minimum variance when compared with all other unbiased estimators.

Thus the least square estimators are best linear unbiased estimators.

Other properties of the least Square fit.

1, the Sum of the residues is any regression model that Contains an intercept β_0 is always Zero.

$$(ie)\quad \Sigma(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$$

2) the Sum of the observed values y_i equals the Sum of the fields values $\hat{y}$.

$$(ie)\quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

3) the Least Square regression line always Passes through the Centeroid of the data.

4) the Sum of the residuals weighted by the corresponding value of the regression Variate always equal.

$$(ie)\quad \sum_{i-1}^{n} x_i\, e_i = 0$$

5) the Sum of the residuals weighted by the corresponding fitted value always Zero, that is

$$\sum_{i=1}^{n} \hat{y}\, e_i = 0.$$

Estimation of σ^2.

the estimate of σ^2 is obtained from the residual (or) error sum of Squares.

$$SS_{Res} = \Sigma e_i^2 = \Sigma (y_i - \hat{y}_i)^2$$

W.K.T

$$\hat{Y}_i = \hat{\beta}_1 , \hat{\beta}_1 x_i$$

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

It can be written as,

$$= \Sigma y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

But,

$$\Sigma y_i^2 - n\bar{y}^2 = \Sigma (y_i - \bar{y})^2 = SS_T$$

is the corrected Sum of Squares of the response observations s

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$$

The residuals Sum of Squares has n-2 d.f because two degrees of freedom are associated with the estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ involved in obtaining $\hat{y}_i$

So an unbiased estimators of σ^2 is $\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$

Where M_{Res} is called the standard error of regression the square root $\sqrt{\hat{\sigma}^2}$ is called the standard error of regression.

Hypothesis Testing on the Slope.

Null hypothesis :

H0 : the Slope equals a Constant Say B0

(ie) $H_0 : \beta_1 = \beta_{10}$

Alternative hypothesis :

H1 : the Slope does not equals a Constant Say B1

(ie) $H_1 : \beta_1 \neq \beta_{10}$

Test Statistic,

under H0,

$$t = \frac{\hat{\beta_1} - \beta_{10}}{\sqrt{MS_{Res} / S_{xx}}}$$

Where,

$$MS_{Res} = \frac{\Sigma y_i^2 - n\bar{y}^2 - \beta_1 S_{xy}}{n-2}$$

$$S_{xy} = \Sigma y_i x_i - \frac{\Sigma y_i \, \Sigma x_i}{n}$$

$$= \Sigma y_i \, (x_i - \bar{x})^2$$

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

Table Value :

Refer Student t-table for (n-2) of f at α% level.

Inference :

If Calculated Value < table Value do not reject H0 (ie) $H_0 : \beta_1 = \beta_{10}$, otherwise reject H0

Hypothesis Testing for intercept β_0:

Null hypothesis:

$$H_0 : \beta_0 = \beta_{00}$$

Alternative hypothesis:

$$H_1 : \beta_0 \neq \beta_{00}$$

Test Statistic:

under H_0,

$$t = \frac{\hat{\beta_0} - \beta_{00}}{\sqrt{MRes\left(\frac{1}{n} + \bar{x}^2/S_{xx}\right)}}$$

Table Value

Refer Studen's t - table for (n-2) d.f at α% level

Inference:

If the Calculated Value < table Value do not reject H_0.

(ie) $H_0 : \beta_0 = \beta_{00}$

otherwise reject H_0.

Estimation By Maximum Likelihood Method :

the method of least Square Can be used to estimate the Parameters in a linear regression model regard sets of the form of the distribution of the errors E.

If the form of the distribution of the errors is known as alternative method of Parameter estimation is method of maximum likelihood.

Consider the data (yi, xi) (i=1,2 ... n)

if we assume that the errors in the regression models are Normally identically distributed $(0, \sigma^2)$. then the observations Y_i in the Sample are normally and identically distributed random Variables with means $\beta_0 + \beta_1 x$ and Variance σ^2 For Simple Linear regression model with normal errors, the Likelihood function is,

$$L(y_i, x_i, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2}$$

$$\exp\left\{ -\frac{1}{2\sigma}^2 (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} \Sigma(y_i - \beta_0 - \beta_1 x_i^2) \right\}$$

the MLE are the parameters Values Say,

$\hat{\beta_0}$, $\hat{\beta_1}$ and $\hat{\sigma}_1^2$ that maximise L or lnL thus,

$$\text{In } L\,(yi,\, xi,\, \beta o,\, \beta_1,\, \sigma^2) - (n/2)\ \text{In} \cdot 2\pi$$

$$-n/2\ \text{In}\,\sigma^2 - \left\{\frac{1}{2\sigma_2}\ \Sigma\,(yi - \beta o - \beta_1\, xi)^2\right\}$$

$$\frac{\partial\, \text{InL}}{\partial \beta o}\,\Big|\, \hat{\beta o},\, \hat{\beta i},\, \hat{\sigma}^2 = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (\bar{y}i - \hat{\beta o} - \hat{\beta i}\ xi)^2 = 0$$

$$\frac{\partial\, \text{InL}}{\partial \beta i}\,\Big|\, \hat{\beta o},\, \hat{\beta i},\, \hat{\sigma}^2 = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (yi - \hat{\beta o} - \hat{\beta i}\ xi) = 0$$

and,

$$\frac{\partial\, \text{InL}}{\partial \sigma^2}\,\Big|\, \hat{\beta o},\, \hat{\beta i},\, \hat{\sigma}^2 = \frac{-n}{2\hat{\sigma}^2} = \frac{-n}{2\hat{\sigma}^2} + \frac{1}{2\sigma} \sum_{i=1}^{n} (y_i - \hat{\beta o} - \hat{\beta}\ x_1)$$

UNIT - II

Multiple Linear Regression Model.

A Regression model that involves more than one regression Variable is called a multiple regression model

the multiple regression model is,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + E$$

this is a multiple linear regression model with two regression Variables.

In general the model,

$$Y - \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_k + E$$

is called a multiple regression model with k regression.

the parameters β_j $(j=1,2 \ldots k)$ are called regression Co-efficient.

Data for multiple Regression Model.

Observation	frequency	Regressions
1	y_1	$x_1, x_2 \cdots x_k$
2	y_2	$x_{11}, x_{12} \cdots x_{1k}$
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1} x_{n2} x_{nk}

the regression model Can be written as,

$$y_1 = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \epsilon_i .,$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \quad (i=1,2 \ldots n) .$$

Estimation of the model Parameters least square estimation of the regression Co-efficient.

the multiple linear regression model is

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon_j \quad (i=1,2 \ldots n)$$

the linear square function is,

$$S(\beta_0, \beta_1, \cdots \beta_K) = \sum_{i=1}^{n} \epsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij})$$

$$\frac{\partial S}{\partial \beta_0}\Big|\hat{\beta_0},\hat{\beta_1},\ldots\hat{\beta_k} = -2\sum_{i=1}^{n}(y_i-\hat{\beta_0}-\sum_{j=1}^{k}\hat{\beta_j}\,x_{ij}) = 0 \rightarrow (1)$$

and

$$\frac{\partial S}{\partial \beta_j}\Big|\hat{\beta_0},\hat{\beta_1},\ldots\hat{\beta_k} = -2\sum_{i=1}^{n}(y_i-\hat{\beta_0} = \sum_{j=1}^{k}\hat{\beta_j}\,x_{ij})\,x_{ij} = 0 \rightarrow (2)$$

when.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix}; \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$(\beta) = \sum_{i=1}^{n}\epsilon_i^2 = \epsilon'\epsilon(Y-X\beta)'\,(Y-X\beta)$$

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

The linear Square estimators must satisfy,

$$\frac{\partial S}{\partial \beta}\Big|\hat{\beta} = -2X'y + 2X'X\hat{\beta} = 0.$$

$$\Rightarrow -2(X'y - X'X\hat{\beta}) = 0$$

$$\Rightarrow X'y\,X'X = X'X\beta'.$$

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

Properties of least Square estimators for multiple linear model.

i) The least Square estimators are unbiased estimators.

We know That,

$$\hat{\beta} = (x'x)^{-1} x'y$$

$$E(\hat{\beta}) = E((x'x)^{-1} xy)$$

$$= E[(x'x)^{-1} x' x\beta' (x'x)^{-1} x'\epsilon]$$

$$= \beta$$

$$\therefore E(\epsilon) = 0.$$

ii) The least Square estimators are linear estimator.

Since the least Square estimators are,

$$\hat{\beta} = (x'x)^{-1} x'y$$

$$\Rightarrow (x'x)^{-1} x [(x\beta + \epsilon)]$$

$$\Rightarrow (x'x)^{-1} x'x\beta + (xx^{-1})^{-1} x' \in$$

$$= \beta + (x'x)^{-1} x' \in$$

iii) Least Square estimates are the last estimates.

Proof :-

We know that

$$\hat{\beta} = (x'x)^{-1} \; x'y$$

$$= (x'x)^{-1} \; x' (x\beta + \epsilon) .$$

$$= [(x'x)^{-1} \; x'x\beta + (x'x)^{-1} x] \epsilon$$

$$= \beta + (xx^{-1}) x' \epsilon$$

$$\hat{\beta} - \beta = (x'x)^{-1} \; x' \epsilon$$

$$v(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta')]$$

$$= E[(x'x)^{-1} x'\epsilon] \quad (x'x)^{-1} x'\epsilon$$

$$\therefore E(\epsilon\epsilon)' = \sigma_{\epsilon}^2$$

$$v(\hat{\beta} = \sigma_{\epsilon}^2 \; (x'x)^{-1} (x'x) \; (x'x)^{-1}$$

$$\Rightarrow v(\hat{\beta}) < v(E)$$

Estimation of σ^2

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i) \Rightarrow \sum_{i=1}^{n} e_i^2 \Rightarrow e'e$$

Sub $e = y - x\hat{\beta}$

$$SS_{Res} = (y - x\hat{\beta})'(y - x\hat{\beta})$$

$$= y'y - \hat{\beta}' y' y - y' x \hat{\beta} + \hat{\beta} x' x \beta$$

$$= y'y - 2\hat{\beta} x' y + \hat{\beta}' x' x \hat{\beta}$$

We know that,

$$((x'x)\hat{\beta}) = x'y$$

$$SS_{Res} = y'y - 2\hat{\beta}' y + \hat{\beta}' x' y$$

$$= y'y - \hat{\beta}' x' y$$

the residual mean square is,

$$MS_{Res} = \frac{SS_{Res}}{n-p}$$

$$\Rightarrow \sigma^2 = MS_{Res}.$$

Assumption in multiple linear regression model.

Some assumption are needed in the model

$Y = X\beta + \epsilon$ for drawing the Satisfied inference.

the following assumption are made

i) $E(\epsilon) = 0$

ii) $E(\epsilon\epsilon') = \sigma^2 I_n$

iii) $\text{Rank}(X) = K$

iv) X is a non-stochastic matrix

v) $\Sigma \sim N(0, \sigma^2 I_n)$

vi) $\lim_{n \to \infty} \left(\frac{X'X}{n} \right) = \Delta$

exists and it is a non-stochastic and non-singula

matrix (with finite elements).

Estimation of parameters.

A general procedure for the estimation of regression co-efficient is to minimize:

$$\sum_{i=1}^{n} M(\epsilon_i) = \sum_{i=1}^{n} M(y_i - x_{i1}\beta_1 - x_{i2}\beta_2 \cdots - x_{in}\beta_{ik})$$

for a suitably chosen function M.

Some examples of choice of M are

$$M(x) = |x|$$

$$M(x) = x^2$$

$M(x) = |x|^p$ is general,

We consider the principal of Least Square which is related to $M(x) = x^2$ and the method of maximum likelihood estimation for the estimation of parameters.

Using Regression Models for forecasting :-

- Forcasting and estimation of Casual effects are Quite different Objectives.

For forecasting,

- $\bar{R}^2$ matters a lot.
- Omitted Variable bias isn't a problem.

* We will not worry about interpretingCo-fficients in foreasting models.

* External Validity is Paramount : the model estimated using historical data must hold into the near future.

* External Validity is Paramount : the model estimated using historical data must be hold into the (near) future.

UNIT - III

Generalized Least Squares :-

Until now we assumed that $\Sigma = \sigma^2 I$ but it can happen that the errors have non-constant variance or are correlated. Suppose instread that $\Sigma = \sigma^2 \Sigma$ where σ^2 is Known but Σ is known - in other words we know that the correlation and relative variance between the errors but we dont know the absolute scale

$$(y - x\beta)^T \Sigma^{-1} (y - x\beta)$$

Which is solved by

$$\hat{\beta} = (x^T \Sigma^{-1} x)^{-1} x^T \Sigma^{-1} y$$

Since we can write $\Sigma = SS^T$, where S is triangular matrix using the choleski Decomposition, we have

$$(y - x\beta)^T S^{-T} S^{-1} (y - x\beta) = (S^{-1} y - S^{-1} x\beta)^T$$

So GLS is like regressing $S^{-1}x$ on $S^{-1}y$.

Furthermore

$$y = x\beta + \epsilon$$

$$S^{-1} y = S^{-1} x\beta + S^{-1}\epsilon$$

$$y' = x'\beta + \epsilon'$$

So we have a new regression equation $y' = x'\beta + \epsilon'$ where if we examine the variance of the new errors. ϵ' we find

$$\text{var}\,\varepsilon' = \text{var}(S^{-1}\varepsilon) = S^{-1}(\text{var}\,\varepsilon)S^{-T}$$
$$= S^{-1}\sigma^2 SS^T S^{-T} = \sigma^2 I$$

So the new variable y' and x' are related by a regression equation which has uncorrelated errors with equal variance. of course, the practical problem is that Σ may not be known.

We find that

$$\text{var}\,\hat{\beta} = (X^T \Sigma^{-1} X)^{-1}\sigma^2$$

Weighted Least Squares:-

Sometimes the errors are uncorrelated, but have unequal variance where the from the inequality is known. Weighted least squares (WLS) can be used in this situation. When Σ is diagonal, the errors are uncorrelated but do not neccessarily have equal variance. We can write $\Sigma = \text{diag}$ $(1/w_1, \cdots 1/w_n)$, where the w_i are the weights so $S = \text{diag}(\sqrt{1/w_1}, \cdots \sqrt{1/w_n})$. So we can regress $\sqrt{w_i}x_i$ on $\sqrt{w_i}y_i$ although the column of ones in the X-matrix needs to be replaced with $\sqrt{w_i}$. cases with low variability should get a high weight, high variabilitya low weight. Some example:

1. Errors proportional to a predictor. $\text{var}(\varepsilon_i) \propto x_i$ suggests $w_i = x_i^{-1}$

2. Y_i are the average of n_i observations then $y_i = \text{var}\,\varepsilon_i = \sigma^2/n_i$ suggests $w_i = n_i$.

The experiment was designed to test certain theories about the nature of ~~a~~ the strong interaction. The cross-section variable is believed to be linearly related to the inverse of the energy. At each level of the momentum, a very large number of observations were taken to that it was possible to accurately estimate the standard of the response (s.d).

Robust Regression:-

In robust statistics robust regression is a form of regression ~~anlax~~ analysis designed to overcome some limitations of traditional parametric seeks and nonparametric methods. Regression analysis is to find the relationship between one or more independent variables and a dependent variable. Certain widely used methods of regression, such as ordinary least squares, have favourable properties in their underlying assumptions are true, but can give misleading results if those assumptions are not true, thus ordinary least squares is said to assumptions by the underlying data-generating process.

In particular least square estimates for regression models are highly sensitive to outliers. while there is no precise definition of an outlier, outliers are observations that do not follow the Pattern of Other observations. This is not normally a pattern if the outlier is simply an extreme observation drawn from the tail of a normal distribution, but if the outlier results from non-normal measurement error or some other violation standard ordinary least Squares assumptions, then it compromises the validity of the regression result if a non-robust requal regression technique is used.

Relationship between Analysis of Variance and regression:-

(or)

Difference between Regression and anova :-

Both the regression and anova are the that statistical methods which are used in order to predict the continuous outcome but in case of the regression, continuous outcome is predicted on basis of the one or more than one continuous Predictor Variable whearreas in case of anova continuous outcome is predicted on basis of the one or more than one categorical predictor Variables.

Regression is a statistical method to establish the relationship between Sets of Variables in order to make predictions of the dependent Variable with the help of independent variables. Anova on the other hand, is a statistical tool applied to unrelated group to find out whether they have a Common mean.

What is Regression Analysis ?

Regression is a very effective statistical method to ~~as~~ establish the relationship between set of Variables. The variables for which the regression analysis is done are the dependent Variable and one or more independent Variable. It is a method to understand the effect on a dependent Variable of one or more than one independent Variable.

(i) Suppose, the example; a paint company uses one of the derivatives of crude solvent & monomers as its raw meterial. We can run a regression analysis between the price of that raw meterial and the price of Brent Crude Prices.

(ii) In this example, the price of the raw metrial is the dependent variable and the price of Brent Prices is the independent variable.

(iii) As the price of solvents and monomers increases and dereases in price with the rise and fall of Brent prices, the Price of the raw meterial is the dependent Variable.

(iv) Similarly for any business decision in order to validate a hypothesis that a particular action will lead to the increases in the profitability of a division can be validated based on the rusult of the regression between the dependent and independent variable.

What is ANOVA?

ANOVA is the short from of analysis of Variance. ANOVA is a statistical tool that is generally used on random variables. It involves group not directly related to each other in order to find out whether there exist any common means.

(i) A simple example to understand this point to run ANOVA for the series of marks of Student from different college in order in to try to find out whether one Student from one school is Better than the other.

(ii) Another example can be if two separate research team is researching different product not related to each other. ANOVA will help to find which one is providing Better results. the three popular techniques of ANOVA are a random effect, fixed effect, and mixed effect.

Generalized Linear Model :-

The generalized Linear model (GLm) is a flexible generalization of ordinary Linear regression that allows for response Variables that have error distribution models other a normal distribution. The GLm generalizes Linear regression by allowing the Linear model to be related to the response Variable Via a Link function and by allowing the magnitude of the Variance of each measurement to be a function of its predicted Value.

Generalized Linear models were formulated by John Nelder and Robert Wedderturn as a way of unifying various other statistical models, including Linear regression, logistic regression and poisson regression. The proposed an alternatively reweighted least squares method for maximum Likelihood estimation ~~esti~~ of the model parameters. Maximum-likelihood estimation remains popular and is the default method of many statistical computing Package, other approaches including Bayesian approaches and least square fits to Variance stabilized responses have been developed.

Comparison to generalized Linear Model:-

The general linear mode (GLm)(2)(3) and generalized linear Model(4)(5) are two commonly used families of statistical methods to relate Some number of continuous and/or Categorical Predictor to single outcome Variable.

The main difference between the two approaches is that the GLm stricly assumes that the residuals will follow a conditionally normal distribution(3) with the GLm Loosens this assumption and allows for a variety of the other distributions from the exponential family for the residuals. of note, the GLm is a special case of the GLm in which the distribution of the residuals follow a conditionally normal distribution.

Non-Linear ~~Regr~~ Regression Model :-

NonLinear regression is a form of regression analysis in which observations data are modeled by a function Which is a non-Linar common combination of the model parrameters and depends on one or more independent Variables. The data are filted by a method of sucessive ~~approxantion~~. approximation.

UNIT-IV

MULTICOLINEARITY - Definition :-

Multicolinearity generally occurs when there are high correlations between two or more predictor variables. In other words, one predictor variable can be used to predictor the other. This creates redundant information skewing the results in a regression model.

For example :-

A person's hight and weight, age and sales price of a car, or year of education and annual income.

An easy way to detect multicolinearity to calculate correlation coefficient forall pairs of predictor variables. If the correlation coefficient, r is exactly $+1$ or -1, this is called predictor perfect multicolinearity. If r is close to or exactly -1 or $+1$. one of the variables should be removed from the model if all possible

It's more common for multicolinearity to rear its ugly head in observational studies, its less common with experimental data.

When the condition is present, it can result in unstable and unreliable regression estimates. Several other problems can interfere with analysis of results, including

(i) The t-statistic will generally be very small and coefficient confidence interval will be very wide. The means that it is horder to reject the null hypothesis.

(ii) The Partial regression coefficient may be an imprecise estimate, standard error may be very wide. This means standard errors may very large.

(iii) Partial regression coefficient may have sign and/or magnitude change as they pass from sample to sample.

(iv) Multicolinearity makes it difficult to gauge the effect of independent variables on dependent variables.

Comparing Regression Models to See the Effect of Multicolinearity:-

The predictors effectively removed the multicolinearity, we could run the same model twice, once with severe multicolinearity and once with moderate multicolinearity.

This provides a great head-to-head comparison and it reveals the classic effects of multicolinearity.

The standard error of the coefficient indicate the precision of the coefficient estimates. Smaller values represent more reliable estimate. In the second model, you can seethat the SE coefficient is smaller for both % Fat and Weight.

Also, % Fat is significant this time, while it was insignificant in the model with severe multicolinearity. Also its sign has switched from +0.005 to -0.005! The % Fat estimate in both models is about the same absolute distance from zero, but it is only significant in the second model because the estimate is more precise

Compare the summary of model statistics between the two models and you'll notice that, S, ~~R~~ R-squared, adjusted R-required, and the other are all identical. Multicolinearity dosen't effect how well the model fits. In fact, if you want to use the model to make predictions, both models produce indentical results for fitted values and prediction interval

Dealing with multicolinearity:-

Multicolinearity 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 Variable is high enough, it Can cause problems when you fit the model and interpret the result.

In Some causes, multicolinearity isn't necessarily a problem, and I'll Show you how to make this determination. I'll work through an example datest which contains multicolinearity to bip bring it all to life.

Multicolinearity based on the 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 Variable Constant. That Last Portion is crucial for our discussion about multicolinearity.

The idea is that you can change the Variable of one independent and not the others. However, when independent Variables are correlated. it indicates that changes in one Variable are associated with shifts in another Variable.

The Stronger the correlation, the more difficult it is to change one Variable without changing another. It becomes difficult for the model to estimate the relationship between each independent Variable and the dependent Variable independent because the independent Variable tend to change in unison.

There are two basic kinds of multicolinearity:

(i) Structural multicolinearity:-

This type occurs when we create a model term using other terms. In other words, it's a byproduct of the model that we specify rather the being present in the data itself.

For example:-

If you square term x to model curvature, clearly there is a correlation between x and x^2.

Data multicolinearity:-

This type of multicolinearity is present in the data itself rather than being an artifact of our model. observational expriments are more likely to exhibit this kind of multicolinearity.

Multicolinearity causes the following two basic types of problems:-

(1) The coefficient estimate can swing based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.

(2) Multicolinearity reduces the precision of the estimate coefficients which weakens the statistical power of your regression model. you might not be able to ~~m~~ trust the p-values to identify independent variables that are statistically significant.

Fix multicolinearity:-

The good news is that it is not always mandatory to fix the multicolinearity. It all depends on the primary goal of the regression model.

The degree of multicolinearity greatly impact the p-values and coefficient but not predictions and goodness-of-fit test. If your goal is to perform the predictions and not neccessary to understand the significance of the independent variable, it is not a mandata to fix the multicolinearity. issue.

How to test multicolinearity?

(1) correlation matrix / correlation plot.

(2) Variation Inflation factor (VIF)

A correlation plat can be used to identify the correlation (or)bivariate relationship between two independent variables VIF is used to identify the correlation of one independent variable with a group of other variables. Hence, it is preferred to use VIF for better understanding.

VIF = 1 → No correlation
VIF = 1 to 5 → Moderate correlation
VIF > 10 → High correlation.

Testing for Multicolinearity with Variance Inflation Factors (VIF)

If you can identify which variables are affected by multicolinearity and the strength of the correlation, youre well on your way to determining whether you need to fix it. Fortunately, there is a very simple test to assess multicolinearity in your regression model.

The variance inflation factor (VIF) identifies correlation independent variables and the strength of that correlation.

statistical software calculates VIF for each independent variable. VIF start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any other. VIFS between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures VIFS greater than 5 represent critical levels of mut multicolinearity where the coefficients are poorly estimated, and the P-values are questionable

Use VIFs to identify correlation between variables and determine the strength of the relationships. most statistical software can display VIFS for you. Assessing VIFS is particularly important for observational studies because these studies are more prone to having multicolinearity

UNIT-V

Models of Time Series

Time Series is statistical data that we arrange and present in a chronological order spreading over a period of time. Time series analysis is a statistical technique dealing with time series data. According to spiegel, "A time series is a set of observations taken at specified times, usually at equal intervals". In statistics, for time series analysis two main categories of models are popular. Let us discuss the models of time series Analysis in details.

In time series quantitative data are arranged in the order of their occurrence and resulting statistical series. The quantitative values are usually recorded over equal time interval such as daily, weekly, monthly, quarterly, half-yearly, yearly or any other measure of time.

Some example are statistics of industrial production in india on a mothly basis, bith-rate figures annually. the yield on ordinary shares, and weekly wholesale Price of rice, etc.

Components of Time Series :-

These is a different kind of forces which influence the time series analysis. Some are continuously effective while others make themselves felt at recurring time and intervals. So our first task is to divide the data and elements into components.

A time series consists of the following four components or basic elements :

1. Basic or secular or long-time trend.
2. Seasonal variations.
3. Business cycles or cyclical movements -
4. Erratic or Irregular fluctuations.

Models of Time Series Analysis :-

The following are the two models which we generally use for the decomposition of time series into its four components. The objectives is to estimate and seperate the four types of variations and to bring out the relative effect of each on the overall behaviour of the time series.

(i) Additive Model (2) Multiplicative model.

(1) Additive Model :-

In the additive Model, we represent a particular observation in a time series as the sum of these four components.

ie., $O = T + S + C + I$

Where O represents the original data. T represents the trend. S represents the Seasonal variations. C represents the cyclical Variations and I represents the irregular Variations. We can writ $y(t) = T(t) + S(t) + C(t) + I(t)$

(2) Multiplicative Model :-

In this model, four components have a multiplicative relationship. So, we represent a Particular observation in a time series as the product of these four components.

ie $O = T \times S \times C \times I$

Where O, T, S, C and I represent the terms as in additive model.

In another way, We can write

$$y(t) = T(t) \times S(t) \times C(t) \times I(t)$$

This model is the most used model in the decomposition of time series. To remove any doubt between the two models, it should be made clear that in multiplicative model

S, C, and I are indices expressed as decimal percentages whereas, in additive model S, C and I are quantitative deviations about a trend that can be expressed as seasonal, cyclical and irregular in nature.

Example:-

If a multiplicative Model.

$T = 500, \; S = 1.4, \; C = 1.20$ and $I = 0.7$

then $O = T \times S \times C \times I$

By substituting the values we get.

$O = 500 \times 1.4 \times 1.20 \times 0.7 = 588$

$O = 588$

If in a additive Model :-

$T = 500, \; S = 100, \; C = 25, \; I = -60$

then $O = 500 + 100 + 25 - 60 = 565$

Growth Curve - Definition:-

Let X be a pxn random matrix Corresponding to the observation. A a $P \times q$ within design matrix with $q \leq P$, B a $q \times k$ parameter matrix, C a $k \times n$ between individual design matrix with rank $(c) + P \leq n$ and Let Σ be a positive definite $P \times p$ matrix, Then

$$X = ABC + \Sigma^{1/2} E$$

definite the growth curve model. Where A and C

are known. B and Σ are unknown and E is a random matrix distributed as $N_{p,n}(0, I_{p,n})$.
This differs from standard MANOVA by the addition of G, a postmatrix $X''^{(2)}$.

Gromportz curve :-

The gompertz curve or gompertz function is a type of Mathematical model for a time series named after Benjamin Gompertz. It is a sigmoid function which describes growth as being slowest at the start and end of a given time period. The right hand or future Value asymptote of the function is approached much more ~~gra~~ generally by the curve than the left-hand (or) lower values asymptote. This is in contrast to the simple logistic function in which both asymptotes are approached by the curve symmentrically It is a special Case of the generalised logistic function. The function Was originally designed to describe human mortality. but since has been modified to be applied in biology. With regard to detailing Populations.

Growth Curve :-

A growth curve is a graphical representation

of how a particular quantity increases over time. Growth Curve are used in statistics to determine the type of growth pattern of the quantity-be it linear, exponential, or cubic. once the type of growth is determind, a business can create a mathematical model to predict ~~future~~ future Sales.

Exponential curve :-

Definition :-

An Exponential function or Curve is a function that grows exponentially, or grows at an increasingly larger rate as you pick larger values of x, and usually takes the form

$Y = a^{x}$, where a is any real number.

$2^{x}, 3^{x}, 10^{x}$ etc are all Examples of Exponential function.

Logistic ~~curve~~ curve :-

A logistic function (or) logistic curve is a common S-shaped curve with equatn.

$$f(z) = \frac{L}{L+e^{-k(x-x_0)}}$$

where x_0 the x Value of the sigmoid's midpoint.

L the curve's maximum value.

k the logistic growth rate or steepness

as the curve

For values of x in the domain of real numbers from $-\infty$ to $+\infty$. the s-curve shown on the right is obtained, with the graph of f approaching L as a x approaches $+\infty$ and approaching zero as x approaches $-\infty$.

The logistic function finds applications in a range of fields, including biology, biomathematics, chemistry, demography, economics

Economics, geoscience, maltematical psychology, Sociology, Political science, linguistics, statistics and artificial natural neuron network.
A generalization of the logistic curve function is the hyperbolastic function of type-I.

AutoRegression :-

A regression model, such as Linear regression, models an output value based on linear combination of input values.

For example:-

$$yhat = b0 + b1 * x1$$

Where yhat is the prediction. b0 and b1 are coefficient

Reg