# UNIT - 4 Assignment

SURESH KUMAR PRAJAPATI

CUSB2302222008

LINEAR MODELS AND REGRESSION ANALYSIS

# Residual & Diagnostics

**Residuals :-** The residuals are equal to the difference between the observations and the corresponding fitted values.

For the regression model,

$$\boxed{Y = X\beta + \varepsilon}$$

If the fitted value of the model are $\hat{y}$, then the residuals is defined as

$$\boxed{e = y - \hat{y}} \quad \text{or} \quad e_i = y_i - \hat{y}_i \; ; \; i = 1, 2, \cdots n$$

For the residuals be, we have.

$E(e) = 0$, $\text{Var}(e) = \sigma^2 (I - H)$, where $H = X(X'X)^{-1}X$ is called Hat matrix. because it transforms $y$ to $\hat{y}$ as $\hat{y} = Xb = X(X'X)^{-1}X'y$ and $I$ is an identity matrix.

If the intercept is include in the mean function, then $\sum_{i=1}^{n} \hat{e}_i = 0$. In the scalar form, the variance of the $i^{th}$ residual is

$\text{Var}(\hat{e}_i) = \hat{\sigma}^2 (1 - h_{ii})$ where $h_{ii}$ is the $i^{th}$ diagonal element of H.

## Types of residuals :-

There are various types of residuals that will be useful in our discussion of regression diagnostics.

(i) Raw or Ordinary residuals :-

$$e_i = Y_i - \hat{Y}_i$$

(ii) Standardized residuals :- (Approx) (variance)

The ~~standardized~~ residuals are standardized based on the concept of residual minus its mean and devided by its standard deviation.

So, the $i^{th}$ standardized residual is given by $z_i = \frac{e_i}{\hat{\sigma}}$ where $\hat{\sigma} = \sqrt{MSE}$ is the estimated error standard deviation.

Also, $E(z_i) = 0$, $Var(z_i) = 1$. So, a large value of $z_i$ (> 3 say) potential indicates an outlier.

(iii) Studentized residuals :- (variance exact)

The $i^{th}$ studentized residuals is defined as

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{z_i}{\sqrt{1-h_{ii}}}$$ where $h_{ii}$ is the $i^{th}$ diagonal element of H.

Remark :- studentized residuals have a mean near 0 and a variance $\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2$ i.e. slightly larger than 1.

In large data sets, the standardized and studentized residuals should not differ dramatically.

(iv) :- Jackknife Residuals or R-student Residuals :-

The Jackknife residuals is defined as

$$r_{(-i)} = r_i \sqrt{\frac{MSE}{MSE_{(-i)}}} = \sqrt{\frac{e_i}{MSE_{(i)}(1-h_{ii})}} = r_i \sqrt{\frac{(n-p-1)-1}{(n-p-1)-r_i^2}}$$

where $MSE_{(-i)}$ is the residual variance computed with the $i^{th}$ observation deleted.

Jackknife residuals have a mean near 0 and a variance $\frac{1}{(n-p-1)-1}\sum_{i=1}^{n} r_{(-i)}^2$ that is slightly greater than 1.

Jackknife residuals are usually the preferred residual for regression diagnostics.

The major assumptions of the model are

1. The relationship b/w the study variable $y$ and explanation variables $X$ is linear.
2. The errors are normally distributed.
3. The mean of error is 0
4. The variance of errors is constant and equals $\sigma^2$
5. The errors are uncorrelated
6. The model contains all predictors related to $E(Y)$
7. The model fits for all data observations.

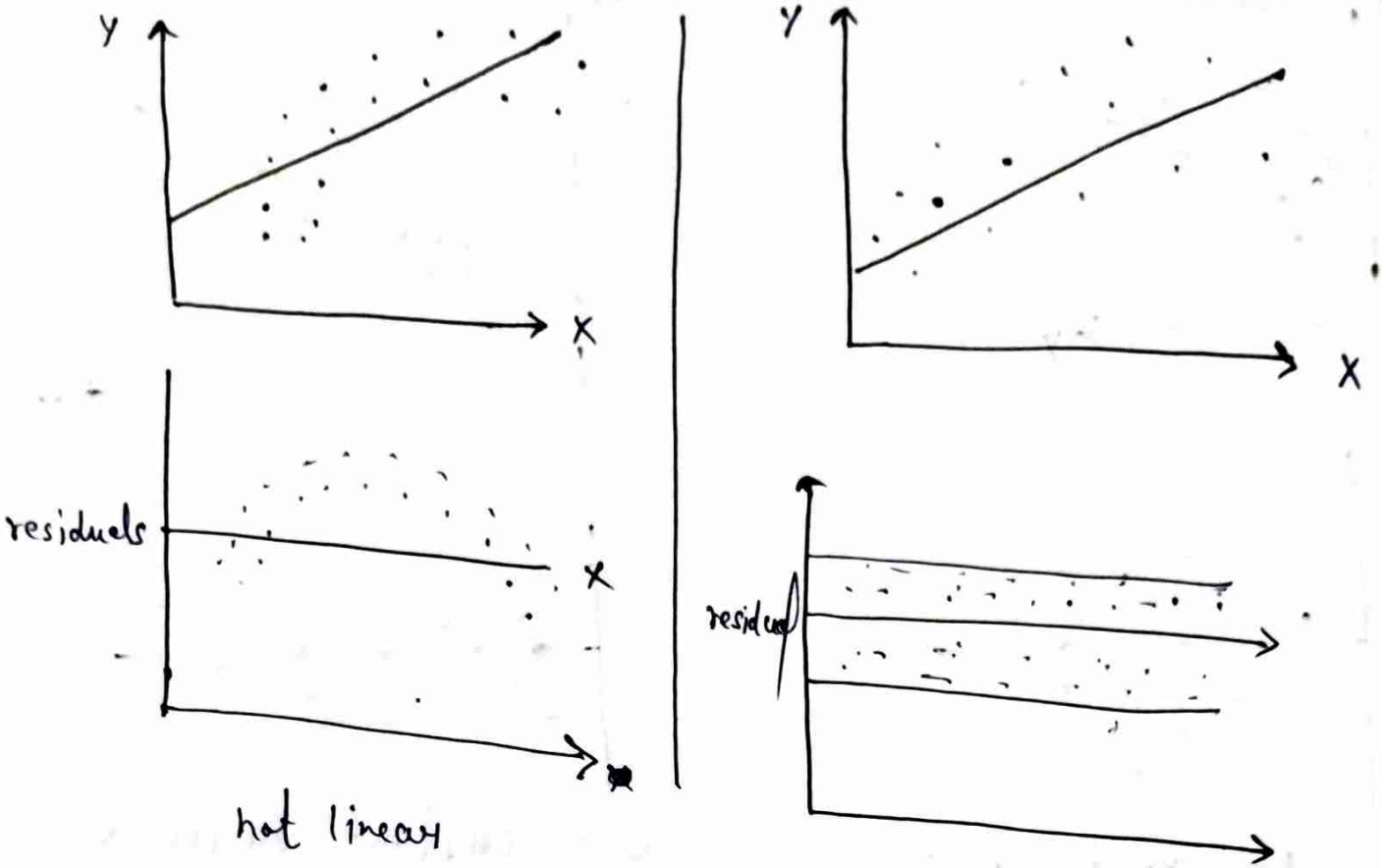
Remark:- The graphical analysis of residuals is very effective way to investigate the adequacy of the fit of a regression model.

Residual plots are best single check for violation of assumptions, such as

1. Residuals vs Fitted values :-

(a) Cheacking Linearity :-

Residuals Vs fitted values plot is used to cheack the linear relationship assumptions. A horizontal line without distinct patterns is an idication for a linear relationship.

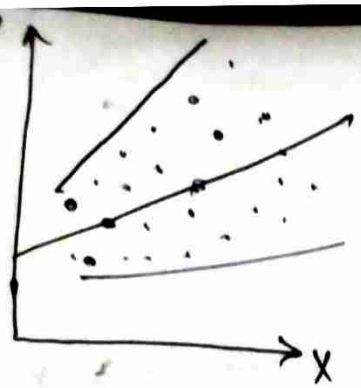


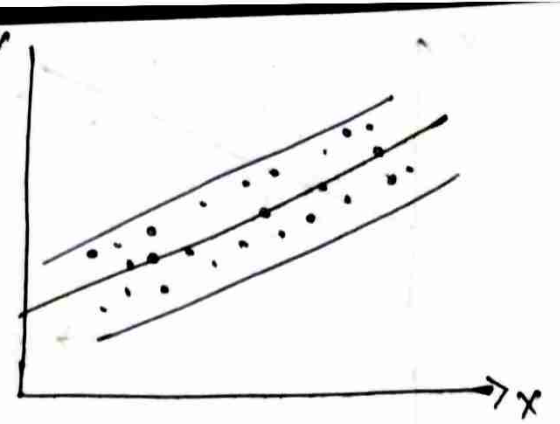

(b) Constant Variance of Errors is constant or Homoscedasticity

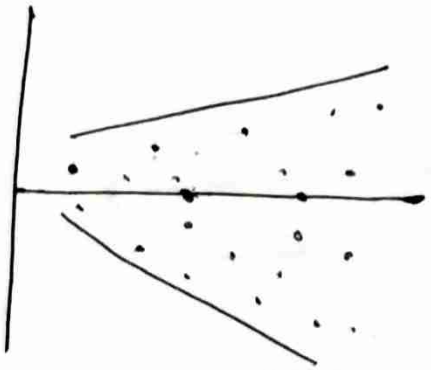
⇒ Residual with fitted values plot is also check the "variance of errors is constant" assumptions.

⇒ If the plot is such that the residuals can be obtained in an "outward opening funnel" then such pattern indicates that the variance of errors is not constant but it is in an increasing function of Y.

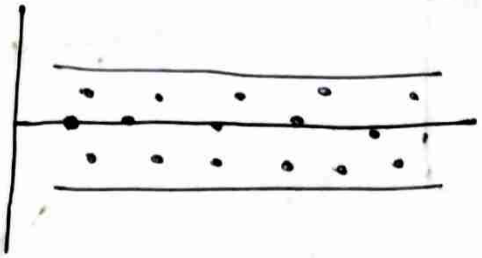
⇒ If plots are such that the residuals can be accommodated in an "inward openning funnel," then such a pattern indicates that the variance of errors is not constant, but it is decreasing function of Y.



Non-constant variance



✓ constant variance

## 2:- Normal Probability Plot of Residuals :-

The assumption of normality of errors is very much needed for the validity of the results for "testing of hypothesis", confident intervals and prediction intervals. There are various plots the cheack the assumptions of normality of errors.

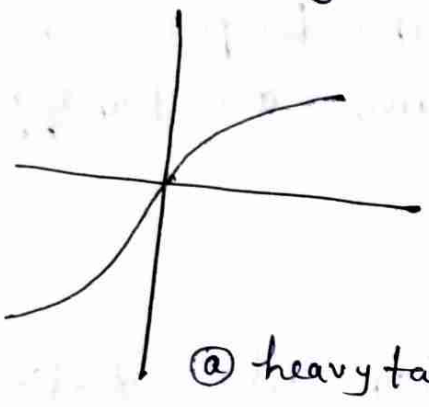
### (i) Normal probability plot :-

The assumption of normality of errors can be checked by examing a normality plot. The normal probability plot is a plot of the ordered standardized residuals versus the so called normal scores.

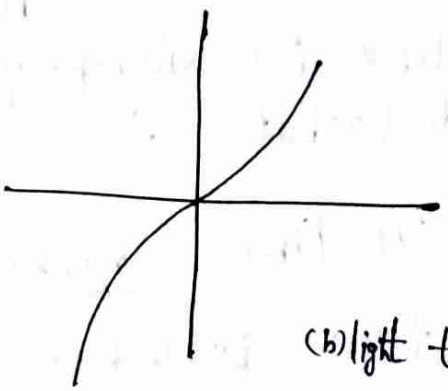
The normal scores are the cumulative probability can be obtained as :

$$P_i = \frac{\left(i - \frac{1}{2}\right)}{n} \quad ; \quad i = 1, 2, \ldots n.$$

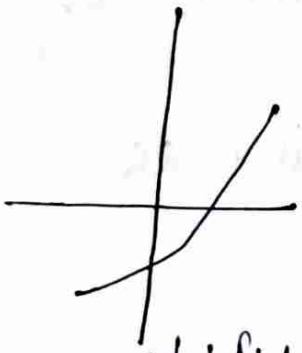
If the residuals $e_1, e_2, \ldots e_n$ are ordered and ranked in increasing order $e_{[1]} < e_{[2]} < \ldots < e_{[n]}$ then plot of $e_{[i]}$ against $P_i$ is called probability plot. If the residuals are normally distributed, then the ordered residuals should be approximately on a diagonal straight line in the plot.



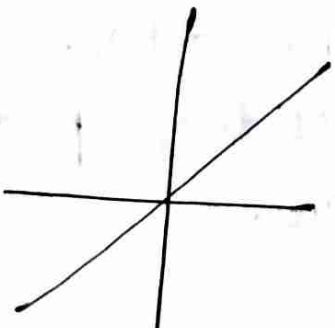
(a) heavy tail



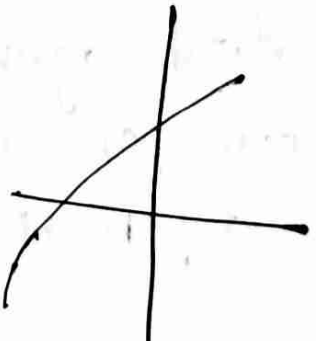
(b) light tail



(d) left skew



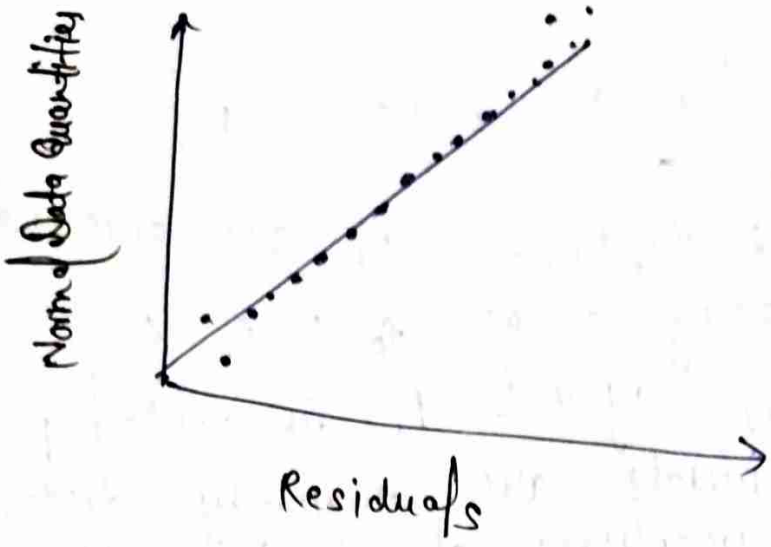
(c) normal



(d) right skew

(ii) QQ plot:-

Quantile-Quantile plot is also used to check the whether the errors are normally distributed. In this plot, the residuals are plotted against their percentage (empirical comulative distribution) point 0 to 1. If the points lie approximately on the straight line and indicate that the underlying distribution is normal.



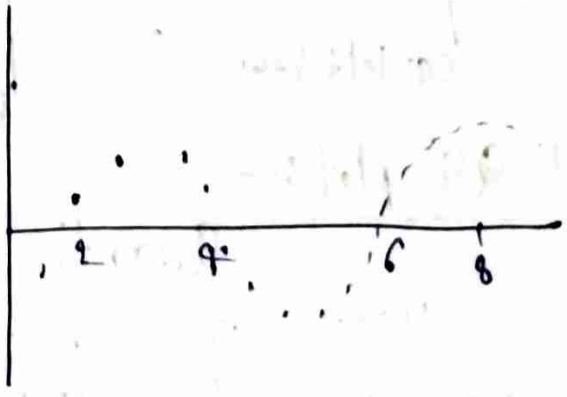

(iii) Histogram and Boxplot :-

Histogram and Boxplot can also be used for diagnosis the errors are normally distributed.

(3) Residuals Vs Time Sequences :-

If the time sequence in which the data were collected is known then residuals can be plotted against time order.

A pattern on a residual plot indicates
- dependent errors.



# Test Involving Residuals :-

1. Test for Normality
   (a) Goodness of fit test such as
      (i) Chi-square test
      (ii) Kolmogorov-Smirnov test.
   (b) Correlation test for Normality :-
   A simple test based on normal probability plot of residuals is discussed which involves testing whether or not the correlation coefficient is indicating linearity and hence normality of error) or not.
   (c) Lilliefors test
   (d) Simple test based on normal probability plot of residuals.

2. Test for Constancy of Error Variance (Homoscedasticity or Heteroscedasticity) :-
   (a) Goldfeld-Quandt (G-Q) test
   (Applicable if heteroscedasticity is related to only one of the explanatory variables)
   (b) White's General Test
   (c) Park Test
   (d) Glejser Test
   (e) Levene Test
   (f) Breusch-Pagan test

(3):- Test for Randomness (Autocorrelation):-

(a) Run test on residual versus time plot data
(b) Durbin-Watson test

(4):- Test for Outliers:-

If chance of getting outlier, assuming other data are given, is small, reject outlier.

Consequences of Heteroscedasticity (Not constant variance):-

(a) Ordinary least squares estimators still linear and unbiased but ordinary least squares estimators are not efficient (No Longer BLUE)
(b) Usual formula give incorrect standard errors least squares.

Remedial measure for Autocorrelation:-

Cochrane-Orcutt Transformation:-

errors term free from autocorrelation, we apply the OLS method to transformed model. Now, the transformed variables will have the desirable BLUE property.

## Multicollinearity :—

The regressors must be independent of each other. But if they are not independent of each other then multicollinearity is said to be present.

The regression parameters are highly are highly affected by the presence of multicollinearity.

## Detection of Multicollinearity :—

There are several methods used for the detection of multicollinearity among which main methods are.

* Use of $R^2$ and t - statistic
* Use of pairwise correlations.
* Use of Auxiliary Regression
* Use of Eigen-Value and condition index.
* Tolerance and Variance inflation factor.

## Use of $R^2$ and T- statistic :—

Generally it is said that a regression model is good if the value of coefficient of determination $R^2$ is high.

It is also a useful tool for detection of multicollinearity. If the value of $R^2$ is high (more than 0.8) and a very few regression parameters comes out to be significant

( Using t-test for individual regression parameters) then multicollinearity is said to be present.

## Use of Pair-Wise Correlation :—

* Used for the detection of multicollinearity.
* Karl Pearson's Product moment correlation coefficient is computed among each pair of regressors.
* Rule of thumb if the pair wise correlation exceeds the value of 0.8 then multicollinearity is said to be present.
* Some researchers prefer the use of partial correlation instead of Pearson's correlation as it removes the effect of other regressors.

## Use of Auxiliary Regression

* Let's consider the regression model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \epsilon_i$$

then the regression equations.

$$X_{1i} = \beta_{10} + \beta_{12} X_{2i} + \beta_{12} X_{3i} + \dots + \beta_{1p} X_{pi} + \epsilon_{1i}$$

$$X_{2i} = \beta_{20} + \beta_{21} X_{1i} + \beta_{23} X_{3i} + \dots + \beta_{2} X_{pi} + \epsilon_{2i}$$

$$\vdots$$

$$X_{pi} = \beta_{10} + \beta_{p1} X_{1i} + \beta_{p2} X_{2i} + \dots + \beta_{p(p-1)} X_{(p-1)i} + \epsilon_{1i}$$

are called as Auxilliary regressions.

## Use of Auxiliary Regression :-

* In this method each of regressor is regressed over remaining regressors.
* For each Auxiliary regression the value of coefficient of determination ($R_j^2$) is computed.
* The value of $R^2$ is computed for the regression model of model of Y on all the regressions.
* Klein's rule of thumb if the $R_j^2$ exceeds the value of $R^2$ then multicollinearity said to be present.

## Eigen - Value and condition index :-

* Another method for detection of multicollinearity.
* In the first the eigen-value for the matrix $X'X$ is computed.
* Then the condition number(K) and condition index (CI) are obtained by.

$$K = \frac{\text{Maximum Eigen Value}}{\text{Minimum eigen value}} ; \quad C.I. = \sqrt{\frac{\text{Max. Eigen Value}}{\text{Min. Eigen Value}}}$$

* As a rule of thumb.
  - If $K < 100$ or $C.I. < 10$ then multicollinearity is absent.

2. If $100 < K < 1000$ or $10 < C.I. < 30$ then moderate to strong multicollinearity is present.

3. If $K > 1000$ or $C.I. > 30$ the severe multicollinearity is present.

## Variance inflation factor and Tolerance :-

* It is another method of detection of multicollinearity.
* The variance inflation factor (VIF) is given by. $VIF = 1/(1 - R_j^2)$

where $R_j^2$ is the coefficient of determination for the auxiliary regression of $x_j$ on remaining regressors.

* As a rule of thumb if VIF exceeds 10 then multicollinearity is said to be present.
* VIF has a drawback that it is unbounded therefore use of tolerance (TOL) is preffered which is defined by.

$$TOL = 1/VIF$$

* The value of TOL lies between 0 and 1.
* As a rule of thumb if the value of TOL is closer to zero then multicollinearity is said to be present.

* Non-linear Regression :-

Models in which the derivatives of the mean function w.r.t. to the parameters depends on one or more of the parameters.

Intrinsically nonlinear models :-

$$Y_i = f(x_i; \theta) + \epsilon_i$$

where $f(x_i; \theta)$ is a nonlinear function relating $E[Y_i]$ to the independent variables $x_i$

* $X_i$ is a $k \times 1$ vector of independent variables.
* $\theta$ is a $p \times 1$ vector of parameters.
* $\epsilon_i$s are iid variable mean zero and variance $\sigma^2$. (sometimes it's normal)

* Non-Linear Least Square :-

The List square of $\theta$, $\hat{\theta}$ is the set of parameters that minimizes the residual sum of square:

$$S(\hat{\theta}) = SSE(\hat{\theta}) = \sum_{i=1}^{n} \{ y_i - f(x_i; \hat{\theta}) \}^2$$

partial derivatives of $S(\theta)$ with respect to each $\theta_j$ and set them 0.

$$\frac{\partial S(\theta)}{\partial \theta_j} = -2 \sum_{i=1}^{n} \{ y_i - f(x_i; \theta) \} \left| \frac{\partial (x_i; \theta)}{\partial \theta_j} \right|_{=0}$$

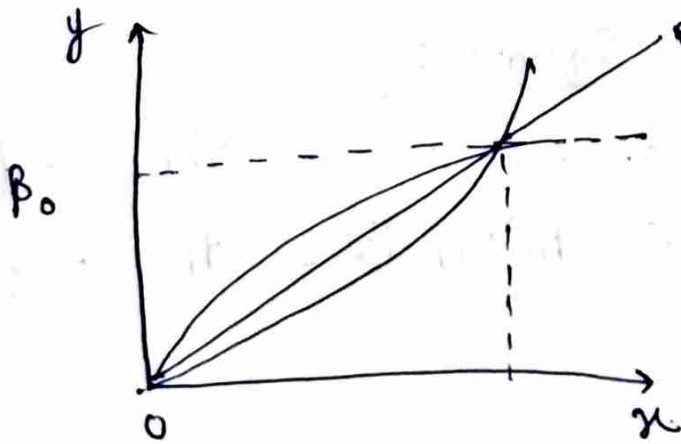
**Transformation to Linearize the model:-**

The basic assumption in linear regression analysis is that relationship between the study variable and explanatory variable is linear. Suppose this assumption is violated can be cheack by scatter plot, matrix scatter diafram, partial regression plot, lack of fit test etc.

The main objective to find the test of hypothesis confidence interval etc.

Some Linearable function are as follows.

① If the curve between Y and X is like as follow:




$y = \beta_0 x^{\beta_1}$ $(\beta_0 > 0, \beta_1 > 0$
$x > 0$

$$\log y = \log \beta_0 + \beta_1 \log x_1$$

Here $\log y = y^*$ $\log \beta_0 = \beta_0^*$ $\log x_1 = x^*$.

Therefore.

$$y^* = \beta_0^* + \beta_1 x^*$$

it is linear model by applying transformation.