GOVT. BILASA GIRLS P.G. (AUTO.) COLLEGE, BILASPUR, C.G.

PAPER-III
'ADVANCE RESEARCH METHODOLOGY'
UNIT-IV
TOPIC
REGRESSION: LINEAR AND MULTIPLE

MANJARY SHARMA ,
ASSISTANT PROFESSOR, PSYCHOLOGY,
GOVT. BILASA GIRLS P.G. (AUTO.). COLLEGE,
BILASPUR,C.G.

2020-2021

REGRESSION: ASSUMPTIONS & TYPES

प्रतिगमन का अर्थ- (प्रतिगमन का सामान्य अर्थ है पीछे जाना)

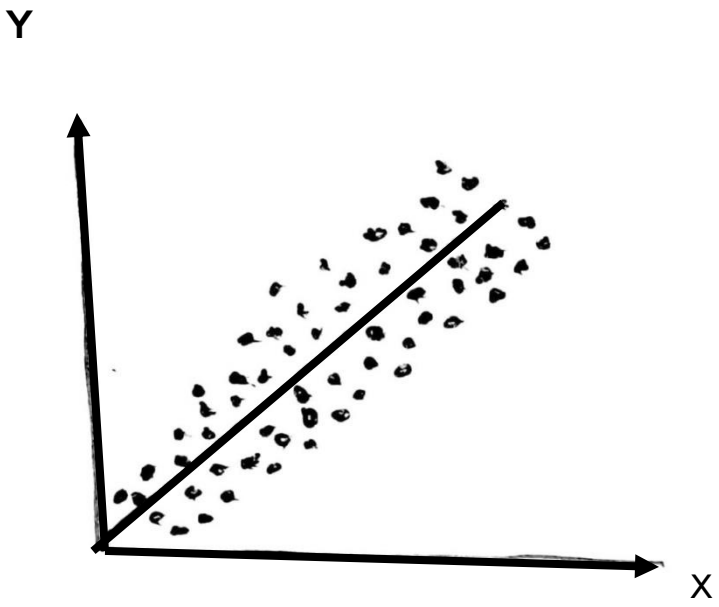
प्रतिगमन विभिन्न संखिकीय प्रक्रियाओं का एक सेट होता है जिसके द्वारा दो या दो से अधिक स्वतंत्र चरों के विभिन्न मानों के आधार पर आश्रित चर के मान के बारे में अनुमान लगाया जाता है|

प्रतिगमन द्वारा स्वतंत्र चर एवं आश्रित चर आपस में आन्किक रूप से कैसे और कितना सम्बंधित है, इसका ज्ञान होता है| इसके द्वारा एक ज्ञात चर (known variable) अर्थात स्वतन्त्र चर (independent variable) में होने वाले इकाई परिवर्तन (unit change) से आश्रित चर (dependent variable) या अनुमानित चर (estimated variable) पर पड़ने वाले प्रभाव का अध्ययन किया जाता है| प्रतिगमन जैसी सांखिकीय विधि का उपयोग फाइनेंस एवं इन्वेस्टिंग, व्यहारात्मक शिक्षा एवं मनोविज्ञान में मुख्य रूप से किया जाता है|

प्रतिगमन से भिन्न, सहसंबंध विधि दो चरों के बीच सम्बन्ध की मात्रा या (extent) का ज्ञान कराता है, इसके द्वारा भाविषय कथन नहीं किया जा सकता| जबकि प्रतिगमन द्वारा स्वतंत्र चर (independent variable) तथा आश्रित चर (dependent variable) के बीच संबंध की शक्ति व उनकी प्रकृति का निर्धारण किया जाता है| स्वतंत्र चरों के विभिन्न मानों के आधार पर आश्रित चर के मान के बारे में अनुमान या भविषय कथन भी किया जाता है| प्रतिगमन में आश्रित चर को परिणाम (outcome/criterion variable) एवं स्वतंत्र चर को कारण (predictor/cause variable) कहा जाता है|

Objective of regression / प्रतिगमन का उद्देश्य-

- प्रतिगमन का सबसे महत्वपूर्ण उद्देश्य भविष्यकथन करना है|
- इसका दूसरा मुख्य उद्देश्य, किसी निश्चित या नियमित (स्वतन्त्र चर) चर के मानों के आधार पर, दुसरे यद्रिश्चिक या अनियमित चर (आश्रित चर) के मानों का अनुमान लगाना होता है|



Regression equation

सामान्य रेखीय प्रतिगमन / Simple linear regression

सामान्य रेखीय प्रतिगमन एक ऐसा मॉडल है जिसके द्वारा एक आश्रित चर तथा एक स्वतंत्र चर के बीच सम्बन्ध का मापन किया जाता है, अर्थात एक स्वतंत्र चर के मानों में परिवर्तन से आश्रित चर में होने वाले परिवर्तन का अनुमान लगाया जाता है| इसे निम्न समीकरण द्वारा प्रदर्शित किया जाता है -

$$Y = bX + a$$

जहाँ, Y= Dependent variable

a= Intercept

b= Slope

X= Inependent variable

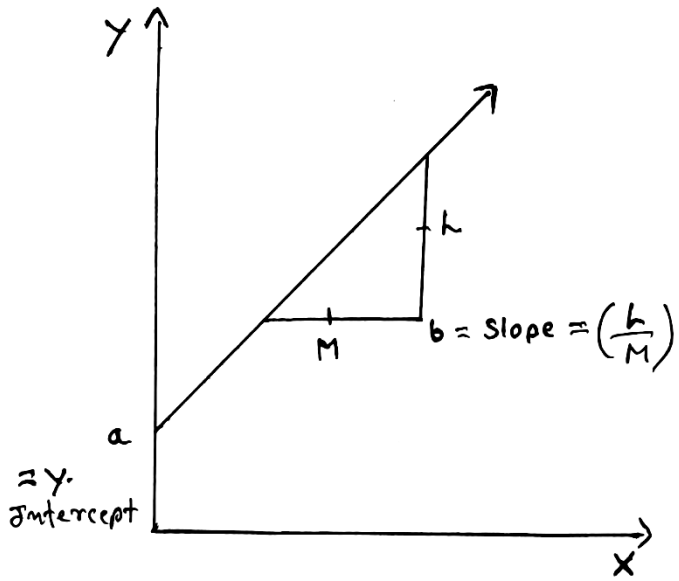
e= Error or residual (अवशिष्ट)

Some common concepts about regression

INTERCEPT 'a'- INTERCEPT 'a', आश्रित चर 'Y' का वह अनुमानित मध्यमान है जो सभी स्वतंत्र चरों के मान को शून्य रखने पर प्राप्त होता है| (Intercept a, is the mean of dependent variable when you set all the independent variables in the model to ZERO). Simple linear regression में एक स्वतंत्र चर, X=0 होने पर, Intercept a, 'X' की उस value (मान) पर 'Y' का अनुमानित मध्यमान होगा| 'X' का मान शून्य नहीं होने पर, Intercept का अपना कोई अर्थ नहीं होता है|

Regression line में intercept वह मूल्य होता है जिस पर regression line Y-axis को Cross (क्रॉस) करती है|

नीचे वर्णित चित्र में 'Y' axis में 'a' intercept है तथा Regression line पर 'b' slope को प्रदर्शित करता है जिसे चित्र में दिखाए अनुसार L/M = 'b' द्वारा प्राप्त किया जा सकता है|



$$Y = a + bX$$

or

$$Y = bX + a$$

SLOPE- किसी प्रतिगमन रेखा का slope या ढाल (Independent variable) 'X' में परिवर्तन के साथ (dependent variable) 'Y' में होने वाले परिवर्तन की दर को प्रदर्शित करता है| (The slope of a regression line 'b' represents the rate of change in 'Y' as 'X' changes. Here, 'Y' is dependent on 'X'. The slope describes the predicted values of 'Y' on given 'X'. It is the change in 'Y' for a unit change in 'X' along the line)

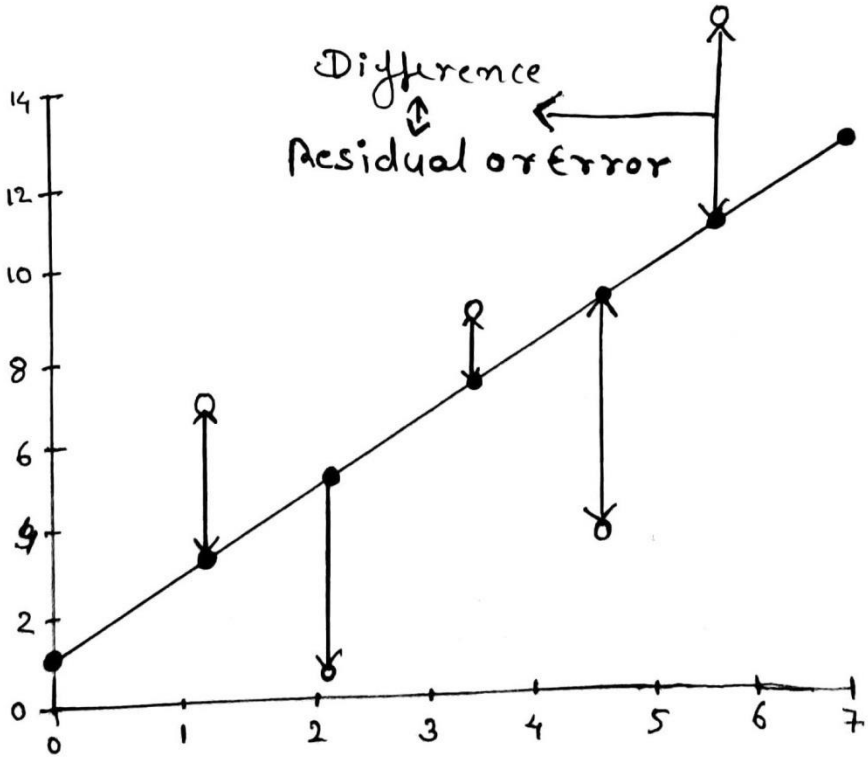
किसी प्रतिगमन रेखा में dependent व independent variable के बीच के सम्बन्ध की सार्थकता की जांच Slope के आधार पर की जाती है| Slope को independent variable 'X' में इकाई परिवर्तन से dependent variable 'Y' में होने वाले परिवर्तन के रूप में प्रदर्शित किया जा सकता है| यदि, slope = 2 है तो हम इसे 2/1 के रूप में मान सकते है, जिससे ये पता चलता है की प्रतिगमन रेखा में आगे बढ़ने पर 'X' में अर्थात स्वतंत्र चर में इकाई परिवर्तन (1 की वृद्धि) होने से चर 'Y' अर्थात आश्रित चर में 2 की वृध्धि होती है|

RESIDUAL या अवशिष्ट= (This is called as errors in prediction) स्वतंत्र चर के मान के आधार पर अनुमानित आश्रित चर का मान (predicted value) एवं आश्रित चर के वास्तविक मान (real value) के अंतर को residual या अवशिष्ट कहते है| उदहारण के रूप में - यदि हम प्रथम मूल्याङ्कन परीक्षा के अंकों के आधार पर द्वितीय मूल्यांकन परीक्षा के अंकों के बारे में अनुमान लगाते है तो यहाँ प्रथम मूल्याङ्कन परीक्षा के अंक, स्वतंत्र चर है तथा द्वितीय मूल्यांकन परीक्षा के अंक, आश्रित चर है| स्वतंत्र चर के मान के आधार पर अनुमानित या रेखीय समीकरण द्वारा अनुमानित, आश्रित चर का मान (predicted value) एवं आश्रित चर के वास्तविक मान (real value) का अंतर को Residual या अवशिष्ट कहते है|

Residual (अवशिष्ट) = Predicted value- Real value

(dependent variable)

नीचे प्रदर्शित चित्र में '●' को आश्रित चर का अनुमानित मान (Predicted value) तथ '○' को वास्तविक मान के रूप में दर्शाया गया है| इन दोनों के अंतर को ही Residual (अवशिष्ट) कहते है|



○ - **Real value of dependent variable (वास्तविक मान)**

● - **Assumend value of dependent variable (अनुमानित मान)**

सामान्य रेखीय प्रतिगमन की शर्तें-

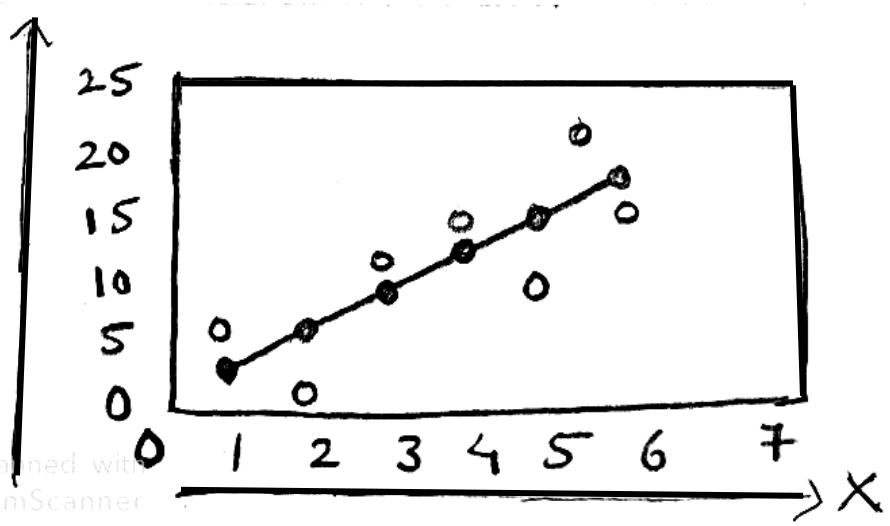
- **Continuous dependent variable-** आश्रित चर को Continuous या निरंतर चर के रूप में होना चाहिए|
- स्वतंत्र चर त्रुटिरहित होना चाहिए|
- **Linearity -** स्वतंत्र चर (predictor) तथा आश्रित चर (outcome variable) के बीच रेखीय संबंध होना चाहिए| रेखियता को scatterplot द्वारा जाँचा (चेक) किया जा सकता है|
- **Residual should be zero value-** Residual या अवशिष्ट को शून्य मान का होना चाहिए| अर्थात आश्रित चर की predicted value, real value के बराबर होनी

चाहिए| स्वतंत्र चर के मान के आधार पर अनुमानित आश्रित चर का मान (predicted value) एवं आश्रित चर के वास्तविक मान (real value) का अंतर शून्य होना चाहिए|

Residual (अवशिष्ट) = Predicted value- Real value= 0

(dependent variable)

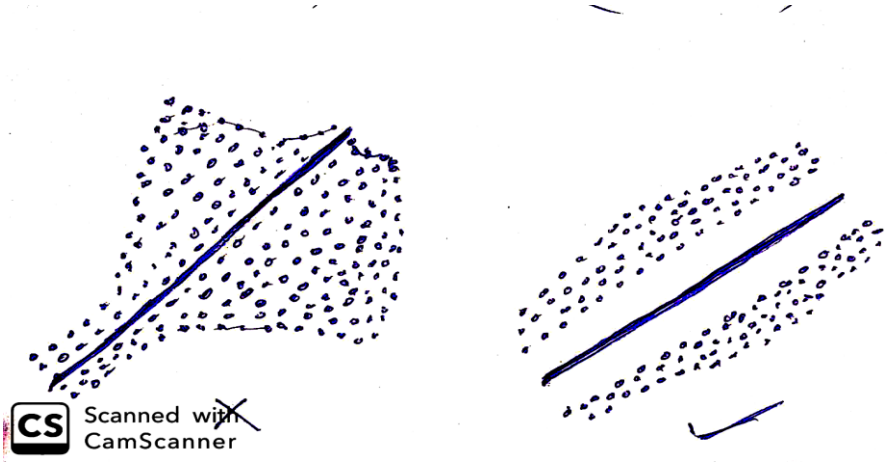
- **Homoscadisticity-** स्वतंत्र चर के प्रत्येक मानों के लिए अवशिष्ट (residual) के प्रसरण का मान सामान होना चाहिए|



Regression line

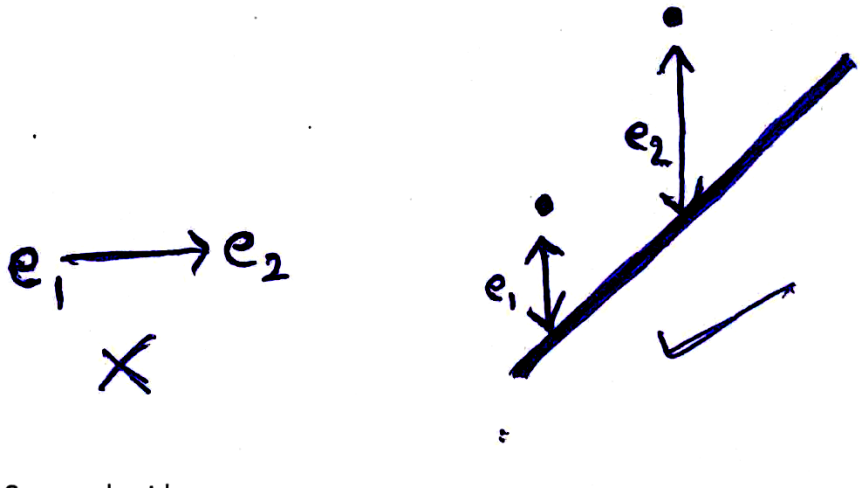
○ Real value

● Predicted value



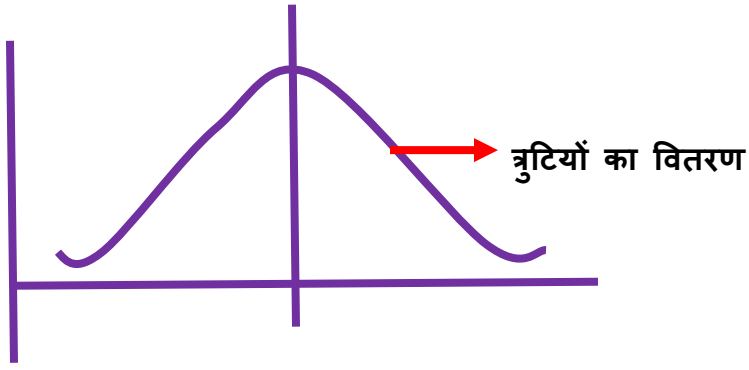
Different variance of residual **Same variance of residual**

- **Independency of the error or residual –** त्रुटियाँ या त्रुटि चर में प्रसरण समान या नियत (constant) होना चाहिए तथा त्रुटियाँ आपस में सहसम्बन्धित नहीं होनी चाहिए|



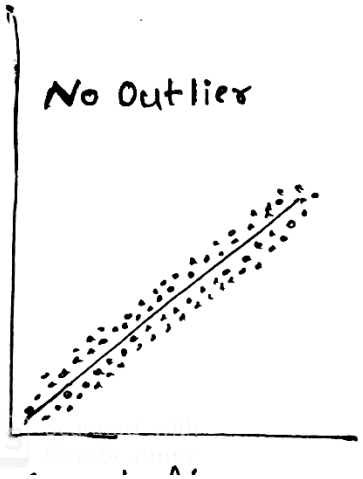
सहसंबंधित त्रुटि **असहसंबंधित त्रुटि**

- **Normality of error term or residual-** त्रुटियाँ या त्रुटि चर सामान्य रूप से वितरित होनी चाहिए| Q-Q Plot या histogram द्वारा प्रसमान्यता (normality) की जाँच की जा सकती है|

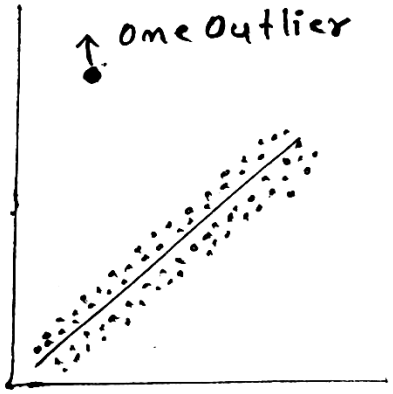


- **Observations are truly independent from each other-** समस्त observations को स्वतंत्र होना चाहिए अर्थात एक दुसरे पर depend नहीं होना चाहिए|
- **Parametric-** समस्त चरों को parametric statistics में होना चाहिए| इसका तात्पर्य है कि सभी चरों को सामान्य वितरित जनसँख्या से होना चाहिए|
- **No-Outliers/ influential cases are present-** An outlier is an observation that is unusually small or unusually large.

  यदि standard residual की absolute value का मान 2 से ज्यादा आता है तो इसका मतलब है की data में outlier मौजूद है| प्रतिगमन के समीकरण को outlier अनुचित तरीके से प्रभावित कर सकता है| सिर्फ एक outlier दो चरों के बीच के सहसंबंध को इतना कम कर देता है, कि शोध का परिणाम अनुचित तरीके से प्रभावित हो जाता है| अतः यह अत्यंत आवश्यक होता है की इस तरह के Outliers, जो अन्य सामान्य data से अत्यधिक भिन्न (different) होता है, को analysis या विश्लेषण से हटाया जाए| इसी प्रकार influential cases शोध के परिणाम को Outliers से भी ज्यादा प्रभावित करते है| अतः scatterplot के द्वारा चेक करके इस तरह के cases को analysis से हटा देना चाहिए|



correlation r = 1



correlation r = .889

बहुरेखीय प्रतिगमन / Multiple linear regression

बहुरेखीय प्रतिगमन, अनिवार्य रूप से इस अपवाद के साथ सामान्य रेखीय प्रतिगमन के समान है कि इसमें दो या दो से अधिक स्वतंत्र चरों का उपयोग किया जाता है| Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model.

इसका गणितीय सूत्र निम्न प्रकार से है:

$$Y = a + bX_1 + cX_2 + dX_3 + Error$$

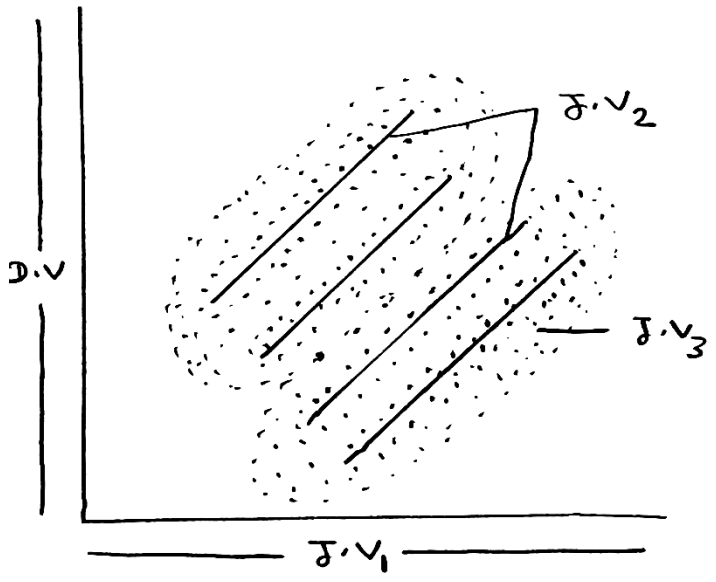
जहाँ, Y= Dependent variable

a= Intercept

b, c, d= Slope

X_1, X_2, X_3 = Inependent variables

e= Error or residual (अवशिष्ट)



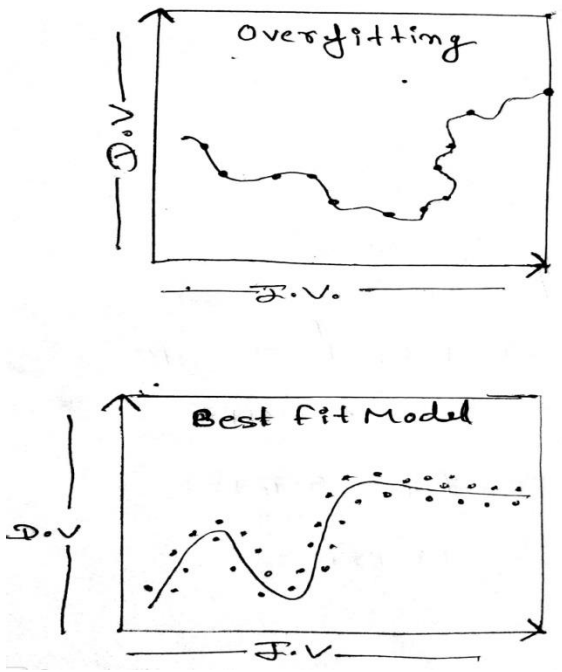
ज.व.1 , ज.व.2 , ज.व.3 = 3 स्वतंत्र चर

बहुरेखीय प्रतिगमन की शर्तें:

बहुरेखीय प्रतिगमन, सामान्य रेखीय प्रतिगमन की तरह ही तमाम शर्तों का अनुसरण तो करता ही है, परन्तु इसके लिए कुछ अन्य अनिवार्य शर्तें भी आवश्यक होती है-

- **Overfitting -** Overfitting मॉडल से तात्पर्य एक ऐसे मॉडल से होता है जिसमे analysis का परिणाम या production, data के एक विशेस सेट के साथ निकटता से मेल खता है या प्रतिगमन रेखा के उप्पर या बिलकुल निकट होता है एवं किसी और data के साथ फिट (fit) नहीं हो पाता और इसके द्वारा आगे के observation के बारे में prediction नहीं किया जा सकता| इससे परिणाम के सामान्यीकरण में कठिनाई आती है|

  overfitting occurs when a statistical model adequately or exactly fit with the underlying structure of the data and may therefore fail to fit additional data or predict future observations reliably.



(It can be used to quantify the relative impacts of two or more independent (the predictor variables) on dependent variable (the outcome variable)).

- **No- Multicollinearity -** स्वतंत्र चरों (predictor variables) में सहसंबंध (correlation) नहीं होना चाहिए| इसे सहसंबंध मैट्रिक्स द्वारा चेक किया जा सकता है|
- **Normally distributed - (Multivariate normality)-** प्रत्येक चरों एवं Observed एवं predicted (कल्पित) मान प्रसामान्य रूप से वितरित होना चाहिए|
- कम से कम दो स्वतंत्र चर होने चाहिए| स्वतंत्र चर Nominal,Ratio,Ordinal एवं Interval स्केल में होना चाहिए|
- **Sample Size -** प्रत्येक स्वतंत्र चर के लिए कम से कम २० केस (case) होने चाहिए|
- अवशिषटो को भी प्रसामान्य रूप से वितरित होना चाहिए|