UNIT-4 6-3-17

Describes the relation between variables

Regression models:- Relationship between one dependent variable & explanatory variables.

1. Use equation to set up relationship
   Numerical dependent (Response) variable.
2. 1 or more Numerical or categorical Independent (Explanatory variables)
3. used mainly for prediction & Estimation

Regression modeling steps:

1. Hypothesize Deterministic component
   Estimate unknown parameters
2. Specify probability distribution of Random Error term.
   Estimate sd of error
3. Evaluate the fitted model.
4. Use model for prediction & Estimation.

Specifying the deterministic component

1. Define the dependent variable & Independent variable

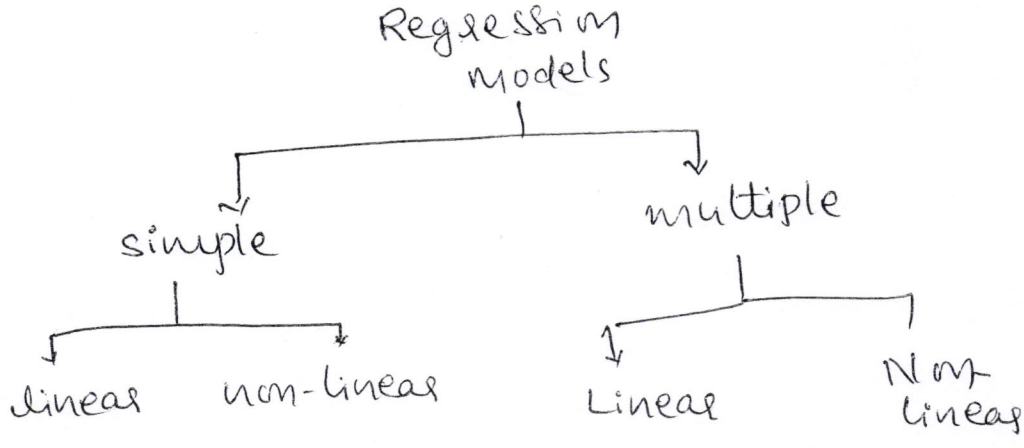
2. Hypothesize Nature of relationship

Expected effects (i,e coefficients's signs)

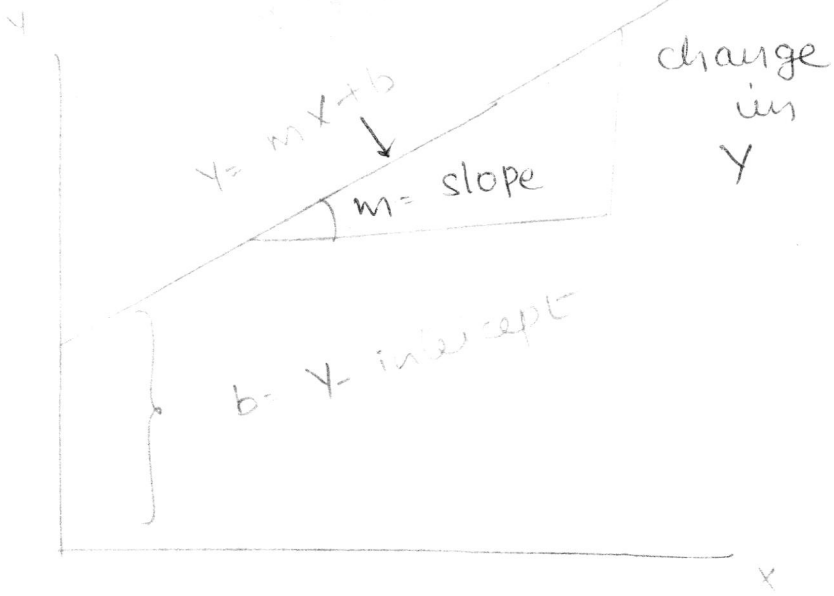
. Functional form (linear or non-linear)

Interactions

Types of Regression models:



Linear equations

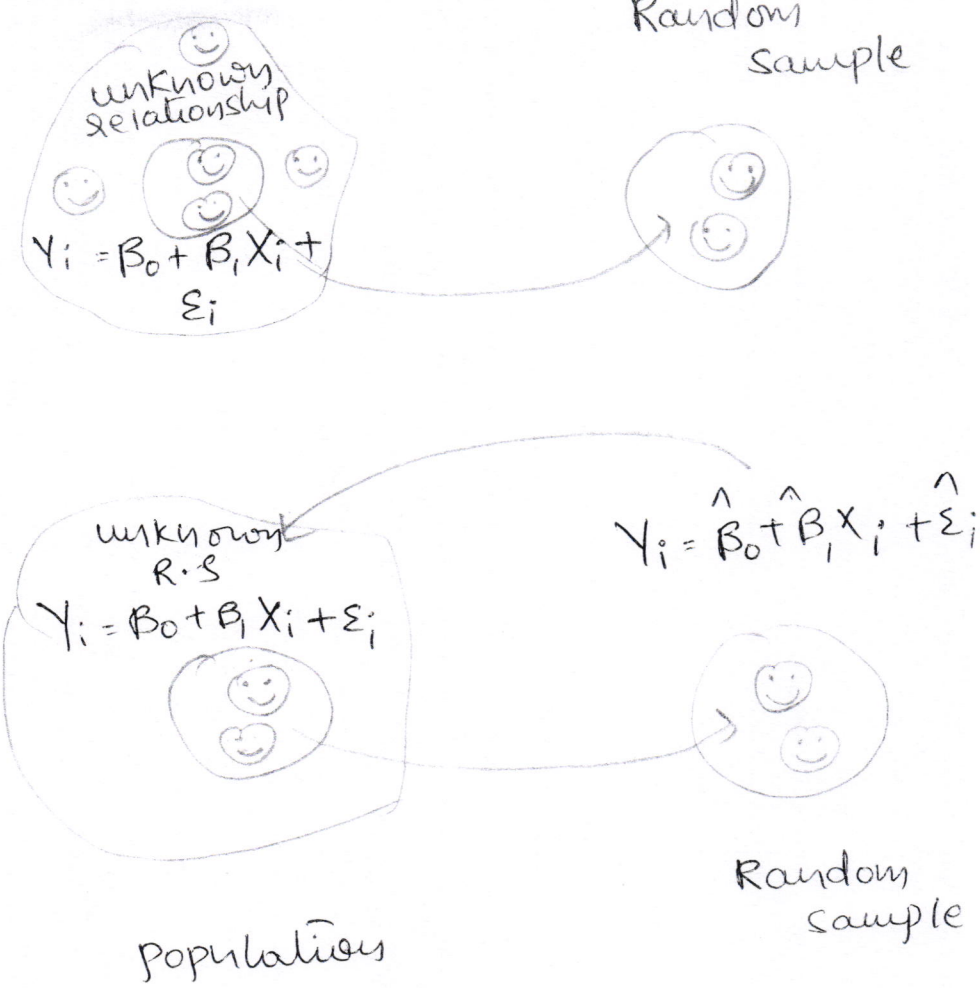


Relationship between variables is a linear function

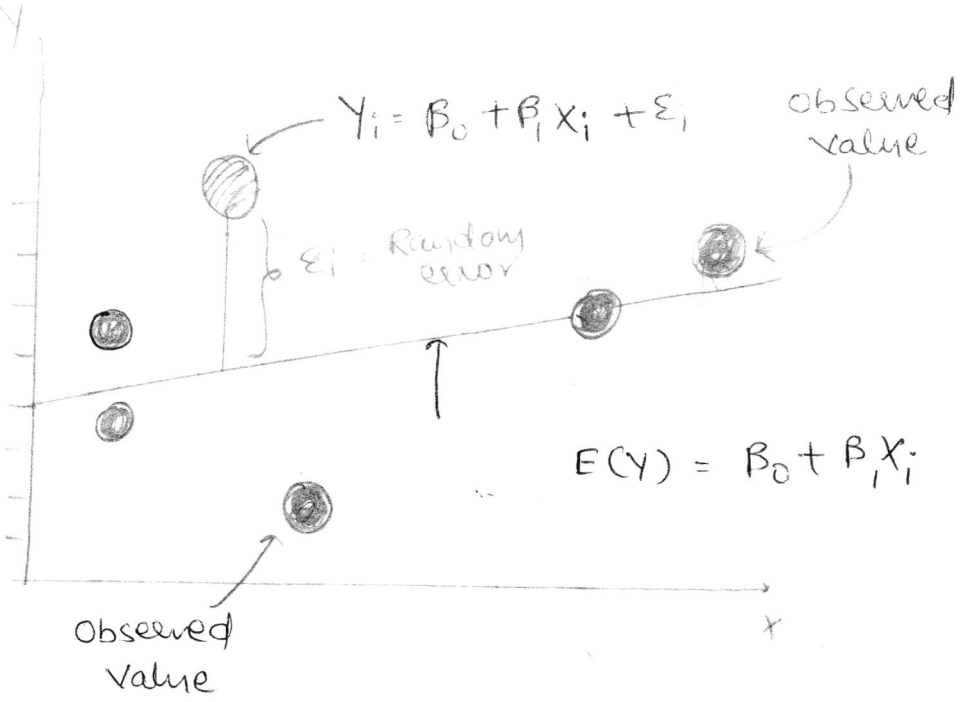
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_1 → population slope
- ε_i → Random error
- Y_i → Dependent (Response variable) e.g. CD+C.
- X_i → Independent (Explanatory) variable e.g. Years s. season)
- β_0 → Population Y-Intercept.

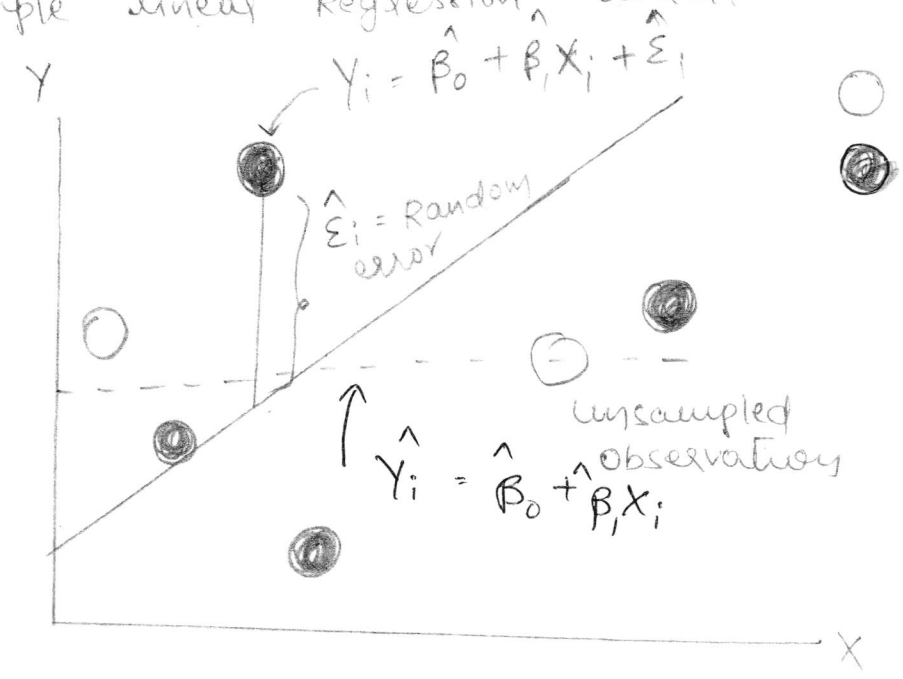
Population & Sample Regression models



Population Linear Regression Model

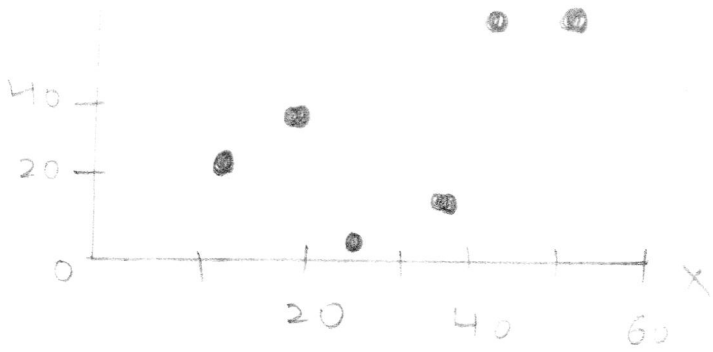


sample linear regression model:-



Scatter plot:

1. Plot of All (X_i, Y_i) pairs
2. Suggests how well model will fit



Thinking challenge:

How would you draw a line through the points? How do you determine which line fits best?

co-relation coefficient

measures the strength of association b/w two variables. The most common correlation coefficient, called the Pearson Product moment correlation coefficient measures the strength of the linear association between variables.

correlation coefficient of a sample	refers to Pearson method. a population
r	ρ or R

How to interpret a correlation coefficient

The sign & the absolute value of a correlation coefficient describe the direction & the magnitude of the relationship between two variables.

1. The value of a correlation coefficient ranges between -1 & 1.
2. The greater the absolute value of a correlation coefficient, the stronger the linear relationship.

curve fitting with linear & non-linear regression

We often think of a relationship between two variables as a straight line. That is, if you increase the predictor by 1 unit, the response always increases by X units. However, not all data have a linear relationship, and your model must fit the curves present in the data.

How do you fit a curve to your data?

→ curve-fitting methods in both linear regression & non-linear regression.

Goal:- To accurately predict the o/p given the input.

difference b/w linear & non-linear equations in R A?

Linear regression requires a linear model

A model is linear when each term is either a constant or the product of a parameter and a predictor variable.

A linear equation is constructed by adding the results for each term. This constrains the equation to just one basic form:

Response = constant + parameter × predict
+ parameter × pre

$$Y = b_0 + b_1X_1 + b_2X_2 -- + b_kX_k.$$

In statistics, a regression equation (or fun
is linear when it is linear in the
parameters. While the equation must b
linear in the parameters, you can tran
the predictor variables in ways that
produce curvature. For instance, you
include a squared variable to produc
a U-shaped curve.

$$Y = b_0 + b_1X_1 + b_2X_1^2 + \ldots$$

This model is still linear in the
parameters even though the predictor
variable is squared. You can also use
log & inverse functional forms that
are linear in the parameters to produce
different types of curves.

Here is an example of a linear
regression model that uses a square
term to fit the curved relationship
between BMI and body fat
percentage.

Simple linear regression example

Problem statement: Last year, 5 randomly selected students took a math aptitude test before they began their statistics course. The Statistics department has 3 equations.

1) What linear regression equation best predicts statistics performance, based on math aptitude scores?

2) If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

3) How well does the regression equation fit the data?

How to find the regression equation:

In the table below, the x_i column shows scores on the aptitude test & y_i column shows statistic grades. The last two rows shows sums & mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	95	85	17	8	289	64
2	85	95	7	18	49	324
3	80	70	2	-7	4	49
4	70	65	-8	-12	64	144
5	60	70	-18	-7	324	49
					730	630

$(x_i - \bar{x})(y_i - \bar{y})$

136

126

-14

96

126

470

The regression equation is a linear equation of the form $\hat{y} = b_0 + b_1 x$.

To conduct a regression analysis, we need to solve for b_0 & b_1. computations are shown below

$$b_1 = \Sigma [(x_i - \bar{x})(y_i - \bar{y}) \Big/ \Sigma (x_i - \bar{x})^2]$$

$$b_1 = 470/730 = 0.644$$

$$b_0 = \bar{y} - b_1 * \bar{x}$$

$$b_0 = 77 - (0.644)(78) = \underline{26.768}$$

Therefore, the regression equation is

$$\hat{y} = 26.768 + 0.644x$$

r

slope = $r \cdot \frac{S_y}{S_x}$

↓

Pearson correla coefficie

How to use the Regression equation?

Once u have the Reg eqn, choose a value for the independent variable (x), perform the computation, & you have an estimated value ($\hat{y}$) for the dependent variable.

In our example, the independent variable is the student's score on the aptitude test.

The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade would be

$$\hat{y} = 26.768 + 0.644x =$$

$$26.768 + 0.644 \times 80$$

$$= 26.768 + 51.52$$

$$= 78.288$$

Extrapolation: When you use a regress
equation, do not use values for the
independent variable that are outsi
the range of values used to crea
the equation. That is called extra
and it can produce unreasonable
estimates.

In this example, the aptitude test sc
used to create the regression equa
ranged from 60 to 95. Therefore, on
values inside that range to estim
statistics grade.

Using values outside that range (les
60 or greater than 95) is problematic

How to find the coefficient of determination:

Whenever you use a regression equation, you should ask how well the equation fits the data. One way to assess fit is to check the coefficient of determination which can be computed from the following formula:

$$R^2 = \left\{ (1/N) * \Sigma \left[(x_i - \bar{x}) * (y_i - \bar{y}) \right] / (\sigma_x * \sigma_y) \right\}^2$$

where 'N' is the no: of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation $\bar{x}$ is the mean value, y_i is the value for observation i, $\bar{y}$ is the mean y value, σ_x is the s.d of x, σ_y is the s.d of y

$$\sigma_x = \text{sqrt}\left[\Sigma (x_i - \bar{x})^2 / N \right] \quad \sigma_y = \text{sqrt}\left[\Sigma (y_i - \bar{y})^2 / N \right]$$

σ_x = sqrt (730/5) = sqrt (146) = 12.083

= sqrt (630/5) = sqrt (126

= 11.225

$$R^2 = \frac{1}{5} * 470 / (12.083 * 11.225)]^2$$

$$= (94 / 135.632)^2 = (0.693)^2 = 0.48$$

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

Residual Analysis in Regression:

Because a linear regression model is not always appropriate for the data. You should assess the appropriateness of the model by defining residuals & examining residual plots.

Residuals:

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Each datapoint has one residual.

Residual = Observed value − Predicted value

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

Both the sum & the mean of the residuals are equal to zero.

That is $\Sigma e = 0$ & $\bar{e} = 0$.

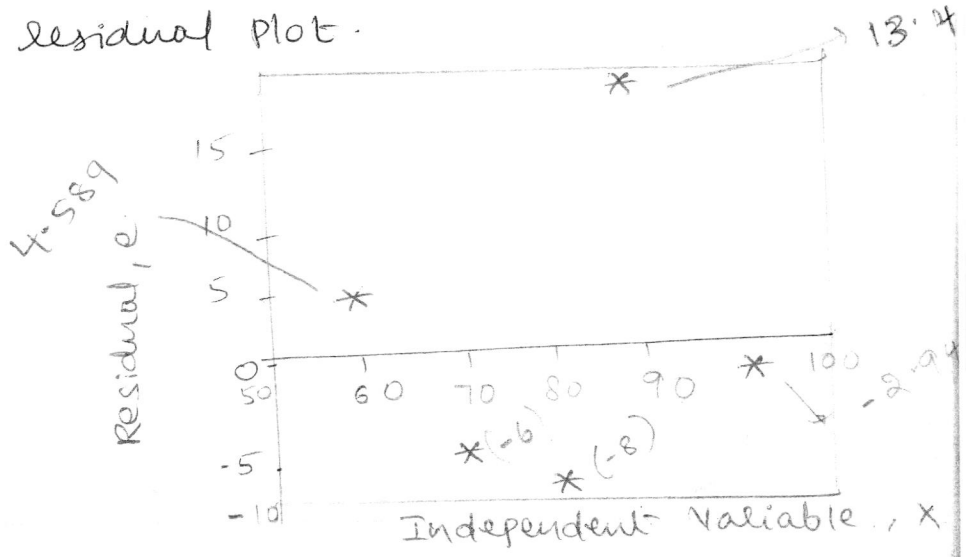
Residual plots

A residual plot is a graph that s the residuals on the vertical axis & th independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizont axis, a linear regression is appropriate for the data, otherwise a non-linear model is more appropriate.

Below, the table shows inputs & o/p's from a simple linear regression analysis.

x	60	70	80	85	95
y	70	65	70	95	85
$\hat{y}$	65.411	71.849	78.288	81.507	87.
e	4.589	-6.849	-8.288	13.493	

And the chart below displays the residu and the independent variable (x) as a residual plot.



The residual plot shows a fairly random pattern.

The first residual is +ve: 4.589

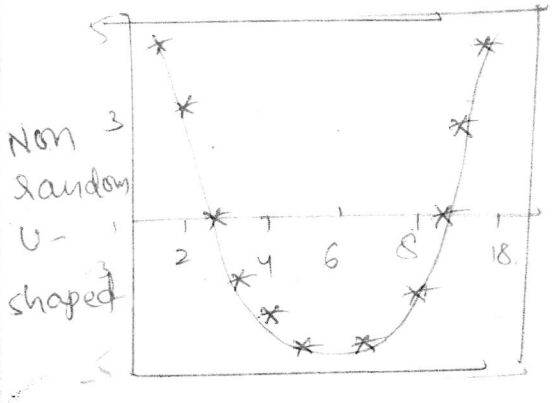
Next two are : -ve: -6.549, -8.288

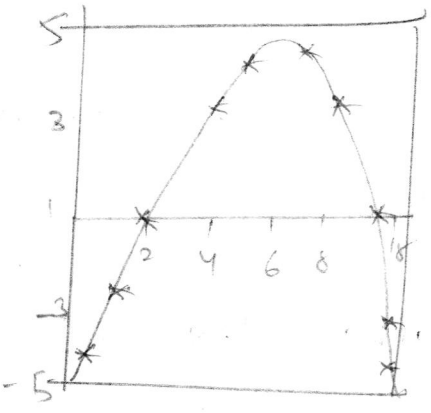
The last residual is -ve -2.945

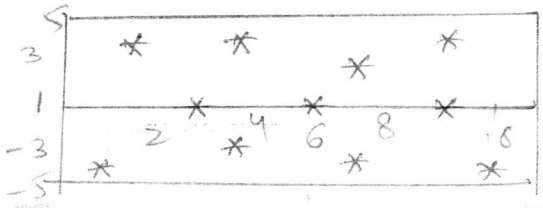
This random pattern indicates that a linear model provides a decent fit to the data.

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model.

Graph patterns - The other plots patterns are non-random (U-shaped & inverted U,) suggesting a better fit for a non-linear model.







Question :-

In the context of R.A, which of the following statements are true?

① When the sum of the residuals is greater than zero, the dataset is non-linear.

2. A random pattern of residuals suppo a linear model.

3. A random pattern of residuals suppo a non-linear model.

A I only

B II only

C III only

D. I & II

E I & III

Solution (B).

A random pattern of residuals suppo linear model.

A non-random pattern supports a non linear model.

The sum of residuals is always zero whether the data set is linear or non-linear.

MULTI-COLLINEARITY

Collinearity: Two independent variables x_1 & x_2 are collinear, when they are correlated with each other.

In a multiple regression study, we assume that the x variables are independent of each other.

we also assume that each x-variable contains a unique piece of info about y

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

In the multiple regression model:

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

we believe that

β_1 = the change in Y for a 1-unit change in x_1, while x_2 is held constant

Also

β_2 = while x_1 is held constant

when collinearity exists:
in that x_1 & x_2 are correlated

B_1 = the change in Y for a 1- change in X_1 with X_2 held constant

B_2 = change X_1 held constant

Effects of multicollinearity:

Variances (& standard errors) of regression coefficient estimators (i.e the b_i) are inflated. This mea that $Var(b_i)$ is too large.

The magnitude of the b_i may be different from what we expect.

The signs of b_i may be opposite o what we expect.

Adding or removing any of the x-va Produce large changes in the values of remaining b_i or their signs.

Sometimes removing a datapoint causes large changes in the value of b_i or their signs.

In some cases, F is significant, but the t-values (for the b_i) may not be significant.

t statistic for β_i

$$= \frac{b_i}{\text{Std Err of } b_i}$$

Tests for Multicollinearity

calculate the correlation coefficient (r) for each pair of the x-variables. If any of the r-values is significantly different from zero, then the independent variables involved may be colinear.

$$r = \frac{\text{covariance}(x_i x_j)}{\sigma_i \sigma_j}$$

Although the 'r' for any two x variables may be too small, these independent variables x_1, x_2 & x_3 may be highly correlated as a group.

2. VIF (Variance Inflation Factor)

Check whether the VIF is too high.

Rule of thumb: collinearity exists if VIF > 5. A VIF of 10, for eg. means $\text{Var}(b_i)$ is 10 times what it should be if no collinearity existed (if no collinearity VIF should be 1).

VIF is more rigorous check for coll
than correlation coefficient.

$$VIF = \frac{1}{(1 - R_i^2)}$$

In the R.M

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 +$$

R_i^2 is obtained from Regressing X
X_2 & X_3 as follows

$$X_1 = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_3 + e$$

Similarly

$$X_2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_3 + e \Rightarrow +$$

$$X_3 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + e \Rightarrow$$

R_3^2

Solutions for M.C

1. Drop the variables causing the problem.
   If using a large no: x-variables stepwise regression could be us
   to determine which of the Va
   to drop.

2.