INTRODUCTION TO REGRESSION

BY BHAVNA KUMARI

HISTORY – The earliest form of regression was the method of least squares , which was published by Legendre in 1805 The term "regression" was coined by Sir Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's

In the 1950s and 1960s, economists used electromechanical desk "calculators" to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression. Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression , regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression

INTRODUCTION : - In statistical modeling , **regression analysis** is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More precisely, regression analysis helps one understand how the typical value of the dependent (or 'criterion variable') changes when any one of the independent varied, while the other independent variables are held fixed.
Most commonly, regression analysis estimates the conditional the dependent variable given the independent variables – that is, value of the dependent variable when the independent variables are fixed. commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, a function of the independent variables called the **regression function** is to be estimated.

In regression analysis, it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable

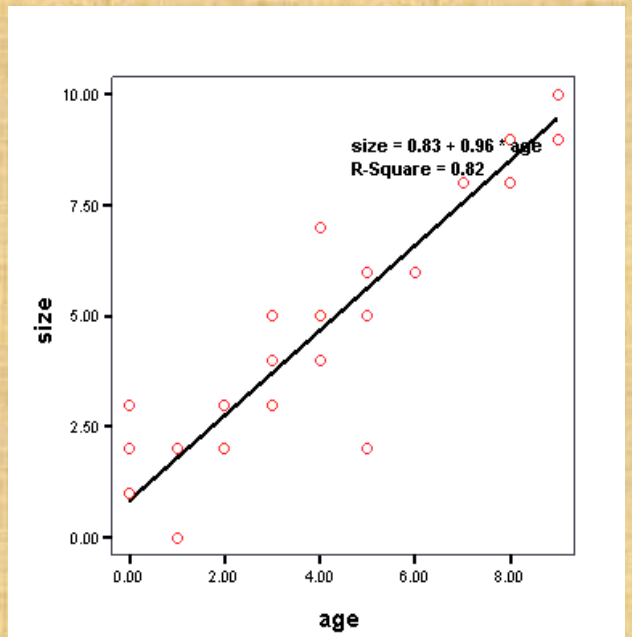
Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.

.

LINEAR REGRESSION :-

If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the curve of regression. If the curve is straight line, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear

The line of regression is a line which gives the best estimate to the value of one variable for any specific value of the other variable . Thus the line of regression is the line of best fit and is obtained by the principle of least squares.



ASSUMPTIONS

1 The sample is representative of the population for the inference inference prediction.

2 The error is a random variable with a mean of zero conditional conditional on the explanatory variables.

3 The independent variables (predictors) are linearly independent, independent, i.e. it is not possible to express any predictor as a linear combination of the others.

4 The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the the variance of the error.

5 The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods methods might instead be used.

PRINCIPLE OF LEAST SQUARE:-

By principle of least square we minimize the error variance to get better estimates. principle of least square is one of the most widely used method to minimize the error variance.

The equation of straight line is

$$Y = a + bx$$

Where y is dependent variable and x is independent variable

We are given n pairs of values on X and Y

X : x_1 x_2.....x_i.....xn

Y : y_1 y_2.....y_i.....yn

We have to fit a straight line

$$Y = a + bx \text{.....(1)}$$

Fitting of straight line means estimating of the constant a and b the equation (1) is mathematical model. The mathematicians consider it exact but the statistician say that the equation (1) is not exact. There is some error in it i.e the equation on (1) is

$$Y = a + bx + \varepsilon \ldots\ldots\ldots\ldots (2)$$

principle of least squares provide a technique through which the constant a and b are to be estimated such that ε is minimized.

1 y on x :- Let the regression line of y on x

$$Y = a + bx \ldots\ldots\ldots (1)$$

where x and y are deviation from their respective means i.e

$$x = X - \bar{x}$$

$$y = Y - \bar{y}$$

we have to estimate a and b

Take sum of (1) from i = 1 to n

$$\Sigma y = na + b\Sigma x$$

or $\Sigma(Y - \bar{y}) = na + b\Sigma(X - \bar{x})$

or $0 = na + b*0$ (since sum of deviation from mean is zero)

or $a = 0$

Now. (1) becomes

$$Y = bx \ldots\ldots\ldots (3)$$

by principle of least squares, the residual sum of squares is

$$S_e^2 = \Sigma(Y - bx)^2$$

Minimizing S_e^2 with respect to b

$$\frac{dS_e^2}{db} = 2\Sigma(Y - bx)(-x)$$

or, $\Sigma(Y - bx)x = 0$

or, $\Sigma yx - b\Sigma x^2 = 0$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{1/n\Sigma(X - \bar{x})\,(Y - \bar{y})}{1/n\,\Sigma(X - \bar{x})^2}$$

$b = Cov(XY)/ V(X)$

b is called regression coefficient of y on x

Now , Substituting $a = 0$

$b = Cov\,(xy) / V(x)$

$Y = 0 + Cov\,(xy) / V(x) * x$

or , $Y - \bar{y} = Cov\,(xy) / V(x) * (X - \bar{x})$ .....(4)

(4) is the equation of regression line of y on x .

2 x on y :- Let the equation of the regression line of x on y be

$X = a + by$.....(1)

x and y are the deviation from their respective means. i.e $x = (X - \bar{x})$ and $y = Y - \bar{y}$ we have to estimate a and b proceeding as above, we get

Again

Regression line of y on x

$$(Y - \bar{y}) = \text{Cov}(xy) / V(x) * (X - \bar{x})$$

Regression line of x on y

$$(X - \bar{x}) = \text{Cov}(xy) / V(y) * (Y - \bar{y})$$

Again

$$(Y - \bar{y}) = r\,\sigma_x\sigma_y/\sigma_y^2 * (X - \bar{x})$$

or,

$$\frac{(Y - \bar{y})}{r\,\sigma_y} = \frac{(X - \bar{x})}{\sigma_x}$$

$$\frac{(X - \bar{x})}{r\sigma_x} = \frac{(Y - \bar{y})}{\sigma_y}$$

Two regression lines are same when $r = +\,-1$

when there is a perfect correlation the two regression line are same.

Regression Coefficients : - ‘ b ‘ the slope of the line of regression of y on x is also called the coefficient of regression of y on x. it represents the increment in the value of dependent variable y corresponding to a unit change in the value of independent variable x . More precisely, we write

byx = Regression coefficient of y on x = $\mu_{11}/\sigma_x^2 = r\,\sigma_y/\sigma_x$

Similarly , the coefficient of regression of y on x indicates the change in the value of variable x to a unit change in the value of variable y and is given by

bxy = Regression coefficient of x on y = $\mu_{11}/\sigma_y^2 = r\sigma_x/\sigma_y$

PROPERTIES OF REGRESSION COEFFICIENTS

- 1 Correlation coefficients is the geometric mean of the regression coefficients.
- 2 If one of the regression coefficients is greater than unity, the other must be less than unity.
- 3 The modulus value of the arithmetic mean of the regression coefficients is not less than the modulus value of the correlation coefficients r.
- 4 Regression coefficients are independent of change of origin but not of scale.

1 What does the coefficient In a regression tell you?

In **regression** with multiple independent variables, the **coefficient tells** you how much the dependent variable is expected to increase when that independent variable increases by one, holding all the other independent variables constant. Remember to keep in mind the units which your variables are measured in.

2 What is the purpose of using a regression analysis?

Regression analysis is used to predict the behavior of an dependent variable based on the behavior of a few/large no. of independent variables(age, height, financial status).

3 Why regression analysis is important?

Regression analysis is one of the most **important** statistical techniques for business applications. It's a statistical methodology that helps estimate the strength and direction of the relationship between two or more variables. The **regression** results show whether this relationship is valid or not.

4 What is the application of regression analysis?

Regressions range from simple models to highly complex **equations**. The two primary **uses** for regression in business are **forecasting** and optimization. In addition to helping managers predict such things as future demand for their products, regression analysis helps fine-tune manufacturing and delivery processes.

5 What is the line of regression?

Linear regression attempts to model the relationship between two variables by fitting a **linear** equation to observed data. A **linear regression line** has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable.

6 What is regression to the mean?

In statistics, **regression** toward (or to) the **mean** is the phenomenon is that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement—and if it is extreme on its second measurement, it will tend to have been closer to the average on its first.

Thank You