e-PGPathshala
Subject : Computer Science
Paper: Data Analytics
Module No 12: CS/DA/12- Data Analysis Foundations – Regression analysis
Quadrant 1 – e-text

1.1 Introduction

The aim of this chapter is to learn the concept of regression and regression analysis. Regression analysis investigates the relationship between variables ie, the relationship between a dependent variable and one or more independent variables. It is used in areas like forecasting, predicting and finding the causal effect of one variable on another.

1.2 Learning Outcomes

- **Learn the fundamentals of regression analysis and types**
- **Understand the difference between regression and correlation**
- **Understand simple linear regression with examples**

1.3 Regression analysis:

1.3.1 Regression

In statistics regression is a measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).

In other words regression is a **statistical technique** to **determine the linear relationship** between two or more variables.

- Regression is primarily used for prediction and causal inference.
- In its simplest (bivariate - one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as *X*, *Y*), for the purpose of determining the empirical relationship between them) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y), as in the formula below:

$$Y = \beta_0+\beta_1X+u$$

- The magnitude and direction of that relation are given by the slope parameter (β_1)
- The status of the dependent variable when the independent variable is absent is given by the intercept parameter (β_0).
- An error term (u) captures the amount of variation not predicted by the slope and intercept terms.
- The regression coefficient (R^2) shows how well the values fit the data.

Regression Vs Correlation

Both Correlation and Regression are statistical tools that deal with two or more variables. Although both relate to the same subject matter, there are differences between the two. The differences, between the two are explained below:

- Correlation analysis is a test of inter-dependence between two variables. Regression analysis gives a mathematical formula to determine value of the dependent variable with respect to a value of independent variable/s.
- Correlation coefficient is independent of choice of origin and scale, but regression coefficient is not so.
- For correlation the values of both the variables have to be random, but this is not so for regression coefficient.
- Correlation determines the strength of the relationship between variables, while regression attempts to describe that relationship between these variables in more detail.

1.3.2 Regression models

Regression models are two types (figure 1): Simple regression model and multiple regression model. Both are divided into linear and nonlinear models.

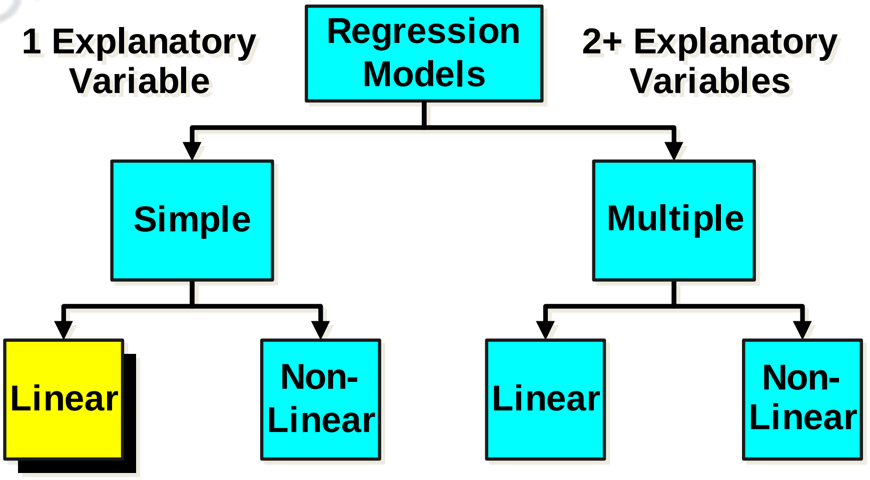
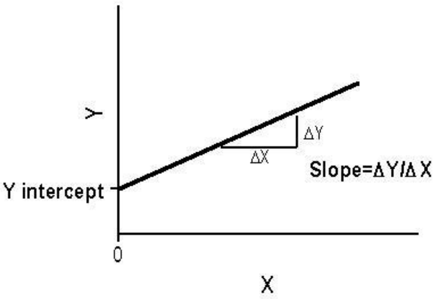


Figure 1. Types of Regression

i) Linear Regression model

Linear regression is a statistical procedure for predicting the value of a dependent variable from an independent variable when the relationship between the variables can be described with a linear model.

A linear regression equation can be written as Y_p= mX + b, where Y_p is the predicted value of the dependent variable, m is the slope of the regression line, and b is the Y-intercept of the regression line.



In statistics, linear regression is a method of estimating the conditional expected value of one variable y given the values of some other variable or variables x. The variable of interest, y, is conventionally called the "dependent variable". The terms "endogenous variable" and "output variable" are also used. The other variables x are called the "independent variables". The terms "exogenous variables" and "input variables" are also used. The dependent and independent variables may be scalars or vectors. If the independent variable is a vector, one speaks of multiple linear regression.

A linear regression model is typically stated in the form $\mathbf{y = \alpha + \beta x \pm \varepsilon}$

The right hand side may take other forms, but generally comprises a linear combination of the parameters, here denoted **α** and **β**. The term ε represents the unpredicted or unexplained variation in the dependent variable; it is conventionally called the "error" whether it is really a measurement error or not. The error term is conventionally assumed to have expected value equal to zero, as a nonzero expected value could be absorbed into **α**. See also errors and residuals in statistics; the difference between an error and a residual is also dealt with below. It is also assumed that is **ε** independent of **x**.

Advantages / Limitations of Linear Regression Model:

- Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results.
- Linear regression is often inappropriately used to model non-linear relationships.

- Linear regression is limited to predicting numeric output.
- A lack of explanation about what has been learned can be a problem.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The variable y is assumed to be normally distributed with mean μ y and variance σ. The least-squares regression line y = b0 + b1x is an estimate of the true population regression line, μ y = β0+ β1x. This line describes how the mean response μ y changes with x. The observed values for y vary about their means μ y and are assumed to have the same standard deviation σ.

Computation of b_0 and b_1 is done as given below:

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

The fitted values b0and b1 estimate the true intercept and slope of the population regression line and is estimated as shown below:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

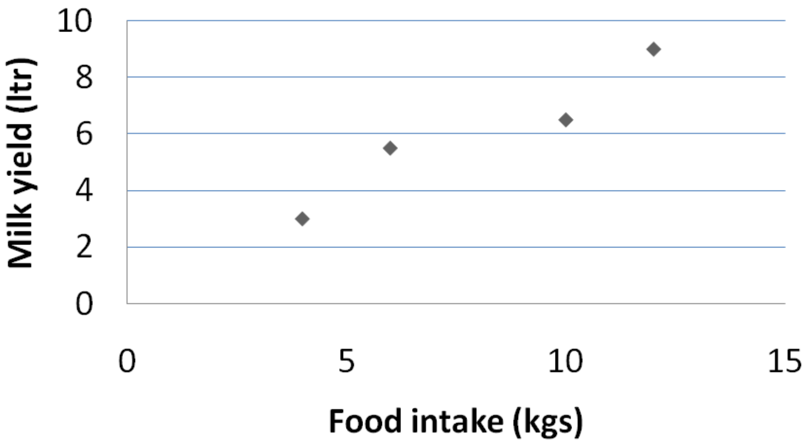
Since the observed values for y vary about their means μ y, the statistical model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression β0 + β1x. The "RESIDUAL" term represents the deviations of the observed values y from their means μ y, which are normally distributed with mean 0 and variance σ. The notation for the model deviations is ε. The interpretation of slope (b1) highlights how Y changes by b1 for each 1 unit increase in X. For Example, if b1 = 2, then Y is expected to increase by 2 for each 1 unit increase in X. Similarly, the Y-Intercept (b0) highlights the average value of Y when X = 0, for example if b0 = 4, then average Y is expected to be 4 when X is 0.

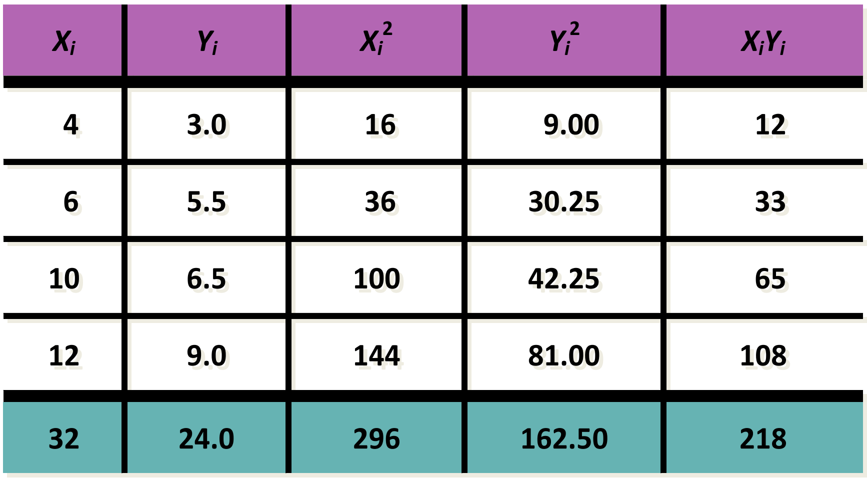
Example: Consider the following data about food intake by cows and milk yield collected from a cattle farm:

Food (kg)	Milk yield (ltrs)
4	3.0
6	5.5
10	6.5
12	9.0

What is the relationship between cows' food intake and milk yield?

Solution: The scatter diagram for the above mentioned data is as shown below:





X_i	Y_i	X_i^2	Y_i^2	X_iY_i
4	3.0	16	9.00	12
6	5.5	36	30.25	33
10	6.5	100	42.25	65
12	9.0	144	81.00	108
32	24.0	296	162.50	218

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{218 - \frac{(32)(24)}{4}}{296 - \frac{(32)^2}{4}} = 0.65$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 6 - (0.65)(8) = 0.80$$

It is understood from Slope (b_1) that Milk yield (*Y*) is expected to increase by .65 ltr for each 1 kg increase in food intake (*X*) and from the **Y-Intercept (b_0),** average milk yield (*Y*) is expected to be 0.8 ltr. when food intake (*X*) is 0.

1.3.3 Sum of squares of error

Residual **sum of squares**. In statistics, the residual **sum of squares** (RSS), also known as the **sum of squared** residuals (SSR) or the **sum of squared errors** of prediction (SSE), is the **sum** of the **squares** of residuals (deviations predicted from actual empirical values of data).

SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0.
The formula for SSE is:

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Where n is the number of observations x_i is the value of the ith observation and 0 is the mean of all the observations. This can also be rearranged to be written as seen in J.H. Ward's paper.

1.3.4 Sum of squares of regression

The sum of squares is a mathematical approach to determining the dispersion of data points. In a regression analysis, the goal is to determine how well a data series can be fitted to a function which might help to explain how the data series was generated. The sum of squares is used as a mathematical way to find the function which best fits (varies least) from the data.

In order to determine the sum of squares the distance between each data point and the line of best fit is squared and then all of the squares are summed up. The line of best fit will minimize this value.

The sum of squares of the residual error is the variation attributed to the error. By comparing the regression sum of squares to the total sum of squares, you determine the proportion of the total variation that is explained by the regression model (R^2, the coefficient of determination).

1.3.5 Total sum of squares

Total sum of squares = sum of squares due to regression + sum of squared errors,

$SST=SSR+SSE$

$SSR = \sum (y - y)$ (measure of explained variation)

$SSE = \sum (y - y)$ (measure of unexplained variation)

$SST = SSR + SSE = \sum (y - y)$ (measure of total variation in y)

1.3.6 Coefficient of Determination

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
- With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

Coefficient of determination. The coefficient of determination (R2) for a linear regression model with one independent variable is:

R2 = { (1 / N) * Σ [(xi - x) * (yi - y)] / (σx * σy) }2

where N is the number of observations used to fit the model, Σ is the summation symbol, xi is the x value for observation i, x is the mean x value, yi is the y value for observation i, y is the mean y value, σx is the standard deviation of x, and σy is the standard deviation of y.

1.3.7 Standard errors of regression

The **standard error** of the estimate is a measure of the accuracy of predictions. Recall that the **regression** line is the line that minimizes the sum of squared **deviations** of prediction (also called the sum of squares **error**).

The Standard Error of a regression is a measure of its variability. It can be used in a similar manner to standard deviation, allowing for prediction intervals.

y ± 2 standard errors will provide approximately 95% accuracy, and 3 standard errors will provide a 99% confidence interval.

Standard Error is calculated by taking the square root of the average prediction error.

Standard Error = $\sqrt{SSE/_{n-k}}$

where n is the number of observations in the sample and k is the total number of variables in the model

1.4 Multiple Regression model

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable from two or more independent variables Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*. The population regression line for *p* explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for *y* vary about their means μ_y and are assumed to have the same standard deviation σ. The fitted values b_0, b_1, ..., b_p estimate the parameters β_0, β_1, ..., β_p of the population regression line.

Since the observed values for *y* vary about their means μ_y, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$. The "RESIDUAL" term represents the deviations of the observed values *y* from their means μ_y, which are normally distributed with mean 0 and variance σ. The notation for the model deviations is ε.

Formally, the model for multiple linear regression, given *n* observations, is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... \beta_p x_{ip} + \varepsilon_i$ for $i = 1,2, ... n$.

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares estimates $b_0, b_1, ... b_p$ are usually computed by statistical software.

The values fit by the equation $b_0 + b_1 x_{i1} + ... + b_p x_{ip}$ are denoted $\hat{y}_i$, and the residuals e_i are equal to $y_i - \hat{y}_i$, the difference between the observed and fitted values. The sum of the residuals is equal to zero.

The variance σ^2 may be estimated by s^2 = , $\dfrac{\sum e_i^2}{n - p - 1}$

also known as the mean-squared error (or MSE). The estimate of the standard error *s* is the square root of the MSE.

1.4.1 Multi-collinearity

Multicollinearity refers to a situation where a number of independent variables in a multiple regression model are closely correlated to one another. Multicollinearity can lead to skewed or misleading results when a researcher or analyst is attempting to determine how well each one of a number of individual independent variables can most effectively be utilized to predict or understand the dependent variable in a statistical model. In general, multicollinearity can lead to wider confidence intervals and less reliable probability values (P values) for the independent variables. Statistical analysts use multiple regression models to predict the value of a specified dependent variable based on the values of two or more independent variables. The dependent variable is also sometimes referred to as the outcome, target or criterion variable. Multicollinearity in a multiple regression model indicates that the collinear independent variables are related in some fashion, although the relationship may or may not be a causal relationship.

One of the most common ways of eliminating the problem of multicollinearity in a study is to first identify collinear independent variables and remove all collinear variables until there is only one remaining. It is also sometimes possible to eliminate multicollinearity by combining two or more collinear variables into a single variable. Statistical analysis can then be conducted to study the relationship between the specified dependent variable and only a single independent variable.

1.4.2 Non linear regression

Nonlinear regression is a regression in which the dependent or criterion variables are modeled as a non-linear function of model parameters and one or more independent variables.

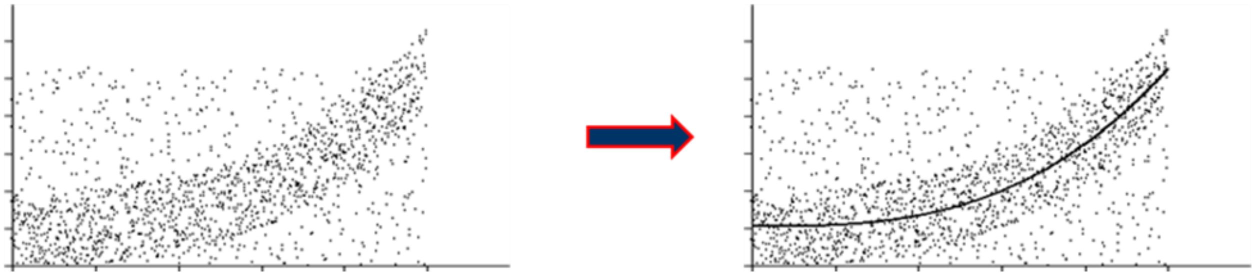


Figure 2. Graphical representation of Non-linear regression

There are several common models, such as Asymptotic Regression/Growth Model, which is given by: b1 + b2 * exp(b3 * x)

Logistic Population Growth Model, which is given by:
b1 / (1 + exp(b2 + b3 * x)), and

Asymptotic Regression/Decay Model, which is given by:
b1 - (b2 * (b3 * x)) etc.

The reason that these models are called nonlinear regression is because the relationships between the dependent and independent parameters are not linear.

There are certain terminologies in nonlinear regression which will help in understanding nonlinear regression in a much better manner. These terminologies are as follows:

Model Expression is the model used, the first task is to create a model. The selection of the model in is based on theory and past experience in the field. For example, in demographics, for the study of population growth, logistic nonlinear regression growth model is useful.

Parameters are those which are estimated. For example, in logistic nonlinear regression growth model, the parameters are b1, b2 and b3.

Segmented model is required for those models which have multiple different equations of different ranges, equations are then specified as a term in multiple conditional logic statements.

Loss function is a function which is required to be minimized. This is done by nonlinear regression.

Assumptions The data level in must be quantitative, the categorical variables must be coded as binary variables.

The value of the coefficients can be correctly interpreted, only if the correct model has been fitted, therefore it is important to identify useful models.

Case Studies
A) http://hrdailyadvisor.blr.com/2014/04/27/regression-analysis-a-case-study/
A nonprofit home healthcare agency has asked “a consultant” whether its CEO is fairly paid relative to the marketplace for similar agencies. The Agency has supplied a database to the consultant, who also has his own survey database of CEO pay.This case will demonstrate how regression data can be used to answer this question.
B) https://www.stat.wisc.edu/courses/st572-larget/handouts06-4.pdf
Multiple Linear Regression Case Study Bret Larget Departments of Botany and of Statistics University of Wisconsin—Madison February 5, 2008 Birds and bats must expend considerable energy to fly. Some bats use echolocation in flight which also requires energy. Other bats eat fruit and do not have the ability to echolocate. Scientists studied energy use of several species of birds and bats to examine the relationship between mass and energy expenditure during flight to see if echolocating bats had a higher cost. Variables are mass (grams), type (factor with levels bird, eBat, and nBat, latter two for echolocating and non-echolocating), and the response energy (Watts).

Summary

- Regression analysis shows us how variation in one variable co-occurs with variation in another.
- An error in regression captures the amount of variation not predicted by the slope and intercept terms.
- For Nonlinear functions any continuous function can be used.