Abstract

Introduction

Regression Analysis is a field of study that deals with approximating the behavior of a system dependent on one or more independent variables. The Observed value is called regressand, endogenous variable, response variable, measured variable, criterion variable or dependent variable. The independent variables are called regressors, exogenous variables, explanatory variables, covariates, input variables or predictor variables. Often, many independent variables are kept constant and the response to a few independent variables is modeled. The regression model could be a system of linear or nonlinear equations, a probability distribution and corresponding confidence intervals, the upper and lower bounds of dependent parameters etc.

The regression model is tested against a set of quantitative data gathered by an empirical study of the system, which represents the system response to changes in the important independent variables. A valid regression model would show acceptable absolute deviation from the recorded behavior. The deviation is measured as a function of the errors (residuals) between the system and model responses to similar stimuli. Some methods of determining regression parameters include: least squares (linear, non-linear, weighted, ordinary, generalized, partial, total, non-negative, regularized or iteratively reweighted least squares), correlation coefficients (Pearson, Spearman’s Rank, Kendall Tau Rank, Goodman and Kruskal’s gamma or Intra-class correlation coefficients), least absolute deviations, maximum likelihood estimation, Berkson’s minimum chi-squared method, Gibbs sampling, convex optimization, generalized method of moments, successive approximation etc.

The underlying assumptions are that the observed data sample is random, has adequate size and is representative of the population; independent variables are independent and measured with no error; ignored parameters are constant etc. If the measured data does not fulfil these assumptions, further error is introduced in the regression model.

All laws of nature have been devised by regression analysis i.e. data collection and curve fitting. The formulae are widely accepted because they generate acceptable error within the required limits of accuracy. The Gravitational law f=GmM/d^2 could be f=(GmM/d^2)exp(-ar) as Laplace proposed in 1790 or f=(GmM/d^2)exp(1+a/r^3)as proposed by Decombes in 1913. After all it was superceded by Einstein’s theory of general relativity which applies in cases of very stong gravitational fields. But newton’s theory had been rigorously tested over so many data samples that it was widely believed to be a law of nature, and applies well in all practical cases. Hence the validity of a regression function is a measure of its accuracy in the desired frame of reference.

A system can have infinite properties. For example, each resistor will have specific acoustical, atomic, chemical, environmental, electrical, magnetic, manufacturing, mechanical, optical, radiological and thermal properties. We tend to ignore majority of the variables and focus only on its resistivity because other parameters are assumed to be constant. The superiority of a regression function depends on its unbiasedness, consistency, efficiency and sufficiency.

Types

Linear

Y is vector of observed values, X is design matrix of independent values, is the vector of effect or regression coefficients. It is also called Parameter vector. is called error term, disturbance term or noise vector. Errors (residuals) have a normal distribution.

Simple

One dependent and one independent variable.

Minimize

Y is n dimensional vector of dependent variable, X is n dimensional vector of independent variable, is the y-intercept.

Non-Linear

Non-linear combination of model parameters.

Exponential function, logarithmic function, trigonometric function, power function, Gaussian function, Lorenz curves.

Polynomial

Multivariate

More than one independent variable, more than one dependent variable.

Qualitative choice models

Predict choices between two or more discrete choices. The alternatives must be finite, mutually exclusive and exhaustive.

Logistic

Model a binary dependent variable e.g. predicting the probability of developing a given disease based on observed characteristics of patient (age, body mass index, blood tests etc.). Independent variables can be real valued, binary valued, categorical valued etc. Logistic regression can be binomial (two choices of dependent variable), ordinal (dependent variable has ordered values) or multinomial (dependent variable has greater than two choices).

Probability

Multinomial Logistic

Categorically distributed dependent variable with K possible values.

Mixed Logit

Probit

Multinomial Probit

Ordered Logit

Ordered Probit

Poisson

Dependent variable has Poisson Distribution (variance is equal to mean) and the logarithm of its expected value can be modeled by a linear combination of unknown parameters.

where ,

The Poisson probability mass function is

Multi-level Models

Fixed Effects Models

Random Effects Model

Mixed Model

Non-Parametric

Semi-parametric

Robust

Quantile regression

Y is a real valued random variable with cumulative distribution function. It can be used to estimate conditional median or other quantiles of the response variable.

The Quantile of Y is

Loss Function is

is the solution of

Isotonic

A non-decreasing free-form line that lies as close to the observations as possible.

Principal Component

Least Angle

Local

Segmented

Errors-In-Variables Models

Independent variable has measurement error

Measured value

Techniques

least squares (

linear

Errors have finite variance and are homoscedastic. Errors are uncorrelated with regressors:

E [

Moore-Penrose Pseudoinverse:

non-linear, weighted, ordinary,

generalized

Errors are correlated or heteroscedastic.

where is the covariance matrix of the errors.

, partial, total, non-negative, regularized or iteratively reweighted least squares,

Instrumental Variables)

Regressors are correlated with the errors ϵ

where E [ and Z is vector or estimators.

Least Absolute Deviations

Minimize sum of absolute errors

It is robust due to resistance to outliers in the data i.e. it gives equal weightage to all observations unlike least squares which gives more weight to larger residuals by squaring them.

Maximum Likelihood Estimation

Parameters are chosen so that Likelihood function is maximized. Parameters are estimated using observations and their assumed uniform distributions.

Bayesian Linear Regression

Errors are independent and identically normally distributed . If a prior distribution is assumed, explicit results are available for the posterior probability distributins of the model parameters. The likelihood function is

The prior distribution is

The posterior probability density distribution is obtained by multiplying the likelihood with the prior probability density distribution:

Case Study

<https://ir.nctu.edu.tw/bitstream/11536/24003/1/000333363000004.pdf>

Data Driven Modeling for Power Transformer Lifespan Evaluation.

Real time lifespan forcasting using remote terminal unit (RTU). Real time condition parameters and historical data monitoring, real time fault diagnosis, remaining life estimation. Logistic regression based on Weibull distribution. Data series from 161kV transformers. Maintenance and Replacement decision support. Oil immersed transformers. Insulating paper and transformer oil undergo electrical and mechanical degradation which is good indicator of transformer’s remaining life. Dissolved gas in oil analysis. Doernenburg ratio method and Rogers diagnosis uses oil gases of for fault analysis of thermal decomposition, partial discharge and arcing. The rapid increase of these gases is strong indicator of degradation of transformer life. Degree of polymerization and tensile strength of insulating paper may indicate the remaining life of the transformer.

Logistic regression:

Continuous, discrete or mixed variables. Predict binary dependent values.

Input: Combustible gases and furfural concentration

Output: Probaility of failure, transformer health, remaining life of transformer.

Probability density function of Weibull random variable:

Weibull distribution is used to estimate failure mean time and estimate time of occurance of different abnormal conditions.

Failure probability:

679 data points collected from 161kV transformers. It included 56 abnormal data sets. Independent variables included 9 combustible gases and 5 furfural concentrations. O2,N2,CO, H2,CH4,C2H2,C2H4,and C2H6,Total Combustible Gases(5-Hydroxymethyl-2-Furaldehyde (5-HMF), 2-Furaldehyde (2-FAL), 2-Furfuryl Alcohol (2-FOL), 2-Acetylfuran (2-ACF), 5-MEF).

Where x=,2-ACF,5-MEF are the condition variables.

Minitab Statistical Software (Minitab Inc. 2011). A p-p plot was used to indicate data compliance with the Weibull distribution (β=3.141, 𝜂=11.67 years, γ=-2.589 years). Hence transformer enters abnormal state after 11.67 years of service with great likelihood of failure. Logistic model for one of the transformers: