Power System Planning Assignment 2

Regression Analysis.

Muhammad Shamaas

2018-MS-EE-4

*University of Engineering and Technology G.T. Road Lahore Punjab Pakistan*

2018msee004@student.uet.edu.pk

**Abstract: T**

***Index Terms*—Regressand, Regressors, Residuals, Regression Coefficients.**

# INTRODUCTION[[1]](#footnote-1)

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 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 minimum deviation from the recorded behavior. The deviation is measured as a function of the errors, also called residuals, disturbance or noise, between the system and model responses to similar stimuli.

The underlying assumptions are that the observed data sample is random, has adequate size and is representative of the population. Also, regressors are independent and measured with no error. 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.

Newton’s Gravitational law states that the Gravitational Force F between two masses M and m, a distance d apart is related to Gravitational Constant G by (1):

(1)

Eq. (2) presents the Gravitational Law proposed by Laplace in 1790. It is an extension of Newton’s Gravitational law (1) with a negative exponential term employing a decay constant .

(2)

Another model proposed by Decombes in 1913 is given in (3) below.

(3)

All of them were superseded by Einstein’s theory of general relativity which applies in cases of very strong gravitational fields. Nevertheless, all of these models apply 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 and infinite regression models. 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 OF REGRESSION MODELS

The regression model could be a system of linear or non-linear equations, a probability distribution and corresponding confidence intervals, the upper and lower bounds of dependent parameters etc.

## A. Linear Regression

The general Linear Regression model is given in (4). 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 the error term. The Errors (residuals) have a normal distribution.

(4)

where

is the regression parameter; are the error terms.

## B. Non-Linear Regression

Non-linear regression involves non-linear combination of model parameters. Some examples include:

### Exponential function:

(5)

are the regression parameters; are the error terms.

### Polynomial function:

are the regression parameters; are the error terms.

### Power function:

### (7)

are the regression parameters; are the error terms.

## C. Logistic Regression

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. The probability p can be expressed as:

(8)

are the regression parameters.

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). If the categorical dependent variable Y has K possible values:

(9)

(10)

are the regression parameters.

## D. Poisson Regression

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.

(11)

where

,

The Poisson probability mass function is

(12)

## E. Quantile Regression

Y is a real valued random variable with cumulative distribution function:

(13)

It can be used to estimate conditional median or other quantiles of the response variable. The Quantile of Y is

(14)

The Loss Function is

is the solution of

(16)

## F. Errors-In-Variables Linear Regression

Independent variable has measurement error .

The measured values are:

(17)

with the linear regression model:

(18)

The correct regression model is:

(19)

(20)

are the regression parameters; are the error terms.

# TECHNIQUES OF REGRESSION

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.

## A. Linear Least Squares

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

E [ (21)

(22)

The Moore-Penrose Pseudoinverse can be used to calculate the regression parameters:

(23)

are the regression parameters; are the error terms.

## B. Generalized Least Squares

Errors are correlated or heteroscedastic.

(24)

where is the covariance matrix of the errors; are the regression parameters.

*C. Method of Instrumental Variables*

If regressors are correlated with the errors ϵ, Instrumental variables can be used to calculate the regression coefficients:

(25)

where E [ and Z is vector or estimators; are the regression parameters.

*D. Partial Differentiation with respect to regression coefficients*

The minimization of

(26)

can be achieved by Partial Differentiation with respect to regression coefficients. The resultant equations are set equal to zero. This results in simultaneous equations which can be solved to calculate the regression coefficients.

*E. Maximum Likelihood Estimation*

In this method, regression parameters are chosen so that the Likelihood function

(27)

are the regression parameters.

is maximized. The Parameters are estimated using observations and their assumed uniform distributions.

*F. Bayes Estimator*

If the errors are independent and identically normally distributed , and a prior distribution is assumed, explicit results are available for the posterior probability distributions of the model parameters.

If the likelihood function is

(28)

and the assumed prior distribution is

(29)

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

(30)

where are the regression parameters.

# CASE STUDY: DATA DRIVEN MODELING FOR POWER TRANSFORMER LIFESPAN EVALUAION

## A. Introduction

Real time transformer lifespan forecasting using remote terminal unit (RTU), Real time condition parameters and historical data monitoring for real time fault diagnosis; Maintenance and Replacement decision support; and remaining life estimation. Insulating paper and transformer oil in Oil immersed transformers undergo electrical and mechanical degradation which is a strong indicator of transformer’s remaining life. Dissolved gas in oil analysis is often used for this purpose. 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. The degree of polymerization and tensile strength of insulating paper may also indicate the remaining life of the transformer. This research uses Logistic regression based on Weibull distribution and Data series from 161kV transformers for power transformer lifespan evaluation.

## B. Logistic Regression Model

Logistic regression is used in the case of Continuous, discrete or mixed independent variables to predict binary dependent values like asset failure. It is the normalized version of a linear regression function.

The input was Combustible gases and furfural concentrations. The output was the Probability of transformer failure which could be used to gauge transformer health and remaining life of transformer.

Weibull distribution was used to estimate failure mean time and estimate time of occurrence of different abnormal conditions. The Probability density function of Weibull random variable is

(31)

where is the shape parameter, 𝜂 is the scale parameter and γ is the location parameter

The failure probability was calculated using the logistic model:

(32)

are the regression parameters.

In this research, 679 data points collected from 161kV transformers. It included 56 abnormal data sets and 623 normal data sets. The Independent variables included concentrations of 9 combustible gases:

and 5 Total Combustible Gases TCG:

5-Hydroxymethyl-2-Furaldehyde (5-HMF), 2-Furaldehyde (2-FAL), 2-Furfuryl Alcohol (2-FOL), 2-Acetylfuran (2-ACF) and 5-MEF.

The Minitab Statistical Software (Minitab Inc. 2011) was used to prepare a p-p plot 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. The Weibull distribution was used to generate the linear regression function:

(33)

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

Hence the Logistic model for failure probability of one of the transformers was developed:

(34)

where t is time of service in years.

# CONCLUSION

T

.

REFERENCES

1. T.

1. [↑](#footnote-ref-1)