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*—Regression is a vital statistical tool for estimating the relationship between variables. With infinite combinations and techniques, it lends a helping hand in all fields of science, finance, economics, politics, medicine, electric power systems and even weather forecast. This research discusses some prominent regression models on the basis of their unbiasedness, consistency, efficiency and sufficiency. A variety of techniques for estimation of regression coefficients is presented that deal with matrix algebra, convex optimization and Bayesian Statistics. Special emphasis is laid on logistic regression by discussing an interesting case study related to regression model for power transformer lifespan and health evaluation using concentrations of dissolved transformer oil gases.**

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

I. 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 [1].

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 [1].

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 [3].

Newton's Gravitational law [2] states that the Gravitational Force F between two masses M and m, a distance d apart is related to Gravitational Constant G by (1):

$$F = \frac{GMm}{d^2} \tag{1}$$

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

$$F = \frac{GMm}{d^2}\ e^{-\alpha d} \tag{2}$$

Another model proposed by Decombes in 1913 [2] is given in (3) below.

$$F = \frac{GMm}{d^2}\ e^{(1+\frac{\alpha}{d^3})} \tag{3}$$

All of them were superseded by Einstein's theory of general relativity which applies in cases of very strong gravitational fields [2]. 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. It is important to ignore a lot of parameters and focus on the most important variables. P. Fritzson has expressed this as:

"At this point, we must be clear about how a system is to be defined. Our first impulse is to point at the pendulum and to say "the system is that thing there." This method, however, has a fundamental disadvantage: every material object contains no less than an infinity of variables, and therefore, of possible systems. The real pendulum, for instance, has not only length and position; it has also mass, temperature, electric conductivity, crystalline structure, chemical impurities, some radioactivity, velocity, reflecting power, tensile strength, a surface film of moisture, bacterial contamination, an optical absorption, elasticity, shape, specific gravity, and so on and on. Any suggestion that we should study all the facts is unrealistic, and actually the attempt is never made. What is necessary is that we should pick out and study the facts that are relevant to some main interest that is already given.” [4].

II. 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. [1]. A few regression models will be discussed here.

A. Linear Regression

The general Linear Regression model [3] is given in (4). Y is vector of observed values, X is design matrix of independent values, β is the vector of effect or regression coefficients. It is also called Parameter vector. ε is the error term. The Errors (residuals) have a normal distribution.

$$Y = \mathrm{X}\,\beta + \alpha + \varepsilon \tag{4}$$

where

$$y_i = \alpha + \beta\mathrm{x}_i + \varepsilon_i, \mathrm{i} = 1, \dots, \mathrm{n}$$

$X, Y, \varepsilon \in \mathbb{R}^n$; β is the regression parameter; ε_i are the error terms.

B. Non-Linear Regression

Non-linear regression involves non-linear combination of model parameters [3]. Some examples include:

1) *Exponential function*:

$$y_i = \beta_0 1 + (\beta_1)^{\mathrm{x}_i} + \dots + (\beta_p)^{\mathrm{x}_i} + \varepsilon_i, \mathrm{i} = 1, \dots, \mathrm{n} \tag{5}$$

$X, Y, \varepsilon \in \mathbb{R}^n$; $\beta_j$are the regression parameters; ε_j are the error terms.

2) *Polynomial function*:

$$y_i = \beta_0 1 + \beta_1 (x_i)^1 + \dots + \beta_p (x_i)^p + \varepsilon_i, i = 1, \dots, n \tag{6}$$

$X, Y, \varepsilon \in \mathbb{R}^n$; $\beta_j$are the regression parameters; ε_j are the error terms.

3) *Power function*:

$$y_i = \beta_0 1 + (x_i)^{\beta_1} + \dots + (x_i)^{\beta_p} + \varepsilon_i, i = 1, \dots, n \tag{7}$$

$X, Y, \varepsilon \in \mathbb{R}^n$; $\beta_j$are the regression parameters; ε_j are the error terms.

C. Logistic Regression

Logistic regression is used to 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.) [5]. The independent variables can be real valued, binary valued, categorical valued etc. The probability p [6] can be expressed as:

$$\mathrm{p}(\mathrm{Y} \mid \mathrm{X}) = \frac{\mathrm{e}^{(\beta_0 + \beta_1 \mathrm{x}_1 + \dots + \beta_n \mathrm{x}_n)}}{1 + \mathrm{e}^{(\beta_0 + \beta_1 \mathrm{x}_1 + \dots + \beta_n \mathrm{x}_n)}} \tag{8}$$

$X \in \mathbb{R}^n$; $\beta_j$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, the probabilities [6] can be expressed as:

$$\mathrm{p}\,(\mathrm{Y} = \mathrm{y}) = \frac{\mathrm{e}^{\beta_y \cdot \mathrm{x}_i}}{1 + \sum_{k=1}^{K-1} \mathrm{e}^{\beta_k \cdot \mathrm{x}_i}} \tag{9}$$

$$\mathrm{p}\,(Y = \mathrm{K}) = \frac{1}{1 + \sum_{k=1}^{K-1} \mathrm{e}^{\beta_k \cdot \mathrm{x}_i}} \tag{10}$$

$X \in \mathbb{R}^n$; $\beta_j$are the regression parameters.

D. Poisson Regression

If the dependent variable has a Poisson Distribution (variance is equal to mean), the logarithm of its expected value [7] can be modeled by a linear combination of unknown parameters as in (11).

$$\log E(Y \mid X) = \theta X \tag{11}$$

where

$$\theta = \alpha + \beta X\ , X, Y \in \mathbb{R}^n;\ \alpha, \theta \in \mathbb{R};\ \beta \in \mathbb{R}^n$$

The Poisson probability mass function [7] is

$$\mathrm{p}(\mathrm{Y} \mid \mathrm{X}\,;\ \theta) = \prod_{i=1}^{m} \frac{e^{Y\theta X} e^{-e^{\theta X}}}{y_i!} \tag{12}$$

E. Quantile Regression

Quantile regression is used when Y is a real valued random variable with cumulative distribution function:

$$F_Y\,(y) = p(Y \le y) \tag{13}$$

It can be used to estimate conditional median or other quantiles of the response variable [8]. The τth Quantile q_τ of Y is

$$q_\tau = \inf\,\{y: F_Y\,(y) \ge \tau\} \tag{14}$$

The Loss Function is

$$L(u) = min_u \left\{ (\tau - 1) \int_{-\infty}^{u} (y - u)\, dF_Y\,(y) + \tau \int_{u}^{\infty} (y - u)\, dF_Y\,(y) \right\} \tag{15}$$

q_τ is the solution of (16).

$$\frac{dL}{du} = 0\ i.e.\,(1 - \tau) \int_{-\infty}^{q_\tau} dF_Y\,(y) - \tau \int_{q_\tau}^{\infty} dF_Y\,(y) = 0 \tag{16}$$

F. Errors-In-Variables Linear Regression

If the independent variable x_i has measurement error η_i, the measured values are expressed as:

$$x_i^* = x_i + \eta_i \tag{17}$$

with the linear regression model:

$$y_i = \alpha + \beta^* x_i^* + \varepsilon_i, i = 1, \dots, n \tag{18}$$

The correct regression model and regression coefficients [9] are given in (19) and (20).

$$y_i = \alpha + \beta x_i + \varepsilon_i, i = 1, \ldots, n \tag{19}$$

$$\beta^* = \frac{Cov\,[x_i, y_i]}{Var\,[x_i]} = \frac{\beta}{1+\frac{\sigma_\eta^2}{\sigma_{x^*}^2}} \tag{20}$$

$X, Y, \varepsilon \in \mathbb{R}^n$; $\beta, \beta^*$are the regression parameters; ε_j are the error terms.

III. 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. [1]. A few techniques will be discussed here.

A. Linear Least Squares

If the errors have finite variance and are homoscedastic; and the errors ϵ_i are uncorrelated with regressors x_i as in (21),

$$\mathrm{E}\,[x_i\, \epsilon_i\,] = 0 \tag{21}$$

the linear least squares given by:

$$\varepsilon^2 = \textstyle\sum_{i=1}^{n} (y_i - \beta\, x_i - \alpha)^2 \tag{22}$$

must be minimized [10].

The Moore-Penrose Pseudoinverse can be used to calculate the regression parameters [10]:

$$\beta = (X^T X)^{-1} X^T Y \tag{23}$$

$X, Y, \varepsilon \in \mathbb{R}^n$; $\beta_j$are the regression parameters; ε_j are the error terms.

B. Generalized Least Squares

If errors are correlated or heteroscedastic, the regression coefficients [11] can be calculated as:

$$\beta = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y \tag{24}$$

where Ω is the covariance matrix of the errors; $X, Y, \varepsilon \in \mathbb{R}^n$; $\beta_j$are the regression parameters.

C. Method of Instrumental Variables

If regressors are correlated with the errors ϵ, instrumental variables z_i can be used to calculate the regression coefficients [14] given the condition in (25) is fulfilled.

$$\beta = (X^T Z (Z^T Z)^{-1} Z^T X)^{-1} X^T Z (Z^T Z)^{-1} Z^T Y \tag{25}$$

where $\mathrm{E}\,[z_i\, \epsilon_i\,] = 0$ and Z is vector or estimators; $X, Y, Z \in \mathbb{R}^n$; $\beta_j$are the regression parameters.

D. Partial Differentiation with respect to regression coefficients

The minimization of (26):

$$\varepsilon^2 = \textstyle\sum_{i=1}^{n} (y_i - f(\beta, x_i))^2 \tag{26}$$

can be achieved by Partial Differentiation of (26) with respect to regression coefficients β_j [6]. The resultant equations are set equal to zero. This results in simultaneous equations which can be solved to calculate the regression coefficients. Such problems are dealt with in the ambit of convex optimization [15].

E. Maximum Likelihood Estimation

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

$$L(\beta | x) = p_\beta(x) \tag{27}$$

$X, Y, \varepsilon \in \mathbb{R}^n$; $\beta_j$are the regression parameters.

is maximized [3]. The regression parameters are estimated using observations and their assumed uniform distributions (uniform, exponential, gamma, Gaussian, Poisson, Bernoulli, binomial etc.) [13].

F. Bayes Estimator

If the errors ε are independent and identically normally distributed $\mathrm{N}(0, \sigma^2)$, and a prior distribution is assumed, explicit results are available for the posterior probability distributions of the model parameters [12].

If the likelihood function is

$$\rho(Y | X, \beta, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp(\frac{-1}{2\sigma^2}\ (Y - X\beta)^T (Y - X\beta)) \tag{28}$$

and the assumed prior distribution is

$$\rho(\beta, \sigma^2) \propto \rho(\sigma^2)\, \rho(\beta | \sigma^2) \tag{29}$$

the posterior probability density distribution (30) is obtained by multiplying the likelihood (28) with the prior probability density distribution (29):

$$\rho(\beta, \sigma^2 \mid Y, X) \propto \rho(\beta | \sigma^2, Y, X)\, \rho(\sigma^2 | Y, X) \tag{30}$$

where $X, Y, \varepsilon \in \mathbb{R}^n$; $\beta_j$are the regression parameters.

IV. 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 (thermal decomposition, partial discharge and arcing); 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 [16]. 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 a rich data series from 161kV transformers for power transformer lifespan evaluation [16].

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 [6].

The Combustible gases and furfural concentrations were used for modeling Probability of transformer failure which could be used to gauge transformer health and remaining life of transformer [16].

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 [17] is

$$f(x;\eta,\beta,\gamma) = \frac{\beta}{\eta}(\frac{x}{\eta})^{\beta-1}e^{-(\frac{x-\gamma}{\eta})^{\beta}}, x \geq 0 \qquad (31)$$

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

The failure probability was calculated using the logistic model [6]:

$$p = \frac{e^{g(x)}}{1+e^{g(x)}} \qquad (32)$$

$$where\ g(x) = \sum_{i=0}^{k} x_i \beta_i$$

$X \in \mathbb{R}^n$; $\beta_j$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).

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) [16]. Hence transformer enters abnormal state after 11.67 years of service with great likelihood of failure [16]. The Weibull distribution was used to generate the linear regression function:

$$g(x) = -4.497 - 0.016[C_2H_4] + 0.685[C_2H_2] - 0.008[CO] + 0.01[TCG] + 0.244[2 - ACF] - 0.564[5 - MEF] \qquad (33)$$

where x=C_2H_4, C_2H_2, CO, TCG,2-ACF,5-MEF are the condition variables.

Hence the Logistic model for failure probability of one of the transformers was developed [16]:

$$p = \frac{1}{1+309.246\ (0.529^{t})} \qquad (34)$$

where t is time of service in years.

V. Conclusion

This research carried out an in-depth study of various interesting regression models with special emphasis on Logistic regression. Some unique techniques for estimation of regression coefficients were presented to deal with errors having finite variance; and correlation of errors with each other or with the regressors. By assuming uniform distributions for the observations, we can minimize sum of squared errors or maximize likelihood function to obtain the optimal regression parameters. The modeling of transformer lifespan using real time data monitoring of dissolved oil gases provided an interesting example for logistic regression modeling. It used a Weibull distribution to estimate the mean failure time; using a huge data set of 161kV transformers. The probability of failure was calculated using a logistic function based on the concentrations of dissolved gases and by cross-verification with unhealthy and out-of-service transformers.

References

[1] "Regression analysis", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Regression_analysis. [Accessed: 26- Oct- 2018].

[2] "Newton's law of universal gravitation", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Newton%27s_law_of_universal_gravitation. [Accessed: 26- Oct- 2018].

[3] X. Yan and X. Su, *Linear regression analysis*. Singapore: World Scientific Pub. Co., 2009, pp. 1-20.

[4] P. Fritzson, *Introduction to Modeling and Simulation of Technical and Physical Systems with Modelica*. IEEE Press, 2018, pp. 15-16.

[5] M. Pohar et al., *Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study*, 1st ed. 2018, pp. 1-20.

[6] "Logistic Regression-Detailed Overview", *towardsdatascience.com*, 2018. [Online]. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc [Accessed: 26- Oct- 2018].

[7] *Poisson regression*, NCSS Statistical Software 2018, pp. 1-6.

[8] R. Koenker, *Quantile regression*, International Encyclopedia of the Social Sciences 2000, pp. 1-11.

[9] "Errors-in-variables models", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Errors-in-variables_models. [Accessed: 26- Oct- 2018].

[10] "Least squares Fitting", *mathworld.wolfram.com*, 2018. [Online]. Available: http://mathworld.wolfram.com/LeastSquaresFitting.html [Accessed: 26- Oct- 2018].

[11] "Generalized least squares", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Generalized_least_squares. [Accessed: 26- Oct- 2018].

[12] "Bayesian Estimation", *nature.com*, 2018. [Online]. Available: https://www.nature.com/scitable/content/bayesian-estimation-46434. [Accessed: 26- Oct- 2018].

[13] J. Watkins, *Maximum likelihood estimation*, 2011. [Online]. Available: ,pp. 1-16.

[14] "Instrumental variables estimation", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Instrumental_variables_estimation. [Accessed: 26- Oct- 2018].

[15] "Convex Optimization", *convexoptimization.com*, 2018. [Online]. Available: http://www.convexoptimization.com/dattorro/convex_optimization.html. [Accessed: 26- Oct- 2018].

[16] C. Trappey et al., *Data Driven Modeling for Power Transformer Lifespan Evaluation*, Systems, Engineering Society of China and Springer-Verlag Berlin Heidelberg 2014, pp. 1-14.

[17] "Characteristics of the Weibull Distribution", *weibull.com*, 2018. [Online]. Available: https://www.weibull.com/hotwire/issue14/relbasics14.htm. [Accessed: 26- Oct- 2018].