A General Study on Regression Analysis (RA)

[1]GEENA.M. G, [2]Dr K. SUGUNA, [3]Dr P.N. RAGUNATH

[1]Research Scholar, Department of Civil &Structural Engineering, Annamalai University Annamalainagar Tamilnadu, India.

[2]Professor, Department of Civil &Structural Engineering, Annamalai University Annamalainagar, Tamilnadu, India.

[3]Professor & Head Department of Civil & Structural Engineering, Annamalai University Annamalainagar, Tamilnadu, India

Abstract
Regression analysis constitutes a collection of statistical methodologies employed for the estimation of relationships between a dependent variable and one or several independent variables. The Regression Analysis (RA) generates a series of regression equations wherein the coefficients signify the relationship between each independent variable and the dependent variable. Regression analysis mainly consist of 2 important function known as the primary and the secondary function. In the primary function first RA is used for predicting and forecasting the places where its uses overlap with that of the machine learning (ML) field. In the secondary function it tries to infer in the association between the dependent variable and the independent variable is examined. This research paper endeavors to elucidate the intricacies of regression analysis comprehensively. The unknown coefficients are determined utilizing the data obtained from experiments or alternative sources, employing Legendres principle of least squares errors. In this document, regression equations have been employed to forecast the ultimate load and ultimate deflection values. Then later the predicted value was compared with the experimental value and the result of it was displayed in further sections.

Keywords: Regression Analysis (RA), Regression Co-efficient, Yield load, Ultimate load.

1. INTRODUCTION

Regression analysis stands out as an incredibly powerful statistical tool that facilitates the exploration and examination of intricate relationships that exist between various variables, allowing researchers and analysts to delve deeply into their interconnections. When the analysis focuses solely on one explanatory variable, this specific approach is distinguished by the term simple regression, which serves as a fundamental building block in the world of statistical analysis. On the other hand, when one employs multiple regression techniques, [1] it opens the door to the inclusion of additional factors, thereby allowing these variables to be analyzed independently yet simultaneously, enriching the overall understanding of the data at hand. This method proves to be immensely valuable as it quantifies the influence of various factors operating concurrently on a single dependent variable, revealing the complex web of interactions that may be at play. Essentially, [2]regression analysis acts as a meticulous procedure for establishing a relationship between known input variables and an output parameter, all underpinned by robust statistical principles that guide the analysis.

The overarching technique in regression involves making certain assumptions about the nature of the relationship that exists between the input parameters and the results, [3] which is often expressed through a mathematical form that includes a series of unknown coefficients representing the strength of

these connections. To uncover these unknown coefficients, which are critical to the accuracy of the model, analysts rely on data obtained from experiments or other reliable sources, [4] employing the sophisticated Legendre's principle of least squared errors as a guiding framework for their calculations. Legendre's principle of least squared errors serves as a versatile and general-purpose curve-fitting technique, [5] which aids in selecting the optimal values for these unknown coefficients, often referred to as regression coefficients, in a manner that maximizes the agreement between the predicted outcomes and the actual target results to the greatest extent possible.

As we delve deeper into the fascinating realm of regression analysis, [6] we encounter a plethora of terms and concepts that are intricately woven into this multifaceted discipline, each deserving of attention and explanation in the following subsections. In fact, [7] regression analysis can be categorized into several distinct types, including but not limited to Linear Regression, which explores linear relationships, Logistic Regression, which is designed for binary outcomes, [8] Ridge Regression, which addresses multicollinearity, Lasso Regression, which performs both variable selection and regularization, Polynomial Regression, [9] which captures non-linear relationships, and [10] Bayesian Regression, which incorporates prior knowledge into the analysis. Each of these categories offers unique insights and methodologies, [11] illustrating the rich diversity and adaptability of regression analysis as a tool for understanding complex data relationships [12].

1.1 Regression

The artful approach employed for molding curves, be they linear or non-linear in their designated form. The aim of regression is to ascertain the elusive coefficients within an equation. The structure of the equation is presumed beforehand in a manner that ideally aligns with the expected connection between the input and the output (Figure 1).

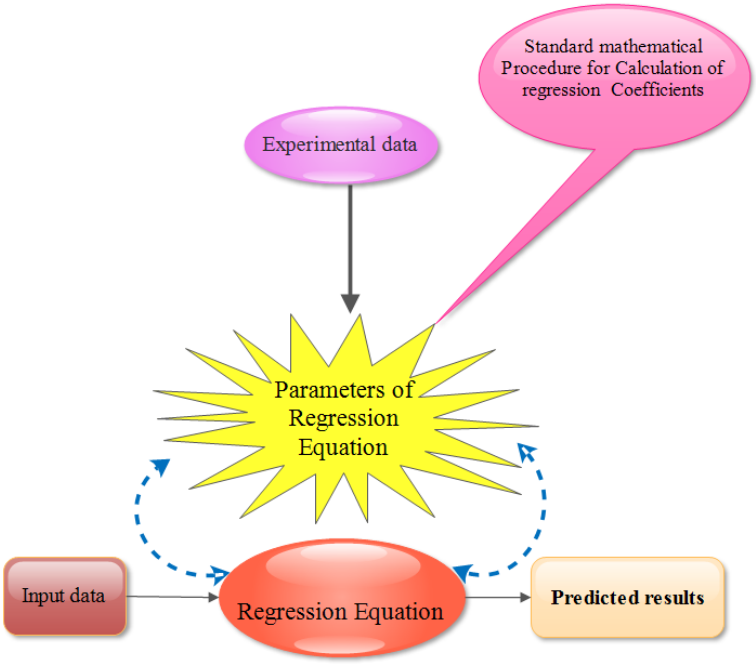

Figure 1 Regression flow process

1.2 Regression Coefficient

The regression coefficient serves as a mysterious element woven into the equation, designed to adjust the input variable or a blend of input variables. By tackling the regression challenge through the lens of minimizing squared errors, every regression coefficient is thoroughly assessed.

1.3 Legendre's Principle of Least Squared Errors

Legendre's principle of minimal squared discrepancies seeks to address the regression challenge by enforcing the condition that the square of the disparity between the actual outcome and the predicted value from the equation must be minimized. This is achieved by calculating the derivative of the square of the error concerning each unknown coefficient in the proposed equation. Each derivative yields a distinct equation, and the total number of equations generated will match the total number of unknown regression coefficients that need to be determined.

1.4 Karl Pearson's Coefficient of Correlation

The correlation coefficient established by Karl Pearson is a numerical value ranging from 0 to 1, which gauges the intensity of the connection between the input variables and their corresponding outcomes. A value near unity for Karl Pearson's correlation coefficient indicates a robust relationship between the inputs and the resultant values. This correlation coefficient remains uninfluenced by the forecasts derived from regression equations; it solely reflects the characteristics of the provided set of inputs and outputs.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (x-\bar{y})^2}} \tag{1}$$

1.5 Sum of Squared Errors (SSE)

The total of squared discrepancies is the aggregation of the squares of the variance between the values anticipated by the regression formula (or another method) and the true outcomes anticipated for the specified input values. A greater SSE signifies a more significant divergence of the predicted values from the anticipated results.

$$SSE = \sum_{l}^{N} (x-\hat{x})^2 \tag{2}$$

1.6 Mean Squared Error (MSE)

Mean squared error is calculated by taking the total of squared discrepancies and dividing it by the count of the summed values. The MSE serves as a superior gauge of error compared to SSE, as it represents the squared error for each individual data point.

$$MSE = \frac{\sum_{1}^{N} (x - \hat{x})^2}{N} \tag{3}$$

1.7 Root Mean Squared Error (RMSE)

The root mean squared error is derived as the square root of the Mean Squared Error. This metric highlights the degree of divergence from the anticipated value, either above or below it. Thus, RMSE serves as a more effective gauge of error in contrast to MSE.

$$RMSE = \frac{\sqrt{\sum_{1}^{N} (x - \hat{x})^2}}{N}$$

1.8 Root Mean Squared Percentage Error (RMSPE)

The root mean squared percentage error (RMSPE) is calculated by taking the square root of the sum of squared percentage discrepancies, divided by the total number of errors considered, and then multiplying by one hundred. The RMSPE can be interpreted without regard to the actual numerical values of the data, as it serves as a normalized metric. In contrast to other error metrics, an RMSE (or MSE or SSE) value of 10 from a dataset with a mean of 15 could indicate a more unfavorable outcome than the same value from a dataset with a mean of 1500. However, since RMSPE is normalized, lower values signify a more accurate fit, while larger values indicate less accuracy (Carpenter and Barthelemy, 1994).

$$RMSPE = \sqrt{\frac{\sum_{1}^{N} \left(\frac{x - \hat{x}}{\bar{x}} \right)^2}{N}} \times 100 \tag{4}$$

2. MULTIVARIATE LINEAR REGRESSION

Multivariate linear regression aids in formulating first degree equations that encompass multiple independent variables. The fundamental structure for multivariate linear regression is,

$$\begin{Bmatrix} \frac{\partial}{\partial a_0} \\ \frac{\partial}{\partial a_1} \\ \frac{\partial}{\partial a_2} \\ \frac{\partial}{\partial a_3} \\ \vdots \\ \frac{\partial}{\partial a_n} \end{Bmatrix} \sum_{i=1}^{K} (P_i - (a_0 + a_1 x_{1i} + a_2 x_{2i} a_3 x_{3i} + \ldots + a_n x_{ni})) = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{Bmatrix} \tag{5}$$

In this context, a0…an represent the coefficients waiting to be uncovered, while x1...xn denote the independent variables. P stands for the dependent variable, reflecting the real outcome for the ith set of input data, and K signifies the number of data sets accessible for regression analysis. By applying the partial derivative operators, equation 5.5 simplifies to,

$$\sum_{i=1}^{N}\begin{bmatrix} 1 & x_{1i} & x_{2i} & x_{3i} & \dots & x_{ni} \\ x_{1i} & x_{1i}^{3} & x_{1i}x_{3i} & x_{1i}x_{3i} & \dots & x_{1i}x_{ni} \\ x_{2i} & x_{2i}x_{1i} & x_{2i}^{2} & x_{2i}x_{3i} & \dots & x_{2i}x_{ni} \\ x_{3i} & x_{3i}x_{1i} & x_{3i}x_{2i} & x_{3i}^{2} & \dots & x_{3u}x_{ni} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ x_{ni} & x_{ni}x_{1i} & x_{ni}x_{2i} & x_{ni}x_{3i} & \dots & x_{ni}x_{ni} \end{bmatrix}\begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{Bmatrix} = \sum_{i=1}^{K}\begin{Bmatrix} P_i \\ P_1P_i \\ P_2P_i \\ \vdots \\ P_nP_i \end{Bmatrix} \quad (6)$$

The equation presented can be unraveled by aggregating the figures of both independent and dependent variables following the necessary procedures.

3. REGRESSION EQUATION FOR STRENGTH

The meticulously organized and comprehensive data that has been carefully compiled and utilized for the intricate and detailed regression analysis, which serves as the backbone of our research findings, is conveniently and clearly laid out for the reader's convenience in Table 1, while the complex and multifaceted regression equations that encapsulate the relationships within the data are thoughtfully and systematically presented in both Table 2 and Table 3, providing a thorough understanding of the analytical process.

Table 1 Data Used for the Regression Analysis

Beam Designation	First Crack load (kN)	Def. @ FCL (mm)	Yield Load (kN)	Def. @ Yeld Load (mm)	Ultimate Load (kN)	Def. @ UL (mm)	Width of Crack (mm)	No. of Cracks	Average Spacing of Cracks (mm)	Spacing of Stirrups	Tensile strength for FRP	E for FRP	Deflection ductility	Energy ductility
200 NS 0	10	0.95	25	2.6	50	9.24	0.4	8	140	100	0	0	3.55	6.74
200 CS 0	12.5	1.05	27.5	2.85	57.5	10.82	0.44	11	128	100	0	0	3.79	9.13
100 NS 0	15	1.12	30	3.2	60	12.1	0.5	13	124	200	0	0	3.78	6.37
100 CS 0	20	1.16	35.5	3.65	65	14.54	0.58	16	116	200	0	0	3.98	7.88

200 NS 3	22.5	1.28	42.5	3.98	90	16.6	0.64	18	108	100	446.9	13965	4.17	8.61
200 CS 3	25	1.34	50	4.18	100.5	18.76	0.72	21	96	100	446.9	13965	4.48	8.59
100 NS 3	27.5	1.48	54.5	4.33	110	20.34	0.8	23	94	200	446.9	13965	4.69	9.89
100 CS 3	30	1.64	60	4.62	120.5	22.5	0.88	25	90	200	446.9	13965	4.87	9.98
200 NS 5	30	1.86	64.5	5.16	130	24.8	0.98	27	88	100	451.5	17365	4.8	10.4
200 CS 5	32.5	2.1	68	5.82	135	26.2	1.2	28	84	100	451.5	17365	4.5	10.05
100 NS 5	32.5	2.46	70	6.28	142.5	28.1	1.34	30	76	200	451.5	17365	4.48	10.32
100 CS 5	35	3.1	72.5	6.85	145	30.4	1.52	33	68	200	451.5	17365	4.43	10.48

Table 2 Data Used for the Regression Analysis for Experimental vs Predictions

Specimen	**Yield load(kN)**		**deflection at YL(mm)**		**Ultimate load(KN)**		**deflection at UL(mm)**		**Energy ductility**		**Deflection ductility**		**Crack width(mm)**	
	Expt	**Pred**	**Expt**	**Pred**	**Expt**	**Pred**	**Expt**	**Pred**	**Expt**	**Pred**	**Expt**	**Pred**	**Expt**	**Pred**
200 NS 0	25	26.125	2.6	2.77	50	50	9.24	10.1975	6.74	7.75	3.55	3.77	0.4	0.365
200 CS 0	27.5	26.125	2.85	2.77	57.5	52.1875	10.82	10.1975	9.13	7.75	3.79	3.77	0.44	0.365
100 NS 0	30	33.4375	3.2	3.465	60	65.15625	12.1	13.63125	6.37	7.815	3.78	3.89	0.5	0.5775
100 CS 0	35.5	33.4375	3.65	3.465	65	65.15625	14.54	13.63125	7.88	7.815	3.98	3.89	0.58	0.5775
200 NS 3	42.5	48.09375	3.98	3.93	90	98.765625	16.6	17.833125	8.61	9.235	4.17	4.4925	0.64	0.65375
200 CS 3	50	48.09375	4.18	3.93	100.5	98.765625	18.76	17.833125	8.59	9.235	4.48	4.4925	0.72	0.65375

100 NS 3	54.5	55.40625	4.33	4.625	110	111.734375	20.34	21.266875	9.89	9.3	4.69	4.6125	0.8	0.86625
100 CS 3	60	55.40625	4.62	4.625	120.5	111.734375	22.5	21.266875	9.98	9.3	4.87	4.6125	0.88	0.86625
200 NS 5	64.5	65.09375	5.16	5.68	130	131.640625	24.8	25.658125	10.4	10.28	4.8	4.4925	0.98	1.15375
200 CS 5	68	65.09375	5.82	5.68	135	131.640625	26.2	25.658125	10.05	10.28	4.5	4.4925	1.2	1.15375
100 NS 5	70	72.40625	6.28	6.375	142.5	144.609375	28.1	29.091875	10.32	10.345	4.48	4.6125	1.34	1.36625
100 CS 5	72.5	72.40625	6.85	6.375	145	144.609375	30.4	29.091875	10.48	10.345	4.43	4.6125	1.52	1.36625

Table 3 Regression Equations

Sl. No	Parameter	Regression	Fitness
1	First Crack Load	0.046SS -0.032fFRP + 0.002E_{FRP}+9.06	0.9769
2	Deflection at First Crack Load	0.004SS -0.008fFRP + 0E_{FRP}+0.475	0.8912
3	Yield Load	0.075SS -0.111 fFRP + 0.005E_{FRP}+18.25	0.9731
4	Deflection at Yield Load	0.007SS -0.014 fFRP + 0.001E_{FRP}+1.99	0.9597
5	Ultimate Load	0.133SS -0.205 fFRP + 0.01E_{FRP}+38.12	0.9597
6	Deflection at Ultimate Load	0.036SS - 0.057fFRP + 0.002E_{FRP}+6.28	0.9768
7	Width of Crack	0.002SS -0.004 fFRP + 0E_{FRP}+0.17	0.9462
8	No. of Cracks	0.045SS -0.052 fFRP +0.002 E_{FRP}+5.25	0.9699
9	Average Spacing of Cracks	(-)0.127SS +0.103 fFRP -0.005 E_{FRP}+146	0.9567
10	Deflection Ductility	2.3E-06+0.001688+0.0012+3.65	0.81
11	Energy Ductility	0.000316+ (-)0.00656+0.00065+7.685	0.87

Note: ***Efrp - Elasticity Modulus of FRP, ffu – Tensile Strength of FRP and Tk – Thickness of FRP***

4. OBSERVATIONS ON THE REGRESSION EQUATIONS

The sophisticated regression equations were meticulously employed for the purpose of forecasting both the ultimate load and the ultimate deflection values, which are critical parameters in structural engineering and material science. A thorough examination of the various measures of fitness associated with the regression indicates that the multivariate linear regression technique possesses the capability to accurately estimate prediction values for an array of factors, including but not limited to yield load, yield deflection, ultimate load, ultimate deflection, deflection ductility, energy ductility, deflection ductility ratio, energy ductility ratio, the total number of cracks present, the maximum crack width observed, and the overall energy absorption capacity of GFRP (Glass Fiber Reinforced Polymer) strengthened reinforced concrete beams. The root mean square error values, which serve as indicators of the model's predictive accuracy, exhibited a range of variability from a minimum of 0.17 to a maximum of 13.76, showcasing the nuanced performance of the regression models across different scenarios.

However, it is important to note that linear regressions, by their very nature, are inherently constrained in their capacity to accurately model exceedingly comprehensive sets of data, particularly because the first order regression parameters endeavor to align themselves with a monotonically varying linear relationship that lacks the necessary curvature to fully encapsulate the complexities of the prediction parameter being analyzed.

In a comparative analysis, the predictions generated from the regression equations were meticulously juxtaposed against empirical experimental values, and the outcomes of this comparison were visually represented in Figures 2 to 8, providing a clear illustration of the efficacy of the regression models in real-world applications.

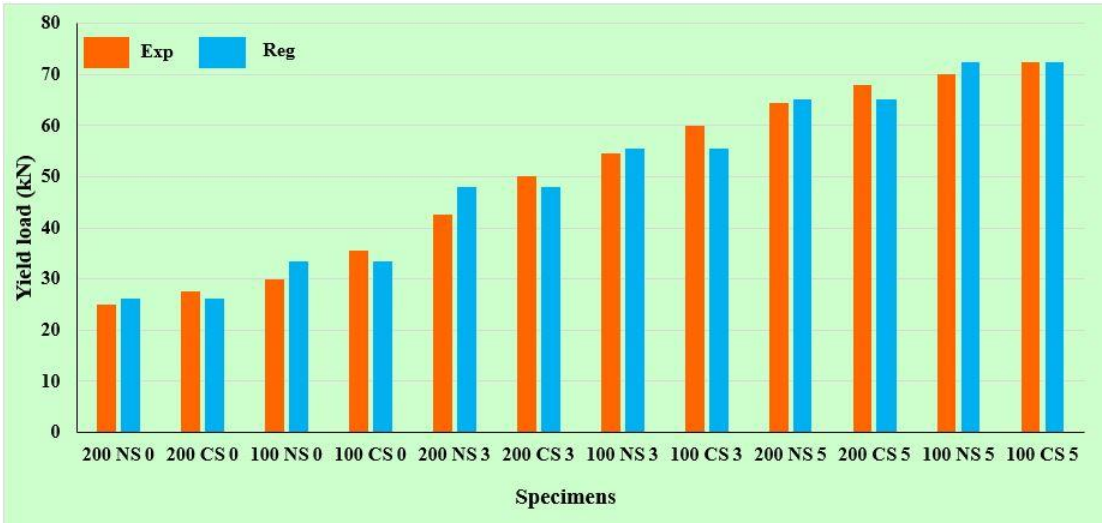

Figure 2 Specimens Predictions for Yield Load (KN)

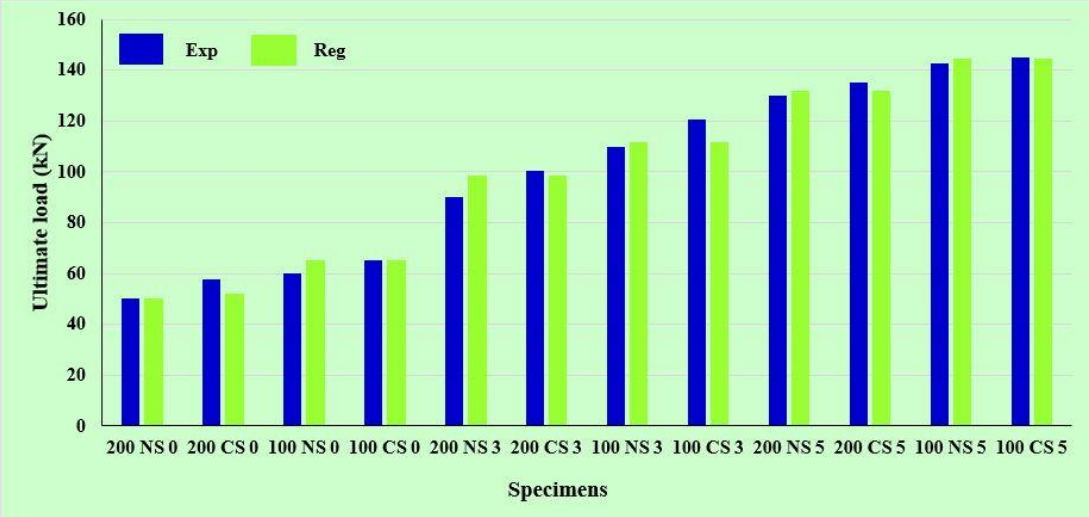

Figure 3 Specimens Predictions for Ultimate Load (KN)

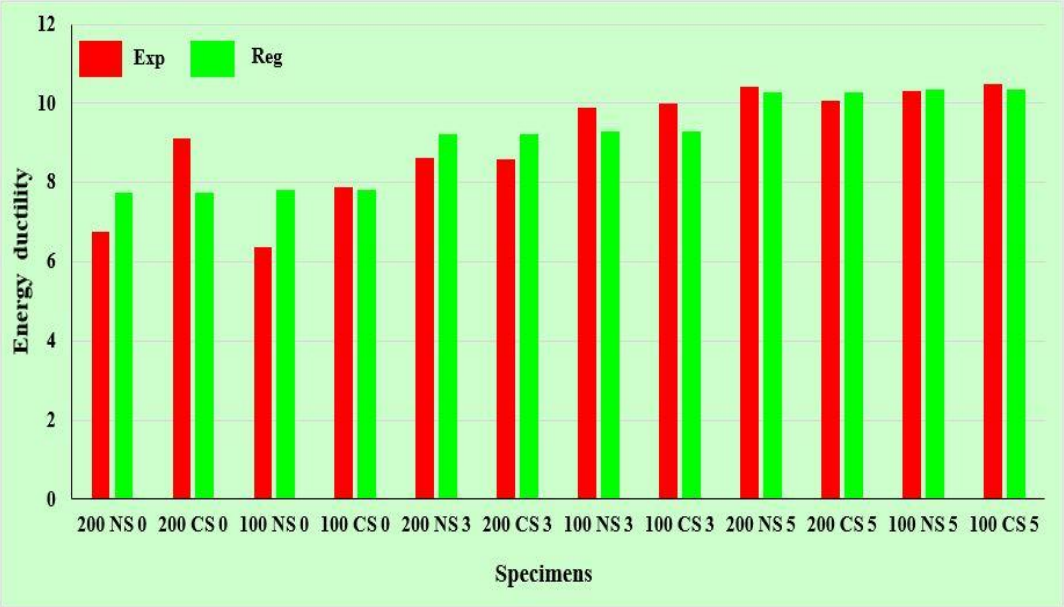

Figure 4 Specimens Predictions for Energy Ductility

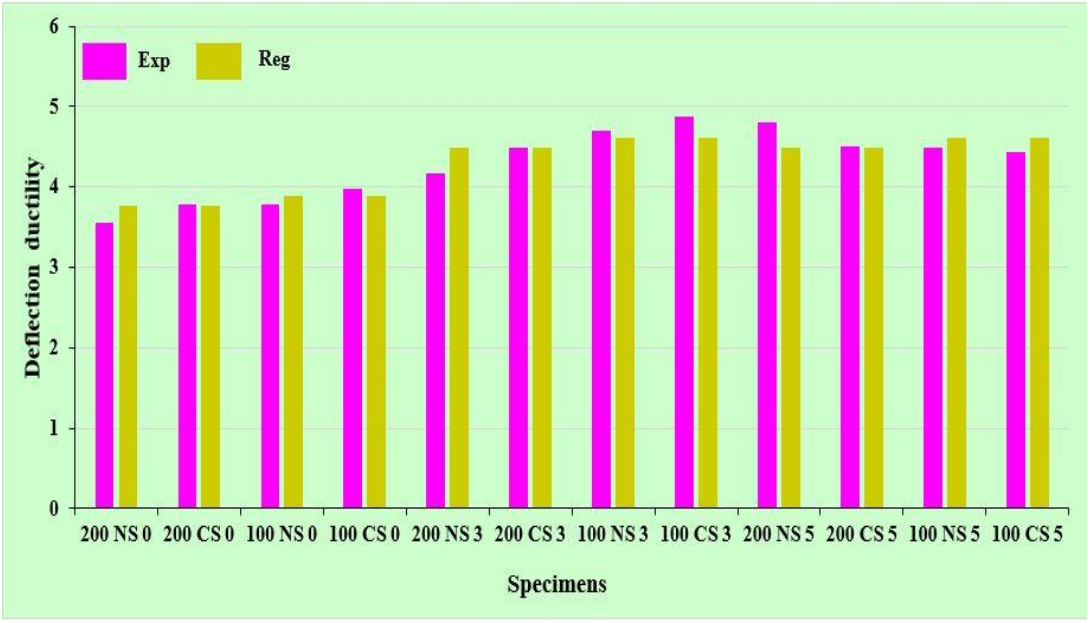

Figure 5 Specimens Predictions for Deflection Ductility

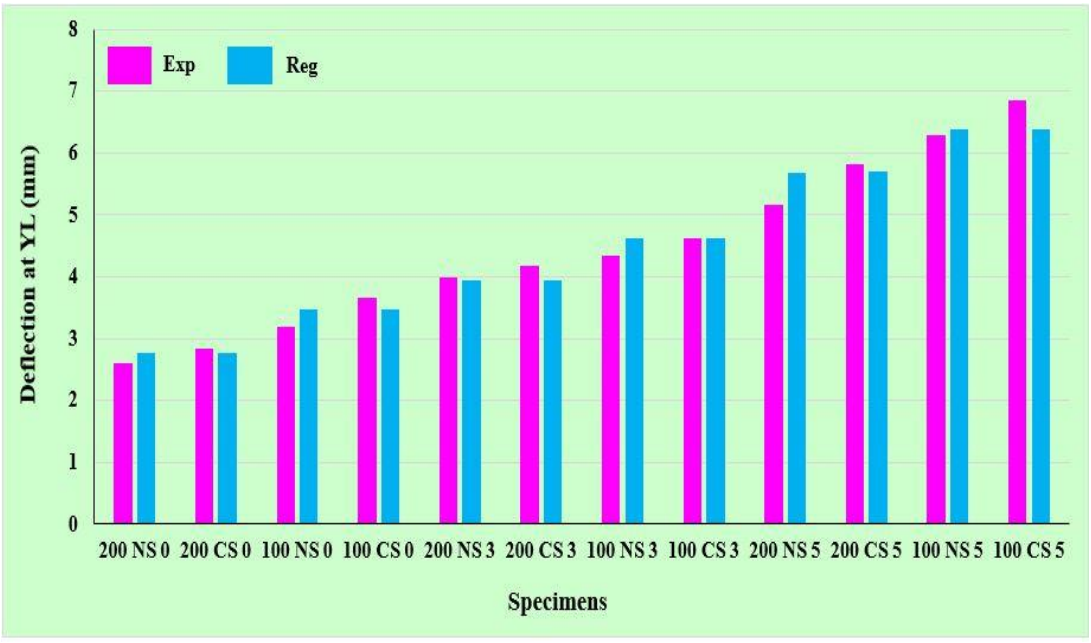

Figure 6 Specimens Predictions for Deflection at YL(mm)

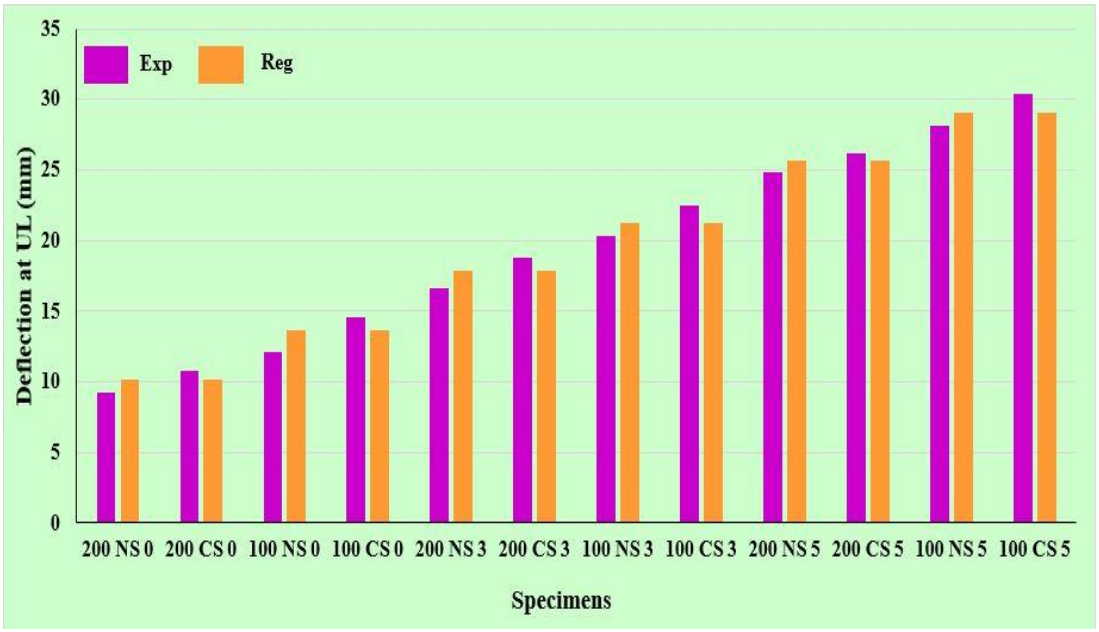

Figure 7 Specimens Predictions for Deflection at UL(mm)

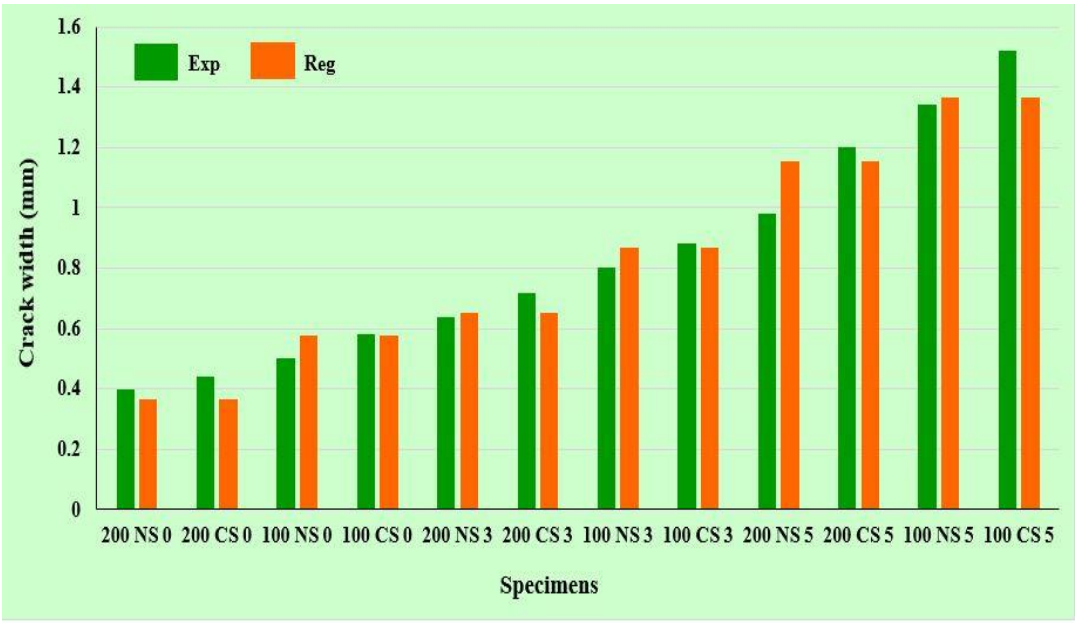

Figure 8 Specimens Predictions for Crack width (mm)

Conclusion

After conducting the Regression analysis, the fitness has been obtained as 0.801. The above observation clearly indicates the validity of the proposed regression equation for the purpose of estimating the performance parameters under both static and cyclic loading conditions. This research tried to explain the general analysis of Regression Analysis and various regression equation have been observed to know the performance of them. From the observation it was found it is posited that the multivariate linear regression possesses the capability to estimate predictive values with an acceptable

degree of precision regarding yield load, yield deflection, ultimate load, ultimate deflection, deflection ductility, and so forth.

REFERENCE

1. Rothman KJ, Greenland S. Modern Epidemiology. 2nd ed. Lippincot-Raven; 1998.
2. Altman DG. Practical Statistics for Medical Research. Chapman & Hall/CRC; 1991.
3. Rosner BA. Fundamentals of Biostatistics. 4th ed. Duxbury; 1995.
4. Draper NR, Smith H. Applied Regression Analysis. Wiley Series in Probability and Statistics; 1998. Applied Regression Analysis.
5. Munro BH. Statistical Methods for Health Care Research. 5th ed. Lippincott Williams & Wilkins; 2005.
6. Alaimo, K., Olson, C. M., and Frongillo, E. A., Jr. (2001). Food insufficiency and American school-aged children's cognitive, academic, and psychosocial development. *Pediatrics, 108*, 44-53.
7. Arum, R. (1998). The effects of resources on vocational student educational outcomes: Invested dollars or diverted dreams? *Sociology of Education, 71*, 130-151.
8. Astone, N.M., and McLanahan, S.S. (1991). Family structure and high school completion. *American Sociological Review, 56*, 309-320.
9. Blake, J. (1989). *Family size and achievement*. University of California Press.
10. Bradley, R. H., and Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*, 371-399.
11. Van Cao, Vui, and Son Quang Pham. "Comparison of CFRP and GFRP wraps on reducing seismic damage of deficient reinforced concrete structures." International Journal of Civil Engineering 17.11 (2019): 1667-1681.
12. Chiew, Sing-Ping, Qin Sun, and Yi Yu. "Flexural strength of RC beams with GFRP laminates." Journal of composites for Construction 11.5 (2007): 497-506.