Journal of Computing, Science &Technology
https://focjournal.unidel.edu.ng/
editorjcst@unidel.edu.ng

ISSN: 3043-6818 Print https://focjournal.unidel.edu.ng

A Systematic Review of Non-Linear Models: Shifting from Parametric to Non-Parametric Models

Agu S. C[1], Obasi E. C. M[2]
[1]Department of Computer Science, Benson Idahosa University, Benin City, Nigeria, [2]Department of Computer Science and Informatics., Federal University Otuoke, Otuoke, Nigeria.
sndyaguu@gmail.com, obasiec@fuotuoke.edu.ng
Corresponding Author's Email sndyaguu@gmail.com

Article Info

Date Received: 16-09-2024
Date Accepted: 02-11-2024

Keywords:

Generalized Additive Models, Polynomial regression, Step function regression, Linear models, Non-linear models, Parametric Models

ABSTRACT

Linear models have been applied in various applications to determine the relationship between the response and the feature variables as well as to facilitate prediction. However, the explanatory and residual conditions assumed by linear model pose huge performance challenges in the face of datasets that do not met the conditions. This study systematical reviewed how the linear models could be shifted to non-linear models using three different methods: Polynomial regression, Step functions, and Generalized Additive Models (GAMs). To build higher performance predictive models, the research uncovered that non-linear model of 3 to 4 degrees of polynomial regression methods are expected to perform better than linear model. And that the coefficient values in linear models are no longer necessary in non-linear models. It further revealed that the step functions improved the global structure of polynomial regressions by breaking the range of the predictor variable X into bins, and converting the continuous variable X into an ordered categorical variable, making it very popular in biostatistics and epidemiology studies. Finally, we found that Generalized Additive Models introduced a complete shift from parametric model to non-parametric model, replacing each linear component with a nonlinear function of the variable x, while maintaining additivity.

1.0 INTRODUCTION

Linear models typically make predictions or draw inferences about a population by finding the relationships between the dependent and independent variables using the Ordinary Least Squares (OLS) method which estimates the best regression coefficients by minimizing the sum of the squares of the errors [1]. Under certain conditions, OLS estimators are the best linear unbiased estimators. Regrettably, these expectations are breached in most situations [2] such as assumption of linear relationship between the predictor and response variables, multivariate normality, uncorrelated error terms, constant variance in error term, non-collinearity among the predictors, no consideration of outliers and high leverage points [3] [4], [5], and these assumption have the effect of reducing the predictive accuracy of the models if they are not met [2].

When the assumptions are not met in a dataset, a non-linear transformation is sought to fix the problem [4].Some non-linear methods such as polynomial regression, piecewise polynomial, cubic spline, natural spline regressions, smoothing spline attempt to extend the linear models to non-linear models but each has its limitations [6]. The objective of this study is to review (a) the extension of linear models to non-linear models with three methods namely polynomial regression, piecewise step function, and Generalized Additive Models (GAMs) and (b) the movement from parametric to non-parametric models.

2.0 Review of Related Literature

In the age of big data, the efficient analysis of vast datasets is paramount, yet hindered by computational limitations such as memory constraints and processing duration. To tackle these obstacles, GAMs approach harnesses the versatile modeling capabilities while alleviating computational burdens and enhancing the precision of parameter estimation [7], providing unified framework for estimation and analysis of high dimensions models [8]. Widely recognized for their ability to create fully interpretable machine learning models for tabular data, training GAMs involves iterative learning algorithms, such as splines, boosted trees, or neural networks, which refine the additive components through repeated error reduction [9]. Incorporating missingness indicators and their interaction terms while maintaining sparsity through l0 regularization, M-GAM, a sparse, generalized, additive modeling approach resolved the problem of maintaining the interpretability of machine learning models in the presence of missing values in a dataset [10].

Some research findings in the past two decades were carried

out using non-linear models of polynomial/step-function regressions across business/scientific fields to curb the linearity assumption issues associated with the linear models in order to improve the models' accuracy: In formulating sustainable biodegradation method for crude oil utilizing native and recombinant microbial strains to lessen the environmental hazard caused by crude oil contamination, polynomial regression with R^2 of 0.863 outperformed multiple linear regression with R^2 of 0.495 [11]. The outcome of polynomial regression was very sensitive to predicting climate change [12]. And the COVID-19 cases were predicted using polynomial based linear regression model while the future cases of COVID-19 were sought to create a predictive model using polynomial regression [13-14]. Seen as a strong substitute for prediction, the ordinary least square method of cubic polynomial regression model was preferred to the linear regression model counterpart [15]. Applying polynomial regression model revealed that the relationship between strains and hole-drilling depth was curvilinear [16]. The study by [17] discovered that hypothetical step function was utilized to predict learner progress. Approximating step function yielded a closed- form solution for optimization matching pursuit algorithm [18].

3.0 EXAMINING EXTENSIONS OF LINEAR MODELS

[2] improved upon OLS using ridge regression, the lasso, and principal components regression, which was obtained by reducing the complexity of the linear model, and hence the variance of the estimates. However linear models were used with the methods which could be enhanced. This study shows how the linearity assumption could be relaxed while still attempting to maintain as much interpretability as possible by examining polynomial regression, step functions, and generalized additive models (GAMs). Polynomial regression and step functions approaches model the relationship between a response Y and a single predictor X in a flexible way while GAMs approach shows that any of the extensions of linear models such as polynomial regression, step functions, splines, or local regression could be seamlessly integrated in order to model a response Y as a function of several predictors $X_1, \ldots, X_p$ [6].

3.1 The Polynomial Regression

The extension of the linear methods by the polynomial regression is stemmed on the addition of extra features and raising each of the features to a power [19] such that the simple linear function $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ is replaced with a polynomial function

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_p \beta_i^d + \epsilon_i \quad (1)$$

According to [6], the interested is no longer in the estimated coefficients $\beta_0, \beta_1, \ldots \beta_d$ Instead, the complete modeled function is considered over the entire instances of the dataset and the recommended degree of the polynomial regression d not exceeding four. Figure 1 shows the result of modeling a polynomial to a 4th degree

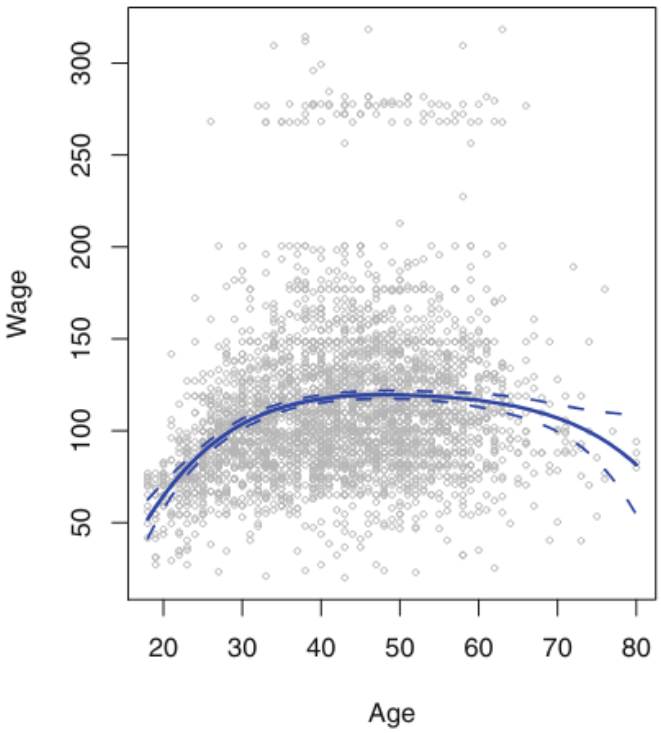

Figure 1: The solid blue curve is the fit of a degree 4 polynomial using least squares method

3.2 Step Functions

A polynomial method has the limitation of enforcing a universal formation on the non-linear function of X. The step functions curb the drawback by breaking the span of X into bins [20] with the cut-points $p_1, p_2, \ldots, p_M$ that are created within the span of X, with the $M + 1$ new features

$$P_0(X), P_1(X), P_2(X), \ldots, P_{M-1}(X), P_M(X) \text{ Eq. (2)}$$

$$\begin{aligned} P_0(X) &= D(X < p_1), \\ P_1(X) &= D(p_1 \le X < p_2), \\ P_2(X) &= D(p_2 \le X < p_3), \\ &\vdots \\ P_{M-1}(X) &= D(p_{M-1} \le X < p_M), \\ P_M(X) &= D(p_M < X), \end{aligned} \quad \text{equ 2}$$

where the indicator function D returns a 1 when the condition is met or 0 otherwise [21] and the varying constants β are estimated for the individual bins Eq. (3)

$$y_i = \beta_0 + \beta_1 P_1 x_i + \beta_2 P_2 x_i + \ldots + \beta_M P_M x_i + \epsilon_i \quad (3)$$

Thus, given a value of X, at most one of $P_1, P_2, \ldots, P_M$ should not be zero. The solid curve in Figure 2 displays the fitted step functions of *age* to the *wage* data using least squares methods [6].

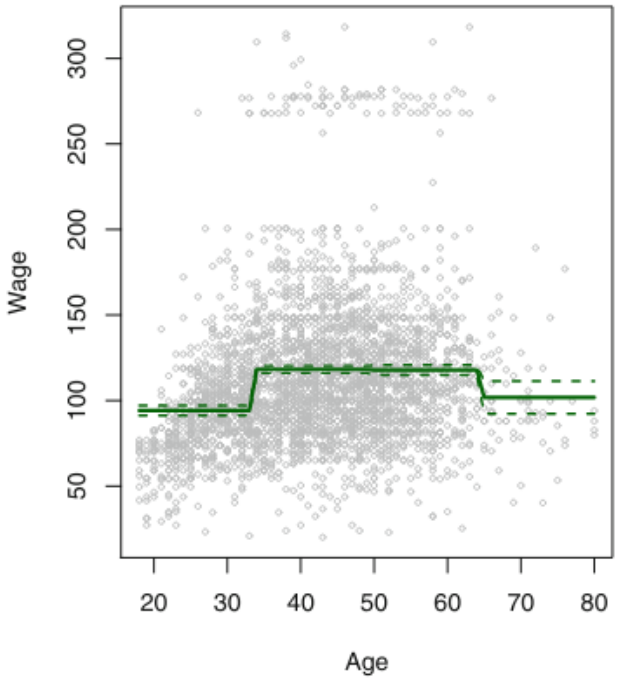

Figure 2: the fitted step functions of *age* to the *wage* data using least squares methods

3.3 Generalized Additive Models (GAMs)

One downside of the polynomial regression and step functions is the use of OLS, which is a parametric method, in estimating their models' fits. GAMs curbs this hitch by initiating a complete shift from parametric model to non-parametric model, thus providing a framework for extending a linear model that allows non-linear functions of each of the variables x, while maintaining additivity. [6-8]. It replaces each linear component $\beta_j x_{ij}$ with a nonlinear function $f_j(x_{ij})$ Eq. (4)

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$$

$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \epsilon_i \qquad (4)$$

[Gareth et al] applied GAMs in modeling wages of employees Eq. 5 estimating $f_1$and $f_2$as smoothing splines with 4 and 5degrees of freedom respectively.

$$wage = \beta_0 + f_1(year) + f_2(age) + \ldots + f_3(education) + \epsilon_i \qquad (5)$$

Applying backfitting techniques, GAMs involved multiple predictors by repeatedly updating the fit for each predictor in turn, holding the others fixed. Figure 3 shows the results of fitting the model in Eq. (5). The left-hand pane shows that holding age and education fixed, wage increase slightly with year. The center panel shows that holding education and year fixed, wage becomes highest for intermediate values of age, and lowest for the very young and very old. The right-hand pane shows that holding year and age fixed, wage increases with education.

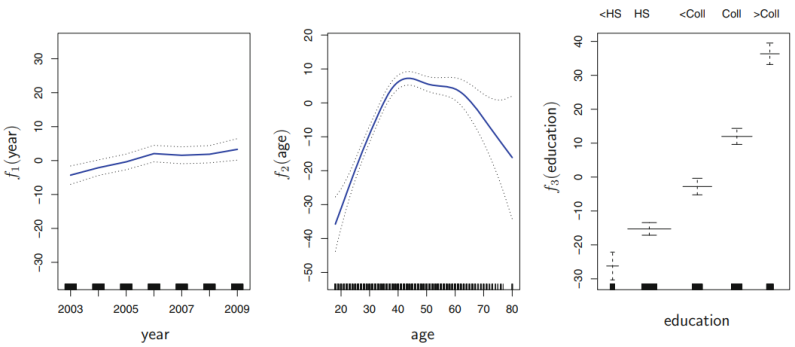

Figure 3. f_1 and f_2 are smoothing splines while f_3 is a step function

4.0 CONCLUSION

Linear models are comparatively easy to understand and implement. They also have advantages over non-linear models with regards to interpretation and inference. Nonetheless, the predictive performance of linear regression is comparatively low because the estimations of linearity assumption are sometimes very low. Hence, non-linear models - polynomial regression, step functions, and GAMs techniques could be deployed to improve on the linear models for better performance of the predictive models.

REFERENCES

[1] S. C. Agu, F. U. Onu, U. K. Ezemagu, and D. Oden (2022), Predicting Gross Domestic Product to Macroeconomic Indicators. Intelligence Systems with Applications. https://doi.org/10.1016/j.iswa.2022.200082

[2] M. Ajona, P. Vasanthi and D. S. Vijayan (2022), Application of multiple linear and polynomial regression in the sustainable biodegradation process of crude oil. Sustainable Energy Technologies and Assessments, https://doi.org/10.1016/j.seta.2022.102797

[3] W. Okba, B. Samir and M. S.Mohamed (2022), The Efficiency of Polynomial Regression Algorithms and Pearson Correlation (r) in Visualizing and Forecasting Weather Change Scenarios. DOI: 10.5772/intechopen.102726

[4] A. Nikhil, A. Saini, S. Panday and N. Gupta (2021), Polynomial Based Linear Regression Model to Predict COVID-19 Cases, in: International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), pp. 66-69, doi: 10.1109/RTEICT52294.2021.9574032.

[5] S. Kuriya and S. Sanjay (2023), Polynomial Regression Model to Predict Future Covid Cases. Journal of Artificial Intelligence and Capsule Network, (5), pp 129-143. DOI: https://doi.org/10.36548/jaicn.2023.2.004

[6] O. A. Isaac, A. A. Adedeji and I. R. Ismail (2012), Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis. Mathematical Theory and Modeling, 2(2), 14-24.
[7] O. Eva (2012). Modelling using polynomial regression. Procedia Engineering, 48 (2012), 500 – 506. doi: 10.1016/j.proeng.2012.09.545
[8] R. Arafiyah, Z.A. Hasibuan and H. B. Santoso (2018), Predicting Learner's Achievement Using Step Functions in: the 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, pp. 1-5, doi: 10.1109/CAIS.2018.8442030.
[9] A. R. Jorge and V. Pierluigi (2023), A solution for the greedy approximation of a step function with a waveform dictionary. *Communications in Nonlinear Science and Numerical Simulation.* p. 106890, https://doi.org/10.1016/j.cnsns.2022.106890
[10] Moving from Linear Models to Non-Linear Models (2024), Retrieved. 10th August, 2024 from https://www.linkedin.com/pulse/moving-from-linear-non-linear-model-agu-sunday-c-phd-cips--j4zme
[11] J. Gareth, W. Daniela, H. Trevor, and T. Robert (2017), An Introduction to Statistical Learning with Application in R, fifth ed., Springer, New York.
[12] Non-Linear Modeling Approaches - Polynomial Regression, Step Functions, Splines and GAMs. Retrieved 14th August, 2024 from https://rpubs.com/ChrisSchmidt/777578
[13] Chapter 8 Step Functions. Retrieved August 26th, 2024 from https://bookdown.org/ssjackson300/Machine-Learning-Lecture-Notes/step.html procedure.
[14] Ermakov, S.M., Leora, S.N. On the Choice of Regression Basis Functions and Machine Learning. Vestnik St.Petersb. Univ.Math. 55, 7–15 (2022). https://doi.org/10.1134/S1063454122010034.
[1] Vivek K and Udisha A (2022). Linear Regression: Assumptions and Limitations https://blog.quantinsti.com/linear-regression-assumptions-limitations/ Accessed 18/12/2024
[2] S. C. Agu, F. U. Onu, U. K. Ezemagu, and D. Oden (2022), Predicting Gross Domestic Product to Macroeconomic Indicators. Intelligence Systems with Applications. https://doi.org/10.1016/j.iswa.2022.200082
[3] Aayush O (2021) Five Obstacles faced in Linear Regression https://towardsdatascience.com/five-obstacles-faced-in-linear-regression-80fb5c599fbc Accessed 18/12/2024
[4] Kalyan (May 29, 2019) Drawbacks (Assumptions) of linear model https://medium.com/@kalyan77k/drawbacks-assumptions-of-linear-model-2e28d1c2f1d2 Accessed 18/12/2024
[5] Peter F. (2022) The Disadvantages Of Linear Regression. https://www.sciencing.com/disadvantages-linear-regression-8562780/. Accesses 18/12/2024
[6] J. Gareth, W. Daniela, H. Trevor, and T. Robert (2017), An Introduction to Statistical Learning with Application in R, fifth ed., Springer, New York.
[7] Lili, L., Bingfan, L., Xiaodi, L., Haolun, S. and Jiguo, C. (2024). Optimal subsampling for generalized additive models on large-scale datasets. 35(15). https://link.springer.com/article/10.1007/s11222-024-10546-x
[8] Asad, H., and Noah, S. (2022). Generalized Sparse Additive Models. Journal of Machine Learning. Research 23 (2022) 1-56
[9] Andreas, M., Julien, S., Harsha, N., David, S., Arber, Z., Rich, C., and Frank, H. (2024). In-Context Learning for Generalized Additive Models. arXiv:2410.04560 [cs.LG]. https://doi.org/10.48550/arXiv.2410.04560
[10] Hayden, M., Jon, D., Margo, S., and Cynthia, R. (2024). Interpretable Generalized Additive Models for Datasets with Missing Values. arXiv:2412.02646 [cs.LG]. https://doi.org/10.48550/arXiv.2412.02646
[11] M. Ajona, P. Vasanthi and D. S. Vijayan (2022), Application of multiple linear and polynomial regression in the sustainable biodegradation process of crude oil. Sustainable Energy Technologies and Assessments, https://doi.org/10.1016/j.seta.2022.102797
[12] W. Okba, B. Samir and M. S.Mohamed (2022), The Efficiency of Polynomial Regression Algorithms and Pearson Correlation (r) in Visualizing and Forecasting Weather Change Scenarios. DOI: 10.5772/intechopen.102726
[13] A. Nikhil, A. Saini, S. Panday and N. Gupta (2021), Polynomial Based Linear Regression Model to Predict COVID-19 Cases, in: International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), pp. 66-69, doi: 10.1109/RTEICT52294.2021.9574032.
[14] S. Kuriya and S. Sanjay (2023), Polynomial Regression Model to Predict Future Covid Cases. Journal of Artificial Intelligence and Capsule Network, (5), pp 129-143. DOI: https://doi.org/10.36548/jaicn.2023.2.004
[15] O. A. Isaac, A. A. Adedeji and I. R. Ismail (2012), Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis. Mathematical Theory and Modeling, 2(2), 14-24.
[16] O. Eva (2012). Modelling using polynomial regression. Procedia Engineering, 48 (2012), 500 – 506. doi: 10.1016/j.proeng.2012.09.545
[17] R. Arafiyah, Z.A. Hasibuan and H. B. Santoso (2018), Predicting Learner's Achievement Using Step Functions in: the 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, pp. 1-5, doi: 10.1109/CAIS.2018.8442030.

[18] A. R. Jorge and V. Pierluigi (2023), A solution for the greedy approximation of a step function with a waveform dictionary. Communications in Nonlinear Science and Numerical Simulation. p. 106890, https://doi.org/10.1016/j.cnsns.2022.106890
[19] Moving from Linear Models to Non-Linear Models (2024), Retrieved. 10th August, 2024 from https://www.linkedin.com/pulse/moving-from-linear-non-linear-model-agu-sunday-c-phd-cips--j4zme
[20] Non-Linear Modeling Approaches - Polynomial Regression, Step Functions, Splines and GAMs. Retrieved 14th August, 2024 from https://rpubs.com/ChrisSchmidt/777578
[21] Chapter 8 Step Functions. Retrieved August 26th, 2024 from https://bookdown.org/ssjackson300/Machine-Learning-Lecture-Notes/step.html procedure.