CURVE FITTING WITH POLYNOMIAL REGRESSION

BY

MARTINS FOLORUNSHO BALOGUN

PSC1707740

FACULTY OF PHYSICAL SCIENCES

DEPARTMENT OF STATISTICS

UNIVERSITY OF BENIN

BENIN CITY

JANUARY, 2023

UNDERTAKING

This project work was carried out by me **MARTINS F. BALOGUN** with **Matriculation number PSC1707740**. I have not copied the work of any other author(s). All works used have duly been cited and acknowledged.

_____ _____

MARTINS F. BALOGUN **DATE**

CERTIFICATION

This is to certify that this project work titled, “Curve fitting with Polynomial Regression” was carried out by **Martins, F. BALOGUN, of** the Department of Statistics, Faculty of physical Sciences, University of Benin, Benin City; under my supervision.

_____ _____

PROF. F.O. OYEGUE Date

Project Supervisor

_____ _____

PROF. C.C. ISHIEKWENE Date

Head of Department

DEDICATION

I dedicate this work to the almighty God why by his infinite wisdom and matchless grace saw me through, throughout the period of my study in this citadel of learning.

ACKNOWLEDGEMENT

The successful completion of this study would not have been possible without the various contributions of certain individuals playing key roles. It is therefore worthwhile acknowledging their support. On this note, I sincerely express my gratitude and appreciation for the invaluable support and encouragement of my ever-dynamic, understanding, father-like and indefatigable supervisor; Professor Francis, O. Oyegue, for his encouragements, suggestions and supervisory roles which can be equated to that of a loving father.

My sincere and immeasurable gratitude also goes to my biological parents Mr. and Mrs. Balogun, for their unending support, concern and relentless prayer towards my endeavor; also my siblings; John, B., Faith, B., Phyllis, B., and others; for their constant counsel, patience, sacrifice, and great financial support, and commitment towards attaining a sound quality education. May God almighty replenish your strength, bless you all abundantly and grant you your various heart desires.

I also express my utmost appreciation to all my lecturers, friends and course mates. A candid appreciation also goes to my amiable mentors; Professor R.I. Okuonghae, Mr. C.O. Odijie, & Dr. P. Osatohanmen.

Finally, permit me to say that it is absolutely impossible to acknowledge all who assisted me in no small measure to make this study in the great University of Benin a reality, to all sundry who gave me aid in any form, thank you so much.

TABLE OF CONTENT

ABSTRACT

In this project work, we look at how the least-square polynomial regression model is used to fit a non-linear relationship between a response variable and an explanatory variable in curve fitting. Finding a mathematical equation or model that best fits a noisy data has experience some drawback in curve fitting. i.e. finding an appropriate fit that best depicts the behaviour of the data. The purpose of this project is to show how the polynomial regression model can be used to show the relationship that exist between two variables; where the linear regression model is inadequate in describing such a relationship.

The method of curve fitting used in this study is the least square polynomial regression method. It is designed in a way that, the model parameters are estimated by minimizing the residual term of the polynomial regression model; and then used the model to find the line that best fits the data points of the data set.

This method was validated by modelling a data extracted from Nigeria Stock Exchange; and the model was able to predict over 80% of the relationship that existed in the data. It was discovered that the inadequacies of the simple linear regression model in describing the relationship that existed in a data set could be easily tackled by fitting a polynomial regression line. This is done by increasing the power of the independent variable to a higher power until we get a best fit.

CHAPTER ONE

INTRODUCTION

1.1 GENERAL INTRODUCTION

In this research study, we are going to look at how the least-squares polynomial regression method is used to model a non-linear relationship between a response and an explanatory variable in curve fitting.

1.2 BACKGROUND OF STUDY

In statistics, regression analysis is simply a process of finding a mathematical equation that best fits a noisy data. According to Wikipedia 2013, the term regression was initially conceptualized by Francis Galton within the framework of inheritance characteristics of fathers and sons; In his famous 1886 paper, an article in which he demonstrated how every characteristics of an individual is inherited by their offspring. Galton examined the average regression relationship between the height of fathers and the height of their sons. He depicted in a chart the average height of the parents and that of the children. He fitted a linear regression line to show the height of the children relative to that of the parents. He observed that the

children with tall parents are also tall, but on an average they are not as tall as their parents. On the contrary, children with short parents are also short, but on an average, they are taller than their parents. Galton noted that, in both cases, the height of the children approached the average of the group. Then, he found the tendency for tall parents to have tall children, but not as tall on average as their parents. Galton called this phenomenon the "Law of Universal Regression for the average heights of adult children tended to "regress" to the mean of the population; where the term regression model was coined.

In Galton's case, the relationship that exists between the height of the parents and that of the children was linear. But there are some other scenarios where a linear relationship may not exist between variables. Such cases were what brought about the non-linear regression method. One of such technique that is commonly used to capture nonlinear associations within the framework of ordinary least square regression (OLS) is the polynomial regression (Cohen et al., 2003).

The polynomial regression method is a type of nonlinear regression method which tells us the relationship between the dependent and independent variable when the dependent variable is related to the independent variable of the *nth* degree. According to Wikipedia, the first design of an experiment for polynomial regression appeared in the work of Joseph-Diez Gergonne; a paper he titled,

Design and Analysis of Polynomial Regression Experiments. The problem Gergonne considered is that, given an observational situation in which a response depends upon a single independent variable, and in which one wishes to estimate the value of the response function and its derivatives at a single point, how should one select the values of the independent variable at which the experiment will be performed, when random errors in the observed responses are expected. He began with a general discussion of the problem of interpolation, viewed both geometrically in terms of points and curves; and algebraically in terms of variables and functions. He observed that even with no errors present, the problem is somewhat indeterminate, but that with sufficiently many observations this would not cause serious difficulty, and one could conveniently fit a simple polynomial model to the data. This method involved fitting a curve that best describes the relationship that exists between the explanatory variable and the response variable. This method also uses the least-squared method to model a non-linear relationship between response variable and an explanatory variable. The best fit line is decided by the degree of the polynomial regression equation. This method, in effect, can be used to fit data curves of virtually any shape (Cohen et al., 2003). The nature of the curve can be studied or visualized by using a simple scatter plot which will give one a better idea about the linearity relationship between the variables and decide accordingly. However, concerns have been raised about its application due to its

potential to run into multi-collinearity if too many power terms are added to the model, or it not having sufficient flexibility to adjust to the varying curves in the slope when few power terms are in the model (Albarran et al., 2011; Marsh and Cormier, 2002). This regression method is used only on the basis that the ordinary least squares OLS regression failed to, or inadequately describe the relationship that exist in a scattered plot of the dataset. Hence, polynomial model is efficient to describing such relationship.

Kolb (1984) describes Curve fitting or the least squares best fit of data, as a technique that is used to determine a mathematical equation that fits a given set of data points in such a way that the deviation of the points from the equation is minimized. It involves constructing a curve, or a mathematical function, that has the best fit to a series of data points, possibly subject to the constraints of the data. It is commonly performed on all kinds of measured data. Sometimes the data is linear, but often higher-order polynomial approximations are necessary to adequately describe the trend in the data. The concept of curve fitting was birthed from the inadequacy of the linear regression analysis to describe a scattered plot that have the form of a curve from closed observation. Then fitting a curve to describe the relationships that exist becomes necessary.

Curve fitting can involve either interpolation, where an exact fit to the data is required, or smoothing, in which a smooth function is constructed that approximately fits the data. This fitted curves show the trends that exist between variables; and it allows us to predict values for which we have taken no data through the methods referred to as; interpolation and extrapolation. Extrapolation over too far a range can be dangerous unless it is certain that the relationship between the variables continues over the entire range. It has played a greater role in Sciences on data visualization, inductive inference and many others. It theoretically describes experimental data with a model that depict the pattern and behaviour of the data. It has been widely used for many research problems over the years; and has experienced some drawbacks. One of the major drawbacks of fitting a curve to a dataset is, finding an appropriate fit that best depicts the behaviour of the data; i.e. the data points to be considered in fitting the curve. These problems have been looked at mainly from a mathematical perspective; and it's often viewed as an exemplar which summarizes the many dimensions and issues associated with inductive inference, including the problem of under-determination and the reliability of inference issue. These problems have a long history in both statistics and philosophy of science.

In the early history of curve fitting, according to Wikipedia, the paper of Gauss (1809) provided the first attempt to place the problem in an error-statistical

framework. The inadequacy of the mathematical approximation perspective, as providing a foundation for inductive inference is brought out by comparing Legendre's (1805) use of least-squares as a curve fitting method with Gauss's empirical modeling perspective.

Glymour (1981) in his work, highlighted the importance of curve fitting and argued that the problem was not well understood in both its philosophical as well as mathematical dimensions. By the mid-1990s, however, the dominating view is that the use of model selection procedures associated with the Akaike Information Criterion (AIC) could address the problem in a satisfactory way by trading goodness-of-fit against simplicity (Forster and Sober (1994), Kukla (1995), Kieseppa (1997), Mulaik (2001)). He reviewed several attempts to justify curve-fitting by using other criteria in addition to goodness of fit, and concluded that there is no satisfactory rationale for curve fitting available to use yet. Among the various attempts that he considered inadequate were two Bayesian procedures based on choosing the curve that is most probable given the evidence, as well as attempts to supplement goodness-of-fit with pragmatic criteria such as simplicity. Despite that condemnation, discussions in philosophy of science in the mid-1990s seem to suggest that using the Akaike Information Criterion (AIC) for model selection, which trades over-fitting against simplicity, does after all, provide a satisfactory solution to the curve fitting problem; Forster and Sober (1994), Kukla

(1995), Kieseppa (1997), Mulaik (2001). Their main argument is that (i) the selection of the best fitting curve is well understood in the sense that least-squares provides the standard solution, and (ii) simplicity can be justified on prediction grounds because simpler curves enjoy better predictive accuracy.

More recently, Hitchcock and Sober (2004) have used the Akaike Information Criterion (AIC) to shed light on the distinction between prediction and accommodation as they relate to over-fitting and novelty. Mayo (1996) argued that for more satisfactory answers one needs to view the curve fitting problem in the context of error-statistical approach; where statistical adequacy provides such a criterion and the associated error probabilities can be used to calibrate the trustworthiness of inductive inference.

This study is specifically carried out to show how we can use the polynomial regression model in curve fitting. This model focuses more on questions of statistical inference such as how much uncertainty is present in a curve that is fitted to data observed with random errors.

1.3 AIM AND OBJECTIVE OF THE STUDY

The aim of this study is to see how the polynomial regression is used to model a non-linear relationship between a response variable and an explanatory variable in fitting a curve to data points. The objectives of this study are as follows;

1. Estimating the model parameters of a polynomial regression model using the least-squares method.

2. Using numerical data to verify polynomial regression model to describe a non-linear relationship in curve fitting.

1.4 SIGNIFICANCE OF THE STUDY

This study shows how the polynomial regression model covers for the inadequacies of the ordinary linear regression model in curve fitting. i.e. using a polynomial regression model to describe a curvilinear relationship that exists a scattered plots.

1.5 LIMITATION OF STUDY

Polynomial regression model becomes less efficient when one or more outliers exist in the fitted curve. It also fails to address the problem of extrapolating from a far range.

1.6 STRUCTURE OF THE STUDY

Chapter one contains the general introduction and the background of the study which entails the aim and objective of the study, the significance of the study, and the limitation of the study. Chapter 2 is dedicated to literature review. In Chapter 3, methods of used for the study are discussed. Chapter 4 entails the Analysis of the

data extracted from the Nigeria Stock Exchange on MTN; and their results displayed and explained using EXCEL analysis tool. Chapter 5 entails a detailed summary, conclusion and recommendation of the study.

1.7 DEFINITIONS OF TERMS

Regression analysis: it is a statistical technique for estimating the relationships among variables.

Model: it is a simplified mathematical expression used to assist calculations and predictions.

Dependent variable: this is the variable that is being measured or tested in an experiment.

Independent variable: this is the variable that is changed or controlled in an experiment.

Curve fitting: this is the process of constructing a curve that has the best fit to a series of data points.

Linear model: this a model that represents the relationship between two quantities and where the degree of the equation is one.

Non-linear model: this is a model in which observational data are modeled by a function which is a non-linear combination of the model parameters and depends on one or more independent variables.

Data Visualization: this is the graphical representation of data and information.

Parameter: this is a variable for which the range of possible values identifies a collection of distinct cases in a problem.

Residual: this is the difference between the observed values and the estimated value of the quantity of interest.

Scattered plot: this is a graph that present the relationship between two variables in a data set. It represents data points on a two-dimensional plane.

Coefficient of determination: this the proportion of the variation in the dependent variable that is predictable from the independent variable.

CHAPTER TWO

LITERATURE REVIEW

2.1. INTRODUCTION

In this chapter, we shall review some literatures on curve fitting and polynomial regression.

2.2. LITERATURE

At times, researchers are interested in understanding change processes that do not conform to a linear pattern. In such instances, the primary focus might involve identifying distinct phases in the curve of a condition under observation. Examples of this scenario may be found in cases where an investigator needs to assess critical points where events or conditions that influences the outcome variable have changed and/or also estimate the magnitude of these changes. Interest in studying the nonlinear associations and change-points of the variables may be driven by, (a) theoretical considerations, (b) an examination of data that shows that the effects of an independent variable say $\boldsymbol{X}$ on a dependent variable say $\boldsymbol{Y}$ vary, depending on the value of $\boldsymbol{X}$ (Pedhazur and PedhazurSchmelkin, 1991), or (c) by the pursuit of a more nuanced understanding of the distinct point(s) where changes occur.

Whatever the reason or influence stimulating this line of inquiry, the greater challenge for researchers lies not in their interest in studying relationships that happen to be nonlinear but in locating suitable and accessible statistical regression methods that can fit the nonlinear data points and also empirically identify the location(s) where the regression line changes direction.

One such technique that is commonly used to capture nonlinear associations within the framework of ordinary least-squared regression (OLS) is the polynomial regression (Cohen, West & Aiken, 2003; Curran, Obeidat, and Losardo, 2010; Pedhazur and PedhazurSchmelkin, 1991).

Polynomial regression, in effect, can be used to fit data curves of virtually any shape (Cohen et al., 2003). However, concerns have been raised about its application due to its potential to run into multi-collinearity if too many power terms are added to the model, or it not having sufficient flexibility to adjust to the varying curves in the slope when few power terms are in the model (Albarran et al., 2011; Marsh and Cormier, 2002).

Another lesser known technique also operating within the OLS framework that can likewise adapt to the twists and turns of a nonlinear relationship is the piecewise regression. Adding to its appeal, piecewise regression is equipped with the capability of empirically estimating the point(s) where the regression line switches

direction. In this regression technique, the change-point(s) are explicit parameters in the model, which may give it an advantage over polynomial regression where the maxima/minima points are merely implied by the model parameters when identifying statistically significant slope change(s) in the regression line. Cudeck and du Toit (2002) explored the possibility of equipping polynomial regression with the capability to empirically estimate meaningful bends in a regression line. To this end, they proposed a re-parameterized quadratic polynomial model that included the local maximum/minimum point of the parabola as a parameter in the model. However, Cudeck and du Toit (2002) proposed model is not as accessible as the better-established piecewise regression approach for locating meaningful points of change in the slope.

Spitzer et al. (2006) developed an algorithm to fit multiple measured curves simultaneously. This algorithm accounts for parameters that are shared by some curves. They created a simulated noisy measurement results to compare the introduced method to the straight forward way of fitting the curves separately. Their analysis of the simulated measurements confirmed that the introduced method yields more accurate parameters compared to the one gained by fitting the measurements separately. The new fitting algorithm was applied to the measurements of Evoked Compound Action Potential (ECAP) of the auditory nerve; and it leads to promising ideas to reduce artifacts generated by the

measuring process. The algorithm uses the relationship between multiple measurement results to increase the accuracy of the parameters. It can be applied to either linear or nonlinear equations.

The paper of Ajao et al. (2012) presented a cubic polynomial least square regression as a method of making cost prediction in business. The study revealed that polynomial regression is a better alternative with a very high coefficient of determination. They recommended that data analysts should endeavor to always plot a simple scattered plot before using any regression model in order to know the type of relationship that exists between the variables of interest.

Bhaumik et al. (2017) presented an algorithm to determine the equation of a hand-drawn curve using polynomial regression. They discussed a technique to extract the data points from an image, which are the two dimensional coordinates of the points in the curve. Using the data captured, polynomial regression was used to fit the curve to an equation in such a way that the error is minimized and at the same time, over-fitting is avoided. The proposed technique successfully identified the degree of the polynomial represented by the curve from the image and also correctly estimate its equation. The equation was subject to the resolution of the image provided and the positioning of the curve in the image. The curve of the predicted equation was visually similar to the given curve, and had the same degree as the given curve.

Michael and Lily (2020) presented a method based on extreme learning machine ELM algorithm for solving nonlinear curve fitting problems. They proposed that the unknown target function is realized by an ELM with introducing an additional linear neuron to correct the localized behavior caused by Gaussian type neurons. Several numerical experiments with benchmark datasets, simulated spectral data and measured data from high energy physics experiments have been conducted to test the proposed method. Accurate fitting was accomplished for various tough curve fitting tasks. Comparing with the results of other methods, the proposed method outperforms the traditional numerical-based technique. This work clearly demonstrates that the classical numerical analysis problem-curve fitting can be satisfactorily resolved via the approach of artificial intelligence.

CHAPTER THREE

METHODOLOGY

3.1. INTRODUCTION

In this chapter, we shall look at the method used in this study. How it is being applied to fitting a polynomial model; and also shows how the parameters of the model is estimated using the least squares approach.

3.2. CURVE FITTING

Curve fitting is a technique that is used to determine a mathematical equation that fits a given set of data points in such a way that the deviation of the points from the equation is minimized. It involve constructing a curve, or a mathematical function, that has the best fit to a series of data points, possibly subject to the constraints of the data. It encompasses methods used in regression analysis; but regression analysis is not necessarily fitting a curve. But both curve fitting and regression analysis try to find a relationship between a dependent and an independent variable. But in the case of regression, there are many more constraints considered in describing the relationship that exist.

The objective of curve fitting is to theoretically describe experimental data with a model and to find the parameters associated with the model, in other to tell the

relationship that exist among the dataset of the variables being investigated. Parameters derived from these models are quantitative estimates of the real system properties. Curve fitting find the estimates of the coefficients which make a function match the data as closely as possible. The best estimates of the parameters are the ones that minimize the error or residual of the model.

Two general approaches for fitting a curve;

1. Least square regression.
2. Interpolation.

But in this study, the focus is on least square regression (polynomial).

3.3. POLYNOMIAL REGRESSION

Polynomial regression is a form of regression in which the relationship between the independent variable say, *X* and the dependent variable say *Y* is modeled as an *pth* degree polynomial in *X*. It fits a nonlinear relationship between the value of *X* and the corresponding conditional mean of *Y*, denoted *E(y | x)*. The polynomial regression models can be used to establish a curvilinear relationship between response and explanatory variable. These models are usually fitted using the method of least squares. This method minimizes the variance of the unbiased estimators of the coefficients, under the conditions of the Gauss–Markov theorem.

In statistics, the Gauss–Markov theorem (or simply Gauss theorem) states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators if the errors in the linear regression model are uncorrelated, have equal variances (homoscedastic), and expectation value of zero.

Polynomial regression models have a shape/degree tradeoff. In order to model data with a complicated structure, the degree of the model must be high, indicating that the associated number of parameters to be estimated will also be high. This can result in highly unstable models. The first degree polynomial equation could also be an exact fit for a single point and an angle; while the third degree polynomial equation could also be an exact fit for two points, an angle constraint, and a curvature constraint. The goal of regression analysis is to model the expected value of a dependent variable *Y* in terms of the value of an independent variable or vector of independent variables *X*.

In general, we can model the expected value of *Y* as a *pth* degree polynomial, yielding the general polynomial regression model:

$$\boldsymbol{y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_p X^p + \varepsilon_i}\ ;\ \text{for } i=1,2,\ldots,n \qquad \textbf{(1)}$$

Conveniently, these model is linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters β_0, β_1, ..., β_p. Therefore, for least squares analysis, the computational and inferential problems of

polynomial regression can be completely addressed using the techniques of multiple regressions. This is done by treating X, X^2, ..., X^p as being distinct independent variables in a multiple regression model.

The polynomial regression model also has the form;

$$\boldsymbol{y_i = \beta_o + \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i \,;\, i = 1(1)n}$$

where,

$y_i \in R$, is the real-valued response for the *ith* observation.

$\beta_o \in R$, is the regression intercept.

$\beta_j \in R$, is the regression slope for the *jth* degree polynomial.

$x_i \in R$, is the predictor for the *ith* observation.

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, is the Gaussian error term or simply the residual.

3.4. FUNDAMENTAL ASSUMPTIONS OF THE POLYNOMIAL REGRESSION MODEL

1. The relationship between the dependent variable *Y* and any independent variable *X* is linear or curvilinear.

2. x_i and y_i are observed random variables.

3. $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ is an unobserved random variables.

4. $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are unknown parameters of the model.

5. $(y_i \mid x_i) \overset{iid}{\sim} N(\beta_o + \sum_{j=1}^{p} \beta_j x_i^j, \ \sigma^2)$ I.e. homogeneity of variance.

3.5. LEAST SQUARE ESTIMATES OF THE POLYNOMIAL REGRESSION PARAMETER.

We estimate the regression parameters by the method of least squares. First we calculate the sum of squared of the error term, $\in_i$ and, then find a set of estimators that minimizes the sum. From equation (1), we will be considering the 2nd order polynomial regression model given as;

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)$$

$$\epsilon_i{}^2 = \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\right]^2$$

$$\sum_{i=1}^{n} \epsilon_i{}^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\right]^2 \quad \textbf{(2)}$$

Let L = $\sum_{i=1}^{n} \epsilon_i{}^2$; To minimize the residual term, we find $\frac{\partial L}{\partial \beta_0} = 0$, $\frac{\partial L}{\partial \beta_1} = 0$, and $\frac{\partial L}{\partial \beta_2} = 0$. That is, the partial derivative of L with respect to $\beta_0, \beta_1, and\ \beta_2$. i.e.

$$\frac{\partial L}{\partial \beta_0} = (-1)(2) \sum_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right] = 0$$

$$\sum_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right] = 0$$

$$\sum_i y_i - n\beta_0 - \beta_1 \sum_i x_i - \beta_2 \sum_i x_i^2 = 0$$

$$\sum_i y_i = n\beta_0 + \beta_1 \sum_i x_i + \beta_2 \sum_i x_i^2 \quad \textbf{(3)}$$

$$\frac{\partial L}{\partial \beta_1} = (-x_i)(2) \sum_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right] = 0$$

$$\sum_i x_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right] = 0$$

$$\sum_i x_i y_i - \beta_0 \sum_i x_i - \beta_1 \sum_i x_i^2 - \beta_2 \sum_i x_i^3 = 0$$

$$\sum_i x_i y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 + \beta_2 \sum_i x_i^3 \quad \textbf{(4)}$$

Also,

$$\frac{\partial L}{\partial \beta_2} = \left(-x_i^2\right)(2)\sum_i \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right] = 0$$

$$\sum_i x_i^2\left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)\right] = 0$$

$$\sum_i x_i^2 y_i - \beta_0 \sum_i x_i^2 - \beta_1 \sum_i x_i^3 - \beta_2 \sum_i x_i^4 = 0$$

$$\sum_i x_i^2 y_i = \beta_0 \sum_i x_i^2 + \beta_1 \sum_i x_i^3 + \beta_2 \sum_i x_i^4 \qquad \textbf{(5)}$$

Bringing equation 3, 4, and 5 together, we have

$$\sum_i y_i = n\beta_0 + \beta_1 \sum_i x_i + \beta_2 \sum_i x_i^2$$

$$\sum_i x_i y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 + \beta_2 \sum_i x_i^3 \qquad \textbf{(6)}$$

$$\sum_i x_i^2 y_i = \beta_0 \sum_i x_i^2 + \beta_1 \sum_i x_i^3 + \beta_2 \sum_i x_i^4$$

Equation (6) can be represented by the matrix:

$$\begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \\ \sum_i x_i^2 y_i \end{bmatrix} = \begin{bmatrix} n & \sum_i x_i & \sum_i x_i^2 \\ \sum_i x_i & \sum_i x_i^2 & \sum_i x_i^3 \\ \sum_i x_i^2 & \sum_i x_i^3 & \sum_i x_i^4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

The expression $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\boldsymbol{Y}$ gives the least square estimates of the parameters $\boldsymbol{\beta}$ of the polynomial regression model.

Alternatively, we can obtain the parameters of the model using the matrix approach. From equation (1), we have;

$$\sum_{i=1}^{n} \epsilon_i{}^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\right]^2$$

In matrix notation, we can rewrite model (1) as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(7)}$$

i.e.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p \\ 1 & x_2 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where response $\boldsymbol{Y}$ and error vector $\boldsymbol{\varepsilon}$ are column vectors of length *n*, vector of parameter $\boldsymbol{\beta}$ is a column vector of length *p+1* and design matrix of $\boldsymbol{X}$ is *n* by *p+1* matrix (with its first column having all elements equal to 1, the second column being filled by the observed values of X_i, etc.). We want to estimate the unknown values of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ **.**

3.6. LEAST SQUARED ERROR APPROACH IN MATRIX FORM

We estimate the regression parameters by the method of least squares. This is an extension of the procedure used in simple linear regression. First, we calculate the

sum of the squared errors and, second, find a set of estimators that minimize the sum. Using equation (7) we obtain for the errors

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}$$

where $\varepsilon_i \sim N(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$, is a multivariate normal. Also, the response vector is a multivariate normal given X: $(Y|X) \sim N(\boldsymbol{Xb}, \sigma^2 \boldsymbol{I}_n)$. In multivariate analysis,

$$\sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon^T \epsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})^{\boldsymbol{T}} (\boldsymbol{Y} - \boldsymbol{X\beta}) \qquad \textbf{(8)}$$

$$= \boldsymbol{Y^T Y} - \boldsymbol{\beta^T X^T Y} - \boldsymbol{Y^T X\beta} + \boldsymbol{\beta^T X^T X\beta}$$

$$= \boldsymbol{Y}' \boldsymbol{Y} - 2\boldsymbol{\beta^T X^T Y} + \boldsymbol{\beta^T X^T X\beta}$$

Since $\boldsymbol{\beta^T X^T Y} \equiv \boldsymbol{Y^T X\beta}$

Let **L** = $\boldsymbol{\epsilon^T \epsilon}$

To find an estimator for $\boldsymbol{\beta}$, we minimize the sum of squares of the errors given as;

$$\boldsymbol{\epsilon^T \epsilon} = \boldsymbol{Y^T Y} - 2\boldsymbol{\beta^T X^T Y} + \boldsymbol{\beta^T X^T X\beta}$$

where the symbol $()^T$ denotes the transpose of the matrix. Here $\boldsymbol{\epsilon^T \epsilon}$ is scalar. We can take the first derivate of this function with respect to the vector $\boldsymbol{\beta}$. Making these equal to **0** (a vector of zeros) we obtain normal equations. i.e.

$$\frac{\partial L}{\partial \beta} = -2\boldsymbol{X^T Y} + \mathbf{2}\boldsymbol{X^T X\widehat{\beta}} = 0$$

This can further be simplified to;

$$2\boldsymbol{X^T X\widehat{\beta}} = 2\boldsymbol{X^T Y}$$

$$\boldsymbol{X^T X\widehat{\beta}} = \boldsymbol{X^T Y} \quad \textbf{(9)}$$

Therefore; use the inverse matrix $\boldsymbol{(X^T X)}^{-1}$, to multiply both sides of equation (9), and we have the least squared estimator for the multiple regression model in matrix form below;

$$\boldsymbol{\widehat{\beta}} = (\boldsymbol{X^T X})^{-1}\, \boldsymbol{X^T Y}$$

Vector $\boldsymbol{\widehat{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

3.7. POLYNOMIAL REGRESSION MODEL AND EVALUATION OF ITS ACCURACY

Polynomial regression is a special case of multiple regression, with only one independent variable *X*. One-variable polynomial regression model can be expressed as;

$$\boldsymbol{y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_p X^p + \varepsilon_i}\ ;\ \text{for } i = 1,2,\ldots,n$$

where *p* is the degree of the polynomial. The degree of the polynomial is the order of the model.

Effectively, this is the same as having a multiple model with $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$, etc.

The mean squared error *MSE* is an unbiased estimator of the variance σ^2 of the random error term and is defined in the equation below;

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(p+1)}$$

where y_i are observed values and $\hat{y}_i$ are the fitted values of the dependent variable *Y* for the *i*th case. Since the mean squared error is the average squared error, where averaging is done by dividing by the degrees of freedom, *MSE* is a measure of how well the regression fits the data. The square root of *MSE* is an estimator of the standard deviation σ of the random error term. The root mean squared error

$RMSE = \sqrt{(MSE)}$;

is not an unbiased estimator of σ, but it is still a good estimator. *MSE* and *RMSE* are measures of the size of the errors in regression and do not give an indication about the explained component of the regression fit.

3.8. COEFFICIENT OF DETERMINATION

This is an important measure of how well the regression model fits the data. It is denoted by R^2 (*R*-square). The coefficient of determination of the multiple regression is similar to the simple linear regression; and it is defined as;

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where *SST* is the total sum of squares and $\bar{y}$ is the arithmetic mean of the *Y* variable. R^2 measures the percentage of variation in the response variable *Y* explained by the explanatory variable *X*. The value of R^2 is always between zero and one, $0 \leq R^2 \leq 1$. An R^2 value of 0.9 or above is very good, a value above 0.8 is good, and a value of 0.6 or above may be satisfactory in some applications, although we must be aware of the fact that, in such cases, errors in prediction may be relatively high. When the R^2 value is 0.5 or below, the regression explains only 50% or less of the variation in the data.

3.9. METHOD OF DATA COLLECTION AND ANALYSIS

The method used in collecting the data used in this project, is the secondary method of data collection. The data used was extracted from the annual records of The Nigeria Stock Exchange.

Data analysis for this study was performed using the MS EXCEL analysis ToolPak.

CHAPTER FOUR

DATA ANALYSIS

4.1. INTRODUCTION

This chapter reviews the results and analyses of the quantitative data from a MTN stocks traded in the year 2021. This data was compiled from the annual records of the Nigeria stocks exchange.

4.2. DATA ANALYSIS

Table 4.2.1: This table shows the volume of MTN stocks traded in Nigeria the year 2021.

MONTHS	MTN	PERCENTAGES
1	23,983,268	5.188151715
2	32,786,581	7.092517852
3	36,498,460	7.89548563
4	36,498,460	7.89548563
5	39,749,820	8.598832186
6	35,328,350	7.642362986
7	37,328,350	8.07501059
8	36,326,350	7.85825414

9	35,928,350	7.772157267
10	45,328,350	9.805601005
11	54,254,563	11.73655334
12	48,259,080	10.43958766
TOTAL	**462,269,982**	**100**
AVERAGE	**38,522,499**	**8**

Source: Nigeria Stock Exchange.

From the table 4.2.1 above, we can deduce that; the total volumes of the MTN stock traded in year 2021 is 462,269,982; and the average volumes traded from the month of January to December is 38,522,499. In January, 5.19% was traded; 7.09% was traded in the month of February. That is, the volume traded increased by approximately 2%. In the month of May, the total volume traded increases by approximately 4%. But there is a 1% drop in volume traded for the month of April compared to the previous month. The highest sales occurred in the month of November, with a total volume of 54,254,563. These deductions can well be visualized through curve fitting.

Fig. 1 below is a scattered plots of the data in the table 4.2.1 above. The plots entail the total volumes of MTN stocks traded in 2021. Every plot in the graph represent the volume traded in each month.

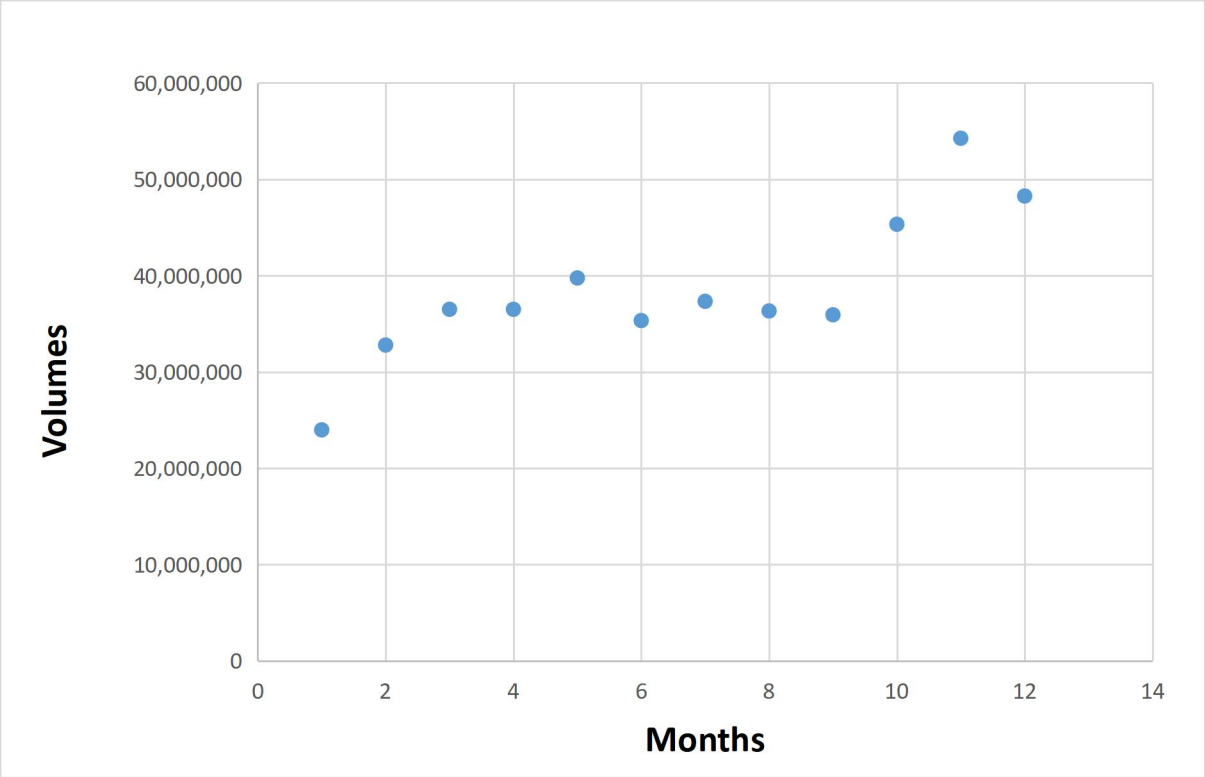

Fig. 1: scattered plots of the volumes of MTN stocks traded per month.

The above scattered plots were created using the MS EXCEL with the following steps:

1. Launch the EXCEL package from the start menu of the PC.
2. Enter the data into the EXCEL spreadsheet.
3. Highlight the cells that contained the data.
4. Click insert on the Menu bar.
5. Click insert scatter plot from the pane.
6. Select scatter plot to create a scattered plot for the data.

From fig.1 above, it is easy to see how the volumes traded increases from month to month. From this graph, we can easily spot that the volumes traded is approximately equal in march and April. Then a sharp drop occurred in June after an increased in the occurred in the month of May; and afterward, the stock begins to consolidate over three months before a sharp increase occurred in the month of October and continued in November before a drop occurred in December.

From the scattered plots above, we can find a relationship that might exist in the graph by fitting a curve or finding a line that best fit the data points. Using the MS EXCEL analysis ToolPak, a linear regression line is fitted on the scattered plots as displayed by Fig.2 below;

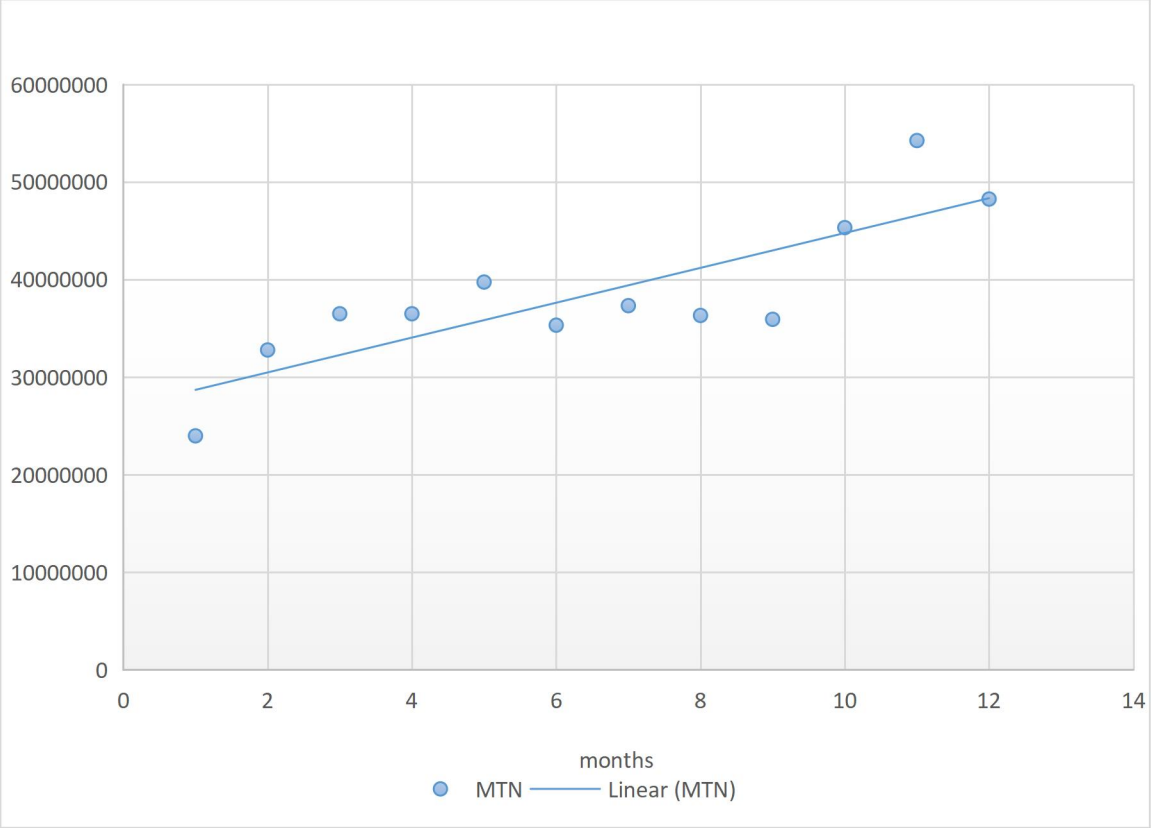

Fig. 2: A linear regression plots of the volumes of MTN stocks traded per month.

In Fig. 2 above, we fitted a linear regression line to check if the relationship that exist among the traded volumes is linear. Obviously, some points deviated from the line of best-fit; with a 68.5% coefficient of determination (R^2). This implies that the linear model predicts 68.5% of the volumes of MTN stocks traded. This results were deducted from the analysis carried out using the MS EXCEL analysis ToolPak. The result of the analysis is shown below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.827939028
R Square	0.685483033
Adjusted R Square	0.654031337
Standard Error	2.120752843
Observations	12

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-8.281290498	3.224828	-2.56798	0.027988
MTN	3.83705E-07	8.22E-08	4.668489	0.000883

The linear regression model is given as;

$$\boldsymbol{y = -8.281290498 + 3.837 * 10^{-07}(X)}$$

But fitting a cubic polynomial regression line to the data set as shown in Fig. 3 below, we have an 80.4% coefficient of determination (R^2). This implies that the polynomial model predicts over 80% of the volumes of MTN stocks traded. Hence, this model is simply the line of goodness-of-fit compared to the ordinary linear regression model. This is the good thing about data visualization. Good and valid deduction can be make easily with lesser efforts.

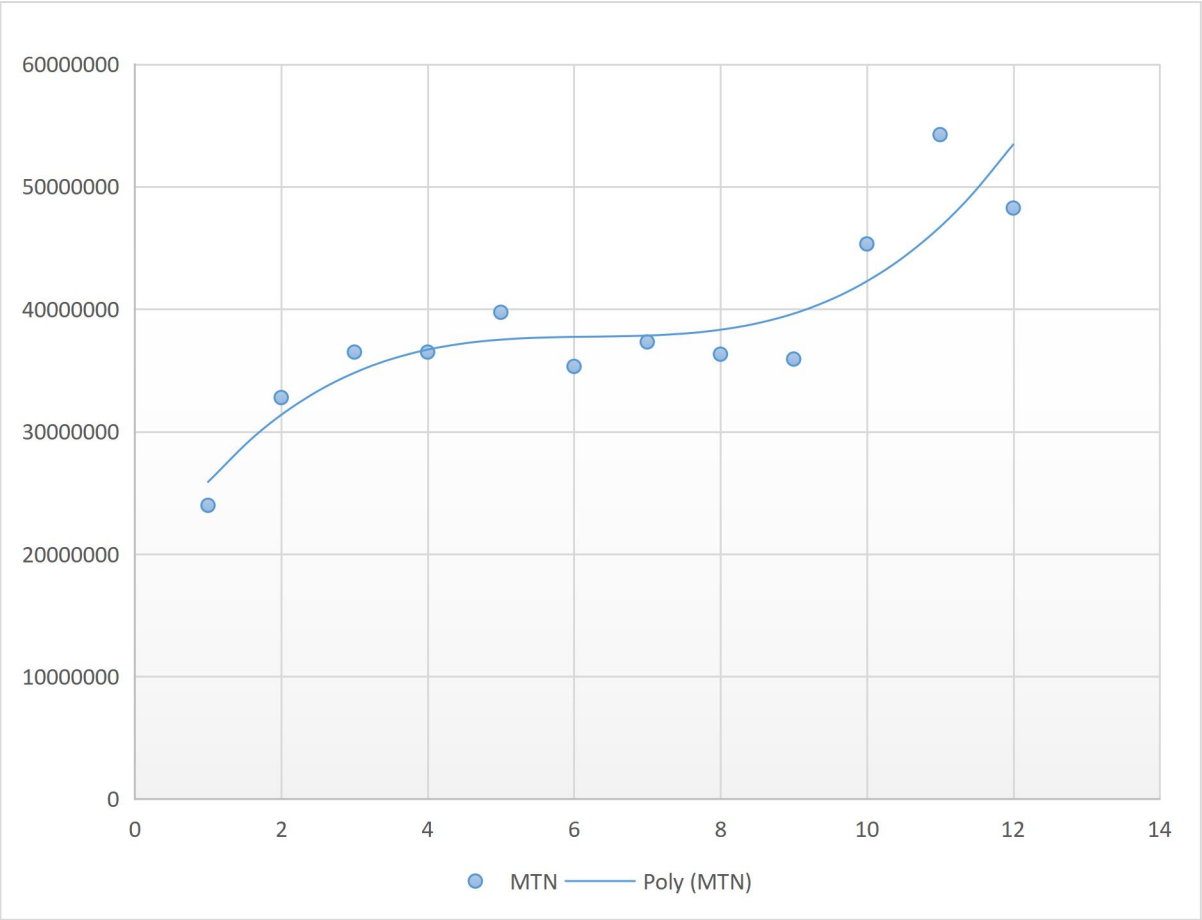

Fig. 3: A polynomial regression plots of the volumes of MTN stocks traded per month.

This results were also deducted from the analysis carried out using the MS EXCEL analysis ToolPak. The result of the analysis is shown below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.845797
R Square	0.804453
Adjusted R Square	0.608637
Standard Error	2.255596
Observations	12

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	34.33552	56.32178	0.609631	0.559019
MTN	-3.4E-06	4.69E-06	-0.71725	0.493629
(MTN) ^ 2	1.05E-13	1.26E-13	0.830207	0.430498
(MTN) ^ 3	-9.3E-22	1.09E-21	-0.85357	0.418164

The cubic polynomial model for this analysis is given as;

$$\boldsymbol{y = 34.33552 - 3.4 * 10^{-06}(X) + 1.05 * 10^{-13}(X^2) - 9.3 * 10^{-22}(X^3)}$$

CHAPTER FIVE

SUMMARY, CONCLUSION, AND RECOMMENDATION

5.1. SUMMARY

This research study has shown how we can make quick and valid deductions through graphical visualization. With the help of curve fitting, we could easily describe the relationship that exist between two variables or quantities; and one of the way of finding this relationship is through regression models.

5.2. CONCLUSION

From my research work, in modelling the volumes of MTN stock traded in 2021, I discovered that the inadequacies of the simple linear regression model in describing the relationship that existed could be easily tackled by fitting a polynomial regression model; and this model shows that a curvilinear relationship exists among the volumes of MTN stocks traded in 2021. Also, this polynomial model has a higher percentage (%) of predicting the volume of the MTN stock compared to that of the simple linear regression model. In conclusion, curve fitting is a great approach in data visualization; and its effects to data analysis should not be under-estimated.

5.3. RECOMMENDATION

In fitting a curve, one should always start with a scattered plot before fitting simple linear curve. The inadequacy of the linear models to predicts to a great extent, the dependent variable, one can now consider fitting a non-linear curve in visualizing the datasets. In other words, data analyst should not be too in a hurry to finding a non-linear relationship in some datasets, where a linear model might exist with a higher coefficient of determination.

REFERENCE

Aiken, L. S., West, S. G., & Reno, R. R. (1991). Multiple regression: Testing and interpreting interactions. Thousand Oaks, CA: Sage Publications.

Allen M. P. (1997). The origins and uses of regression analysis. In Understanding regression analysis. Plenum Press, New York, and London. DOI: 10.1007/978-0-585-25657-3_1.

Allison, P. D. (1999). Multiple regression: A primer. Thousand Oaks, CA: Pine Forge Press.

Bhaumik, C., Ajay, V., & Swati, M. (2017). "Finding Best Fit for Hand-Drawn Curves Using Polynomial Regression". International Journal of Computer Applications, 174(5):20-23, Doi:10.5120/ijca2017915390.

Box, G.E.P., and Draper, N.R. (1987). Empirical Model-Building and Response Surfaces. John Wiley & Sons, New York.

Cohen, J., & Cohen, P. (1983). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (2d ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, J., West, S.G., Aiken, L. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: L. Erlbaum Associates.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. Psychological Methods, 7, 41–63. doi:10.1037/1082-989X.7.1.41.

Cudeck, R., & du Toit, S. H. C. (2002). A version of quadratic regression with interpretable parameters. Multivariate Behavioral Research, 37, 501–519. doi:10.1207/S15327906MBR3704_04.

Darshana, S., & Maura, A. (2019). "Polynomial Regression and Response Surface Methodology: Theoretical Non- Linearity, Tutorial and Applications for Information Systems Research". Australasian Journal of Information Systems.

Fan, J., & Irene, G. (1996). "1.1 From linear regression to nonlinear regression. Local Polynomial Modelling and Its Applications". Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Glymor, C. (1981). "Theory and Evidence". Princeton University Press, NJ.

Gupta, S.C. (2013). "Fundamentals of Statistics". Seventh Edition. Himalaya Publishing House PVT. LTD., Mumbai, India.

Harvey, J.M., & Lennart, A.R. (1987). "Fitting Curves to Data Using Nonlinear Regression: A Practical and Nonmathematical Review". The FASEB Journal.

Hitchcock, C., & Sober, E. (2004). "Prediction Versus Accommodation and The Risk of Overfitting". British Journal for The Philosophy of Science 55(1):1-34.

Isaac, O.A., Adedeji, A.A., & Ismail, I.R. (2012). "Polynomial Regression Model of Making Cost Prediction in Mixed Cost Analysis". Mathematical Theory and Modeling, Vol.2, No.2.

Jeffrey, R.E., & Mark, E.P. (1993). "The Use of Polynomial Regression Equations as an Alternative to Difference Scores in Organizational Research". Academy of Management Journal, Vol. 36, No. 6, 1577-1613.

Kopf, D. (2015). The Discovery of Statistical Regression. PRICEONOMICS. Retrieved from https://priceonomics. com/the- discovery-of-statistical-regression/ Accessed on 12.25.2016.

Lonnie, M. (1998). "Non-local Behavior in Polynomial Regressions". The American Statistician (American Statistical Association) **52** (1): 20–22.

Michael, L., & Lily, D. (2020). "A Novel Method of Curve Fitting Based on Optimized Extreme Learning Machine". Pg 849-865, Applied Artificial Intelligence, an International Journal Volume 34.

Mohiuddeen, K., & Kanishk, S. (2020). "Regression Model for Better Generalization and Regression Analysis". International Conference on Machine Learning and Soft Computing. Pages 30-33.

Ostertagová, E., Frankovský, P., & Ostertag, O. (2016). "Application of Polynomial Regression Models for Prediction of Stress State in Structural

Elements". Global Journal of Pure and Applied Mathematics. ISSN 0973-1768 Volume 12, Pp. 3187-3199.

Ostertagová, E. (2012). "Modelling using polynomial regression". Procedia Engineering, vol. 48, pp. 500-506.

Wikipedia the free encyclopedia (2016). "History of statistics". Retrieved from https://en.wikipedia.org/wiki/ History of statistics Accessed on 20.12.2016

Zapata, A. (2019). "The Ability of Polynomial and Piecewise Regression Models to Fit Polynomial, Piecewise, and Hybrid Functional Forms". City University of New York.