A STUDY ON FRACTIONAL POLYNOMIAL REGRESSION

BY

Musa Uba MUHAMMAD

M.Sc/SCIE/10297/2010-2011

A THESIS SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES, AHMADU BELLO UNIVERSITY, ZARIA.

IN PARTIAL FULFILLMENT OF THE REQUIMENT FOR THE AWARD OF A MASTER OF SCIENCE DEGREE IN STATISTICS, DEPARTMENT OF MATHEMATICS, AHMADU BELLO UNIVERSITY, ZARIA, NIGERIA.

JUNE, 2015

DECLARATION

I declare that the work in this thesis entitled “A Study on Fractional polynomial Regression” has been carried out by me in the Department of Mathematics. The information derived from the literature has been duly acknowledged in the text and a list of references provided. No part of this thesis was previously presented for another degree or diploma at this or any other Institution.

---------- ----------

Muhammad Musa Uba Date

CERTIFICATION

This thesis entitled "A study on Fractional Polynomials Regression" by Musa Uba Muhammad met the regulations governing the award of the degree of Master of Science (Statistics) of the Ahmadu Bello University (A.B.U.) Zaria, and is approved for its contribution to scientific knowledge and literary presentation.

Prof. O. E. Asiribo..... ……

Chairman, Supervisory Committee Signature Date

Dr. H. G. Dikko……. ……

Member, Supervisory Committee Signature Date

External Examiner ……. ……..

Member, Supervisory Committee Signature Date

Dr. Babangida Sani …….. ……..

Head of Department Signature Date

Prof. A. H. Zoaka……. ……..

Dean, Postgraduate School Signature Date

DEDICATION

This thesis is dedicated to my beloved parents, Alhaji Uba Muhammad D. Ajingi, and Hajiya Ramlatu Muhammad, for their care, guidance, support and prayers. May the Almighty Allah (S.W.T) make Al-jannatul firdaus be their final destination, ameen.

ACKNOWLEDGEMENTS

All praises are due to Allah, the Lord of the universe, the Compassionate and the Merciful for making it possible for me to work under this field of study. May the peace and blessing of Allah be upon Prophet Muhammad (S. A. W.), his Companions, members of his family as well as entire brothers and sisters in Islam.

I extremely appreciate the efforts of my humble supervisors in persons of Prof. O. E. Asiribo and Dr. H. G. Dikko, Sirs on behalf my family I thank you all for your fatherly guidance, supervision, care, concern and above all the feeling of belonging I enjoyed from you. May Almighty Allah blesses and rewarded you abundantly. In the same vein, I would like to thank my lecturers in persons of Dr. Ibrahim Abdullahi Fagge and Dr.C. N. Nnamani for their tremendous contributions, the PG coordinator in person of Dr. Abubakar Yahaya as well as the entire members of staff of the Department of Mathematics, Ahmadu Bello University, Zaria, for their positive contributions.

My immediate families deserve special commendation for their advice, concern and support. Particularly, my dear wife not only her love, but patience, understanding and care for me. In addition, I would like to thank my brothers; Garba, Shazali, Murtala, Muawiya, Bashir, and sisters; Marariya, Fatima (Hajjan), Fatima (Ummi), and A'isha (Hajiyan Gaya). May Almighty Allah bless us all, ameen.

Finally, let me register my sincere appreciations to my dear colleagues of the Department of Statistics, in particular, as well as the entire staff and the Management of Kano University of Science and Technology, Wudil, for their supports and encouragements. Furthermore, my sincere appreciation goes to my friends, well-wishers and students.

ABSTRACT

This research work was carried out on fractional polynomial regression model, by fitting continuous covariates and grouped covariates. A new median algorithm method for grouping continuous covariate was proposed. A generalized linear model was fitted for the fractional polynomial using two different approaches on a set of experimental design data to study the effect of nitrogen fertilizer(x_1) and manure(x_2) on cowpea variety. When a proposed method by Royston and Altman was used, the algorithm for the selection of factors with significant effect converged at $\varphi(1, 1)$.On the other hand, when the study applied ordinary fractional polynomial regression model, it was observed that, for fertilizer the algorithm for the selection of factors with significant effect converged at $\varphi(x_1, 3)$, for the manure, the algorithm for the selection of factors with significant effect converged at $\varphi(x_2, -2)$.

TABLE OF CONTENTS

CHAPTER ONE

GENERAL INTRODUCTION

1.0 Introduction

It is common in Statistics to be interested in a simple approximation for smoothing relationships between variables and such relationships may be known but complicated, or completely unknown. Research work includes the collection and analysis of data on one or more variables. Often multiple regression analyses are used to model such data sets which may include only linear terms in the covariate(s). In most applications, the choice of the model building is based on simple linear effect modeling approach, but the linearity assumption may be questionable.

According to Sauerbrei and Royston (2010) to avoid this strong assumption, researchers often apply cutpoints to categorize the variable, implying regression models with step functions and this simplify the analysis and interpretation of the result. In many research studies, covariate(s) encountered are continuous and most regression models constructed for such data type include only linear terms in the covariate(s) but if curvature is suspected between the outcome variable Y and a covariate X, the model may be extended to include a quadratic term.

Royston and Sauerbrei (2008) stated that in most applications, a choice is made between linear and quadratic, with cubic or higher order polynomials being rarely used. It has been recognized that conventional low order polynomials do not always fit the data well. Higher order polynomials tend to fit the data more closely but may fit badly at the extremes of the observed range of X (Royston and Altman, 1994). Also, polynomials do not have asymptotes and cannot fit data where limiting behavior is expected (McCullagh and Nelder, 1989). In an attempt to obtain acceptable models, Box and Tidwell (1962) developed an approximate linearization of each variable in a

multiple-regression model giving $\sum_{i=1}^{n} \beta_i f(x_i)$. They concentrated on power transformations of the *X*'s and showed how to estimate the powers iteratively. Royston and Altman (1994) stated that models with more than one *X*-variable have considerable difficulties in estimating their powers reliably. They believed that estimation of the precise power(s) is unnecessary because the likelihood surface is usually nearly flat near maximum, but in any case *Y* may not be linear in X^p.

In this research work, fractional polynomial regression model given by Royston and Altman (1994) will be studied under normal errors. The fractional polynomials are models whose power terms are restricted to a small predefined set of integer and non-integer values. The powers selected encompass that of the conventional polynomials (Royston and Sauerbrei, 2008).

1.1 Purpose of the Study

Most of the existing method on fractional polynomial models focused on fitting models to psychological and pharmacokinetic experimental data. Little has been done on agronomic data although; Nelder (1966) introduced and applied the inverse polynomial model on fertilizer trials, while Salawu (2007) applied the inverse polynomial model at quadratic variable on fertilizer response of three rice varieties.

This research focuses on fitting all the power set of a fractional polynomial model on Pmain aim is to observe how well the fractional polynomial model fit the data using normal errors regression analysis when the covariates are continuous or are grouped.

1.2 Significance of the Study

The main significance of the study is to present how to fit a fractional polynomial model and assessing the fitted model using deviance difference test. Attempt was made

to show if the fractional polynomial models are more efficient compared to the conventional polynomial models in fitting data. Furthermore, effort to come-up with a new median algorithms method for grouping continuous covariates was pursued. It will serve as a reference material to researchers and scientist who may wish to undertake similar study.

1.3 Aim and Objectives of the Study

The study is aimed at comparing fractional polynomial models to conventional polynomial models using grouped and ungrouped continuous covariates with a view to achieving the following objectives to;

1. Fit a generalized linear model fractional polynomial models for all predefined set of powers {-2, 1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3} with categorized covariate and compared with a non-categorized covariate fit. Powers selection is based on Royston and Altman (1994) Algorithms.
2. Compare the fitted models using the deviance difference measure, and
3. Propose a new median algorithms method for grouping continuous covariates.

1.4 Methodology

The methodology used in this research work are fractional polynomial for normal error regression models, median method for grouping continuous covariates and the deviance method for checking model adequacy.

1.5 Scope of the Study

This research focuses on fitting all pre-defined set of power of a fractional polynomial model on an agronomic set of data with continuous covariate(s) and grouped

covariate(s).The data was obtained from Data Processing Unit, Institute of Agricultural Research, Ahmadu Bello University, Zaria.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This chapter seeks to review literature on the works of different scholars with regards to techniques applied in areas of conventional polynomial regression models, fractional polynomial regression model and issues of assessing model adequacy.

2.1 Polynomial Regression Models for Continuous Covariates

A basic choice in modeling is between parametric and non-parametric models. Parametric models such as polynomials are easy to fit and the risk function may be written down concisely, but they may fit the data badly and give misleading inference. On the other hand, non-parametric models may fit the data well but difficult to interpret due to fluctuations in the fitted curves. The risk function is usually impossible to write down concisely (Royston *et al*., 1999). Polynomial regression entails an inherent trade-off between accuracy and efficiency. As the degree of the polynomial increases, the accuracy of the model increases up to a certain point, however the time and space needed increases as well (Stronger and Stone, 2006). Bremer (2012) stated that in practice, we usually start with models of degree one, and if transformations on the predictor or the response are insufficient then we consider models of degree two. Higher degree models should be avoided unless the context from which the data is coming explicitly calls for one of these models. To decide on the appropriate degree of a polynomial regression model, two different strategies are possible. One can start with a linear model and include higher order terms one by one until the highest order term becomes non-significant. This method is generally called Forward Variable selection. On the other hand one could start with a high order model and exclude the non-

significant highest order terms one by one until the remaining highest order term becomes significant. This method is generally referred to as Backward Variable selection. In general, the two methods may not lead to the same model. For polynomial models, these methods are likely overpowered, since we can restrict our attention to first and second order polynomial models (Bremer, 2012). It is possible to select the predictor functions more carefully as curve linear functions of X to avoid this problem.

One problem that is always encountered in regression model building is nonlinearity in the relation between the outcome variable and continuous or ordered predictors. Traditionally, such predictors are entered into stepwise selection procedures as linear terms or as dummy variables obtained after grouping, though the assumption of linearity may be incorrect (Royston and Sauerberi, 2008). Categorization introduces problems of defining cutpoint(s) (Altman *et al.*, 1994), overparametrization and loss of efficiency; Lagakos (1988). In any case, a cutpoint model is an unrealistic way to describe a smooth relationship between a predictor and a response variable. An alternative approach is to keep the variable continuous and allow some form of nonlinearity. Hitherto, quadratic or cubic polynomials have been used, but the range of curve shapes afforded by conventional low-order polynomials is limited (Royston and Sauerbrei, 2008). Box and Tidwell (1962) proposed a method of determining a power transformation of a predictor. A more general family of parametric models, proposed by Royston and Altman (1994), is based on fractional polynomial (FP) functions and can be traced back to Box and Tidwell's (1962) approach. Royston and Altman (1994) presented the FP functions which encompass conventional polynomials as a special case where one, two or more terms of the form x^p are fitted, the exponent's p being chosen from a small, preselected set of integer and non-integer values.

For non-parametric regression and scatter plot smoothing are other methods in modeling continuous covariates other than linear and FP functions. For a function of x with the global-influence property, the fit at a given value x_0 of x may be relatively unaffected by local perturbations of the response at x_0, but the fit at points distant to x_0 may be affected, perhaps considerably. This property may be regarded by proponents of local-influence models as a fatal flaw (Royston and Sauerbrei, 2008). Conventional polynomial regression is a popular nonparametric regression technique due to its attractive asymptotic properties, in particular at the border of the support. For fully observed responses, a local polynomial regression estimate of $m(x_0)$ is obtained by estimating a polynomial in x with weighted ordinary least squares. Each unit is weighted depending on its distance in x to the design point of interest (focal value) x_0, thereby making the procedure local (Karlsson *et al.*, 2009). According to Royston and Sauerbrei (2008), a rigorous definition of the global-influence property has not been framed, but such models are usually parametric in nature. Examples include polynomials, nonlinear models such as exponential and logistic functions, and fractional polynomials developed by Royston and Altman (1994). By contrast, functions with the local-influence property, including regression splines (de Boor, 2001), smoothing splines (Green and Silverman, 1994), and kernel-based scatter-plot smoothers such as locally weighted scatter plot smoothers ”LOWESS” (Cleveland and Devlin, 1988), are typically nonparametric in character. Perturbation of the response at x_0 usually greatly affects the fit at x_0 but hardly affects it at points distant to x_0. One key argument favoring functions with global influence is their potential for use in future applications and datasets (Royston and Sauerbrei, 2008). Without such an aim, functions with local influence might appear the more attractive (Hand and Vinciotti, 2003). According to Royston and Sauerbrei (2008) fractional polynomial functions

retain the global-influence property; they are much more flexible than polynomials. Further, they stated that low-dimensional fractional polynomial curves may provide a satisfactory fit where high-order polynomials fail (Royston and Altman, 1994). Fractional polynomials are intermediate between polynomials and nonlinear curves. They may be seen as a good compromise between ultra-flexible but potentially unstable local-influence models and the relatively inflexible conventional polynomials (Royston and Sauerbrei, 2008).

Royston and Altman (1994) stated that the nonparametric and spline smoothers are powerful and flexible tools which indeed impose few limitations on the functional form. However, the fitting process can be computationally intensive even though the methods are successful for describing data; a major drawback is that, since they use local models, they do not yield simple equations for prediction. According to Royston and Altman (1994), the cubic spline may be seen as the link between conventional polynomials and the modern methods of nonparametric smoothing. Splines were originally developed in the 1920s for interpolation (Whittaker, 1923), much later, the smoothing spline was developed as a method for fitting curves to data (Reinsch, 1967; Silverman, 1985). Bremer (2012) stated that splines are piecewise polynomial functions that satisfy certain "smoothness" criteria. These criteria are often continuity and possibly continuity of the derivative(s). The points where different polynomial pieces are joined together are called the knots of the spline. A knot is placed at each data point and a parameter is used to control the degree of smoothing. A simpler variant is the regression spline (Poirer, 1973), which has no smoothing parameter and which uses a small number of knots (Royston and Altman, 1994). In another study by Bremer (2012) stated that if the positions of the knots in a spline are known, the fitting of a spline function to a data reduces to a nonlinear regression problem. If the positions are not

known, the problem becomes more complicated. It is not easy to decide how many knots to use and how they should be placed. In general, each piece of the spline should be kept as simple as possible to avoid over fitting the data. A special case of spline functions are the piecewise-linear functions. As with splines, it can be assumed piecewise linear functions to be continuous, or discontinuous at the knots and if the scatter plot shows some periodic behavior, then trigonometric functions (sine, cosine) may be reasonable to include in the model. The trigonometric terms can have varying amplitudes and frequencies.

2.2 Fractional Polynomial Modeling

Various attempts have been made to devise more acceptable models. Box and Tidwell (1962) developed an approximate linearization of each variable in a multiple-regression model given as $\sum_{i=1}^{n} \beta_i f(x_i)$. They concentrated on power transformations of X_s and showed how to estimate the powers iteratively. The opinion by Royston and Altman (1994) led to development of fractional polynomials regression with continuous covariates which they termed to be easy to understand, parsimonious, simple and quick to fit using standard multiple regression software. Though, they did not limit generalization of their technique to continuous covariates, discrete covariates as well can be utilized (Royston and Altman, 1994; Royston and Sauerbrei, 2008).

Most of the curves used in the description of human growth are non-linear, but few are linear. For example, in some early papers on the analysis of longitudinal growth studies, Count (1942, 1943) modeled skull growth in American children and stature of Chinese children from 3 month to 7 years using $\beta_0 + \beta_1 x + \beta_2 \ln x$, where X is age. Wingerd (1970) compared the Count model with a conventional quadratic and with Y= $\beta_0 + \beta_1 x + \beta_2 x^{1/2}$. Concluding that, models with more than two terms are required

for such data (Berkey and Reed, 1987). Nelder (1966) suggested the system of inverse polynomials of the form X/Y = $\sum_{i=1}^{n} \beta_i x_i$. for example a quadratic model gives X/Y $= \beta_0 + \beta_1 x + \beta_2 x^2$, which lead to 1/Y $= \gamma_0 + \gamma_1 x + \gamma_2 x^{-1}$. Nelder (1966) used the model to describe the relationship between plant yield and fertilizer concentration.

Royston and Altman (1994) developed a parametric nonlinear polynomial regression models that incorporate a set of restricted small predefined powers of integers and non-integer values called *S* to be used in their technique for fitting data of continuous covariates. These powers ranges from -2 to 3 and their choice is to justify the inclusion of conventional polynomials and also the models of Count (1942, 1943), Nelder (1966) and Wingerd (1970) as subset of the family *S*. They also developed their technique to include multivariable covariates and of recent Royston and Sauerbrei (2008) developed the technique to include interaction effect. The context of Royston and Altman (1994) technique was aimed at modeling trend in a response variable *Y* in terms of covariate(s) with restriction to generalized linear models (GLMs) and proportional hazards regression models (Cox models). Their examples on GLM are of normal and binomial error models, where the link functions are taken as $g(\mu) = \mu$ and $g(\mu) = \ln \mu/(1-\mu)$ respectively.

For Cox models, they used standard formulation whereby the hazard function $\lambda(t; X)$ is factored as $\lambda_0(t)\exp\eta$, $\lambda_0(t)$ being the base-line hazard function. Stone (1985) stated that a flexible but tractable model function is the additive predictor $\eta = f_0 + \sum_{i=1}^{n} f_i x_i$, where f_0 is a constant term and f_i (i>0) is a function of X_i and a set of parameters. For example, Hastie and Tibshirani (1986, 1990) used locally linear smoothers such as LOWESS as the functions $f_i(X_i)$. In justifying their approach,

Royston and Altman (1994) stated that the linear predictor in a GLM is an additive predictor with $fi(Xi) = \beta iXi$ for each i. Also, that a model incorporating a quadratic polynomial in X_i has $f_i(X_i) = \beta_{i1}X_i + \beta_{i2}X_i^2$ and if each of the components. Hence, they suggested the use of fractional polynomials, which are simple non-local functions, as $f_i(X_i)$ because these functions produced models whose additive predictor, is linear.

Royston and Sauerbrei (2008) acknowledged that Box and Tidwell (1962) were the first to propose the systematic use of power functions of the form $\beta_0 + \beta_1 x^\tau$, where τ is any real number. Likewise, Mostller and Tukey (1977) made extensive use of power transformations in data analysis, where they described them as re-expression of the (scale of) x. Box and Tidwell (1962) also discussed quadratic functions in x^τ of the form $\beta_0 + \beta_1 x^\tau + \beta_2 x^{2\tau}$.

Royston and Sauerbrei (2008) stated that the Box-Tidwell quadratic functions and their fractional polynomial of order two (FP2) classes of models have some members in common (for example, $\tau = \pm 0.5, \pm 1$), but in general they differ. Also, they related their approach to multi-exponential functions $\beta_0 + \beta_1 \exp(\tau_1 x) + \beta_2 \exp(\tau_2 x) + \ldots$ which provide another link with FP functions. That is, if $z = \exp x$, then a multi-exponential function with m terms is seen to resemble the FP_m function $\beta_0 + \beta_1 z^{\tau 1} + \beta_2 z^{\tau 2} + \ldots.$ They concluded that the critical difference between Box-Tidwell, multi-exponential functions and FP_s is that with FP_s the powers are restricted to belong to the set S.

There have been extensive applications of the FP models proposed by Royston and Altman (1994). First, they applied their method to six data sets which illustrate the use of the method in three types of regression analysis: normal errors, logistic and Cox Proportional hazards they generally extend those of the original researchers. The first

two data sets are taken from studies to develop *X*-specific reference centiles for certain physical measurements in humans, *X* being gestational age in the first and age since birth in the second data set. In the third data set, the concentration of C-peptide (an insulin-related protein) was predicted from two continuous covariates in a study of childhood diabetes. The fourth and fifth data sets exemplified logistic regression. The fourth is from a study of *in vitro* fertilization (IVF) which aimed at estimating the serum oestrogen concentration *X* at which the probability of pregnancy was maximal. The fifth is from a clinical trial of two treatments for bone marrow cancer; the probability of patient survival was related to several continuous and categorical covariates. The final data set which illustrates Cox regression arose from a randomized clinical trial of two treatments for leg ulcers in which one aim was to construct a predictive score for the healing time as a function of several covariates. For the first, second and fourth data sets there was one covariate while the third, fifth and sixth data sets had more than one covariate.

For the first data set Royston and Altman (1994) concluded that *Y* [*Y* = ln(mandible length)] seems linear in 1/*X* and among the conventional polynomials, the cubic produced a good fit as that of $\phi_1(X;-1)$ but it failed when the *X* covariate was extrapolated to *X* = 34 weeks for it behaved wildly for *X*>28. However the fractional polynomial gave sensible results, fitting the additional points reasonably well. While for the second data they observed visual evidence of positive skewness. They followed Isaacs *et al.* (1983) and took the response variable *Y* to be $\sqrt{Y}$ in order to eliminate the skewness. The best fractional polynomial for order 1 was $\phi_1(X;0)$, with a higher deviance gain indicating a significantly better fit than that of the straight line model $\phi_1(X;1)$. For degree 2, the best fractional polynomial has $\phi_2(X_2;$ -2, 2) and among the

conventional polynomials, the quartic was required to give a fit that is comparable with that of $\phi_2(X_2; -2, 2)$. Royston and Altman (1994) stated that the quartic shows the usual waviness and end effects that are often associated with high degree polynomials. The same data was fitted by Isaacs *et al.* (1983) using a quadratic in $X^{½}$ or in Royston and Altman notation $\phi_2(X; 0.5, 1)$ and compared with a cubic model $\phi_3(X_2; 1, 2, 3)$ as an alternative but was rejected by them because of the clinical implausible rise near $X = 6$ years.

The third data Royston and Altman (1994) considered was a diabetes data from a study of 84 children, an unspecified subset of 43 children from the same data was first used by Hastie and Tibshirani (1990) where they used the data to exemplify generalized additive modeling. The aim was to investigate the dependence of *Y*, the logarithm of C-peptide concentration (a protein secreted together with insulin), on age (X_1) and base deficit (a measure of acidity). Royston and Altman (1994) applied multivariable fractional polynomial of degree 1 and 2 to the data. The model selected for $m = 1$ was $\phi_1(X_1; -1)$ and $\phi_1(X_1; 0)$. While for degree 2 ($m = 2$), $\phi_2(X_2; -1)$ and $\phi_2(X_2; 0)$ were the best fit for the data.

The fourth data considered by Royston and Altman (1994) was only one of the covariate (E2; plasma oestradiol concentration). They stated that graphical methods are important in the assessment of model fit, most basically (in the case of a single covariate) a plot of *Y* against *X* with the fitted line superimposed. Further, they stated that plots are difficult to construct if the response is binary. Royston (1992) recommended a modified version of LOWSS (omitting iterative reweighting) with a bandwidth of 0.8 for producing reasonably smooth plots of binary data. However, applying fractional polynomial in a logistic regression to the data it revealed that the

relationship between the LOWESS-estimated probability of pregnancy and the E2 concentration X is non-monotonic. The best powers for $m = 1$ and $m = 2$ are -2 and $\tilde{p} = (-1, -1)$ with $m = 2$ the gain was higher as they expected but the base-line model $\phi_1(X; 1)$ was a poor fit and the conventional quadratic model $\phi_2(X; 1, 2)$ was not a good fit. With Royston (1992) approach a model which was a quadratic in ln(X) that is $\phi_2(X; 0, 0)$ was obtained but Royston and Altman (1994) claimed that Royston (1992) approach is not as well fitted as theirs, visually.

MacLennan *et al*. (1988) described the fifth Medical Research Council trial of treatments for myelomatosis, a type of leukemia. Patients were randomly allocated to receive a new four-drug regime (ABCM) or conventional therapy (M7). The main outcome variable was the survival time (in days) from randomization to death. The following covariates were measured at randomization: age (X_1), haemoglobin (X_2), creatinine (X_3), serum β_2-microglobulin or SB2 (X_4), calcium (X_5), albumin (X_6) and immunoglobulin-M (IgM) (X_7). Royston and Altman (1994) analyzed this data using multivariable logistic fractional polynomial regression, they adjusted X_7 due to non-positive values, and so to ensure positivity they transformed X_7 to IgM + 0.1. They defined $Y = 1$ if the patient died between 0 and 913 days, $Y = 0$ otherwise. They build a logistic probability of death, Pr($Y = 1|\mathbf{X}$), in terms of$X_1, \ldots, X_7$ and three categorical (binary) covariates: treatment (Z_1), severity index ≥ 2 (Z_2) and severity index ≥ 3 (Z_3). The final model that was obtained from their technique was $\phi_2(X_4; -2, 3)$, X_5, X_6, $Z_1$and Z_2; the remaining variables were omitted by stepwise elimination compared to $m_4 = 1$, $\tilde{p}_4 = 1$ using their deviance difference values of 14.14 and 9.14 respectively. But they further investigated the bias associated with the data with reason that high values of SB2 (X_4) are known to be associated with poor prognosis, and in fact all patients with

values above 40 mg l^{-1} died. Using partial residual plot for the investigation revealed that in the pure region of Y = 1 where SB2 > 40 the residuals are all positive and tracked the fitted function and a further investigation showed that the m_4 = 3 model ϕ_3 (X_4; 0, 3, 3) was apparently a better fit than $\phi_2(X_4; -2, 3)$. However, they stated that the new model $\phi_3(X_4; 0, 3, 3)$ has even more abrupt behavior, turning upwards rapidly when SB2 > 40. They became concerned about the end effects, so they repeated the stepwise analysis omitting the 15 patients with SB2 > 40. This time SB2 appeared as $\phi_3(X_4; -1)$, with an associated gain of 15.10. They concluded that a larger sample size would be needed to settle the question of which fractional polynomial is the more appropriate for the true function may lie between the two.

The last data they used was on clinical trial of two treatments of leg ulcer (Smith *et al*., 1992), the response variable Y was the number of days from diagnosis to complete healing. Royston and Altman, (1994) fitted multivariable fractional polynomial Cox regression to the data. Nine covariates were studied out of which seven are continuous covariates (initial ulcerated area (X_1), number of months since onset of the ulcer (X_2), age (X_3), diastolic blood pressure (X_4), height (X_5), ankle pressure (X_6), body weight (X_7)) and two categorical covariates (presence or absence of deep vein involvement (Z_1) and treatment difference (Z_2). The X_2 variable has several zeros so they transformed it by adding 1 to each data point and they fitted the each continuous covariate separately. The stepwise selection procedure selected a model which comprises of $\phi_1(X_1; 0.5)$, ϕ_1 (X_2; 0), $\phi_1(X_3; -2)$, X_4 and Z_1. The gains for X_1, X_2 and X_3 are 3.69, 13.20 and 2.94 respectively, so the greatest improvement in fit occurs through X_2 rather than X_1.

In the study of estimating the relationship between body mass index (BMI) and mortality Wong *et al*. (2011) applied the fractional polynomial logistic regression

model with three covariates (BMI, age and smoking status). They compared the approach with other commonly used regression models and they found out that the multiple fractional polynomials (MFP) model presented a better fit in terms of the shape pattern exhibited. A J-shaped pattern for women and a U-shaped pattern for men were observed and they concluded that MFP approach developed by Royston and Altman (1994) provides a robust alternative to categorization or conventional linear-quadratic models for BMI, which limit the number of curve shapes. Royston and Altman (1994) applied the multivariable fractional polynomial modeling to data on prognostic factors for breast cancer survival and diagnostic indicators for malignant breast tumors. They concluded that the fractional polynomial approach demonstrate superiority in contrast with conventional procedures in which the variables are first categorized.

Geweke and Petrella (2012) showed that regular fractional polynomials can approximate regular cost, production and utility functions and their first two derivatives on closed compact subsets of the strictly positive orthant of Euclidean space arbitrarily well. These functions therefore can provide reliable approximations to demand functions and other economically relevant characteristics of tastes and technology. Using canonical cost function data, they showed that full Bayesian inference for these approximations can be implemented using standard Markov chain Monte Carlo methods.

2.3 Categorization of Covariates

Maxwell and Delaney (1993) posed the question; why have researchers continued to ignore methodologists' advice not to dichotomize their measures? Measurements of continuous variables are made in all branches of medicine, aiding in the diagnosis and

treatment of patients. In medical research, such continuous variables are often converted into categorical variables by grouping values into two or more categories. It seems that the usual approach in clinical and psychological research is to dichotomize continuous variables, whereas in epidemiological studies it is customary to create several categories, often four or five, allowing investigation of a possible dose–response relation (Royston *et al*., 2005). As noted by Royston *et al*. (2005) categorization is done to make the analysis and interpretation of results simple. Furthermore, clinical decision-making often requires two classes, such as normal / abnormal cancerous / benign, treat / do not treat, and so on. Although necessary and sensible in clinical settings, in a research context such simplicity is gained at a high cost, and may well create problems rather than solve them. As noted by Weinberg (1995), 'alternative methods that make full use of the information at hand should indeed be preferred, where they make sense'. Such approaches include different types of splines, and fractional polynomials (Hastie and Tibshirani, 1990; Royston and Altman, 1994).

Dichotomization is widespread in clinical studies (Del Priore *et al.*, 1997), but the reasons for its popularity are largely a matter for speculation. There is to be a general need in clinical practice to label individuals as having or not having an attribute (such as 'hypertensive', 'obese', 'high' PSA), often preliminary to determining diagnostic or therapeutic procedures. Unfortunately, this attitude perhaps affects the way in which research is done (Royston *et al*., 2005). However, a similar liking for reducing data to two groups has been observed in other fields including psychology (MacCallum *et al*., 2002) and marketing (Irwin and McClelland, 2003). As it is so common, many researchers may feel that this is in some sense the recommended approach. They may be inexperienced in analyzing continuous variables, and may be unaware of the

considerable range of suitable methods of analysis. Also, they may simply prefer more familiar and easier analyses. Additionally, among those who are more comfortable with regression there may be concerns about assuming a linear relation between the explanatory variable and the outcome. Such an automatic assumption may be wrong, and is neither necessary nor desirable (Royston *et al*., 2005).

Various perceived advantages of dichotomizing continuous explanatory variables have been advanced, but they generally cannot be supported on statistical grounds (MacCallum *et al*., 2002). The most common argument seems to be simplicity. Forcing all individuals into two groups is widely perceived to greatly simplify statistical analysis and lead to easy interpretation and presentation of results. A binary split leads to a comparison of groups of individuals with high or low values of the measurement, leading in the simplest case to a t test or χ^2 test and an estimate of the difference between the groups (with its confidence interval). In the context of a regression model with multiple explanatory variables the advantage is not as clear, although the regression coefficient (or odds ratio) for a binary variable may be easier to understand than that for a change in one unit of a continuous variable. Likewise the analysis of a single binary variable is much easier than that of a multi-category variable, which necessitates the creation of several dummy variables and for which there are several possible coding options and analysis strategies.

Such relative simplicity may be illusory, however. Even if there are good reasons to suppose that there is an underlying grouping, dichotomization at the median will not reveal it (MacCallum *et al*., 2002). The same study considered various other weak or false arguments that may be put forward in support of dichotomization. For example, investigators may argue that because the analysis of a dichotomized variable is

conservative, if a significant relation is found we can expect that the underlying relation is a strong one. They may also argue that dichotomization makes sense when the measurement is recorded imprecisely, and would provide a more reliable measure. This argument is incorrect—dichotomization will reduce the correlation with the (unknown) true values (MacCallum *et al*., 2002). Not only are many of the perceived advantages illusory, dichotomization comes at a cost (Royston *et al.*, 2005*).*

Grouping may be seen as introducing an extreme form of rounding, with an inevitable loss of information and power. When a normally distributed predictor is dichotomized at the median, the asymptotic efficiency relative to an ungrouped analysis is 65 per cent (Lagakos, 1988). Dichotomizing is effectively equivalent to losing a third of the data, with a serious loss of power to detect real relationships. If the predictor is exponentially distributed, the loss associated with dichotomization at the median is even larger efficiency is only 48 per cent (Lagakos, 1988). Discarding a high proportion of the data is regrettable when many research studies are too small and hence underpowered. It seems likely that many who do this are unaware of the implications (MacCallum *et al*., 2002). Furthermore, dichotomization may increase the probability of false positive results (Austin and Brunner, 2004). When the true risk increases (or decreases) monotonically with the level of the variable of interest, the apparent spread of risk will increase with the number of groups used. With just two groups one may seriously underestimate the extent of variation in risk (Breslow and Day, 1980). Put differently, when individuals are divided into just two categories, considerable variability may be subsumed within each group.

Furthermore, the cutpoint model is unrealistic, with individuals close to but on opposite sides of the cutpoint characterized as having very different rather than very similar

outcome. It would be expected that the underlying relation with outcome to be smooth but not necessarily linear, and usually but not necessarily monotonic (Royston *et al*., 2005). Using two groups makes it impossible to detect any non-linearity in the relation between the variable and outcome. Lastly, if regression is being used to adjust for the effect of a confounding variable, dichotomization of that variable will lead to residual confounding compared with adjustment for the underlying continuous variable (Cochran, 1968; Becher, 1992; Brenner and Blettner, 1997). Further issues arise when more than one explanatory variable is dichotomized.

Several approaches are possible for determining cutpoint. For a few variables there are recognized cutpoints which are widely used (e.g. >25 kg/m^2 to define 'overweight' based on body mass index). For some variables, such as age, it is usual to take a 'round number', an elusive concept which in this context usually means a multiple of five or ten. Another possibility is to use the upper limit of the reference interval in healthy individuals. Otherwise the cutpoint used in previous studies may be adopted. In the absence of a *prior* cutpoint the most common approach is to take the sample median. However, using the sample median implies that different studies will take different cutpoints so that their results cannot easily be compared (Royston *et al*., 2005). For example, in prognostic studies in breast cancer, Altman *et al,* (1994) found 19 different cutpoints used in the literature to dichotomize S-phase fraction. The median cutpoint was used in 10 studies. The range of the cutpoints was 2.6–12.5 per cent cells in S-phase, whereas the range of 5 'optimal' cutpoints was 6.7–15.0 per cent. Incidentally, Royston *et al*. (2005) noted that moving the cutpoint to a higher value leads to higher mean values of the variable in both groups.

The arbitrariness of the choice of cutpoint may lead to the idea of trying more than one value and choosing that which, in some sense, gives the most satisfactory result. Taken to extremes, this approach leads to trying every possible cutpoint and choosing the value which minimizes the P-value (Royston *et al*., 2005) or perhaps maximizes an estimate such as the odds ratio (Wartenberg and Northridge 1991). In practice, the search may be restricted to, say, the central 80 or 90 per cent of observations (Miller and Siegmund, 1982; Altman *et al*., 1994). The cutpoint giving the minimum P-value is often termed 'optimal', but it is optimal only in a narrow sense, and is unlikely to be optimal beyond the sample analysed (Altman *et al*., 1994). Because of the multiple testing the overall type I error rate will be very high, being around 25–50 per cent rather than the nominal 5 per cent (Miller and Siegmund, 1982; Altman *et al*., 1994; Lausen and Schumacher, 1996; Hilsenbeck and Clark, 1996). Also, the cutpoint chosen will have a wide confidence interval and will not be clinically meaningful. Crucially, the difference in outcome between the two groups will be over-estimated, perhaps considerably, and the confidence interval will be too narrow. It is possible to correct the P-value for multiple testing (Miller and Siegmund, 1982; Altman *et al*., 1994; Lausen and Schumacher, 1996; Hilsenbeck and Clark, 1996).

In addition, different types of shrinkage factor can be applied to correct for the bias and confidence intervals with the desired coverage using bootstrap resampling (Schumacher *et al*., 1997; Hollander *et al*., 2004). However, it is not clear which shrinkage factor is best, and the approach is complex and little used so far. Almost all studies using optimal cutpoints derive the cutpoint using univariate analysis and then use the resulting binary variable in multivariable analysis. Unless adjustment is made the results will be severely misleading (Altman *et al*., 1994). Mazumdar *et al*. (2003) extended the method of searching for a cutpoint for one specific predictor by adjusting

in a multivariable model for other predictors known to be important. In particular, if a model reduction algorithm is used, the dichotomized predictor may lead to other, more influential variables being displaced. This data-dependent approach to analysis should be avoided. The strategy has been used frequently in some research.

Royston and Altman (2005) stated that to evaluate the significance level and the hazard ratio (HR) associated with an 'optimal' cutpoint, Faraggi and Simon (1996) suggested an approach based on twofold cross-validation. The main feature is that the cutpoint used to classify an observation is 'optimally' selected from a subset that excludes the observation. The algorithm may be summarized as follows. The data set is divided at random into two approximately equal subsets. The 'optimal' cutpoint is determined within each subset and is used to dichotomize observations in the other subset. With this procedure, three usually different 'optimal' cutpoints are estimated. The approach defines a single dichotomization for all patients and is used for calculating the HR and P-value. The 'optimal' cutpoint from the original data is retained for later use.

Mazumdar *et al*. (2003) stressed that if the underlying clinical setting is truly multivariable, the cutpoint search should incorporate other important variables. The same point was made earlier by Faraggi and Simon (1996). In epidemiological language, one should adjust for such variables in some way. However, Mazumdar *et al*. (2003) gave no suggestions or comments on how to determine the adjustment model. Mazumdar *et al*. (2003) proposed modification of the Faraggi–Simon method which is to search for the three cutpoints as before, but adjusting for these other variables. Assuming in a simulation study that other correlated variables influence the outcome, they showed that their modification improves power and reduces bias in the estimated hazard ratio and the cutpoint when the true model has a cutpoint. In practice, there is

often more than one continuous explanatory variable in a regression analysis. The effect of dichotomization of two X variables will depend on the correlation coefficients between them and the response (*Y*), and cannot easily be predicted. Under some conditions, the inclusion of two dichotomized correlated variables can lead to a spurious relation between an *X* variable and *Y* (Maxwell and Delaney, 1993; MacCallum *et al.,* 2002). This is especially likely to occur when the partial correlation between one X variable and Y is close to zero. Also, this scenario can lead to spuriously significant interactions between X variables (MacCallum *et al.*, 2002). These findings suggest that regression models with two or more dichotomized continuous explanatory variables could be seriously misleading, both in respect of which variables are significant in the model, and perhaps also with respect to the overall predictive ability. If some of the cutpoints were selected using a data-dependent method, problems would worsen (Royston *et al*., 2005). In this research work we seek to make use of the most frequent methods in categorizing continuous variables which is the median method and compare results obtained with non-categorization of the same continuous variable using fractional polynomial regression.

2.4 Model Adequacy Check

The problem of assessing model adequacy is historically old and has generated much research related to certain statistical models. The modeling process encompasses several steps that begin with a clear statement of the objectives of the model, assumptions about the model boundaries, appropriateness of the available data, design of the model structure, evaluation of the simulations, and providing feedback for recommendations and redesign processes.

Model testing is commonly used to prove the rightness of a model and the tests are typically presented as evidences to promote its acceptance and usability (Sterman, 2002). However, the understanding and acceptance of the wrongness and weaknesses of a model strengthens the modeling process. In systems thinking, the understanding that models are wrong and acceptance of the limitations of our knowledge is essential in creating an environment in which we can learn about the complexity of systems (Sterman, 2002). When (developing and) evaluating a model, one should incorporate important variables despite the foreseeable confines of our scientific knowledge and current modeling techniques. The adequate tests of the model should be designed to evaluate the model and identify weaknesses that need to be addressed. Adequate statistical analysis is an indispensable step especially for predictive models (Tedeschi, 2006).

The evaluation of model adequacy is an essential step of the modeling process because it indicates the level of precision and accuracy of the model predictions. This is an important phase either to build up confidence on the current model or to allow selection of alternative models. Forrester (1961) emphasized that the validity of a mathematical model has to be judged by its sustainability for a particular purpose; that means, it is a valid and sound model if it accomplishes what is expected of it. Shaeffer (1980) developed a methodological approach to evaluate models that consisted of six tasks: (a.) model examination, (b.) algorithm examination, (c.) data evaluation, (d.) sensitivity analysis, (e.) validation studies, and (f.) code comparison studies.

Montgomery (2005) stated that the decomposition of the variability in the observations through an analysis of variance (ANOVA) identity is purely algebraic relationship. He stated that however, the use of the partitioning to test formally for no differences in

treatment means requires that certain assumptions be satisfied. Specifically, these assumptions are that the model adequately describes the observations.

For example; $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

Where Y_{ij} is the response variable, μ is the overall mean, τ_i is the ith treatment effect and ε_{ij} is the random error.

In addition, that the errors are normally and independently distributed with mean zero and constant but unknown variance σ^2. Further, he stated that if these assumptions are valid, the analysis of variance procedure is an exact test of the hypothesis of no difference in treatment means. In practice, however, these assumptions will usually not hold exactly. Consequently, it is usually unwise to rely on the analysis of variance until the validity of these assumptions has been checked. Violations of the basic assumptions and model adequacy can be easily investigated by the examination of residuals. He concluded that examination of the residuals should be an automatic part of any analysis of variance and if the model is adequate, the residuals should be structure-less; that is, they should contain no obvious pattern.

Almini *et al.* (2006) stated that the four measures that predominate many measures of model adequacy are R^2 (coefficient of determination), R^2 – Adjusted (adjusted coefficient of determination), press (prediction error sum of squares) and R^2 - Prediction (prediction coefficient of determination). Dent and Blackie (1979) stated that several quantitative approaches involving statistical analysis were used to evaluate model adequacy and a linear regression between observed and predicted values is commonly used. They further stated that the hypothesis for the test is, the regression passes through the origin and has a slope of unity.

Almini *et al.* (2006) stated that although R^2 is a useful measure of goodness of fit, it has some limitations (Montgomery 2005): R^2 can increase by adding terms to the model, the magnitude of R^2 depends on the range of variability in the regressor variable, and R^2 does not measure the magnitude of the slope of the regression line. Draper and Smith (1998) stated that R^2 does not tell whether:

(a) the independent variables are a true cause of the changes in the dependent variable.
(b) omitted – variable bias exists, the correct regression was used.
(c) the most appropriate set of independent variables has been chosen.
(d) there is collinearity present in the data on the explanatory variables.
(e) and the model might be improved by using transformed versions of the existing set of independent variables.

Nonetheless, the use of the least-square method to derive a linear regression of observed on model-predicted values for model evaluation has little interest since the predicted value is useless in evaluating the mathematical model; therefore, the R^2 is irrelevant since one does not intend to make predictions from the fitted line (Mitchell, 1997). This conforms to the statement by Almini *et al.* (2006) that the limitations about R^2 mentioned by Montgomery (2005) call for the use of alternative measures. He further stated that the R^2 – adjusted is preferred to R^2 because its value only increases if a variable that is being added to the model reduces the residual mean square. Hence, unlike R^2, R^2 – adjusted does not increase if irrelevant terms are being added to the model.

Kavalseth (1985) proposed the use of resistant coefficient of determination (r_r^2), which uses the medians instead of the means and results in a coefficient that is highly resistant to outliers or extreme data points. Tedeschi (2006) stated that the ρ coefficient measures the linear relationship by measuring how far observations deviate from the

best-fit regression. In other words, it measures the amount of overall data variation due to between-subject variability. Therefore, it does not distinguish between situations in which observed and model-predicted values strongly agree and those in which a strong linear relationship exists but the measurements do not agree.

Masoud and Rahim (2010) compared the performance of the ordinary least squares (OLS), least median squares (LMS), least trimmed (sum of) squares (LTS), maximum estimation (ME) and minimum maximum estimate (MME), they found out that the LMS and the LTS where highly resistant to outliers in both the response variable and the explanatory variables (leverage points). They stated that R^2 is unreliable where there are outliers in the data sets and if used it should be considered as a reference method since only one outlier can change mean value so the OLS has a breakdown point of 0%, which is the smallest fraction of contamination that can cause an estimator of parameter to take on values arbitrarily far from an estimation of the parameter based on uncontaminated data. On the other hand, they said the median is not so sensitive; it is resistant to gross errors and has a 50% breakdown point.

Almini *et al.* (2006) stated that the prediction error sum of squares, PRESS and R^2-prediction which are useful measures of model adequacy proposed by Allen (1971), is used to evaluate the performance of the model in predicting the responses in new and future experiments. Moreover, PRESS is very useful when comparing models because a model with a small value of PRESS is generally better than a model with a large value of PRESS. Similarly, R^2-Prediction, which is based on PRESS, is a measure of the predictive performance of a model.

Lin (1989) introduced the concordance correlation coefficient (CCC) also known as reproducibility index. The CCC statistic is suitable for continuous variables whereas

the kappa statistic is appropriate for discrete variables (Cohen, 1960, 1968). The concordance correlation coefficient proposed by Lin (1989) is limited to two variables, he argued that, even though the agreement is often evaluated by using the Pearson correlation coefficient, the paired t-test, the least square analysis of slope (=1) and intercept (=0), the coefficient of variation, none of these can fully assess the desired reproducibility characteristics. He further explained that the Pearson correlation coefficient only measures precision of a linear relationship, not accuracy. Both the paired t-test and least squares analysis can falsely reject (accept) the hypothesis of high agreement when the residual error is very small (large). The coefficient of variation and the intraclass correlation coefficient often assume that the two readings by two observers are interchangeable. Barnhart *et al.* (2002) stated advantages of the CCC that:

(a) The index is based on the differences between the observations made by two observers on the same subject.

(b) It evaluates the agreement between two readings by measuring the variation from the 45^{o} line through the origin.

(c) The CCC has good intuitive interpretation because it includes components of both precision (degree of variation) and accuracy (degree of location or scale shift).

The limitation of the Lin (1989) CCC statistic is in the context of comparing two fixed observers or variables. Barnhart *et al.* (2002) developed an extension of the CCC called the overall concordance correlation coefficient, OCCC for assessing agreement among multiple fixed observers they stated that the OCCC turns out to be equivalent to the generalized CCC (Lin, 1989, 2000; King and Chinchilli, 2001) when the squared distance function is used. King *et al.* (2007) also developed an extension of Lin (1989),

CCC for repeated measures, where they stated that the $\rho_{c,rm}$ is an aggregated index. They stated that $\rho_{c,rm}$ that is the repeated measures CCC that can handle either few or many repeated measurements, has a variance that can be estimated in a straightforward manner by U-statistic methodology, and performs well with small samples. However, their approach was still based on two responses say X and Y in the presence of repeated measurement. The CCC was also used by Vonesh *et al.* (1996) in assessing the goodness of fit in generalized linear and nonlinear mixed-effect models.

Tedeschi (2006) presented the modeling efficiency statistic (MEF), in which he stated that the method is similar to ρ, which is interpreted as the proportion of variation explained by the fitted line whereas the MEF statistic is the proportion of variation explained by the line $Y = f(X_1, \ldots, X_p)$. He further stated that the statistic has been extensively used in hydrology models, but can certainly be used in biological models. He concluded that in a perfect fit, both the ρ (correlation coefficient) and the MEF would result in a value equal to one and that the upper bound of MEF is one and the (theoretical) lower bound of MEF is negative infinity. Loague and Green (1991) stated that if MEF is lower than zero; the model predicted values are worse than the observed mean. The MEF statistic may be used as a good indicator of goodness of fit (Mayer and Butler, 1993).

The mean square error (MSE) and the mean square error of prediction (MSEP) are probably the most common and reliable estimate to assess the precision and measure the predictive accuracy of a model respectively. Tedeschi (2006) stated that the MSE assesses the precision of the fitted linear regression using the difference between observed values (Y_i) and regression-predicted values ($\hat{Y}_i$). While MSEP consist of the difference between observed values (Y_i) and model – predicted values ($f(X_i,\ldots,X_p)_i$)

rather than regression-predicted value. Lehmann and Casella (1998), stated that two or more statistical models may be compared using their MSEs as a measure of how well they explain a given set of observations: the unbiased model with the smallest MSE is generally interpreted as best explaining the variability in the observations and is called the best unbiased estimator or MVUE (minimum variance unbiased estimator). In addition, DeGroot (1980), stated that both linear regression techniques such as analysis of variance estimate the MSE as part of other analysis and use the estimated MSE to determine the statistical significance of the factors or predictors under study. The goal of experimental design is to construct experiments in such a way that when the observations are analyzed, the MSE is close to zero relative to the magnitude of at least one of the estimated treatment effects.

Some drawbacks of the MSE and MSEP analysis are notable. Mitchell and Sheehy (1997) stated that the MSEP (or its root) removes the negative sign and weights the deviation by their squares, thus giving more influence to larger data points, and does not provide any information about model precision.

Tedeschi (2006) stated that when each pair of the data ($f(X_1,\ldots, X_p)_i$, Y_i) is mutually independent (that is, the outcome of one pair does not depend on the outcome of another pair), and the model is independent (that is, the parameters of the model were derived from independent experiments and were not adjusted to the current experiment being predicted). The MSEP estimate is a reliable measure of model accuracy. Nonetheless, the reliability will decrease as *n* decreases. Therefore, the estimate of MSEP variance has an important role. Tedeschi (2006) further stated that a different scenario arises when the model parameters are adjusted to the data set; the real MSEP will be underestimated because the model will reproduce more closely the data that have been modeled than it would for the entire population of interest.

Berger (1985), stated that the mean squared error (MSE) conflicts with losses derived from utility functions and is convex everywhere, whereas most losses derived from utility theory have concave tails (and may be concave everywhere). There are however, some scenarios where mean squared error can serve as a good approximation to a loss function occurring naturally in an application. Sergio and Joan (2001) stated that like variance, MSE has the disadvantage of heavily weighting outliers. They said it is because of the squaring of each term, which effectively weights large errors more heavily than small ones. This property, undesirable in many applications, has led researchers to use alternatives such as the mean absolute error, or those based on the median.

Huang and Zhang (2008) stated that for testing a parametric versus nonparametric covariate effect, the likelihood ratio test (LRT) is a natural choice. The LRT has been popular in situations where we need to compare two nested models. However, extending the LRT to testing the adequacy of a parametric covariate effect is not straightforward. A considerable amount of work has been done in constructing likelihood ratio based test statistics for comparing parametric versus nonparametric covariate effects. Depending on how the nonparametric alternatives were specified and what types of smoothing techniques were used, a number of versions of likelihood ratio based testing procedures have been proposed.

By converting the complicated nonparametric regression in exponential families to a normal theory regression problem, Brown *et al*. (2010) developed a method of fitting regression models in exponential families. They have proposed an approach that uses a mean matching variance stabilizing transformation on the data. Using an R-square measure, a goodness of fit test was developed by Cameron and Windmeijer (1997). The authors have applied this method to a class of exponential family regression models,

including logit, probit, Poisson, Geometric, Gamma and Exponential. The R-square measure considered here is based on the Kullback-Leibler divergence which measures the proportionate reduction in uncertainty due to the inclusion of predictor variables. Pan and Lin (2005) developed methods for checking the adequacy of generalized linear mixed models based on the cumulative sums of residuals over covariates or predicted values of the response variable developed by Su and Wei (1991), which can handle the case of non-replication. This test is designed to detect the inaccuracy of the mean function, even if the variance of the response variable is misspecified. Claeskens and Hjort (2008) discussed order selection tests and Neyman smooth-type tests to assess model adequacy in general and in particular for the generalized linear model. Considering a generalized partially linear model, Hardle *et al*. (1998) developed a test statistic to decide between a parametric and a semi-parametric model. The generalized linear model is perturbed here with a nonlinear function, and using two examples the authors tested the null hypothesis of a parametric model versus the semi-parametric alternative. Addressing the problem of over dispersion and under dispersion in count data, Sellers and Shmuelli (2010) introduced a method of fitting Poisson models. This method was based on the Conway-Maxwell-Poisson distribution. Using the exponential family properties, the leverages, Deviance and Pearson residuals were also computed and diagnostic methods presented. Several approaches for assessing the goodness of fit of logistic regression models have been proposed.

Tsiatis (1980) proposed a goodness of fit test based on partitioning the space of covariates into distinct regions, and using a score statistic for the coefficients for the grouping variable. A strategy for determining the groups was not indicated. Hosmer and Lemeshow (1980) considered a goodness of fit test for the multiple logistic regression models by using the chi-square test statistic for contingency tables. The

method used different ways of calculating the expected frequencies with some predefined grouping strategies. Hosmer *et al*. (1988) developed another goodness of fit statistic for logistic models when the estimated probabilities are small.

Many procedures for selecting variables have been proposed. Often they do not lead to the same solution when applied to the same problem. There seems to be no consensus among modelers as to the advantages and disadvantages of the various procedures. All procedures are criticized, but for different reasons (Sauerbrei *et al*., 2007). According to Royston and Sauerbrei (2008) there are two main types of strategy for variable selection. Sequential strategies, such as forward elimination, stepwise or backward elimination procedures, are based on a sequence of tests of whether a given variable should be added to the current model or removed from it, or selection should stop. A nominal significance level for each of these tests is chosen in advance and largely determines how many variables end up in the model. In the second type, the all-subsets strategies, all 2^k possible models are fitted and the best model is chosen by optimizing an information criterion derived from the likelihood function such as the Akaike information criteria (AIC) and Bayesian information criteria (BIC).

Royston and Altman (1994) used the deviance and likelihood methods in selecting the best fractional polynomial model where a model with a small deviance or large log likelihood value is best. They extended the deviance method to deviance difference which is the difference in deviance of two fractional polynomials of varying degrees and they term the difference as gain (G). A model with the largest gain is considerable the best fit.

In this research the model adequacy model to be adopted are those used by Royston and Altman (1994) which include, deviance difference or gain (G), log likelihood function, AIC and BIC the reason is that these measures are adequate and simple to interpret.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter seeks to explain the methodology used in the research. The study going to present the fractional polynomial for normal error regression models, the median method for categorizing continuous covariates and the deviance method for parameter estimation and checking adequacy of the fitted model.

3.1 Normal Error Model

For an individual with response *y*, the multiple linear regression model with normal errors $\varepsilon \sim N(0,\sigma^2)$ and covariate vector $X = (x_1, x_2,...,x_k)$ with *k* variables, may be written as $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon = X\beta + \varepsilon$ (3.1)

The linear predictor or 'index', $\eta = \beta_0 + X\beta$ is an important quantity in multivariable modeling and equation (3.1) is called the normal error model (Royston and Sauerbrei, 2008).

3.2 Fractional Polynomials

A fractional polynomial of degree *m* is defined to be the function

$$\phi_m(X;\xi,P) = \xi_o + \sum_{j=1}^{m} \xi_j X^{(p_j)} \quad (3.2)$$

where m is a positive integer, p = $(p_1, p_2,..., p_m)$ is a real-valued vector of powers with$p_1 < \ldots < p_m$ and $\xi = (\xi_0, \xi_1,..., \xi_m)$ are real valued-coefficients. The round bracket notation signifies the Box-Tidwell transformation (Royston and Altman, 1994),

$$X^{(p_j)} = \begin{cases} X^{p_j} if p_j \neq 0 \\ \ln X \; if p_j = 0, \end{cases} \quad (3.3)$$

as distinct from the more familiar Box-Cox transformation of a response variable (Box and Cox, 1964), namely $Y^{(\lambda)} = (Y^{\lambda} - 1)/\lambda$ for $\lambda \neq 0$, $Y^{(0)}$ = ln*Y*. A conventional polynomial of degree *m* has $p_j = j$ for $j = 1, \ldots, m$ and $\xi_m \neq 0$.

Royston and Altman (1994) gave an extension of equation (3.3) to the case of equal powers, that is $m > 1$ and $p_i = p_j$ for at least one pair of distinct indices (i, j), $1 \leq i, j \leq m$

For$m = 2$, $(i, j) = (1, 2)$ and p = (p_1, p_1), we have

$$\phi_2(\mathbf{X};\xi,\mathbf{p}) = \xi_0 + (\xi_1 + \xi_2)\mathbf{X}^{(p_1)} \tag{3.4}$$

a fractional polynomial of degree 1, not 2. Hence, the lim $\phi_2(\mathbf{X};\xi,\mathbf{p}) = \xi_0 + (\xi_1 + \xi_2)\mathbf{X}^{(p_1)}$ it as p_2 tends to p_1.

$$\xi_0 + \xi_1\mathbf{X}^{(p_1)} + \xi_2\mathbf{X}^{(p_1)}(\mathbf{X}^{(p_2 - p_1)} - 1)/(p_2 - p_1) \tag{3.5}$$

Above expression (3.5) is obtained from standard form

$$\phi_2(\mathbf{X};\xi^*,\mathbf{p}) = \xi^*{}_0 + \xi^*{}_1\mathbf{X}^{(p1)} + \xi^*{}_2\mathbf{X}^{(p2)} \tag{3.6}$$

where :

$\xi_o = \xi_o^*, \xi_1 = \xi_1^* + \xi_2^*, and \xi_2 = \xi_2^*(p_2 - p_1).$

$\lim_{p_2 \to p_1} \mathbf{X}^{(p_2 - p_1)-1} = \mathbf{X}^{-1} = \ln \mathbf{X}$ hence, equation (3.5) becomes (3.7) as

$$\xi_0 + \xi_1\mathbf{X}^{(p_1)} + \xi_2\mathbf{X}^{(p_1)} \ln \mathbf{X} \tag{3.7}$$

which is a three parameter family of curves. The generalization of equation (3.7) for $m>2$ and $p_1 = \ldots = p_m$, can be expressed as:

$$\xi_0 + \xi_1\mathbf{X}^{(p_1)} + \sum_{j=2}^{m} \xi_j\mathbf{X}^{(p_1)}(\ln \mathbf{X})^{j-1} \tag{3.8}$$

For arbitrary powers $p_1 = \ldots = p_m$, equation (3.2) is combined with equation (3.8) and set $H_0(\mathbf{X}) = 1$, $p_0 = 0$ an extended form;

$$\phi_m(\mathbf{X};\xi,\mathbf{p}) = \sum_{j=0}^{m} \xi_j H_j(\mathbf{X}) \tag{3.9}$$

where for $j = 1, \ldots, m$

$$\boldsymbol{X}^{(p_j)} = \begin{cases} \boldsymbol{X}^{(p_j)} if p_j \neq p_{j-1}, \\ H_{j-1}(\boldsymbol{X}) \ln \boldsymbol{X} \, if p_j = p_{j-1}. \end{cases} \quad (3.10)$$

Royston and Altman (1994) stated that the recurrence relation in equation (3.10) for H_j (**X**) in terms of H_{j-1}(X)when$p_j = p_{j-1}$ is a representation of the functional part of equation (3.11) and it makes computer evaluation of fractional polynomials straight forward. H_j(X) can be written as a vector function **H** (**X**) = (H_0, H_1, . . .,H_m) and equation (3.12) and (3.13) are the full (and most concise) definition of a fractional polynomial of degree *m*.

3.3 Fractional Polynomials with Multiple Covariates

Suppose that prior reasoning or preliminary modeling have identified a set of *k* continuous covariates **X** and *c* categorized covariates **Z** that are wanted in a final model. Then, a satisfactory combination of power vectors $\tilde{\mathbf{p}}_1,\ldots,\tilde{\mathbf{p}}_k$ and degrees $m_1,\ldots,m_k$ are what to be found. The adopted procedure is iterative which was proposed by Royston and Altman (1994), it is closely related to backfitting method of variable selection from a model. The ordering of the covariates is random however; the procedure steps are as follows;

1. Fit the multiple regression model whose linear predictor is

$$\eta_1 = \xi_0 + \sum_{j=1}^{m_1} \xi_{1j} \mathbf{H}_{1j}(\mathbf{X}_1) + \sum_{j=2}^{k} \beta_j \mathbf{X}_j + \sum_{j=1}^{c} \gamma_j \mathbf{Z}_j,$$

   where all the relationships between *Y* and the covariates are initially taken as straight line, except possibly that between *Y* and $\mathbf{X}_1$.

2. Fix the functions $H_{1j}(X_1)$ (but not the coefficients $\xi_{11},\ldots,\xi_{1m_1}$) and fit the model

$$\eta_2 = \xi_0 + \sum_{j=1}^{m_2} \xi_{2j}\mathbf{H}_{2j}(\mathbf{X}_2) + \sum_{j=1}^{m_1} \xi_{1j}\mathbf{H}_{1j}(\mathbf{X}_1) + \sum_{j=3}^{k} \beta_j \mathbf{X}_j + \sum_{j=1}^{c} \gamma_j \mathbf{Z}_j ,$$

re-estimating the ξ_{1j}, β_j and γ_j and obtain m_2, $\tilde{p}_2$ and $\mathbf{H}_2$ ($\mathbf{X}_2$).

3. Continuing in the same manner of step two, the first iteration will be complete when η_k is reached. The model will comprise of fractional polynomial terms only; it will no longer include the term $\sum \beta_j \mathbf{X}_j$.

The three steps given above are for the first iteration, the second iteration follows the same steps. Convergence is achieved when the fractional polynomial functions $\mathbf{H}_1(\mathbf{X}_1)$, . . .,$\mathbf{H}_k(\mathbf{X}_k)$ does not change from one iteration to the next.

3.4 Deviance Measure of Model Fitness

The deviance *D* is one of the measures of assessing the adequacy of model fit. Royston and Altman (1994) used *D* in selecting variables that are to be included in a fitted model and in determining the adequacy of the final fitted model. The *D* method uses the likelihood based on the assumption that the fitted model is by maximum likelihood. The log-likelihood can be expressed in terms of the mean parameter μ and the log-likelihood ratio which is the scaled deviance expressed as

$$\mathbf{D^*(y;\hat{\mu})} = -2\{l(\hat{\mu}; y) - l(\mu_{\max}; y)\} \qquad (3.11)$$

where, $l(\hat{\mu}; y)$ is the log-likelihood under the model; $l(\mu_{\max}; y)$ is the log-likelihood under the maximum achievable (saturated) model.

For generalized linear models, the scaled deviance can be expressed as

$$\mathrm{D}^*(y; \hat{\mu}) = \frac{1}{\phi}\mathrm{D}(\mathbf{y}; \hat{\mu}) \qquad (3.12)$$

where, D($y; \hat{\mu}$)is the residual deviance for the model and it's the sum of individual deviance contributions and ϕ is the dispersion parameter.

Royston and Altman (1994) stated that for a given *m*, the best power vector $\tilde{\mathbf{p}} = (\tilde{p}_1,...,\tilde{p}_m)$ is that associated with the model with the highest likelihood or equivalently with the lowest deviance *D*. Thus $\tilde{\mathbf{p}}$ may be regarded as the maximum likelihood estimate (MLE) of **p** over the restricted parameter space based on *S*. simply they defined *D* = – 2(log-likelihood) which does not include the log-likelihood of the saturated (maximum achievable) model.

Suppose the elements of **p** are allowed to vary continuously, rather than being restricted to *S*. Then $\phi_m(\mathbf{X};\mathbf{p})$ is a non-linear model with parameters **p** and ξ. Let $\hat{\mathbf{p}}$ be the full MLE of **p**. Using an apparent expression,

$$D\ (m, \mathbf{p}) - D\ (m, \hat{\mathbf{p}}\ ) \sim \chi^2_{df=m} \qquad (3.13)$$

meaning that the quantity of equation (16) asymptotically has a chi-square distribution with *m* degrees of freedom (DF). Further, as $D(m, \tilde{p}) \geq D(m, \hat{p})$ the statistic $D(m, \mathbf{p}) - D(m, \tilde{p}\ )$ provides an (asymptotically conservative) test of a given value of **p**. Royston and Altman (1994) stated that it may be used as a guide to assess the adequacy of conventional polynomial of degree *m* against fractional polynomial alternatives of the same degree.

Also, it is convenient to use the deviance $D(1, 1)$ associated with the straight line model $\phi_1(\mathbf{X}; 1)$ that is $m = 1$, $p = 1$ as a base line for reporting the deviances of other models. Hence, a *gain G* for a model can be defined on a set of data as the deviance for $\phi_1(\mathbf{X}; 1)$ minus that for the model in question:

$$G = G(m, \mathbf{p}) = D(1, 1) - D(m, \mathbf{p}) \qquad (3.14)$$

Since *G* moves in the opposite direction to *D*, a *larger* gain indicates a *better* fit (Royston and Altman, 1994).

3.5 Working Rule for using Deviance

Royston and Altman (1994) presented four working rules for using the deviance measure of model selection and adequacy test. These rules are as follows;

1. Specifically, when $m = 1$, the criterion $D(1, 1) - D(m, \hat{p}) > \chi^2_{1;0.90}$ (the 90th percentile of chi-square with 1 DF, that is 2.7) gives a monotonic test of linearity with a significance level of about 10% (that is H_0: $p = 1$ vs H_1: $p \neq 1$.
2. Models with values of $\mathbf{p}$ should be chosen such that $D(m, \mathbf{p}) - D(m, \tilde{\mathbf{p}}) < \chi^2_{m;0.90}$ as the best fitting among those of degree *m*.
3. In deciding whether model(s) with degree *m* are adequate or whether degree $m + 1$ is required, two extra parameters (a power and a regression coefficient) are estimated when *m* is increased by 1. Therefore, $D(m, \hat{\mathbf{p}}) - D(m + 1, \hat{\mathbf{p}})$ is asymptotically distributed as chi-square on 2 DF when the degree *m* is adequate. Hence, a suggested criterion as a rule for preferring models with degree $m + 1$ to those with degree *m* is $D(m, \mathbf{p}) - D(m, \tilde{\mathbf{p}}) > \chi^2_{m;0.90}$.
4. For normal error models with small samples of say $n < 100$, appropriate critical points from F_{v_1,v_2}-distribution should be used instead of the χ^2-distribution for comparing models with degree *m*, each with a constant (ξ_0), the DF for *F* are $v_1 = m$ and $v_2 = n - 2m - 1$; for comparing degree *m* with $m + 1$, the DF for *F* are $v_1 = 2$ and $v_2 = n - 2m - 3$.

3.6 Median Method of Categorizing Continuous Covariates

The median method will be applied in determining a cutpoint for categorizing the continuous covariates in the data to be used in the research work. The procedure is described below.

1. Obtain the median of each variable to be categorized
2. Group the variable into two base on the median where a data point is classified say low if it is less than or equal to the median and high if greater than the median. Let X^* be the median value then;

   $X \leq X^*$ is categorized as 0 and 1 otherwise.

In this research we propose a different median categorization method. The procedures are presented below.

1. First ordered the data for the covariate of interest.
2. Split the data for the covariate X of interest into four approximately equal subsets (X_1, X_2, X_3, X_4).
3. Obtain the median of each subset $_{(i= 1, 2, 3, 4)}$ Med_1, Med_2, Med_3 and Med_4.
4. . Obtained J = mean of Med_2 and Med_3
5. Combine the four subsets into one covariate.
6. Group the covariate (X) into more than two groups base on the median (J) where a data point is classified say low if it is less than the median, equal to the median and high if greater than the median. (i.e $X<J$ is categorized as 1, $X = J$ or , $X = J$ *to some interval* is categorized as 2 and $X>J$ or $X \geq$ given number is categorized as 3*)* depending on the number of groups one may have.

Strength and weakness of our median method

1. Simplicity and easier to compute.

2. It reduces the chance of spurious relationship between variable of interest (MacCallum *et al*., 2002).

The weakness is that, researchers should avoid arbitrary selection of cutpoint for grouping with very wide or too narrow confidence intervals.

Note This study adopts grouping of the covariate to be polytomous and not dichotomous that is having more than two groups. Since X was divided into four subsets the grouping was done at four intervals between the lower class boundary and the upper class boundary after grouping points that are $\leq J$ as 1.

The justification for our median categorization algorithm is to improve on the existing one and compared to uncategorized covariate when both are applied to a fractional polynomial regression analysis. From literature it was observed that categorization of continuous variable have been criticized extensively (see Lagakos, 1988, Faraggi and Simon, 1996, Austing and Brunner, 2004, Royston *et al*., 2005, etc.). The existing median categorization method group a continuous variable into two but ours is more than two, this is done based on Breslow and Day (1980) view that "when the true risk increases (or decreases) monotonically with the level of the variable of interest, the apparent spread of risk will increase with the number of groups used. With just two groups one may seriously underestimate the extent of variation in risk." This means that when individuals are divided into just two groups, considerable variability may be subsumed within each group. The idea of the algorithm developed was based on Faraggi and Simon (1996) algorithm which is an approach based on twofold cross-validation. Their method was used to dichotomize the observations but these study group observations into more than two groups in order to reduce the chances of spurious relationship (MacCallum *et al*., 2002), avoid arbitrary selection of cutpoint for

grouping and obtain cutpoint which does not have wide confidence interval and produce estimates whose confidence interval will not be too narrow.

3.7 Data Description

The data used for this research work is an experimental design data for determining the combined effect of nitrogen fertilizer and manure on the yield of two cowpea varieties. The experiment was designed as a 2x3x4 randomized complete block design. The fertilizer rates are 15kg/ha, 30kg/ha, 45kg/ha and 60kg/ha, three manure rates which are 5kg/ha, 10kg/ha and 15kg/ha, and two varieties of cowpea. The fertilizer rates are coded as 1, 2, 3 and 4, the manure rates are coded as 1, 2 and 3 while the qualitative variable variety is coded 1 and 2. The data was obtained from Data Processing Unit, Institute of Agricultural Research, Ahmadu Bello University, Zaria. This study analyzed the data by fitting a multiple fractional polynomial model (MFP) with fertilizer, manure and varieties as factors .The data is presented in the appendix. And the Statistical package used for the purpose of this research is STATA 9.1.

CHAPTER FOUR

ANALYSIS AND DISCUSSION OF RESULTS

4.0 Introduction

This chapter seeks to present results from the analyses of data and interpretation of results for the normal error fractional polynomial regression for the data sets described in chapter three and presented in the appendix. The data set was analyzed as a generalized linear model (GLM), using two different approaches.

4.1: Generalized Linear Model Multivariable Fractional polynomial Regression Results;

Table 4.1 presents the selection algorithm for the variables i.e fertilizer rate, manure rate and cowpea variety effects on cowpea yield where fertilizer and manure rates are quantitative (continuous) variable and it is observed that the algorithm converged after 2 cycles with a final deviance of 130.202 while the deviance for the model with all terms untransformed is 127.968 which is less than the final deviance when the terms in the model has been transformed.

Table 4.1: Effect of Fertilizer and Manure on Variety of Cowpea Yield

Variables	Models	Deviance	Dev. diff.	P-value	Powers
Fertrate	null vs. FP1	130.202	6.203	0.013*	0 vs. 1
	Lin vs. FP3	129.345	0.857	0.355	1 vs. 3
	Final	130.202			
Manure	null vs. lin.	130.202	25.845	0.0001*	0 vs. 1
	Final	130.202			
Variety	null vs. lin.	127.968	2.233	0.135	0 vs. 1
	Final	130.202			

Algorithm converged after 2 cycles. Untransformed model terms deviance = 127.968

Also, from the Table it can be observed that variety as a factor is not significant; hence, it is excluded from the model because among all contending fractional polynomial and conventional polynomials none of the fitted powers was significant. While for fertilizer and manure rates are significant at $p = 1$, their highest deviance difference or gain G are 6.203 and 25.845 respectively with a p-value of 0.013 and 0.0001 respectively which are less than $\alpha = 5\%$ significance level. This implies that fertilizer rate and manure rate significantly improve cowpea yield at $p = 1$ which is a linear fit while cowpea variety which is not significant and not included in the model does not improve the yield of cowpea.

Table 4.2 presents the parameter estimation results from the generalized linear multiple fractional polynomial regression for cowpea yield with fertilizer and manure rates. It can be observed that approximately 35.7 and 10 was subtracted from fertilizer and manure rates respectively before transformation to improve the scaling of the regression coefficients. The coefficients estimated for fertilizer rate (0.0073) and manure rate (0.0649) are significant since their p-values of 0.013 and 0.0001 respectively are less than $\alpha = 5\%$ significant level. Hence, the final model formulation is given below as:

$$Y_{cp} = 0.4148 + 0.0073FR + 0.0649MR \qquad (4.1)$$

where, Y_{cp} is cowpea yield, FR is fertilizer rate and MR is the manure rate.

Table 4.2: Parameter Estimation for Cowpea Yield from Fertilizer and Manure Effects

Variables	Coefficients	Std. Error	Z-Value	P-Value	[95% Conf. Interval]
Fertrate	0.0073444	0.0029477	2.49	0.013	0.001567 - 0.131218
Manure	0.0649062	0.0121089	5.36	0.0001	0.041173 - 0.088639
Constant	1.324375	0.0494346	26.79	0.0001	1.227485 - 1.421265

Fertilizer rate transformation = fertrate –37.5

Manure rate transformation = manure – 10

Log likelihood = – 65.1008

AIC = 1.418766

BIC = – 402.6663

Approximately 35.7 and 10 (the median of fertilizer and manure) was subtracted from Fertilizer and Manure Rates respectively before transformation to improve the scaling of the regression coefficients.

Also, from the same Table 4.2 it can be observed that the adequacy of the final model measured by the log likelihood (– 65.1008), AIC (1.4188) and BIC (–402.67) shows that the final fitted model is adequate because the values of each of the measure statistics are small.

Table 4.3 presents the selection algorithm for the grouped (dichotomous) variables fertilizer rate, manure rate and cowpea variety effects on cowpea yield where fertilizer and manure rates are quantitative (continuous) variable. Therefore, it is observed that the algorithm converged at zero cycles with a deviance for the model with all terms untransformed is 137.388.

Table 4.3: Effect of Grouped Fertilizer and Manure on Variety of Cowpea Yield (Dichotomous)

Variables	Models	Deviance	Dev. diff.	P-value	Powers
Fertrate	null vs.FP1	139.415	4.385	0.036*	0 vs 1
	Final	139.415			
Manure	null vs. lin.	139.415	17.748	0.001*	0 vs 1
	Final	139.415			
Variety	null vs. lin.	137.388	2.027	0.155	0 vs 1
	Final	139.415			

Algorithm converged after 2 cycles. Untransformed model terms deviance = 137.388

Also, from Table 4.3 it can be observed that variety as a factor is not significant; hence, it is excluded from the model because among all contending fractional polynomial and conventional polynomials, none of the fitted powers were significant. On the other hand, fertilizer and manure rates are significant at $p = 1$, their highest deviance difference or gain G are 4.385 and 17.748 respectively with a p-value of 0.013 and 0.0001 respectively which are less than $\alpha = 5\%$ significance level. This implies that fertilizer rate and manure rate significantly improve cowpea yield at $p = 1$ which is a linear fit while cowpea variety which is not significant and not included in the model does not improve the yield of cowpea.

Table 4.4 presents the parameter estimation results from the generalized linear multiple fractional polynomial regression for cowpea yield with grouped fertilizer and manure rates. The coefficients estimated for fertilizer rate (0.21625) and manure rate (0.47813)

are significant since their p-values of 0.036 and 0.0001 respectively are less than $\alpha = 5\%$ significant level. Hence, the final model formulation is given below as:

$$Y_{cp} = 1.05688 + 0.21625FR + 0.47813MR \quad (4.2)$$

where, Y_{cp} is cowpea yield, FR is fertilizer rate and MR is the manure rate.

Table 4.4: Parameter Estimation for Cowpea Yield from Grouped Fertilizer and Manure Effects (Dichotomous)

Variables	Coefficients	Std. Error	Z-Value	P-Value	[95% Conf. Interval]
Fertrate	0.21625	0.1037292	2.08	0.037	0.0129445 - 0.419555
Manure	0.478125	0.1100214	4.35	0.0001	0.2624870-0.6937630
Constant	1.056875	0.0820051	12.89	0.0001	0.8961479 - 1.217602
Log likelihood = – 69.70745931					
AIC = 1.514739					
BIC = – 400.4686					

Also, from the same Table 4.4 it can be observed that the adequacy of the final model measured by the log likelihood (– 69.7075), AIC (1.51474) and BIC (–400.47) shows that the final fitted model is adequate because the values of each of the measure statistics are small.

Table 4.5 presents the selection algorithm for the grouped (polytomous) variables fertilizer rate, manure rate and cowpea variety effects on cowpea yield where fertilizer and manure rates are quantitative (continuous) variable and it is observed that the algorithm converged after 2 cycles with a final deviance of 137.475 while the deviance for the model with all terms untransformed is 135.407 which is less than the final deviance when the terms in the model has been transformed.

Table 4.5: Effect of Grouped Fertilizer and Manure on Variety of Cowpea Yield (polytomous)

Variables	Models	Deviance	Dev. diff.	P-value	Powers
Fertrate	null vs.FP1	137.475	6.325	0.012*	0 vs 1
	Final	137.475			
Manure	null vs. lin.	137.475	18.078	0.001*	0 vs 1
	Final	137.475			
Variety	null vs. lin.	135.407	2.069	0.150	0 vs 1
	Final	137.475			

Algorithm converged after 2 cycles. Untransformed model terms deviance = 135.407

Also, from the Table 4.5 it can be observed that variety as a factor is not significant; hence, it is excluded from the model because among all contending fractional polynomial and conventional polynomials none of the fitted powers were significant. While fertilizer and manure rates are significant at $p = 1$, their highest deviance difference or gain G are 6.345 and 18.078 respectively with a p-value of 0.012 and 0.0001 respectively which are less than $\alpha = 5\%$ significance level. This implies that fertilizer rate and manure rate significantly improve cowpea yield at $p = 1$ which is a linear fit while cowpea variety which is not significant and not included in the model does not improve the yield of cowpea

Table 4.6 presents the parameter estimation results from the generalized linear multiple fractional polynomial regression for cowpea yield with grouped fertilizer and manure rates. It can be observed that approximately 1.75 was subtracted from fertilizer rate

before transformation to improve the scaling of the regression coefficients. The coefficients estimated for fertilizer rate (0.15583) and manure rate (0.47813) are significant since their p-values of 0.0012 and 0.0001 respectively are less than $\alpha = 5\%$ significant level. Hence, the final model formulation is given below as:

$$Y_{cp} = 0.89230 + 0.15583FR + 0.47813MR \quad (4.3)$$

where, Y_{cp} is cowpea yield, *FR* is fertilizer rate and *MR* is the manure rate.

Table 4.6: Parameter Estimation for Cowpea Yield from Grouped Fertilizer and Manure Effects (polytomous)

Variables	Coefficients	Std. Error	Z-Value	P-Value	[95% Conf. Interval]
Fertrate	0.1558333	0.0619223	2.52	0.012	0.0344679 - 0.2771987
Manure	0.2478125	0.1089154	4.39	0.0001	0.2646545- 0.6915953
Constant	1.165000	0.0628824	18.53	0.0001	1.0417530 - 1.288247

Fertilizer rate transformation = fertrate –1.75

Log likelihood = – 68.73754

AIC = 1.494532

BIC = – 400.949

Approximately 1.75 and 2 was subtracted from Fertilizer and Manure Rates respectively before transformation to improve the scaling of the regression coefficients.

Also, from the same Table 4.6 above, it can be observed that the adequacy of the final model measured by the log likelihood (– 64.7777), AIC (1.4120) and BIC (–402.81) shows that the final fitted model is adequate because the values of each of the measure statistics are small.

4.2: Generalized Linear Model Fractional polynomial Regression Results;

Table 4.7 presents algorithms selection for variables fertilizer and manure rates on cowpea yield where fertilizer and manure rates are quantitative (continuous) variable and it is observed that algorithm converged with the deviance for the model with all terms of 129.33. The parameter estimation results from the generalized linear model fractional polynomial regression for cowpea yield with fertilizer and manure rates. It can be observed that, approximately 10 was subtracted from manure rate, (x^2-14.0625) and (x^3-52.734) from fertilizer rate, where x = fertilizer rate/10. The coefficients estimated for manure rate is (0.0649062), fertilizer rates 1 and 2 are (0.0029155) and (0.0020366) respectively. It has been observed that only manure rate is significant at p-value of 0.0001, while fertilizer rates are not significant at 5% significance level. Hence the final model formulation is given below as:

$$Y_{cp}=0.73832+0.0649MR-0.00204x^2 - 0.00292x^3. \qquad (4.4)$$

Where Y_{cp} is cowpea yield, x is fertilizer rate / 10 and MR is manure rate.

Table 4.7: Effect and Parameter Estimation for Cowpea Yield from Fertilizer and Manure (ungrouped)

Variables	Coefficients	Std Error	Z-value	P-value	[95% conf. interval]
Fertrate 1	-0.0029155	0.0282081	-0.10	0.918	-0.0582 – 0.0524
Fertrate 2	0.0020366	0.0043960	0.47	0.641	0.0065 – 0.0106
Manure	0.0649062	0.0121197	5.36	0.001*	0.0412 – 0.0887
Constant	1.268136	0.0782632	16.20	0.001*	1.1147 – 1.4215

Deviance: 129.33 Best powers fertrate models fit are 2, 3.

Log likelihood = -64.6667

AIC = 1.4301

BIC = -398.2984

Approximately 10 was subtracted from manure rate, (x^2-14.0625) and (x^3-52.734) from fertilizer rate, where x = fertilizer rate/10, to improve the scaling of the regression coefficients.

Also from same Table 4.7 above, it can be observed that the adequacy of the model measured by the Log likelihood (-64.6667), AIC (1.4301) and BIC (-398.29), shows that the fitted model is adequate because the value of each of the measure statistics is small.

Table 4.8 presents the selection algorithm for the variables manure and fertilizer rates on cowpea yield, where manure and fertilizer rates are quantitative (continuous) variable and it is observed that algorithm converged with the deviance for the model with all terms of 130.17. The parameter estimation results from the generalized linear model fractional polynomial regression for cowpea yield with manure and fertilizer rates. It can be observed that approximately 37.5 was subtracted from fertilizer rate,(x^{-2}-*1)* and *($x^{-2}lnx$)* from manure rate, where *x* = manure rate/10. The coefficients estimated for fertilizer rate is (0.0073444), manure rates 1 and 2 are (-0.790328) and (-0.7318464) respectively. It has been observed that the manure rate 1 and fertilizer rate are significance at p-values of 0.024 and 0.013 respectively, while manure rate 2 is not significant at 5% significant level. Hence the final model formulation is given below as:

$$Y_{cp} = 1.7939 + 0.0073FR - 0.7903x^{-2} - 0.7318x^{-2}\ln x. \qquad (4.5)$$

where Y_{cp} is cowpea yield, *x* is manure rate / 10 and *FR* is fertilizer rate.

Table 4.8: Effect and Parameter Estimation for Cowpea Yield from Manure and Fertilizer (ungrouped)

Variables	Coefficients	Std Error	Z-value	P-value	[95% conf. interval]
Manure 1	-0.790328	0.3490893	-2.26	0.024*	-1.474530 - -0.1061257
Manure 2	-0.7318464	0.4033906	-1.81	0.070	-1.522478 – 0.0587847
Fertrate 1	0.0073444	0.0029632	2.48	0.013*	0.0015366 – 0.0131523
Constant	1.3359380	0.0860746	15.52	0.001*	1.1672340 – 1.5046410

Deviance: 130.17 Best powers manure rate models fit are -2, 2.

Log likelihood = -65.08666028

AIC = 1.439305

BIC = -398.1084

Approximately 37.5 was subtracted from manure rate, (x^{-2}-1) and ($x^{-2}lnx$) from fertilizer rate, where x = fertilizer rate/10, to improve the scaling of the regression coefficients

Also from same Table 4.8 above, it can be observed that the adequacy of the model measured by the Log likelihood (-65.08666028), AIC (1.439305) and BIC (-398.1084), shows that the fitted model is adequate because the value of each of the measure statistics is small.

Table 4.9 presents algorithms selection for variables grouped fertilizer and manure rates on cowpea yield where fertilizer and manure rates are quantitative (continuous) variable and it is observed that algorithm converged with the deviance for the model with all terms of 129.38.Hence the final model formulation is given below as:

$$Y_{cp} = 1.2379 + 0.3245MR - 0.6697x^{-2} - 2.2516x^{-2}lnx. \quad (4.6)$$

where Y_{cp} is cowpea yield, x is fertilizer rate / 10 and MR is manure rate.

Table 4.9: Effect and Parameter Estimation for Cowpea Yield from Grouped Fertilizer and Manure (polytomous)

Variables	Coefficients	Std Error	Z-value	P-value	[95% conf. interval]
Fertrate 1	-0.6696712	0.3406439	-1.97	0.049*	-1.337321 – - 0.0020216
Fertrate 2	-2.251586	1.848929	-1.22	0.2223	-5.875420 – 1.3722480
Manure	0.3245312	0.0606118	5.35	0.001*	0.2057343 – 0.4433282
Constant	1.255817	0.1311328	9.58	0.001*	0.9988014 – 1.512833

Deviance: 129.38 Best powers fertrate models fit are -2, -2.

Log likelihood = -64.6881204

AIC = 1.431003

BIC = -398.2888

Approximately 2 was subtracted from manure rate, (x^{-2}-0.327) and (x^{-2}lnx-0.183) from fertilizer rate, where x = fertilizer rate/10, to improve the scaling of the regression coefficients.

Also from same Table 4.9, it can be observed that the adequacy of the model measured by the Log likelihood (-63.5983), AIC (1.4279) and BIC (-394.24), shows that the fitted model is adequate because the value of each of the measure statistics is small.

Table 4.10 presents the selection algorithm for the grouped variables manure and fertilizer rates on cowpea yield, where manure and fertilizer rates are quantitative (continuous) variable and it is observed that algorithm converged with the deviance for the model with all terms of 129.53.

The parameter estimation results from the generalized linear model fractional polynomial regression for cowpea yield with manure and fertilizer rates. It can be observed that, approximately 1.75 was subtracted from fertilizer rate, *(x^{-2}-0.25)* and*(x^{-2}ln(x)-0.173)* from manure rate, where *x* = manure rate. The coefficients estimated for fertilizer rate is (0.15583), manure rates 1 and 2 are (-1.13220) and (-2.92739) respectively. It has been observed that the manure rate 1 and fertilizer rate are significance at p-values of 0.001 and 0.009 respectively, while manure rate 2 is not significant at 5% significant level. Hence the final model formulation is given below as:

$$Y_{cp} = 1.2865 + 0.1558FR - 1.1322x^{-2} - 2.29273x^{-2}\ln x. \qquad (4.7)$$

where Y_{cp} is cowpea yield, *x* is manure rate / 10 and *FR* is fertilizer rate.

Table 4.10: Effect and Parameter Estimation for Cowpea Yield from Manure and Fertilizer (polytomous)

Variables	Coefficients	Std Error	Z-value	P-value	[95% conf. interval]
Manure 1	-1.1322038	0.2993962	-3.78	0.001*	-1.719009 - -0.5453973
Manure 2	-2.9273860	1.608140	-1.82	0.069	-6.079283 – 0.2245117
Fertrate 1	0.1558333	0.0597333	2.61	0.009*	0.0387583 – 0.2729084
Constant	1.3359380	0.0857854	15.57	0.001*	1.167801 – 1.504074

Deviance: 129.53 Best powers manure rate models fit are -2, -2.

Log likelihood = -64.76352478

AIC = 1.432573

BIC = -398.2547

Approximately 1.75 was subtracted from manure rate, (x^{-2}-0.25) and (x^{-2}ln(x)-0.1733) from fertilizer rate, where x =manure rate, to improve the scaling of the regression coefficients

Also from same Table 4.10 it can be observed that the adequacy of the model measured by the Log likelihood (-64.7635), AIC (1.4326) and BIC (-398.255), shows that the fitted model is adequate because the value of each of the measure statistics is small.

4.3 Discussion

Tables 4.1 through Table 4.10 are presentations of the individual results from the analyses of the data set described in chapter three. This set of data was analyzed as fractional polynomial regression with covariates as grouped and ungrouped, continuous covariates

For the experimental design data, nitrogen fertilizer rates, manure rates on cowpea varieties were analyzed using two different approaches, where fertilizer and manure rates were analyzed, and cowpea variety as a factor considered in the analysis was naturally coded because it is a qualitative factor

Table 4.1 and Table 4.2 present the result of the multiple fractional polynomial regression analysis for fertilizer and manure rates and it was observed that fertilizer and manure rates are significant at 5% level with a p-value of 0.013 and 0.0001 respectively but variety was not significant at 5% level because its p-value of 0.135 is greater than 5% significance level. The best fitted power for fertilizer rate and manure rate is $p = 1$ which is a linear fit. Also, from the same Table 4.1, it was observed that the algorithm for the selection of factors with significant effect converged after 2 cycles and the untransformed model terms deviance is 127.968 which is less than the final deviance of 130.202 obtained after the transformation of the model. This implies that the transformed model is a better fit because of the achieved deviance difference or gain *G* of 7.060 for fertilizer effect and 25.845 for manure effect. In Table 4.2, the parameter estimation was presented for fertilizer rate and manure rate and it was observed that the coefficients 0.007344 and 0.0649062 respectively for both factors are significant at 5% level with a p-value of 0.013 and 0.0001 respectively. Approximately 35.7 and 10 were subtracted from fertilizer and manure rates respectively before transformation to improve the scaling of the regression coefficients and the log likelihood value of – 65.101, AIC value of 1.419 and BIC value of – 402.67 are small hence the fitted model is adequate.

Tables 4.3 and 4.4, presented the result of the grouped (dichotomous) multiple fractional polynomial regression analysis for fertilizer and manure rates it was observed that the algorithm for the selection of factors with significant effect converged after 2 cycles and the untransformed model terms deviance is 137.388 which is less than the final deviance of 139.415 obtained after the transformation of the model. This implies that the transformed model is a better fit because of the achieved deviance difference or gain *G* of 4.476 for fertilizer effect and 17.748 for manure effect. In Table 4.4 the

parameter estimation was presented for fertilizer rate and manure rate and it was observed that the coefficients 0.21625 and 0.47813 respectively for both factors are significant at 5% level with a p-value of 0.036 and 0.0001 respectively. With the log likelihood value of –69.7075, AIC value of 1.15474 and BIC value of – 400.469 are small hence the fitted model is adequate.

Tables 4.5 and 4.6 presented the results of the grouped (polytomous) multiple fractional polynomial regression analysis for fertilizer and manure rates and it was observed that fertilizer and manure rates are significant at 5% level with a p-value of 0.012 and 0.0001 respectively but variety was not significant at 5% level because its p-value of 0.150 is greater than 5% significance level. The best fitted power for fertilizer rate and manure rate is $p = 1$ which is a linear fit. Also, from the same Table 4.5, it was observed that the algorithm for the selection of factors with significant effect converged after 2 cycles and the untransformed model terms deviance is 135.407 which is less than the final deviance of 137.475 obtained after the transformation of the model. This implies that the transformed model is a better fit because of the achieved deviance difference or gain *G* of 6.458 for fertilizer effect and 18.437 for manure effect. In Table 4.6 the parameter estimation was presented for fertilizer rate and manure rate and it was observed that the coefficients 0.15583 and 0.47813 respectively for both factors are significant at 5% level with a p-value of 0.012 and 0.0001 respectively. Approximately, 1.75 was subtracted from fertilizer rate before transformation to improve the scaling of the regression coefficients and the log likelihood value of –68.737, AIC value of 1.495 and BIC value of – 400.95 are small hence the fitted model is adequate.

Table 4.7 presented the result of fractional polynomial regression analysis for fertilizer and manure rates as $\varphi(x_{1,}3)$, it was observed that only manure is significant at 5% level with p-value of 0.0001 but fertilizers were not significant at 5% level. The best power

for fertilizer rates are 2 and 3, while manure rate is 1. In same Table, parameter estimation was presented for both factors and it was observed that the coefficients 0.0649,-0.0292 and 0.00204 for manure, fertrate 1 and fertrate 2 respectively were obtained. Approximately 10 was subtracted from manure rate, $(x^2-14.0625)$ and $(x^3-52.734)$ from fertilizer rate, where x = fertilizer rate/10, to improve the scaling of the regression coefficients and the Log likelihood (-63.5983), AIC (1.4279) and BIC (-394.24), shows that the fitted model is adequate. Also, when manure and fertilizer rates were analyzed, in Table 4.8 as $\varphi(x_{2,}-2)$, it was observed that the manure rate 1 and fertilizer rate were significance at p-values of 0.024 and 0.013 respectively, while manure rate 2 is not significant at 5% significant level. The best power for manure rates are -2 and -2 while fertilizer rate is 1. In same Table, parameter estimation was presented for both factors and it was observed that the coefficients for fertilizer rate is (0.0073444), manure rates 1 and 2 are (-0.790328) and (-0.7318464) respectively. Approximately 37.5 was subtracted from fertilizer rate, $(x^{-2}-1)$ and $(x^{-2}lnx)$ from manure rate, where x = manure rate/10, to improve the scaling of the regression coefficients and the Log likelihood (-65.08666028), AIC (1.439305) and BIC (-398.1084).

Table 4.9 presents the result of grouped (polytomous) fractional polynomial regression analysis for fertilizer and manure rates as $\varphi(x_{1,}-2)$, it was observed that , fertrate 1 and manure are significant at 5% level with p-value of 0.049 and 0.0001 respectively, but fertrate 2 is not significant at 5% level. The best power for fertilizer rates are -2 and -2, while manure rate is 1. In the same Table, parameter estimation was presented for both factors and it was observed that the coefficients 0.32453,-2.25159 and -0.66967 for manure, fertrate2 and fertrate1respectively were obtained. Approximately 2 was subtracted from manure rate, $(x^{-2}-0.327)$ and $(x^{-2}lnx-0.183)$ from fertilizer rate, where x = fertilizer rate, to improve the scaling of the regression coefficients and the Log

likelihood (-64.6881), AIC (1.4310) and BIC (-398.28). This showed that the fitted model is adequate. Also, when manure and fertilizer rates were analyzed, in Table 4.10 as $\varphi(x_{2,}\text{-}2)$, it was observed that the manure rate 1 and fertilizer rate are significance at p-values of 0.001 and 0.009 respectively, while manure rate 2 is not significant at 5% significant level. The best power for manure rates are -2 and -2 while fertilizer rate is 1. In same table, parameter estimation was presented for both factors and it was observed that the coefficients for fertilizer rate is (0.15583), manure rates 1 and 2 are (-2.92739) and (-1.13220) respectively. Approximately 1.75 was subtracted from fertilizer rate, $(x^{-2}\text{-}0.25)$ and$(x^{-2}lnx - 0.173)$ from manure rate, where x = manure rate, to improve the scaling of the regression coefficients and the Log likelihood (-64.76352), AIC (1.43257) and BIC (-398.2547).

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.0 Introduction

This chapter presents the summary, conclusion and recommendations based on the results obtained in chapter four.

5.1 Summary

The main objective of this research work is to fit a fractional polynomial regression model with continuous covariate and grouped covariate. The essence is to compare the performance of the fit for continuous covariate and grouped covariates. The generalized linear model was fitted for the fractional polynomials. Two different approaches were used. The first approach is Royston and Altman method, while the second approach is ordinary fractional polynomials fit.

From the fitted fractional polynomial regression models it was observed that the median algorithm method for grouping continuous covariates that was proposed gave a better results compared to the continuous covariate. For the experimental design data, the effect of nitrogen fertilizer, manure and cowpea variety on cowpea yield presented in table 4.1 through table 4.10 when MFP regression proposed by Royston and Altman was applied the algorithm for selection factors with significant effects converged at $\varphi(1,, 1)$ with final deviance of 127.97, both fertilizer and manure rates are significant at 5% level with P-value of 0.029 and 0.001 respectively. On the other hand variety was not significant at 5% level. When fertilizer was considered as an independent variable one i.e (x_1), the algorithm for the selection of factors with significant effects converged at $\varphi(x_{1,} 3)$ with model terms deviance of 127.08.The model for selection of factors with significant effects converged at $\varphi(x_{2,} -2)$ with model terms deviance of 130.17 for the manure.

5.2 Conclusion

Based on the observations above, the study conclude that the median grouping method (polytomous) for grouping continuous covariate did not performed badly, since it gave most significant result compare to ungrouped and grouped (dichotomous) continuous covariates. For the Fractional polynomials regression, the continuous covariates produced the gain (G) of 3.09.When multivariable fractional polynomials regression was used, the gain (G) of 6.20 and 25.85 were produced. In fact, most data analyst always grouped their treatment levels before analysis except otherwise. Therefore grouping could be done adequately depending on the method one obtained his cutpoint or carrying out the grouping of the continuous covariate.

5.3 Recommendation

Based on our observations from chapter four, the study recommends the following:

(i) Fractional polynomial regression model fit the data well because it is open in the sense that a set of pre-defined powers are available which the best powers among other contending powers can be selected.

(ii) The use of FP modeling in experimental design data too, since programs for interaction effect have been developed in latest versions of STATA, though, was not studied in this research work so as to extend its applicability.

(iii) The median algorithm method of grouping continuous covariate is recommend because in this research work it showed a contending strength with covariate that is not grouped.

5.4 Contribution to Knowledge

The contributions to knowledge from this research work are as follows;

1. The comparison between fractional polynomial regression model with continuous covariate and grouped covariate is achieved.
2. Median algorithm method of grouping continuous covariates has developed.
3. Fitted a fractional polynomial regression model in analyzing experimental design data has been successfully executed.

5.5 Suggestion for Further Research

Extension of fractional polynomial regression model in fitting experimental design data with interaction effects is suggested.

REFERENCES

Allen, D. M. (1971).“Mean Square Error of Prediction as a Criterion for Selecting Variables”. *Technometrics* 13, 3, pp. 469-475.

Almini, A. A., Kulahci, M. and Montgomery, D. C. (2006). “Checking the Adequacy of fit of Models from split-plot Designs”. *Journal of Quality Technology* 38. pp. 58 – 190.

Altman, D. G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994). The dangers of using ‘optimal’ cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 86, pp. 829–835.

Austin, P. C. and Brunner, L. J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine*, 23, pp.1159 –1178.

Barnhart, H. X.; Haber, M.; and Song, J. (2002). “Overall Concordance Correlation Coefficient for Evaluating Agreement among Multiple Observers”. *Biometrics* 58, pp.1020 - 1027

Becher, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine* 11, pp.1747 –1758.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.pg. 60.

Berkey, C. S. and Reed, R. B. (1987). A model for describing normal and abnormal growth in early childhood. *Human Biology* 59, pp. 973 – 987.

Box, G. E. P. and Tidwell, P. W. (1962).Transformation of the independent variables. *Technometrics*4, pp. 531 – 550.

Bremer, M. (2012).Polynomial Regression Models. MATH 261A-SPRING 2012 Lecture. Retrieved 2013 from http://book.analysis3.com/Polynomial-Regression-Models-download-wii65.pdf

Brenner, H. and Blettner, M. (1997). Controlling for continuous confounders in epidemiologic research. *Epidemiology* 8, pp. 429–434.

Breslow, N. E. and Day, N. E. (1980).*Statistical Methods in Cancer Research*, vol. 1. IARC

Scientific Publications: Lyon.

Brown, L. D., Cai, T. T. and Zhou, H. H. (2010). Nonparametric regression in exponential families. *The Annals of Statistics* 38, pp. 2005 – 2046.

Cameron, A. C. and Windmeijer, F. A. G. (1997). Nonparametric regression in exponential families. *Journal of Econometrics* 77, pp. 329 – 342.

Claeskens, G. and Hjort, N. L. (2008).*Model Selection and Model Averaging*. Cambridge Press, New York.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association 83(403),pp. 596 – 610.*

Cochran, W. G. (1968). The effectiveness of adjustment by sub classification in removing bias in observational studies. *Biometrics* 24, pp. 295 –313.

Cohen, J. (1960). "A coefficient of Agreement for Nominal Scales". *Educational and Psychological Measurement* 20, pp. 37-46.

Cohen, J. (1968). "Weighted kappa: normal scale Agreement with Provision for Scaled Disagreement or Partial credit. *Psychological Bulletin* 70, pp. 213-220.

Count, E. W. (1942). A quantitative analysis of growth in certain human skull dimensions. *Human Biology* 14, pp. 143 – 165.

Count, E. W. (1943). Growth patterns of human physique: an approach to kinetic anthropometry. *Human Biology* 15, pp. 1 – 32.

de Boor, C. (2001), *A practical guide to splines*, Springer.

DeGroot, M. H. (1980). *Probability and Statistics*.2nd ed. Addison-Wesley.
.

Dent, J. B. and Blackie, M. J. (1979) "*Systems Simulation in Agriculture*". Applied Science, London.

Draper, N. R. and Smith, H. (1998).*Applied Regression Analysis* .Wiley-Interscience. ISBN 0-417-17082-8.

Faraggi, D. and Simon, R. (1996).A simulation study of cross-validation for selecting an

Optimal cutpoint in univariable survival analysis. *Statistics in Medicine* 15, pp. 2203–2213.

Forrester, J. W. (1961). *Industrial Dynamics*. MIT Press, Cambridge.

Geweke, J. and Petrella, L. (2012).Likelihood-based Inference for Regular Functions with Fractional Polynomial Approximations. Retrieved April 15, 2013 from: www.censoc.uts.edu.au/pdfs/geweke_papers/gp_working_4b.pdf.

Green, P. J. and Silverman, B. W. (1994).*Nonparametric Regression and Generalized Linear Models*, Vol. 58 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Hand, D. J. and Vinciotti, V. (2003), Local versus global models for classification problems: fitting models where it matters. *The American Statistician*, 57, pp. 124-131.

Hardle, W., Mammen, E. and Muller, M. (1998). Testing parametric versus semi parametric modeling in generalized linear models. *Journal of the American Statistical Association*93, pp. 1461 – 1474.

Hastie, T. J. and Tibshirani, R. J, (1986).Generalized additive models (with discussion). *Statistical Science* 1, pp. 297 – 318.

Hastie, T. J. and Tibshirani, R. J, (1990).*Generalized Additive Models*. London: Chapman and Hall.

Hilsenbeck, S. G. and Clark, G. M. (1996). Practical P-value adjustment for optimally selected cutpoints. *Statistics in Medicine* 15, pp. 103 –112.

Hollander, N., Sauerbrei, W. and Schumacher, M. (2004). Confidence intervals for the effect of aprognostic factor after selection of an 'optimal' cutpoint. *Statistics in Me-dicine* 23, pp. 1701–1713.

Hosmer, D. W., Jr., and Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics: Theory and Methods, Part A* 9: pp. 1043–1069.

Hosmer, D. W., Lemeshow, S. and Klar, J. (1988). Goodness-of-fit testing for the logistic

regression model when the estimated probabilities are small. *Biometrical Journal*pp.11,911 – 924.

Huang, M. and Zhang, D. (2008). Testing polynomial covariate effects in linear and generalized linear mixed models. Statistics Survey 2, pp. 154 – 169.

Irwin, J. R. and McClelland, G. H. (2003).Negative consequences of dichotomizing continuous Predictor variables. *Journal of Marketing Research* 40, pp. 366 –371.

Isaacs, D., Altman, D. G., Tidmarsh, C. E., Valman, H. B. and Webster, A. D. B. (1983). Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges fir IgG, IgA, IgM. *Journal of Clinical Pathology* 36, pp. 1193 – 1196.

Karlsson, M., Cantoni, E. and de Luna, X. (2009).Local polynomial regression with truncated or censored response. *Journal of Economic Letter*, C14, pp. 1 – 18.

Kavalseth, T. O. (1985). "Cautionary note about R^2".*The American Statistician* 39, pp. 279-285

King, T. S.and Chinchilli, V. M. (2001). A Generalized Concordance Correlation Coefficient for Continuous and Categorical data. *Statistics in Medicine* 20: pp. 2131-2147.

King, T. S.; Chinchilli, V. M.; Carrasco, J. L. (2007)."A Repeated Measures Concordance Correlation Coefficient". *Statistics in medicine* 20, pp. 2131-2147.

Lagakos, S. W. (1988). Effects of mismodelling and mismeasuring explanatory variables on Testsof their association with a response variable. *Statistics in Medicine* 7, pp. 257 – 274.

Lausen, B. and Schumacher, M. (1996).Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics and Data Analysis* 21, pp. 307 –326.

Lehmann, E. L. and Casella, G. (1998).*Theory of Point Estimation*. 2nded. New York: Springer. MR1639875 (http://www.ams.org/mathscinet-getitem?mr=1639875).ISBN 0-387-98502-6.

Lin, L. (1989). “A Concordance Correlation Coefficient to Evaluate Reproducibility”. *Biometrics*45, pp.255-268.

Lin, L. (2000). A Note on the Concordance Correlation Coefficient. *Biometrics* 56: pp. 324-325.

Loague, K. and Green, R. E. (1991).”Statistical and Graphical Methods for Evaluating Solute Transport Models: Overview and Application”. *Journal of Contaminant Hydrology* 7, pp. 51-73.

MacLennan, I. C. M., Kelly, K., Crockson, R. A., Cooper, E. H., Cuzick, J. and Chapman, C. (1988). Results of the MRC myelomatosis trials for patients entered since 1980. *Hematology Oncology* 6, pp. 145 – 158.

MacCallum, R. C., Zhang, S., Preacher, K. J. and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables.*Psychological Methods* 7, pp. 19–40.

Masoud, Y. and Rahim, M. (2010) “The Effect of Outliers on Robust and Resistant Coefficient of Determination in the Linear Regression Models”.*International Journal of Academic Research*. Vol 2 No. 3. pp. 133-138.

Maxwell, S. E. and Delaney, H. D. (1993). Bivariate median-splits and spurious statistical Significance. *Psychological Bulletin* 113, pp. 181–190.

Mayer, D.G. and Butter, D.G. (1993) “Statistical Validation”. *Ecological modeling* 68, pp. 21 – 32.

Mazumdar, M., Smith, A. and Bacik, J. (2003) Methods for categorizing a prognostic variable in a multivariable setting. *Statistics in Medicine* 22, pp. 559 –571.

McCullagh, P. and Nelder, J. A. (1989).*Generalized Linear Models*, 2nd edn, p.16. London; Chapman and Hall.

Miller, R. and Siegmund, D. (1982).Maximally selected chi-square statistics.*Biometrics* 38, pp. 1011–1016.

Mitchell, P. L. (1997). “Misuse of Regression for Empirical Validation of Models”. *Agricultural Systems* 54, pp. 313-326.

Mitchell, P. L. and Sheehy, J. E. (1997). "Comparison of Predictions and observations to assess Model Performance: a method of Empirical validation". Kluwer Academic Publishers, Boston. MA, pp. 437-451.

Montgomery, D. C. (2005).*Design and Analysis of Experiments*, 5th ed. John Wiley and Sons New York, pp. 76-86.

Mosteller, F. and Tukey, J. W. (1977b).*Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, MA.

Nelder, J. A. (1966). Inverse polynomials a useful group of multi-factor response functions. *Biometrics* 22, pp. 128 – 141.

Pan, Z. and Lin, D. Y. (2005).Goodness-of-fit methods for generalized linear mixed models. *Biometrics* 61, pp. 1000 – 1009.

Poirer, D. J. (1973).Piecewise regression using cubic splines. *Journal of American Statistical Association* 68, pp. 515 – 524.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerical Mathematics*, 10, pp. 177 – 183.

Royston, P. (1992). The use of cusums and other techniques in modeling continuous covariates in logistic regression. *Statistics in Medicine*, 11, pp. 1115 – 1129.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics*, 43(3): pp. 429–467.

Royston, P., Ambler, G. and Sauerbrei, W. (1999).The use of fractional polynomials to model continuous risk variables in epidemiology .*International Journal of Epidemiology* 28**,** pp. 964 – 974.

Royston, P., Altman, D. G. and Sauerbrei, W. (2005). Dichotomizing continuous predictors in multiple regression: a bad idea. Wiley Interscience (www.interscience.wiley.com)

Royston P, and Sauerbrei W. (2008). *Multivariable Model-Building*: *A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Continuous Variables*.

Wiley: New York.

Salawu, I. S. (2007).Transformed inverse model at quadratic variable in fertilizer response. *Asian Journal of Agricultural Research* 1 (2): pp. 80 – 85.

Sauerbrei, W. and Royston, P. (2010). Continuous variables: to categorize or to model? In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, the Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

Sauerbrei, W., Royston, P. and Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* 26, pp. 5512 – 5528.

Schumacher, M., Hollander, N. and Sauerbrei, W. (1997). Resampling and cross-validation techniques: a tool to reduce bias caused by model-building? *Statistics in Medicine* 16, pp. 2813 –2827.

Sellers, K. F. and Shmuelli, G. (2010).A flexible regression model for count data. *The Annals of Applied Statistics* 4, pp. 943 – 961.

Sergio, B. and Joan, C. (2001)."Oriented Principal Component Analysis for Large margin Classifiers". *Neural Networks*, vol. 14, No. 10, pp. 1447-1461.

Shaffer, D. L. (1980). A model evaluation methodology applicable to environmental Assessment C models. Ecological Modeling 8, pp. 275 – 295.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistics Society* B, 47, pp. 1 – 52.

Smith, J. M., Dore, C. J., Charlett, A. and Lewis, J. D. (1992). A randomized trial of Biofilm dressing for venous leg ulcers. *Phlebology*, 7, pp. 108 – 113.

Sterman, J. D. (2002). "All models are wrong: Reflections on Becoming a System Scientist. *System Dynamics review* 18, pp. 501 – 531.

Stone, C. J. (1985). Additive regression and other nonparametric models.*Annals of Statistics*,

13, pp. 689 – 705.

Stronger, D. and Stone, P. C. (2006). Polynomial Regression with Automated Degree: A Function Approximator for Autonomous Agents. *Technical Report* UT-AI-TR-06-329, pp. 1–12. http://www.cs.utexas.edu/˜{stronger,pstone}

Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics* 29, pp. 535 – 545.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association* 86, pp. 420 – 426.

Tedeschi, L. O. (2006). "Assessment of the adequacy of mathematical Models".*Agricultural Systems* 89, pp. 225-247.

Tsiatis, A. A. (1980). A note on a goodness of fit test for the logistic regression model. *Biometrika* 67, pp. 250 – 251.

Vonesh, E. F., Chinchilli, V. M. and Pu, K. (1996).Assessing the goodness of fit in generalized linear and nonlinear mixed-effect models. *Biometrics*. V52: pp. 572 – 587.

Wartenberg, D. and Northridge, M. (1991).Defining exposure in case-control studies: a new approach. *American Journal of Epidemiology* 133, pp. 1058 –1071.

Weinberg, C. R. (1995). How bad is categorization? *Epidemiology* 6, pp. 345 –347.

Whittaker, E. T. (1923). On a new method of graduation. *Proceedings of Edinburgh Mathematical Society*, 41, pp. 63 – 75.

Wingerd, J. (1970). The relation of growth from birth to two years to sex, parental size and Otherfactors, using Rao's method of the transformed time scale. *Human Biology*, 42, pp. 105 – 131.

Wong, E. S., Wang, B. C. M., Garrison, L. P., Alfonso-Cristancho, R., Flum, D. R., Arterburn, D. E. and Sullivan, S. D. (2011). Examining the BMI-mortality relationship using fractional Polynomials .*Biomedical Central Medical Research Methodology* 11(175), pp. 1471 – 2288

Appendix: Experimental Design Data

Cowpea yield	FertRate	Variety	ManureRt	PtgmanR	DcgfertR	DcgmanR	PtgfertR
0.07	15	1	5	1	0	0	1
0.09	30	1	10	2	0	0	1
0.49	45	1	15	3	1	1	2
0.58	60	1	5	1	1	0	3
0.93	15	1	10	2	0	0	1
1.42	30	1	15	3	0	1	1
0.98	45	1	5	1	1	0	2
1.38	60	1	10	2	1	0	3
2.04	15	1	15	3	0	1	1
1.1	30	1	5	1	0	0	1
1.51	45	1	10	2	1	0	2
2.27	60	1	15	3	1	1	3
0.42	15	1	5	1	0	0	1
0.93	30	1	10	2	0	0	1
0.51	45	1	15	3	1	1	2
0.91	60	1	5	1	1	0	3
0.89	15	1	10	2	0	0	1
1.04	30	1	15	3	0	1	1
1.02	45	1	5	1	1	0	2
0.93	60	1	10	2	1	0	3
1.78	15	1	15	3	0	1	1
0.6	30	1	5	1	0	0	1
1.58	45	1	10	2	1	0	2
2.27	60	1	15	3	1	1	3
0.18	15	1	5	1	0	0	1
1.04	30	1	10	2	0	0	1
0.62	45	1	15	3	1	1	2
0.91	60	1	5	1	1	0	3
1.8	15	1	10	2	0	0	1
1.53	30	1	15	3	0	1	1
0.96	45	1	5	1	1	0	2
1.36	60	1	10	2	1	0	3
2.4	15	1	15	3	0	1	1
0.93	30	1	5	1	0	0	1
1.51	45	1	10	2	1	0	2
2.41	60	1	15	3	1	1	3
0.95	15	1	5	1	0	0	1
1.86	30	1	10	2	0	0	1
1.94	45	1	15	3	1	1	2
1.28	60	1	5	1	1	0	3
1.41	15	1	10	2	0	0	1
1.33	30	1	15	3	0	1	1

1.61	45	1	5	1	1	0	2
1.46	60	1	10	2	1	0	3
1.92	15	1	15	3	0	1	1
1.3	30	1	5	1	0	0	1
1.24	45	1	10	2	1	0	2
2.41	60	1	15	3	1	1	3
0.79	15	2	5	1	0	0	1
1.4	30	2	10	2	0	0	1
2.1	45	2	15	3	1	1	2
0.53	60	2	5	1	1	0	3
1.95	15	2	10	2	0	0	1
1.43	30	2	15	3	0	1	1
0.82	45	2	5	1	1	0	2
1.61	60	2	10	2	1	0	3
2.36	15	2	15	3	0	1	1
1.92	30	2	5	1	0	0	1
1.63	45	2	10	2	1	0	2
2.58	60	2	15	3	1	1	3
0.1	15	2	5	1	0	0	1
1.57	30	2	10	2	0	0	1
1.9	45	2	15	3	1	1	2
0.82	60	2	5	1	1	0	3
1.4	15	2	10	2	0	0	1
1.41	30	2	15	3	0	1	1
1.1	45	2	5	1	1	0	2
0.99	60	2	10	2	1	0	3
1.82	15	2	15	3	0	1	1
1.44	30	2	5	1	0	0	1
1.52	45	2	10	2	1	0	2
2.11	60	2	15	3	1	1	3
0.21	15	2	5	1	0	0	1
0.98	30	2	10	2	0	0	1
1.47	45	2	15	3	1	1	2
1.59	60	2	5	1	1	0	3
0.62	15	2	10	2	0	0	1
0.95	30	2	15	3	0	1	1
1.24	45	2	5	1	1	0	2
1.48	60	2	10	2	1	0	3
0.61	15	2	15	3	0	1	1
1.41	30	2	5	1	0	0	1
1.57	45	2	10	2	1	0	2
1.62	60	2	15	3	1	1	3
1.58	15	2	5	1	0	0	1
1.34	30	2	10	2	0	0	1
1.66	45	2	15	3	1	1	2
1.65	60	2	5	1	1	0	3

1.43	15	2	10	2	0	0	1
1.3	30	2	15	3	0	1	1
1.6	45	2	5	1	1	0	2
2.04	60	2	10	2	1	0	3
1.19	15	2	15	3	0	1	1
1.21	30	2	5	1	0	0	1
1.3	45	2	10	2	1	0	2
1.69	60	2	15	3	1	1	3

KEY:

Fertilizer: 1- 15 kg/ha, 2- 30 kg/ha, 3- 45 kg/ha, 4- 60 kg/h**a**

Manure: 1- 5 kg/ha, 2- 10kg/ha and 3- 15kg/ha

Variety: 1- variety 1 and 2- variety 2.

PtgfertR: 1- < 37.5, 2- 37.5 to 52.5, 3- > or = 53.

PtgmanR: 1-< 10, 2- = 10, 3- > 10.

DcgfertR: 0- < or = 37.5, 1 > 37.5

DcgmanR: 0- < or = 10, 1- > 10.