Study on Varied Aspects of Linear/Nonlinear Models and Their Application in Agriculture

Yasmeena Ismail
(2015-614-D)

Division of Agricultural Statistics
Faculty of Horticulture

Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir

2018

Study on Varied Aspects of Linear/Nonlinear Models and Their Application in Agriculture

Yasmeena Ismail
(2015-614-D)

Thesis

Submitted to
The Faculty of Horticulture

Sher-e-Kashmir
University of Agricultural Sciences &Technology of Kashmir in partial fulfilment of requirement for the award of the degree of

Doctor of Philosophy in Statistics

2018


Dedicated
To
My Parents
Mr Mohammad Ismail
Mrs Jameela
And
My Nephew
Saim Abrar

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Faculty of Horticulture, Division of Agricultural Statistics

Certificate – I

This is to certify that the thesis entitled, **"Study on Varied Aspects of Linear/Nonlinear Models and Their Application in Agriculture"** submitted in partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy in Statistics**, to the **Faculty of Horticulture, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir** is a record of bonafide research work carriedoutby **Ms.Yasmeena Ismail(Regd. No. 2015-614-D)** under my supervision and guidance. No part of the thesis has been submitted for any other degree or diploma.

It is further certified that information received during the course of investigation has duly been acknowledged.

(Prof.S.A.Mir)
Chairman
Advisory Committee

Endorsed

Head,
Division of Agricultural Statistics,
SKUAST-Kashmir, Shalimar

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Faculty of Horticulture, Division of Agricultural Statistics

Certificate – II

We, the members of the Advisory Committee of **Ms.Yasmeena Ismail(Regd. No. 2015-614-D)**, a candidate for the degree of **Doctor of Philosophy in Statistics**, have gone through the manuscript of the thesis entitled, **"Study on Varied Aspects of Linear/Nonlinear Models and Their Application in Agriculture"** and recommend that it may be submitted by the student in partial fulfilment of the requirements for the award of the degree.

Advisory Committee

Chairman

Prof. S.A. Mir
Prof. & Head
Division of Agricultural Statistics,
SKUAST-Kashmir

Dr. NageenaNazir
Assistant Professor,
Division of Agricultural Statistics,
SKUAST-Kashmir

Dr. M. S. Pukhta
Associate Professor,
Division of Agricultural Statistics, SKUAST-Kashmir

Prof. M.H. Wani
Registrar,
SKUAST-Kashmir

Dean's Nominee

Prof. Shabir Ahmad Wani
Prof.& Head,
Division of Agri-Economics

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Faculty of Horticulture, Division of Agricultural Statistics

Certificate – III

This is to certify that the thesis entitled, **"Study on Varied Aspects of Linear/Nonlinear Models and Their Application in Agriculture"** submitted by **Ms.Yasmeena Ismail(Regd. No. 2015-614-D)**,to the **Faculty of Horticulture, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir** in partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy in Statistics**was examined and approved by the Advisory Committee and External Examiner on ……

Chairman
Advisory Committee

External Examiner

Head,
Division of Agricultural Statistics

Dean,
Faculty of Horticulture,
SKUAST-Kashmir

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Faculty of Horticulture, Division of Agricultural Statistics

Name of the student	:	**Yasmeena Ismail**
Registration No.	:	2015-614-D
Major subject	:	Statistics
Minor subjects	:	Agri. Economics and Marketing
Major advisor	:	**Prof. S.A. Mir** Prof.& Head Division of Agricultural Statistics, SKUAST-Kashmir
Title of the Thesis	:	**"Study on Varied Aspects of Linear/Nonlinear Models and Their Application in Agriculture"**

ABSTRACT

The linear model is defined as the functional relationship between the response variable and the covariates. The basic linear regression models assume the response to be normally distributed but there are situations in which the assumption of normality is violated and in these situations it is necessary to fit the data by some alternative approach we have studied such situations and have extended the basic linear models to generalized linear models. In generalized linear models we have studied the models where the response is from an exponential family of distributions in particular Gamma and Inverse Gaussian distributions. Similar is the case with the linear mixed models when the response is non-normal we fit the models in such situations by the generalized linear mixed models. Further, we have studied different estimation methods of fitting generalized linear mixed models. The results of the estimation methods are compared in terms of average relative bias, average squared relative bias, average absolute bias and average squared deviation. Numerical results on the real horticultural data highlights estimation method which fits the data the best than other methods.

Also the basic nonlinear mixed effects model is extended to allow heteroscedastic correlated within group errors. Library nlme() of R software is used to fit the extended nonlinear mixed effects model. It has been shown that the estimation and computational methods of simple nonlinear mixed effects models can be applied to the extended model and decomposition of variance, covariance structure of within group errors into two independent components: a variance structure and a correlation structure. For long term trend analysis we do not have

the exact information about the data distribution. In such situation the application of parametric models may not come up with the appropriate results .We have studied the nonparametric regression models for the long term trend analysis. The numerical results on the real horticultural data has been compared on the basis of Finite Sample Corrected AIC (AIC_C), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Max Absolute Percentage Error (MaxAPE), Max Absolute Error (MaxAE).

The above discussed methods are illustrated practically with the utilization of SAS and R software. Various functions were developed to fit the extended generalized linear models and extended generalized linear mixed effects models viz GLMLG(), GLMLIG(), GLMMEST(), Relative Bias() and Absolute Bias(). The nlme() function of R software is used to fit the extended nonlinear mixed effects model through examples on the real data sets. The function developed in SAS software consist of number of SAS procedures viz PROC LOESS, PROC SPLINES, PROC PRINT. All these functions are run on the real horticultural data sets.

Key words: Generalized linear models, Gamma and Inverse Gaussian distributions, Generalized linear mixed effects model, Nonlinear mixed effects model, Nonparametric regression, SAS/R software.

Signature of the Student

Dated_____

Signature of Major Advisor

Dated_____

ACKNOWLEDGEMENT

First of all, I thank the Almighty Allah "the most merciful and gracious" for His blessings.

Firstly, I express my thanks and gratitude to my supervisor and mentor ***Prof. S. A. Mir****, Head, Division of Agricultural Statistics, SKUAST-Kashmir, Shalimar whose supervision and mentoring has been instrumental in completing my Research. Prof. S. A. Mir's supervision has significantly eased my efforts throughout my Research work. His in-depth knowledge about the subject, meticulous scrutiny, timely scholarly advice and approach have helped me to a very great extent to successfully complete my Research work.*

I would take this opportunity to thank the University Vice Chancellor Prof. Nazeer Ahmad for establishing the state-of-the-art Computer Lab equipped with latest facilities in the Division of Agricultural Statistics for the Scholars to carry out the research work.

I am also thankful to the worthy members of my Advisory Committee Dr. NageenaNazir, Assistant Professor, Division of Agri-Statistics; Dr. M. S. Pukhta, Associate Professor, Division of Agri-Statistics, SKUAST-Kashmir; Prof. M. H. Wani, Registrar, SKUAST-Kashmir; Prof. Shabir Ahmad Wani, Prof. & Head, Division of Agri-Economics(Dean P. G Nominee)for their timely help and support.

I also thank the faculty members Dr. T. A. Raja, Associate Professor, Division of Agri-Statistics; Dr. ShowkatMaqbool, Assistant Professor, Division of Agri-Statistics; Dr. Imran Khan, Assistant Professor, Division of Agri-Statistics; Dr. B. S. Dhekale, Assistant Professor, Division of Agri-Statistics; Dr Mushtaq Ahmad Bhat, Assistant Professor, Division of Agri-Statistics;Mr. MohdShafi, Technical Assistant, Division of Agri-Statisticswho directly or indirectly have lent their helping hand throughout my research work.

My sincere thanks to our Research Group members Dr Subzar Ahmad Mir, Dr TabassumMushtaq, Mrs Rumana Majid, Ms UzmaMajeed, Mr Immad Shah and all other Scholars in the department for their support, encouragement and valuable suggestions. I will always remember our productive group discussions and the time full of humor we spent together.

A big thank you to my best friends Ms Fozia Amin, Ms.Sehra Zahoor, Ms Monisa Aslam Darvesh and Ms Umaira Sidiq for helping and supporting me during the difficult times.

I thank Mr Jan Mohammad Qadri, Miss Masrat, Mr Ghulam Mohi-u-Din Rather, Mr Mohammad Maqbool Mir and Mr Gulzar Ahmad Sheikh office staff of Division of Agricultural Statistics, SKUAST-Kashmir for their cooperation.

I am thankful to Dr Nasreen Fatima, Assistant Professor, Division of Plant Pathology; Miss Insha, Division of Floriculture for their support in the course of study.

I am highly thankful to ARIS and Library Staff members of SKUAST-K for their constant encouragement valuable suggestions and generous help during preparation of this manuscript.

Very sincere and special thanks to my beloved sister Mrs Rehana Ismail who supported me during my Thesis compilation and is a source of happiness and inspiration for me.

It is needless to mention the name of my brother Zahoor Ahmad who helped me through multiple ways throughout my Research work and I very much thank him. I would mention my little cousin Daniya and my grandfather Mr Mohammad Sultan for their love they have shown.

I dedicate this Thesis and Research work to my parents Mr Mohammad Ismail and Mrs Jameela whose inspiration, motivation and great support has helped me to successfully achieve the important and major career Milestone. My parent's unconditional love and prayers have made this Research work possible and with much ease.

Lastly, special thanks to Mr. M. Rafiq and Mr Shahid Sultan of M/s Universal Computers, Shalimar for composing this manuscript beautifully and giving it a final shape in the shortest possible time.

Yasmeena Ismail

Place : Shalimar, Srinagar

Dated :

S

Chapter 1

INTRODUCTION

1.1 Linear Models

Linear models play a central part in modern statistical methods. On the one hand, these models are able to approximate a large amount of metric data structures in their entire range of definition or at least piecewise. On the other hand, approaches such as the analysis of variance which model effects such as linear deviations from a total mean, have proved best their flexibility (Radhakrishna Rao and Toutenburg 1999). A model provides a theoretical framework for better understanding of a phenomenon of interest. Thus a model is a mathematical construct that we believe may represent the mechanism that generated the observations at hand. The postulated model may be an idealized oversimplification of the complex real-world situation, but in many such cases, empirical models provide useful approximations of the relationships among variables. These relationships may be either associative or causative.

In simple linear models, one attempts to model the relationship between two variables, for example, yield and plant height, income and number of years of education, height and weight of people, length and width of envelopes, temperature and output of an industrial process, altitude and boiling point of water, or dose of a drug and response. For a linear relationship, we can use a model of the form

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1.1.1)$$

Where,

y is the dependent or response variable

x is the independent or predictor variable

ε is the error term in the model

In this context, error does not mean mistake but is a statistical term representing random fluctuations, measurement errors, or the effect of factors outside of one's control. The linearity of the model in (1.1.1) is an assumption. One typically adds other assumptions about the distribution of the error terms, independence of the observed values of y, and so on. Using observed values of $x$and$y$, we estimateβ_0 and $\beta_1$and make inferences such as confidence intervals and tests of hypothesis for β_0 and β_1. We may also use the estimated model to forecast or predict the value of $y$for a particular value of$x$, in which case a measure of predictive accuracy may also be of interest.

The response y is often influenced by more than one predictor variable. For example, the yield of a crop may depend on the amount of nitrogen, potash, and phosphate fertilizers used. These variables are controlled by the experimenter, but the yield may also depend on uncontrollable variables such as those associated with weather. A linear model relating the response y to several predictors has the form

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon \qquad (1.1.2)$$

The parameters β_0, β_1, … β_k are called regression coefficients. As in (1.1.1), ε provides for random variation in y not explained by the $x$variables. This random variation may be due partly to other variables that affect y but are not known or not observed.

The model in (1.1.2) is linear in the β parameters; it is not necessarily linear in the x variables. Thus models such as

$$y = \beta_0 + \beta_1 x_{\cdot 1} + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 \sin x_2 + \varepsilon$$

are included in the designation linear model.

Classical linear model and least squares began with the work of Gauss and Legendre (Stigler 1981) who applied the method to the astronomical data. Their

data were usually measurements of continuous quantities such as positions and magnitudes of the heavenly bodies and, at least in the astronomical investigation, the variability in the observations was largely the effect of measurement error. The normal, or Gaussian, distribution was viewed as a mathematical construct developed to describe the properties of such errors; later in nineteenth century same distribution was used to describe the variation between the individuals in as biological population in respect of the character such as height, an application quite different in kind from its use for describing measurement error, and leading to the numerous applications of linear models. Some continuous measurements encountered in practice have non-normal error distributions, and the class of generalized linear models has been introduced for the analysis of such data. Generalized linear models permit us to study patterns of systematic variation in much the same way as ordinary linear models are used to study the joint effect of treatments and covariates. Generalized Linear Model (GLM) include as special cases, linear models and analysis of variance models, logit and probit models for quantal responses, log-linear models and multinomial response models for counts and some commonly used models for survival data (McCullagh and Nelder 1989).The basic principle behind the generalized linear model is that the systematic component of the linear model can be transformed to create an analytical framework that closely resembles the standard linear model but accommodates a wide variety of non-normal and non-interval measured outcome variables. The Gauss-Markov assumptions that underlie linear model theory require that the error component be distributed independently with mean zero and constant variance. If the outcome variable is drawn from a non-normal distribution, then these assumptions often cannot be met and serious errors of estimation efficiency occur, although the linear model is robust to mild deviations. Generalized linear models employ a “link function” which defines the relationship between the systematic component of the data and the outcome variable in such a way that asymptotic normality and consistency of variance are no longer required.Generalized linear models do not differ in any important way from

regular linear models in terms of the process of the model specification except that a link function is included to accommodate non-continuous and possibly bounded outcome variables(Gill 2000). Generalized linear models require uncorrelated cases. Time series and spatial problems can be accommodated but not without additional and sometimes complicated enhancements. Also, there can be only one error term specified in the model. While the distribution of this error term is no longer required to be asymptotically normal with constant variance (as in the linear model), approaches such as cell means models with "stacked" error terms are excluded in the basic framework. Finally, generalized linear models are inherently parametric in that the form of the likelihood function is completely defined by the researcher. Relaxation of this requirement through the use of smoothers leads to the more flexible but more complicated form referred to as generalized additive models (Hastie and Tibshirani 1990).

Generalized linear models are an extension of classical linear models introduced by Nelder and Weddeburn in 1972.They showed that regression and analysis of variance methods could be applied to any response variable whose distribution belongs to the exponential family (Stroup and Kachman, 1994). The development of the theory of the generalized linear model is based upon the exponential family of distributions. Fisher (1934) developed the idea that many commonly applied probability mass functions and probability density functions are really just special cases of a more general classification he called the exponential family(Fisher 1934). The basic idea is to identify a general mathematical structure to the function in which uniformly labeled sub-functions characterize individual differences. The label "exponential family" comes from the convention that sub-functions are contained within the exponent component of the natural exponential function. This is not a rigid restriction as any sub-function that is not in the exponent can be placed there by substituting its natural logarithm. The primary payoff to re-parameterizing a common and familiar function into the exponential form is that the isolated sub-functions quite naturally

produce a small number of statistics which compactly summarize even large data sets without any loss of information. Specifically, the exponential family form readily yields sufficient statistics for the unknown parameters. A sufficient statistic for some parameter is one which contains all the information available in a given data set about the parameter(Gill 2000). Three components specify a generalized linear model: A random component identifies the response variable Y and its probability distribution;a systematic component specifies explanatory variables used in a linearpredictor function; and a link function specifies the function of $E(Y)$ that themodel equates to the systematic component. The random component of a GLM consists of a response variable Y with independent observations $(y_1, y_2, ..., y_N)$ from a distribution in the natural exponential family. This family has probability density function or mass function of form

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)] \qquad (1.1.3)$$

The value of the parameter θ_i may vary for $i = 1,2,...,N$ depending on values of explanatory variables. The term $Q(\theta)$ is called the naturalparameter. The systematic component of a GLM relates a vector $(n_1, n_2, ..., n_N)$ to the explanatory variables through a linear model. Let x_{ij} denote the value of predictor $j(j = 1,2,...,p)$ for subject i. Then

$$\eta_i = \sum_j \beta_j x_{ij}; i = 1,..., N$$

This linear combination of explanatory variables is called the linear predictor.

The third component of a GLM is a link function that connects the random and systematic components. Let $\mu_i = E(Y_i), i = 1,..., N.$ The model links μ_i to η_i by $\eta_i = g(\mu_i)$, where the link function g is a monotonic, differentiable function. Thus, g links $E(Y_i)$ to explanatory variables through the formula

$$g(\mu_i) = \sum_j \beta_j x_{ij}, i = 1,..., N$$

The link function $g(\mu) = \mu$, called the identity link, has $\eta_i = \mu_i$. It specifies a linear model for the mean itself. This is the link function for ordinary regression with normally distributed Y. The link function that transforms the mean to the natural parameter is called the canonical link. For it, and $Q(\theta_i) = \sum_j \beta_j x_{ij}$.

Usually the same link function is used for all observations. Then, the canonical link function is that function which transforms the mean to a canonical location parameter of the exponential dispersion family member. With the canonical link function, all unknown parameters of the linear structure have sufficient statistics if the response distribution is a member of the exponential dispersion family and the scale parameter is known. However, the link function is just an artifact to simplify the numerical methods of estimation when a model involves a linear part, that is, to allow the iterative weighted least squares (IWLS) algorithm to work. For strictly nonlinear regression models, it loses its meaning (Lindsey 1974).The canonical links for the common GLMs are shown below:

Family	Link	Variance Function
Normal	$\eta = \mu$	1
Poisson	$\eta = \log\mu$	μ
Binomial	$\eta = \log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	μ^2
Inverse Gaussian	$\eta = \mu^{-2}$	μ^3

The GLM approach is attractive because it (1) provides a general theoretical framework for many commonly encountered statistical models; (2) simplifies the implementation of these different models in statistical software, since essentially the same algorithm can be used for estimation, inference and

assessing model adequacy for all GLMs. In the GLM framework, it is customary to use a quantity known as deviance to formally assess model adequacy and to compare models. Deviance statistics are identical to those obtained using likelihood ratio test statistics.Deviance is an important idea associated with a fitted GLM. It can be used to test the fit of the link function and linear predictor to the data, or to test the significance of a particular predictor variable (or variables) in the model. Let Y_{ik} be the response observed on the k^{th} replicate of the i^{th} distinct combination of covariate values $(x_{i1},...,x_{ip}), k = 1,...,n_i$ and $p = 1,...,k$. Here $n \equiv \sum_{i=1}^{K} n_i$ and $K \le n$. K is the maximum number of parameters that we can estimate from these data. Let $\mu_i \equiv E[Y_{ik}]$. The model of interest specifies a distribution for Y_{ik} and $g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j \equiv \eta_i$ where $p \le K$ is the number of parameters to be estimated. In other words, this model constrains the means μ_i to lie on the surface given by η_i.In the GLM setting we define the saturated model as the GLM with the same distribution and link function as the model of interest, but with $g(\mu_i) = \psi_i$ for $i = 1,...,K$. In other words, the saturated model allows a different mean response $(\mu_i = g^{-1}(\psi_i))$ for each group of replicates, and hence has *K* parameters to be estimated. We will denote this vector of parameters Ψ .we can think of the saturated model as having the most general possible mean structure for the data since the means μ_i are unconstrained. The saturated model is also referred to as the full model or maximal model. Let $L_S(\psi; y)$ and $L(\beta; y)$ be the likelihoods corresponding to the saturated and proposed model, respectively. We know that $L_S(\psi; y) \ge L(\beta; y)$ since the model of interest is a special case of the saturated model. Comparing $L_S(\psi; y)$ and $L(\beta; y)$ or equivalently $l_S(\psi) \equiv \log L_S(\psi; y)$ and $l(\beta) \equiv \log L(\beta; y)$ is one means of assessing how well our assumed

link function and form of the linear predictor fit the data. We define the deviance or likelihood ratio statistic, D as

$$D = 2[l_S(\hat{\psi}) - l(\hat{\beta})], \qquad (1.1.4)$$

Where $\hat{\psi}$ and $\hat{\beta}$ are the MLEs of the saturated and proposed model, respectively. If the proposed model describes the data nearly as well as the saturated model, then asymptotically

$$D \sim \chi^2_{K-p},$$

Where, K and p are the number of parameters in the saturated and proposed models, respectively. If the proposed model is poor, D will be larger than predicted by the χ^2_{K-p} distribution.

1.2 Non-parametric Regression

Since linear models assume:

- a linear relationship of y to the $x's$.
- that the conditional distribution of y is, except for its mean, everywhere the same, and that this distribution is a normal distribution.
- that observations are sampled independently.

These are strong assumptions, and there are many ways in which they can go wrong. For example: As is typically the case in time-series data, the errors may not be independent; the conditional variance of y (the 'error variance') may not be constant; the conditional distribution of y may be very non-normal — heavy-tailed or skewed. Nonparametric regression analysis relaxes the assumption of linearity, substituting the much weaker assumption of a smooth population regression function $f(\cdot)$. The cost of relaxing the assumption of linearity is much greater computation and, in some instances, a more difficult-to-understand result. The gain is potentially a more accurate estimate of the regression function.

Removing the nonparametric assumptions from the classical linear models we get the nonparametric regression model. The general nonparametric regression model is written as:

$$y_i = f(x_i') + \varepsilon_i \qquad (1.2.1)$$
$$= f(x_{i1}, x_{i2}, ..., x_{ik}) + \varepsilon_i$$

Where, f is left unspecified. Moreover, the object of nonparametric regression is to estimate the regression function $f(\cdot)$ directly, rather than to estimate parameters. Compared to the linear model (1.1.1), the nonparametric regression model is more flexible.One has to choose an appropriate value of some smoothing parameters that typically control the smoothness of the estimator. Most methods of nonparametric regression implicitly assume that $f(\cdot)$ is a smooth, continuous function.Nonparametric simple regression is called scatter plot smoothing, because the method passes a smooth curve through the points in a scatter plot of y against x. Scatter plots are (or should be) omnipresent in statistical data analysis and presentation. Nonparametric regression are very flexible but their statistical precision decreases greatly if several explanatory variables are included in the model. The latter caveat has been appropriately termed the curse of dimensionality. Let $X, X_1, X_2, ..., X_n$ be independent and identically distributed R^d- valued random variables with uniformly distributed in the hypercube$[0,1]^d$. Denote the expected supremum-norm distance of X to its nearest neighbor in $X_1, ..., X_n$ by $d_\infty(d,n)$, i.e., set

$$d_\infty(d,n) = E\left\{\min_{i=1,...,n} \left\| X - X_i \right\|_\infty \right\} \qquad (1.2.2)$$

Here $\left\| x \right\|_\infty$ is the supremum norm of a vector $x = (x^{(1)}, ..., x^{(d)})^T \in R^d$ defined by

$$\left\| x \right\|_\infty = \max_{l=1,...,d} \left| x^{(l)} \right|$$

Therefore,

$$d_{\infty}(d,n) = \frac{d}{2(d+1)} \cdot \frac{1}{n^{1/d}}. \qquad (1.2.3)$$

For dimension $d = 10$ or $d = 20$ this lower bound do not approach to zero even if the sample size is very large. So for most values of x one only has data points (X_i, Y_i) available where X_i is not close to x. Butat such data points $m(X_i)$ will, in general, not be close to $m(x)$ even for asmooth regression function. The only way to overcome the curse of dimensionality is to incorporate additional assumptions about the regression function besides the sample. This is implicitly done by nearly all multivariate estimation procedures, including projection pursuit, neural networks, radial basis function networks, trees, etc.

The intention of variable selection is to choose an appropriate subset of variables, $X_r = (X_{j1},...,X_{jr})^T \in X = (X_1,...,X_d)^T$, from the set of all variables that could potentially enter the regression. Of course, the selection of the variables could be determined by the particular problem at hand, i.e, we choose the variables according to insights provided by some underlying economic theory. This approach, however, does not really solve the statistical side of our modeling process. The curse of dimensionality could lead us to keep the number of variables as low as possible. On the other hand, fewer variables could in turn reduce the explanatory power of the model. Thus, after having chosen a set of variables on theoretical grounds in a first step, we still do not know how many and, more importantly, which of these variables will lead to optimal regression results. Therefore, a variable selection method is needed that uses a statistical selection criterion. Vein (1994) has proposed to use the integrated square error (ISE) to measure the quality of a given subset of variables. In theory, a subset of variables is defined to be an optimal subset if it minimizes the integrated squared error

$$ISE(X_r^{opt}) = \min_{X_r}(ISE(X_r))$$

Where, $X_r \subset X$.In practice, the ISE is replaced by its sample analog, the multivariate analog of the cross validation function. After the variables have been selected, the conditional expectation of Y on X_r is calculated by some kind of standard nonparametric multivariate regression technique such as the kernel estimator. The kernel estimate of a regression function takes the form

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)} \qquad (1.2.4)$$

If the denominator is nonzero, and 0 otherwise. Here the bandwidth $h_n > 0$ depends only on the sample size n, and the function $K : R^d \to [0, \infty)$ is called a kernel. Usually $K(x)$ is "large" if $\|x\|$ is "small", therefore the kernel estimate again is a local averaging estimate. The essential idea behind local averaging is that, as long as the regression function is smooth, observations with x-values near a focal x_0 are informative about $f(x_0)$. In local averaging we move a narrow class interval (called a window) continuously over the data, averaging the observations that fall in the window. We can calculate $f(x)$ at a number of focal values of x, usually equally spread within the range of observed x-values, or at the (ordered) observations, $x(1), x(2), ..., x(n)$. We can employ a window of fixed width w centered on the focal value x_0, or can adjust the width of the window to include a constant number of observations, m. These are the m *nearest neighbors* of the focal value. Problems occur near the extremes of the $x's$. For example, all of the nearest neighbors of $x(1)$ are greater than or equal to $x(1)$, and the nearest neighbors of $x(2)$ are almost surely the same as those of $x(1)$, producing an artificial flattening of the regression curve at the extreme left, called *boundary bias*. A similar flattening occurs at the extreme right, near $x(n)$. In addition to the obvious flattening of the regression curve at the left and right, local averages can

be rough, because $\hat{f}(x)$ tends to take small jumps as observations enter and exit the window. The kernel estimator (described shortly) produces a smoother result.Local averages are also subject to distortion when outliers fall in the window, a problem addressed by robust estimation. The optimality of a kernel estimate can be extended using a local polynomial kernel estimate. Similarly to the partitioning estimate, notice that the kernel estimate can be written as a solution of the following minimization problem:

$$m_n(x) = \arg\min_c \sum_{i=1}^{n} (Y_i - c)^2 K_{h_n}(x - X_i) \qquad (1.2.5)$$

To generalize this, choose functions $\phi_0,...,\phi_M$ on R^d, and define the estimate by

$$m_n(x) = \sum_{l=0}^{M} c_l(x)\phi_l(x),$$

Where, $(c_0(x),...,c_M(x)) = \arg\min_{(c_0,...,c_M)} \sum_{i=1}^{n} \left(Y_i - \sum_{l=0}^{M} c_l \phi_l(X_i) \right)^2 K_{h_n}(x - X_i)$

The most popular example for estimates of this kind is the local polynomial kernel estimate, where the $\phi_l(x)$'s are monomial of the components of x. For simplicity we consider only $d = 1$. Then $\phi_l(x) = x^l (l = 0,1,...,M)$, and the estimate m_n is defined by locally fitting polynomial to the data. If $M = 0$, then m_n is the standard kernel estimate. If $M = 1$, then m_n is the so called locally linear kernel estimate.

Index models play an important role in econometric. An index is a summary of different variables into one number, e.g. the price index, the growth index, or the cost of living index. It is clear that by summarizing all the information contained in the variables $X_1,...,X_d$ into one "single index" term we will greatly reduce the dimensionality of a problem. Models based on such an single index models of the following form:

$$E(Y \mid X) = m(X) = g\{v_\beta(X)\}, \tag{1.2.6}$$

Where, $g(\bullet)$ is an unknown link function and $v_\beta(\bullet)$ an up to β specified index function. The estimation can be carried out in two steps. First, we estimate β. Then, using the index values for our observations, we can estimate g by nonparametric regression. Note that estimating $g(\bullet)$ by regressing the Y on $v_\beta(X)$ is only a one-dimensional regression problem. In many applications a canonical partitioning of the explanatory variables exists. In particular, if there are categorical or discrete explanatory variables we may want to keep them separate from the other design variables. Note that only the continuous variables in the nonparametric part of the model cause the curse of dimensionality.

Consequently, researchers have tried to develop models and estimators which offer more flexibility than standard parametric regression but overcome the curse of dimensionality by employing some form of dimension reduction. Such methods usually combine features of parametric and nonparametric techniques. As a consequence, they are usually referred to as semi-parametric methods. Nonparametric simple regression forms the basis, by extension, for nonparametric multiple regression, and directly supplies the building blocks for a particular kind of nonparametric multiple regression called additive regression. Nonparametric regression can be even used when the distribution of the data is not known (Fox 2005).

The nonparametric approach to estimating a regression curve has four main purposes. First, it provides a versatile method of exploring a general relationship between two variables. Second, it gives prediction of observations yet to be made without reference to a fixed parametric model. Third, it provides a tool for finding spurious observations by studying the influence of the isolated points. Fourth, it constitutes a flexible method of substituting for missing values or interpolating between adjacent X-values. The flexibility of the method is

extremely helpful in a preliminary and exploratory statistical analysis of the data set. If a prior model information about the regression curve is not available, the nonparametric analysis could help in suggesting simple parametric formulation of the regression relationship (Härdle and Linton 1994).

1.3 Linear Mixed Effect Model

The linear mixed model (LMM) is very flexible and capable of fitting a large variety of datasets. It is widely used for repeated measures data or longitudinal studies where data are grouped. The form of the LMM that we use is that of Laird and Ware (Laird and Ware 1982) which can be considered an extension of the classical linear model. The mixed model is well-known in statistics. Eisenhart identified three types of linear models: the fixed effects, random effects, and mixed model. The general form of the mixed model is

$$Y = X\beta + Zu + e \tag{1.3.1}$$

Where,

Y is the vector of observations,

X is a matrix of known constants associated with the fixed effects,

β is a vector of fixed effects,

Z is a matrix of known constants associated with the random effects.

u is a vector of random model effects, and

e is a vector of random errors.

The joint distribution of the random models effects and errors is

$$\begin{bmatrix} u \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right)$$

Thus $E(y) = X\beta$ and $\text{var}(y) = ZGZ' + R$. For the purposes of connecting the mixed model and the generalized linear model, it is useful to note that the

conditional expectation of y, $E(y/u) = X\beta + Zu$, and the conditional variance is $\text{var}(y/u) = R$. Mixed models typically incorporate specific random effects that explain the additional between variations in the data not explained by the fixed part of the model. Fixed effects have levels that are of primary interest and would be used again if the experiment were repeated. Random effects have levels that are not of primary interest, but rather are thought of as a random selection from a much larger set of levels. Subject effects are almost always random effects, while treatment levels are almost always fixed effects. The mixed model may be used in a variety of applications. Split-plot models, models for multilocational experiments, quantitative genetics models, etc., are common application of mixed models. The main limitation of the mixed model is the requirement of normally distributed errors (Stroup and Kachman, 1994). The generalized linear mixed model drops this requirement. Generalized linear mixed models (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Alternatively, we could think of GLMMs as an extension of generalized linear models (e.g., logistic regression) to include both fixed and random effects (hence mixed models). A GLMM consists of the following components:

1. The random effects $U_1,...,U_n$, consider the following exponential family of distributions:

$$f(y_{ij} \mid \underline{u}_{ij}, \theta, \phi) = \exp\left\{ \frac{\left[y_{ij}\theta_{ij} - b(\theta_{ij}) \right]}{\phi} + c(y_{ij}, \phi) \right\}, \qquad (1.3.2)$$

Where $\underline{u}_{ij} = (u_{i1},..., u_{ik})$ are variates from normally distributed k-dimensional random vectors

$$U_i \sim N(0, D)$$

D isthe variance-covariance matrix and $\mu_{ij} = E[y_{ij} \mid U_i = \underline{u}_{ij}] = b'(\theta_{ij})$.

The variance of the observations, conditional on the random effects, is given by $\text{var}[y_{ij} \mid U_i = \underline{u}_{ij}] = A_i^{1/2} R_i A_i^{1/2}$. The diagonal matrix A_i contains the variance functions of the model, which express the variance of a response Y_{ij} as a function of its mean μ_{ij}. The matrix R_i is the variance-covariance matrix for the random effects.

2. The linear mixed effects model is defined as:

$$\eta_i = X\beta + Zu$$

for the fixed effects parameter vector $\beta = (\beta_{1`}, \ldots \beta_p)^T$ and random effects vector. Here $X = (x_{i1}, \ldots, x_{ip})^T$ and $Z = (z_{i1}, \ldots, z_{ik})^T$ are both covariates.

3. A link function g,

$$g(\mu_{ij}) = \eta_{ij} \qquad ; i = 1,2,\ldots,n; j = 1,2,\ldots,n_i \tag{1.3.3}$$

completes the model.

Most estimation methods for β and u of GLMMs rest on some form of likelihood principle, and numerical methods are needed in most cases to obtain the estimates. There are two types of numerical algorithms to solve for log-likelihood of (1.3.2) and (1.3.3). The first type is based on Taylor series and hence these algorithms are known as linearization methods. The series expansions give an approximate model based on pseudo-data, with fewer non-linear components. This computation of the linear approximation must be repeated several times until convergence is reached, according to some criterion. Schaben-berger and Gregoire gave several algorithms based on Taylor series for clustered data (Schabenberger and Gregoire 1996). These fitting techniques based on linearizations are usually doubly iterative. The GLMM is first approximated by a linear mixed model based on current values of the covariance parameter estimates. Then the resulting linear mixed model is fitted, forming an iterative process. At convergence, the new

parameter estimates are used to update the linearization, generating a new linear mixed model. The process stops when parameter estimates, for successive fits of the linear mixed model, change only within a specified tolerance. The second type of algorithm is based on integral approximations. The log-likelihood of the GLMM is first approximated before the numerical optimization. Various techniques exist to compute the approximation: Laplace and quadrature methods, Monte Carlo integration, and Markov chain Monte Carlo methods. The advantage of these integral approximation methods is that they give an actual objective function for the optimization step. This allows for likelihood ratio tests among nested models, and the computation of likelihood-based fit statistics. The estimation requires only a single iterative process.

Difference between LMMs and GLMMs is that the response variables can come from different distributions besides Gaussian. In addition, rather than modelling the responses directly, some link function is often applied, such as a log link. The interpretation of GLMMs is similar to GLMs; however, there is an added complexity because of the random effects. On the linearized matrix (after taking the link function), interpretation continues as usual. However, it is often easier to back transform the results to the original metric. For example, in a random effects logistic model, one might want to talk about the probability of an event given some specific values of the predictors. Likewise in a Poisson (count) model, one might want to talk about the expected count rather than the expected log count. These transformations complicate matters because they are nonlinear and so even random intercepts no longer play a strictly additive role and instead can have a multiplicative effect. For parameter estimation, because there are not closed form solutions for GLMMs, we must use some approximation. The generalized linear mixed model provides a unifying framework for linear models. Depending on one's perspective, it allows the extension of generalized linear models to accommodate random effects, or it allows the extension of mixed model methods to accommodate non-normal errors. The generalized linear model, the

mixed model, and the traditional linear model are all special cases of the GLMM. Inference on the GLMM involves straightforward extension of methods used for generalized and mixed linear models. These methods can be used for models with correlated errors, non-scalar link functions, and in principle, can be extended to more general distributions for random model effects.

1.4 Nonlinear Mixed Effects Models

The basic idea of nonlinear models is the same as that of linear models, namely to relate a response Y to a vector of predictor variables $X = (x_1,...,x_k)^T$. Nonlinear model is characterized by the fact that the prediction equation depends nonlinearly on one or more unknown parameters. Whereas linear model is often used for building a purely empirical model, nonlinear model usually arises when there are physical reasons for believing that the relationship between the response and the predictors follows a particular functional form. A nonlinear model has the form

$$y_i = f(x_i, \theta) + \varepsilon_i \; ; i = 1,2,...,n \qquad (1.4.1)$$

Where, the Y_i are responses, f is a known function of the covariate vector $x_i = (x_{i1}, x_{i2},...,x_{ik})^T$, the parameter vector $\theta = (\theta_1, \theta_2,..., \theta_p)^T$ and ε_i are random errors. The ε_i are usually assumed to be uncorrelated with mean zero and constant variance. There are situations where in the nonlinear models either some, or all, of the predictor variables are fixed and random effects such models are known as nonlinear mixed effects models. The Non-linear Mixed Effect Model (NLMM) is a statistical approach that describes the expected trajectory (for example with a parametric nonlinear regression function defined by a mechanistic model), but also accounts for a correlation structure in the data by modeling the between and with-in subject variability using random effects. They can be regarded either as an extension of linear mixed-effects models in which the conditional expectation of the response given the random effects is allowed to be a nonlinear function of the

coefficients, or as an extension of nonlinear models for independent data(Bates, Bates and Watts 1988) in which random effects are incorporated in the coefficients to allow then to vary by group, thus inducing correlation within the groups. In the NLMM approach, we assume that the between subject differences are due to the variations in the mechanistic parameters over the population of subjects. The NLMM (also known as hierarchical nonlinear model) incorporates the mechanistic model, and the between and within variation in the statistical framework appropriate for describing the longitudinal data. Non-Linear Mixed Effects (NLME) models include both fixed effects, which are parameters associated with an entire population or with certain repeatable levels of experimental factors, and random effects, which are associated with individual experimental units drawn at random from a population. Random effects account for spatial and temporal correlation by defining the covariance structure of the model's random components and by using this structure during parameter estimation. NLME models provide an efficient statistical method for explicitly modeling hierarchical stochastic structure. Growth models can be calibrated by predicting random components from tree- or plot-level covariates when a new subject is available and is not used in the fitting of the model by using the empirical best linear unbiased predictors (EBLUPs).

The nonlinear mixed-effects model for repeated measures proposed by Lindstrom and Bates can be thought of as a hierarchical model(Lindstrom and Bates 1990). At one level the j^{th}observation on the i^{th}group is modeled as

$$y_{ij} = f(\phi_{ij}, v_{ij}) + \varepsilon_{ij}, \qquad i = 1,...,M,; j = 1,...,n_i \quad (1.4.2)$$

Where

M is the number of groups,

n_i is the number of observations on the ith groups,

f is a general, real-valued, differentiable function of a group-specific

parameter

Vector ϕ_{ij} and covariate vector v_{ij}, and ε_{ij} is a normally distributed within-group error term. The function f is nonlinear in at least one component of the group-specific parameter vector ϕ_{ij}, which is modeled as

$$\phi_{ij} = A_{ij}\beta + B_{ij}b_i, \qquad b_i \sim N(0,\Psi), \tag{1.4.3}$$

Where

β is a p -dimensional vector of fixed effects

b_i is a q -dimensional random effects vector associated with the ith group (not varyingwith j)

ψ is avariance-covariance matrix .

The matrices A_{ij} and B_{ij} are of appropriate dimensions and depend on the group and possibly on the values of some covariates at the jthobservation. This model is a slight generalization of that described in Lindstrom and Bates (1990) in that A_{ij} and B_{ij} can depend on j. This generalization allows the incorporation of "time-varying" covariates in the fixed effects or the random effects for the model. It is assumed that observations corresponding to different groups are independent and that the within-group errors ε_{ij} are independently distributed as $N(0,\sigma^2)$ and independent of the b_i. The assumption of independence and homoscedasticity for the within –group errors can be relaxed.

Because f can be any nonlinear function of ϕ_{ij}, the representation of the group-specific coefficients ϕ_{ij} could be chosen so that A_{ij} and B_{ij} are always simple incidence matrices. However, it is desirable to encapsulate as much modeling of the ϕ_{ij} as possible in this second stage, as this simplifies the

calculation of the derivatives of the model function with respect to β and b_i, used in the optimization algorithm.

We can write (1.4.2) and (1.4.3) in matrix form as

$$y_i = f_i(\phi_i, v_i) + \varepsilon_i,$$

$$\phi_i = A_i\beta + B_i b_i, \tag{1.4.4}$$

for $i = 1,...,M$ where

$$y_i = \begin{bmatrix} y_{i1} \\ . \\ . \\ . \\ y_{in_i} \end{bmatrix}, \phi_i = \begin{bmatrix} \phi_{i1} \\ . \\ . \\ . \\ \phi_{in_i} \end{bmatrix}, \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ . \\ . \\ . \\ \varepsilon_{in_i} \end{bmatrix}, f_i(\phi_i, v_i) = \begin{bmatrix} f(\phi_{i1}, v_{i1}) \\ . \\ . \\ . \\ f(\phi_{in_i}, v_{in_i}) \end{bmatrix}, v_i = \begin{bmatrix} v_{i1} \\ . \\ . \\ .. \\ v_{in_i} \end{bmatrix},$$

$$A_i = \begin{bmatrix} A_{i1} \\ . \\ . \\ . \\ A_{in_i} \end{bmatrix}, B_i = \begin{bmatrix} B_{i1} \\ . \\ . \\ . \\ B_{in_i} \end{bmatrix} \tag{1.4.5}$$

Different methods have been proposed to estimate the parameters in the NLME model. Some of the estimation methods are described below:

1.4.1 Likelihood estimation

Because the random effects are unobserved quantities, maximum likelihood estimation in mixed-effects models is based on the marginal density of the responses *y*, which, for a model with Q levels of nesting, is calculated as:

$$p(y \mid \beta, \sigma^2, \Psi_1, ..., \Psi_Q) = \int p(y \mid b, \beta, \sigma^2)\, p(b \mid \Psi_1, ..., \Psi_Q)\, db, \tag{1.4.6}$$

Where $p(y \mid \beta, \sigma^2, \Psi_1, ..., \Psi_Q)$ is the marginal density of y, $p(y \mid b, \beta, \sigma^2)$ is the conditional density of y given the random effects b, and the marginal

distribution of b is $p(b \mid \Psi_1,..., \Psi_Q)$. For the NLME model (1.4.2), expressing the random effects variance-covariance matrix in terms of the precision factor Δ, so that $\Psi^{-1} = \sigma^{-2}\Delta^T\Delta$, provides the marginal density of y as

$$p(y \mid \beta, \sigma^2, \Delta) = \frac{|\Delta|^M}{(2\pi\sigma^2)^{(N+Mq)/2}} \prod_{i=1}^{M} \int \exp\left\{ \frac{\|y_i - f_i(\beta, b_i)\|^2 + \|\Delta b_i\|^2}{-2\sigma^2} \right\} db_i \,, \tag{1.4.7}$$

Where $f_i(\beta, b_i) = f_i[\phi_i(\beta, b_i), v_i]$.

Because the model function f can be nonlinear in the random effects, the integral in (1.4.6) generally does not have a closed-form expression. To make the numerical optimization of the likelihood function a tractable problem, different approximations to (1.4.6) have been proposed. Some of these methods consist of taking a first-order Taylor expansion of the model function f around the expected value of the random effects (Sheiner and Beal 1980);(Vonesh and Carter 1992), or around the conditional (on Δ) modes of the random effects (Linstorm and Bates, 1990). Gaussian quadrature rules have also been used (Davidian and Gallant 1992).

We describe three different methods for approximating the likelihood function in the NLME model. The first, proposed by Lindstorm and Bates (1990), approximates (1.4.7) by the likelihood of a linear mixed-effects model. We call this the LME approximation. It is the basis of the estimation algorithm currently implemented in the `nlme` function. The second method uses a Laplacian approximation to the likelihood function, and the last method uses an adaptive Gaussian quadrature rule to improve the Laplacian approximation. The LME, Laplacian, and adaptive Gaussian approximations have increasing degrees of accuracy, at the cost of increasing computational complexity.

1.4.2 Lindstorm and Bates Algorithm

The estimation algorithm described by Lindstorm and Bates (1990)

alternates between two steps, a penalized nonlinear least squares (PNLS) step, and a linear mixed effects (LME) step, as described below. We initially consider the alternating algorithm for the single-level NLME model (1.4.2).

In the PNLS step, the current estimate of Δ (the precision factor) is held fixed, and the conditional modes of the random effects b_i and the conditional estimates of the fixed effects β are obtained by minimizing a penalized nonlinear least squares objective function

$$\sum_{i=1}^{M}\left[\left\|y_i - f_i(\beta, b_i)\right\|^2 + \left\|\Delta b_i\right\|^2\right] \tag{1.4.8}$$

The LME step updates the estimate of Δ based on a first-order Taylor expansion of the model function f around the current estimates of β and the conditional modes of the random effects b_i, which we will denote by $\hat{\beta}^{(w)}$ and $\hat{b}_i^{(w)}$, respectively. Letting

$$\begin{aligned}\hat{X}_i^{(w)} &= \frac{\partial f_i}{\partial \beta^T}\Big|_{\hat{\beta}^{(w)},\hat{b}_i^{(w)}}, \hat{Z}_i^{(w)} = \frac{\partial f_i}{\partial b_i^T}\Big|_{\hat{\beta}^{(w)},\hat{b}_i^{(w)}}, \\ \hat{w}_i^{(w)} &= y_i - f_i(\hat{\beta}^{(w)},\hat{b}_i^{(w)}) + \hat{X}_i^{(w)}\hat{\beta}^{(w)} + \hat{Z}_i^{(w)}\hat{b}_i^{(w)}\end{aligned} \tag{1.4.9}$$

the approximate log-likelihood function used to estimate Δ is

$$l_{LME}(\beta,\sigma^2,\Delta \mid y) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{M}\left\{\log\left|\sum_i(\Delta)\right| + \sigma^{-2}\left[\hat{w}_i^{(w)} - \hat{X}_i^{(w)}\beta\right]^T\sum_i^{-1}(\Delta)\left[\hat{w}_i^{(w)} - \hat{X}_i^{(w)}\beta\right]\right\} \tag{1.4.10}$$

Where $\sum_i(\Delta) = I + \hat{Z}_i^{(w)}\Delta^{-1}\Delta^{-T}\hat{Z}_i^{(w)^T}$. This log-likelihood is identical to that of a linear mixed-effects model in which the response vector is given by $\hat{w}^{(w)}$ and the fixed and random –effects design matrices are given by $\hat{X}^{(w)}$ and $\hat{Z}^{(w)}$, respectively.

Lindstorm and Bates (1990) also proposed a restricted maximum likelihood estimation method for Δ, which consists of replacing the log-likelihood in the

LME step of the alternating algorithm by the log-restricted-likelihood

$$l^{R}_{LME}(\sigma^2,\Delta \mid y) = l_{LME}(\hat{\beta}(\Delta),\sigma^2,\Delta \mid y) - \frac{1}{2}\sum_{i=1}^{M}\log\left|\sigma^{-2}\hat{X}_i^{(w)^T}\sum_{i}^{-1}(\Delta)\hat{X}_i^{(w)}\right| \quad (1.4.11)$$

Note that, because $\hat{X}_i^{(w)}$ depends on both $\hat{\beta}^{(w)}$ and $\hat{b}_i^{(w)}$, changes in either the fixed effects model or the random effects model imply changes in the penalty factor for the log-restricted-likelihood (1.4.11). Therefore, log-restricted-likelihoods from NLME models with different fixed or random effects models are not comparable.

The algorithm alternates between the PNLS and LME steps until a convergence criterion is met. Such alternating algorithms tend to be more efficient when the estimates of the variance-covariance components (Δ and σ^2)are not highly corrected with the estimates of the fixed effects (β). Pinheiro in 1994 has shown that, in the linear mixed-effects model, the maximum likelihood estimates of Δ and σ^2 are asymptotically independent of the maximum likelihood estimates of β. These results have not yet been extended to the nonlinear mixed-effects model (1.4.2)(Pinheiro and J.C.Bates 2000).

Lindstorm and Bates (1990) only use the LME step to update the estimate of Δ. However, the LME step also produces updated estimates of β and the conditional modes of b_i. Thus, one can iterate LME steps by re-evaluating (1.4.9) and (1.4.10) or (1.4.11 for the log-restricted-likelihood)at the updated estimates of β and b_i, as described in Wolfinger and O'connell (1993). Because the updated estimates correspond to the values obtained in the first iteration of a Gauss-Newton algorithm for the PNLS step, iterated LME steps will converge to the same values as the alternating algorithm, though possibly not as quickly.

Wolfinger and O'connell (1993) also shows that, when a flat prior is assumed for β, the LME approximation to the log-restricted-likelihood (1.4.11) is equivalent to a Laplacian approximation (Tierney and Kadane 1986) to the

integral (1.4.6). The alternating algorithm and the LME approximation to the NLME log-likelihood can be extended to multilevel models. For example, for an NLME model with two levels of nesting, the PNLS step consists of minimizing the penalized nonlinear least-squares function

$$\sum_{i=1}^{M}\left\{\sum_{j=1}^{M_i}\left[\left\|y_{ij}-f_{ij}(\beta,b_i,b_{ij})\right\|^2+\left\|\Delta_2 b_{ij}\right\|^2\right]+\left\|\Delta_1 b_i\right\|^2\right\} \tag{1.4.12}$$

To obtain estimates for the fixed effects β and the conditional (on Δ_1 and Δ_2) modes of the random effects b_i and b_{ij}.

Letting

$$\hat{X}_{ij}^{(w)}=\frac{\partial f_{ij}}{\partial \beta^T}\bigg|_{\hat{\beta}^{(w)},\hat{b}_i^{(w)},\hat{b}_{ij}^{(w)}},\hat{Z}_{i,j}^{(w)}=\frac{\partial f_{ij}}{\partial b_i^T}\bigg|_{\hat{\beta}^{(w)},\hat{b}_i^{(w)},\hat{b}_{ij}^{(w)}},\hat{Z}_{ij}^{(w)}=\frac{\partial f_{ij}}{\partial b_{ij}^T}\bigg|_{\hat{\beta}^{(w)},\hat{b}_i^{(w)},\hat{b}_{ij}^{(w)}},$$

$$\hat{w}_{ij}^{(w)}=y_{ij}-f_{ij}(\hat{\beta}^{(w)},\hat{b}_i^{(w)},\hat{b}_{ij}^{(w)})+\hat{X}_{ij}^{(w)}\hat{\beta}^{(w)}+\hat{Z}_{i,j}^{(w)}\hat{b}_i^{(w)}+\hat{Z}_{ij}^{(w)}\hat{b}_{ij}^{(w)}$$

$$\hat{X}_i^{(w)}=\begin{bmatrix}\hat{X}_{i1}^{(w)}\\ \cdot\\ \cdot\\ \cdot\\ \hat{X}_{iM_i}^{(w)}\end{bmatrix},\hat{Z}_i^{(w)}=\begin{bmatrix}\hat{Z}_{i,1}^{(w)}\\ \cdot\\ \cdot\\ \cdot\\ \hat{Z}_{i,M_i}^{(w)}\end{bmatrix},\hat{w}_i^{(w)}=\begin{bmatrix}\hat{w}_{i1}^{(w)}\\ \cdot\\ \cdot\\ \cdot\\ \hat{w}_{iM_i}^{(w)}\end{bmatrix} \tag{1.4.13}$$

The approximate log-likelihood function used to estimate Δ_1 and Δ_2 in the two-level NLME models is

$$l_{LME}(\beta,\sigma^2,\Delta_1,\Delta_2\mid y)=-\frac{N}{2}\log(2\pi\sigma^2)$$

$$-\frac{1}{2}\sum_{i=1}^{M}\left\{\log\left|\sum_i(\Delta_1,\Delta_2)\right|+\sigma^{-2}\left[\hat{w}_i^{(w)}-\hat{X}_i^{(w)}\beta\right]^T\sum_i{}^{-1}(\Delta_1,\Delta_2)\left[\hat{w}_i^{(w)}-\hat{X}_i^{(w)}\beta\right]\right\}$$

Where $\sum_i(\Delta_1,\Delta_2)=I+\hat{Z}_i^{(w)}\Delta_1^{-1}\Delta_1^{-T}\hat{Z}_i^{(w)^T}+\oplus_{j=1}^{M_i}\hat{Z}_{ij}^{(w)}\Delta_2^{-1}\Delta_2^{-T}\hat{Z}_{ij}^{(w)^T}$ and $\oplus$ denotes the

direct sum operator. The corresponding log-restricted-likelihood is

$$l_{LME}^{R}(\sigma^2,\Delta_1,\Delta_2 \mid y) = l_{LME}\left(\hat{\beta}(\Delta_1,\Delta_2),\sigma^2,\Delta_1,\Delta_2 \mid y\right)$$

$$-\frac{1}{2}\sum_{i=1}^{M}\log\left|\sigma^{-2}\hat{X}_i^{(w)T}\sum_{i}^{-1}(\Delta_1,\Delta_2)\hat{X}_i^{(w)}\right|$$

This formulation can be extended to multilevel NLME models with an arbitrary number of levels.

The alternating algorithm is the only estimation algorithm used in the nlme function. It is implemented for maximum likelihood and restricted maximum likelihood estimation in single and multilevel NLME models.

1.4.3 Laplacian Approximation

Laplacian approximation are used frequently in Bayesian inference to estimate marginal posterior densities and predictive distributions (Tierney and Kadane,1986; (Leonard, Hsu and Tsui 1989). These techniques can also be used for approximating the likelihood function in NLME models.

We consider initially the single-level NLME models. The integral that we want to estimate to obtain the marginal distribution of y_i in (1.4.6) can be written as:

$$p(y_i \mid \beta,\sigma^2,\Delta) = \int (2\pi\sigma^2)^{-(n_i+q)/2}\left|\Delta\right|\exp\left[-g(\beta,\Delta,y_i,b_i)/2\sigma^2\right]db_i$$

Where $g(\beta,\Delta,y_i,b_i) = \left\|y_i - f_i(\beta,b_i)\right\|^2 + \left\|\Delta b_i\right\|^2$, the sum of which is the objective function for the PNLS step of the alternating algorithm defined in (1.4.7), Let

$$\begin{aligned}
&\hat{b}_i = \hat{b}_i(\beta,\Delta,y_i) = \arg\min_{b_i} g(\beta,\Delta,y_i,b_i),\\
&g'(\beta,\Delta,y_i,b_i) = \frac{\partial g(\beta,\Delta,y_i,b_i)}{\partial b_i},\\
&g''(\beta,\Delta,y_i,b_i) = \frac{\partial^2 g(\beta,\Delta,y_i,b_i)}{\partial b_i \partial b_i^T}
\end{aligned} \qquad (1.4.14)$$

and consider a second-order Taylor expansion of g around $\hat{b}_i$

$$g(\beta,\Delta,y_i,b_i) \cong g(\beta,\Delta,y_i,\hat{b}_i) + \frac{1}{2}\left[b_i-\hat{b}_i\right]^T g''(\beta,\Delta,y_i,\hat{b}_i)\left[b_i-\hat{b}_i\right] \quad (1.4.15)$$

(The linear term in the expansion vanishes because $g'(\beta,\Delta,y_i,\hat{b}_i)=0$)

The Laplacian approximation is defined as:

$$\begin{aligned} p(y\mid\beta,\sigma^2,\Delta) &\cong (2\pi\sigma^2)^{-\frac{N}{2}}|\Delta|^M \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{M} g(\beta,\Delta,y_i,\hat{b}_i)\right] \\ &\quad\times\prod_{i=1}^{M}\int (2\pi\sigma^2)^{\frac{q}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[b_i-\hat{b}_i\right]^T g''(\beta,\Delta,y_i,\hat{b}_i)\left[b_i-\hat{b}_i\right]\right\}db_i \\ &= (2\pi\sigma^2)^{-\frac{N}{2}}|\Delta|^M \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{M} g(\beta,\Delta,y_i,\hat{b}_i)\right]\prod_{i=1}^{M}\left|g''(\beta,\Delta,y_i,\hat{b}_i)\right|^{-\frac{1}{2}} \end{aligned}$$

The Hessian

$$g''(\beta,\Delta,y_i,\hat{b}_i) = \frac{-\partial^2 f_i(\beta,b_i)}{\partial b_i \partial b_i^T}\Big|_{\hat{b}_i}\left[y_i - f_i(\beta,\hat{b}_i)\right] + \frac{\partial f_i(\beta,\hat{b}_i)}{\partial b_i}\Big|_{\hat{b}_i}\frac{\partial f_i(\beta,b_i)}{\partial b_i^T}\Big|_{\hat{b}_i} + \Delta^T\Delta$$

involves second derivatives of f but, at $\hat{b}_i$, the contribution of

$$\frac{\partial^2 f_i(\beta,b_i)}{\partial b_i \partial b_i^T}\Big|_{\hat{b}_i}\left[y_i - f_i(\beta,\hat{b}_i)\right]$$

is usually negligible compared to that of $\partial f_i(\beta,b_i)/\partial b_i\Big|_{\hat{b}_i}\partial f_i(\beta,b_i)/\partial b_i^T\Big|_{\hat{b}_i}$ ((Bates and Watts 1980)Therefore, we use the approximation

$$g''(\beta,\Delta,y_i,\hat{b}_i) \cong G(\beta,\Delta,y_i) = \frac{\partial f_i(\beta,b_i)}{\partial b_i}\Big|_{\hat{b}_i}\frac{\partial f_i(\beta,b_i)}{\partial b_i^T}\Big|_{\hat{b}_i} + \Delta^T\Delta \quad (1.4.16)$$

This approximation is similar to that used in the Gauss-Newton algorithm for nonlinear least squares and has the advantage of requiring only the first order partial derivatives of f with respect to the random effects. These are usually available as a by-product of the estimation of $\hat{b}_i$, which is a penalized least squares problem, for which standard and reliable code is available.

The modified Laplacian approximation to the log-likelihood of the single-level NLME model (1.4.2) is then given by

$$l_{LA}(\beta,\sigma^2,\Delta,|\, y) = -\frac{N}{2}\log(2\pi\sigma^2) + M\log|\,\Delta\,| - \frac{1}{2}\left\{\sum_{i=1}^{M}\log\left|G(\beta,\Delta,y_i)\right| + \sigma^{-2}\sum_{i=1}^{M} g(\beta,\Delta,y_i,\hat{b}_i)\right\} \tag{1.4.17}$$

Because $\hat{b}_i$ does not depend on σ^2, for given β and Δ the maximum likelihood estimate of σ^2 (based upon l_{LA}) is

$$\hat{\sigma}^2 = \hat{\sigma}^2(\beta,\Delta,y) = \sum_{i=1}^{M} g(\beta,\Delta,y_i,\hat{b}_i)/N$$

We can profile l_{LA} on σ^2 to reduce the dimension of the optimization problem, obtaining

$$l_{LA\,p}(\beta,\Delta) = -\frac{N}{2}\left[1+\log(2\pi)+\log(\hat{\sigma}^2)\right] + M\,\log|\,\Delta\,| - \frac{1}{2}\sum_{i=1}^{M}\log\left|G(\beta,\Delta,y_i)\right|$$

If the model function f is linear in the random effects, then the modified Laplacian approximation is exact because the second-order Taylor expansion in (1.4.15) is exact when $f_i(\beta,b_i) = f_i(\beta) + Z_i(\beta)b_i$.

There does not yet seem to be a straightforward generalization of the concept of restricted maximum likelihood to NLME models. The difficulty is that restricted maximum likelihood depends heavily upon the linearity of the fixed effects in the model function, which does not occur in nonlinear models. Lindstrom and Bates (1990) circumvented that problem by using an approximation to the model function *f* in which the fixed effects, β, occur linearly. This cannot be done for the Laplacian approximation,unless we consider yet another Taylor expansion of the model function, what would lead us back to something very similar to Lindstrom and Bates approach.

The Laplacian approximation generally gives more accurate estimates than the alternating algorithm, as it uses an expansion around the estimated random effects only, while the LME approximation in the alternating algorithm uses an expansion around the estimated fixed and random effects. Because it requires solving a different penalized nonlinear least squares problem for each group in the data and its objective function cannot be profiled on the fixed effects, the Laplacian approximation is more computationally intensive than the alternating algorithm. The algorithm for calculating the Laplacian approximation can be easily parallelized, because the individual PNLS problems are independently optimized.

One of the important assumptions of the classical linear models is that the variance of each disturbance terme_t, conditional on the chosen values of the explanatory variables, is some constant equal to σ_t^2. This is the assumption of homoscedasticity, that is, equal variance.If the variance σ_t^2 of the errors in the nonlinear mixed effects modelis known to depend onx_t, viz.

$$\sigma_t^2 = \frac{\sigma^2}{\psi^2(x_t)}$$

Then the situation is termed as heteroscedasticity. Heteroscedasticity does not destroy the un-biasedness and consistency properties of OLS estimators, but they are no longer efficient, not even asymptotically. This is to be remedied using weighted least squares. And correlation between the members of series of observations is called auto-correlation. The classical linear model assumes that such auto-correlation does not exist in the disturbancesu_t. If such correlation does exist in the nonlinear mixed effects model then the errors are said to be auto–correlated errors. The covariance $Cov(u_t, u_{t+h})$ of the time series depends only on the gap $h$and not on the position of t in time. In consequence, the variance-covariance matrix Γ_n of the disturbance vector $u = (u_1, u_2, ..., u_n)'$ of order $(n \times 1)$

. We Will have banded structure with typical element $\gamma_{ij} = \gamma(i-j)$, where $\gamma(h)$ is the auto covariance function of the process, viz.

$$\gamma(h) = Cov(u_t, u_{t+h});\ h = 0, \pm 1, \pm 2, \ldots$$

The appropriate estimator, were Γ_n known, would be the generalized nonlinear least square estimator.

The agricultural data present in categorical form is being analyzed by using the less precise techniques.In order to analyze the categorical data in a more precise way the present study entitled "Study on Varied Aspects of Linear/Nonlinear Models and their Application in Agriculture" shall be taken with the following objectives:

To study the:

- Extension of linear models allowing the response from exponential family of distributions.
- Extension of linear mixed models allowing non-normal response.
- Extension of basic nonlinear mixed effect model to incorporate the heteroscedasticity and within-group correlated errors
- Non-parametric approach of linear regression model.
- Development of functions on above procedure utilizing SAS and R-software.

1.5 Data Sets

1.5.1 Floricultural Data

Floricultural data set comprises of three varieties of Tulip (*Tulipa* sp.) viz *Hollandia, Carribean Parrot* and *Red Beauty* on which four treatments *Trichoderma. harzianum, T. viride, carbendazim and Gliocladium virens*were applied for the management of *Fusarium oxysporum* Schlecht f.sp.*tulipae*

andeach treatment was replicated three times. The dataset was taken from Division of Floriculture, SKUAST-K. The dataset named Tulipdata() in R Software.It consists of eight variables in total.Data set has 36 rows and 11 columns. The first three columns corresponds to Variety, Treatment, Replication and rest are the variables observed. These variables were named EMER (days taken to sprouting of tulip bulbs), BUD, PHT (plant height of tulip), SLT (scalp length of tulip), DM (diameter of flower), LA (leaf area), DF (duration of flower), LP (leaves per plant).

1.5.2 Pathological Data

Data generated on the mortality of *Venturia inequilis* organism responsible for apple scab by Division of plant pathology. The trial was conducted at three different locations Shalimar (*shalimar*), Zakura (*zakura*), Kpunnil (*kunnil*). Four different chemicals myclobutanil (*Mylobutanil*), hexaconazole (*Hexaconazole*), fenarimol (*Fenarimol*) and defenaconazole (*Defenaconazole*) were used for the control of the disease,given at four different concentrations (.001, 0.01, 0.1, 1.0ml). The dataset was named venturiadata()for analysis in R Software. It has 48 rows and 4 columns. The column names are Location, Treat, Conc., Mortality.

1.5.3 Horticultural Data

Data of three Apple varieties (*Gala Red Lum, Fuji Zehn Aztec, Golden Clone B*) on yieldand trunk cross-sectional area, was collected from High Density Plantation at SKUAST-K Shalimar Campus. The data was named Applehpd()for analysis in R Software. The data has 33 rows and 3 columns. The column names are tree, tca(trunk cross sectional area), yield.

Another data set consists of long term data of apple (*Malus domestica*) from 1974-2015 on area, production and productivity trends.The data was collected from Directorate of Horticulture. And the data was named appleforecast()for analysis in SAS software. The data consists of 42 rows and 4 columns. The column names are time, area, production and productivity.

Different graphical curves for the observed characters were obtained by using the R/SAS Software. This provides the full information on the distribution of the data.

1.6 Utilization of R and SAS software in the study

R Software is an integrated suite of software facilities for data manipulation, simulation, calculation and graphical display. It handles and analyzes data very effectively and it contains a suite of operators for calculations on array and matrices. In addition, it has the graphical capacities for sophisticated graphs and data displays. R Software is available in Windows and Machintosh versions, as well as in various flavors of Unix and Linx.

The R project was started by Robert Gentlemen and Ross Ihaka from the Statistics Department in the University of Auckland in 1995. The software has quickly gained a widespread audience. It is currently maintained by the R core development team a hardworking, international group of volunteer developers.

R Software is an interpreted language, not a compiled one, meaning that all commands typed on the keyboard are directly executed without requiring to build a complete program like in most computer languages (C, Fortran, Pascal,…).R's syntax is very simple and intuitive. For instance, a linear regression can be done with the command lm(y~x) which means "fitting a linear model with y as response and x as predictor". When R Software is running, variables, data, functions, results, etc, are stored in the active memory of the computer in the form of objects which have a name. The user can do actions on these objects with operators (arithmetic, logical, comparison,,..)and functions (which are themselves objects). The arguments in R Software can be objects ("data", formulae, expressions…), some of which could be defined by default in the function; these default values may be modified by the user by specifying options. An R function may require no argument: either all arguments are defined by default(and their values can be modified with the options), or no argument has been defined in the

function. The functions available to the user are stored in a library localized on the disk in a directory called R_HOME/library (R_HOME is the directory where R Software is installed). This directory contains packages of functions, which are themselves structured in directories. The package named base is in a way the core of R Software and contains the basic functions of the language, particularly, for reading and manipulating data. Each package has a directory called R with a file named like the packages (for instance, for the package base, this is the file R_HOME/library/base/R/base). This file contains all the functions of the package.

SAS is a sophisticated computer package containing many components. Originally the letters SAS stand for Statistical Analysis System. SAS software provides comprehensive statistical tools for a wide range of statistical analyses, including analysis of variance, categorical data analysis, cluster analysis, multiple imputation, multivariate analysis, nonparametric analysis, power and sample size computations, psychometric analysis, regression, survey data analysis, and survival analysis. SAS has been utilized in studies, like, nonlinear mixed models, generalized linear models, correspondence analysis, and robust regression. For over three decades, SAS software has been used by programmers, analysts and scientists to manipulate and analyze data. SAS (pronounced as sass) is used around the world in 110 countries at over 40000 sites by more than 4.5 million users.

In both the software's functions available for all practical purposes of data analysis and graphics. However, in some specific situations data analyst is forced to develop his own functions according to his requirements. Consequently, a few functions have been developed according to requirement of this study. A brief summary and detailed codes of these functions have been reported here.

1.6.1 GLMLG (data)

This is a function developed in R Software. It takes the argument data which pertains to the data set to be analyzed.Where the data represents the data or

data frame. It returns the Deviance Residuals, Coefficients, Null deviance, Residual deviance and AIC of Gamma distribution. The codes of the function are as follows:

```
GLMLG=function (data)
{
Tulipdata=read.table("clipboard",header=T)
Names(Tulipdata)
gmdl1<-glm(EMER~.^2,family=Gamma(link=log),Tulipdata)
fit1= summary(gmdl1)
gmdl2<-
glm(EMER~.^2,family=gaussian(link=identity),Tulipdata)
fit2=summary(gmdl2)
gmdl3<-glm(BUD~.^2,family=Gamma(link=log),Tulipdata)
fit3=summary(gmdl3)
gmdl4<-
glm(BUD~.^2,family=gaussian(link=identity),Tulipdata)
fit4=summary(gmdl4)
list(summaryStatisticsofgmdl=fit1,fit2,fit3,fit4)#
gives the null deviance, residual deviance and AIC
}
```

1.6.2 GLMLIG (data)

This function is also developed in R. It also takes the argument data which is to be analyzed. Itreturns the Deviance Residuals, Coefficients, Null deviance, Residual deviance and AIC of Inverse Gaussian distribution. The codes of the function are as follows:

```
>GLMLIG=function (data)
{
Tulipdata=read.table("clipboard",header=T)
Names(Tulipdata)
gmdl5<-
glm(PHT~.^2,family=inverse.gaussian(link="identity"),Tu
lipdata)
fit5= summary(gmdl5)
gmdl6<-glm(EMER~.^2,family=gaussian(link=identity),
Tulipdata)
fit6=summary(gmdl6)
gmdl7<-
glm(SLT~.^2,family=inverse.gaussian(link="identity"),Tu
lipdata)
fit7=summary(gmdl7)
gmdl8<-glm(BUD~.^2,family=gaussian(link=identity),
Tulipdata)
fit8=summary(gmdl8)
list(summaryStatisticsofgmdl=fit5,fit6,fit7,fit8)#
gives the null deviance, residual deviance and AIC
}
```

1.6.3 GLMMEST(data)

This function which is also developed in R returns AIC, BIC, deviance, scaled residuals, random effects, fixed effects, correlation of fixed effects for different estimation methods. The codes of the function are as follows:

```
>GLMMEST(data)

{

Venturiadata=read.table("clipboard",header=T

names(Venturiadata)

fitlmer=lmer(mortality~conc+treat+(1|location),data=Ven
turiadata,REML=FALSE)

A=summary(fitlmer)

fitglmer=glmer(mortality~conc+treat+(1|location),family
=poisson,data=Venturiadata)

B=summary (fitglmer)

fitglmmPQL=glmmPQL(mortality~conc+treat,random=~1|locat
ion,family=poisson,data=Venturiadata)

C=summary(fitglmmPQL)

glmmlassofit<-
glmmLasso(mortality~as.factor(conc)+as.factor(treat),rn
d=list(location=~1),family=poisson(link=log),data=data,
lambda=12,final.re=TRUE)

D=summary(glmmlasso)

List (summarystatisticsofA,B,C,D)

}
```

1.6.4 Relative Bias (est,act)

This is another function developed in R software. It takes two augments viz., estimated value of a particular estimator and actual value for the variable of interest. It returns the average relative bias and average squared relative bias of all the estimators of all the four estimation methods of generalized linear mixed models. The codes of the function follows as:

```
Rb<-function(est,true)
{
M<-length(est)
Arb<-formatC((est-true/m),digit=2)
Asrb<-formatC(sum((est-true)^2)/m),digit=2))
List(AverageRelativeBias=Arb,AverageSquaredRelativeBias
=Asrb)
}
```

1.6.5 Absolute Bias(est,act)

This is another function developed in R software. It also takes two augments viz., estimated value of a particular estimator and actual value for the variable of interest. It returns the average absolute bias and average squared deviation of all the estimators for all the four estimation procedures of generalized linear mixed model. The codes of the function are as follows:

```
AB<-function(est,true)
{
M<-length(est)
AAB<-formatC((est-true),digit=2)
AASB<-formatC(sum((est-true)^2),digit=2)
list(AverageAbsoluteBias=AAB,AverageSquaredDeviations=A
ASB)
}
```

Two functions were also developed in SAS to obtain the nonparametric estimate of a model. The function consists of number of statements in SAS, utilizing the SAS procedure PROC LOESS, PROC SPLINES, PROC PRINT. The

functions returns bandwidth, AICC, residual sum of squares, predicted values.

The function nlme() of R Software is used for fitting of the extended linear mixed effects model. The codes of the function is as follows:

```
>getInitial(yieldk~SSlogis(tca,Asym,xmid,scal)Applehpd)

>nls(yieldkg~ SSlogis(tca, Asym, xmid, scal),apple)

> fm1apple.lis <-nlsList(yieldkg ~ SSlogis(tca, Asym, xmid, scal) | tree,data = Applehpd)

> fm1apple.nlme <- nlme(fm1apple.lis)

> fm1apple.nlme

>summary(fm1apple.nlme)

> fm1apple.nls <- nls(yieldkg ~ logist(tca, Asym, xmid, scal),data = Applehpd, start = c(Asym =141.28, xmid =41.55, scal =14.33))

>summary(fm1apple.nls)

>pairs(fm1apple.nlme)

> fm2apple.nlme <- update(fm1apple.nlme, random = Asym ~ 1)

>anova(fm1apple.nlme, fm2apple.nlme)

>plot(fm1apple.nlme)

>qqnorm(fm2apple.nlme, abline = c(0,1))

>fm3apple.nlme<update(fm2apple.nlme,weights=varPower())

>summary(fm3apple.nlme)

>intervals(fm3apple.nlme)

>anova(fm2apple.nlme,fm3apple.nlme)
```

```
>Variogram(fm3apple.nlme,form=~tca)

>plot(Variogram(fm3apple.nlme,form=~tca,maxDist=70))

>fm4apple.nlme<-
update(fm3apple.nlme,corr=corAR1(o.242))

>fm5apple.nlme<-
update(fm3apple.nlme,corr=corARMA(p=0,q=2))

>fm6apple.nlme<-
update(fm3apple.nlme,corr=corARMA(p=1,q=1))

>anova(fm4apple.nlme,fm5apple.nlme)

>anova(fm4apple.nlme,fm6apple.nlme)

>summary(fm6apple.nlme)

>intervals(fm6apple.nlme)

>plot(Variogram(fm6apple.nlme,form=~tca,maxDist=70))
```

1.7 Brief Resume of Work Done in India and Abroad

Henderson et.al (1959) suggested the linear mixed model, which enabled to model correlation in the data.Harville and Mee (1984) proposed a mixed model procedure for analyzing ordered categorical data. Beitler and Landis (1985) proposed mixed model for categorical data from unbalanced designs which was directly analogous to a two-way ANOVA model for quantitative data. They showed an extension of the fitting constants method developed to estimate model variance components based on appropriate reductions in sums of squares. The resulting variance component estimators were incorporated into the covariance structure of a general linear models Wald statistic to test for treatment differences. Gilmour *et al*., (1985) suggested the analysis of binomial data by a generalized linear mixed models. McLean*et al*.,(1991) suggested a unified approach to mixed linear models.

Nelder and Wedderburn (1972) proposed the iterative weighted linear regression technique can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is provided for these models using log-likelihoods. The generalized linear models developed give a consistent way of linking together the systematic elements in a model with the ransom elements. Too often one meets complex systematic linear models only in connection with normal errors and probit analysis may be used but that seems to have little to do with the linear regression theory. The complementary set of probability distributions would be introduced in the usual way, including the use of transformations of data to attain desirable properties of the errors. The difficult problem of discussing how far transformations can produce both linearity and normality simultaneously now disappears because the models allow two different transformations to be used, one to induce the linearity of the systematic component and one to induce the desired distribution in the error component. The systematic use of log-likelihood-ratios (or, equivalently, differences in deviance) extends the ideas of analysis of variance to other distributions and produces an additive decomposition for the sequential fit of the model. To appreciate the simplicity that this can produce it is only necessary to look at the algebraic complexities arising from the attempts to analyze contingency tables by extensions of the Pearson Chi square approach. Jørgensen (1983) stated that the class of generalized linear models can be extended to allow for correlated observations, nonlinear models and error distributions not of the exponential family form. The extended class of models include a number of important examples, particularly of the composite transformational type. Large-sample inference and maximum likelihood estimation for the extended class of generalized linear models is obtained and the analysis of deviance is generalized to the extended class of models. Calculation of likelihood estimate for a general likelihood by Fisher's scoring method and a related method is considered. Pierce

and Schafer (1986) estimated the residuals in generalized linear models. Liang and Zeger (1986) suggested the longitudinal data analysis using generalized linear models. Williams (1987) in his paper proposedthe one step approximation, derived by (Pregibon 1981), for the changes in the deviance of a generalized linear model when a single case is deleted from the data. Three process of model fitting are distinguished: (i) model selection, (ii) parameter estimation and (iii) prediction of future values. In distinguishing these three processes, it is not assumed that an analysis consists of the successive application of each just once. Models that are selected to fit the data are usually chosen from a particular class and, if the model fitting process is to be useful, this class must be broadly relevant to the kind of data under study. An important characteristic of generalized linear models is that they assume independent (or at least uncorrelated) observations. More generally, the observations may be independent in blocks of fixed known sizes. This assumption of independence is characteristic of the linear models of classical regression analysis, and is carried over without modification to the wider class of generalized linear models. Having selected a particular model, it is required to estimate the parameters and to assess the precision of the estimates. In the case of generalized linear models, estimation proceeds by defining a measure of goodness of fit between the observed data and the fitted values generated by the model. The parameter estimates are the values that minimize the goodness of fit criterion. To be useful, predicted quantities need to be accompanied by measures of precision. These are ordinarily calculated on the assumption that the set-up that produced the data remains constant, and that the model used in the analysis is substantially correct. For an account of prediction as a unifying idea connecting the analysis of covariance and various kinds of standardization see (Lane and Nelder 1982).This approximation suggests a particular set of residuals which can be used, not only to identify outliers and examine distributional assumptions, but also to calculate measures of the influence of single cases on various inferences that can be drawn from the fitted model using likelihood ratio statistics (McCullagh and Nelder, 1989). Zeger and Karim (1991) in their article described

that the generalized linear random effects model in a Bayesian framework and use a Monte Carlo method, the Gibbs sampler, can overcome thenumerical integration of random effects to evaluate likelihoods. The resulting algorithm is flexible to easily accommodate to the changes in the number of random effects and in their assumed distribution when warranted. Lee and Nelder (1996) described hierarchical generalized linear models, which allows random effects to be not normal distributed. Cai and Tsai (1999)proposed the concept of diagnostics for non-linearity in generalized linear models. Dey*et al*.,(2000) introduced a Bayesian perspective ofgeneralized linear models.

Altman (1992) has presented kernel and nearest-neighbor regression estimators for the nonparametric regression. These are the local versions of univariate location estimators, and so they can readily be introduced to beginning students and consulting clients who are familiar with such summaries as the sample mean and median.Livanis *et al*. (2009) have re-examined the use of the primal production function framework using nonparametric regression techniques. Specifically, in this paper they have demonstrated how a nonparametric regression based on a kernel density estimator can be used to estimate a production function using data on corn production from Illinois and Indiana. Nonparametric results have been compared to common parametric specifications using the Nadaraya-Watson kernel regression estimator. The parametric and nonparametric forms have also been compared in terms of describing the true technology of the firm by obtaining measures of the elasticity of scale and the marginal physical product through nonparametric estimation of the gradient of the production surface. Finally, the elasticities of substitution have been compared between both parametric and nonparametric representations. Charytanowicz *et al*.(2015) have proposed a method for determining the soil pore size distribution, constituting the subject of the presented investigations. A research study has been conducted using image analysis algorithms, and in turn, nonparametric statistical techniques. The purpose of this investigation is to discover the relationship between the pore size

and volume of the corresponding pores. They have presented the algorithm which is based on the theory of statistical kernel estimators. This frees it of assumptions in regard to the form of regression function. The approach is claimed to be universal, and can be successfully applied for many tasks in data mining, where arbitrary assumptions concerning the form of regression function are not recommended. Shenoy *et al.*(2015) developed a method for building nonparametric stochastic models of multivariate distributions from large data sets. The motivation is stochastic optimization based on time series forecasting models. The proposed non-parametric stochastic modeling approach is based on multiple quantile regressions with inter-quantile smoothing. The models are built using ADMM optimization approach scalable to large datasets. As an application example, the paper considers forecasting of the loads in the electrical power grid. The forecasted load is used for the electricity procurement in the day-ahead power market. The stochastic optimization trades the costs of advance and spot procurements of the electricity. This problem is currently important because the random variability in the grid power load increases with integration of renewable generation.

Breslow and Clayton (1993) suggested that PQL tends to underestimate somewhat the variance components and (in absolute values) fixed effects when applied to clustered binary data, but the situation improves rapidly for binomial observations having denominators greater than one. Groll and Tutz (2014)presented an approach for fitting of the generalized linear mixed models includes an L1-penalty term that enforces variable selection and shrinkage simultaneously. A gradient ascent algorithm has been proposed that allows to maximize the penalized log-likelihood yielding models with reduced complexity. In contrast to common procedures the utility of the present approach is that it can be used in high-dimensional settings where a large number of potentially influential explanatory variables is available. The method has been investigated in simulation studies and illustrated by use of real data sets.Wolfinger and

O'connell(1993) introduced a psedo-likelihood approach for generalized linear mixed model. Stroup and Kachman (1994) an overview of the methodology for generalized mixed linear models. Relevant background, estimating equations, and general approaches to interval estimation and hypothesis testing was presented. Breslow and Lin (1995) derived general expressions for the asymptotic biases in three approximate estimators of regression coefficients and variance component, for small values of the variance component, in generalised linear mixed models with canonical link function and a single source of extraneous variation.McCulloch (1997) suggested the maximum likelihood algorithm for generalized linear mixed models. Booth and Hobert (1998, 1999) gave the concept of standard errors of prediction in generalized linear models and proposed two new implementations of the EM algorithm for maximum likelihood fitting of generalized linear mixed models. Both methods used random (independent and identically distributed) sampling to construct Monte Carlo approximations at the E-step. One approach involved generating random samples from the exact conditional distribution of the random effects by rejection sampling, using the marginal distribution as a candidate. The second method used a multivariate t importance sampling approximation. In many applications the two methods were complementary. Monte Carlo approximation using random samples allowed the Monte Carlo error at each iteration to be assessed by using standard central limit theory combined with Taylor series methods. Specifically, a sandwich variance estimate was constructed for the maximizer at each approximate E-step. This suggested a rule for automatically increasing the Monte Carlo sample size after iterations in which the true EM step was swamped by Monte Carlo error. Rabe-Hesketh *et al*.(2002) proposed adaptive quadrature for multilevel models. It has been showed that adaptive quadrature works well in problems where ordinary quadrature fails. Furthermore, even when ordinary quadrature works, adaptive quadrature is often computationally more efficient since it requires fewer quadrature points to achieve the same precision.

Chen*et al.* (2002) proposed a Monte Carlo EM algorithm using a rejection sampling scheme to estimate the fixed parameters of the linear predictor, variance components and the semi nonparametric density. Sinha (2004)developed a technique for finding robust maximum likelihood (RML) estimates of the model parameters in GLMM's, which appears to be useful in down weighting the influential data points when estimating the parameters. The asymptotic properties of the robust estimators were investigated under some regularity conditions. Small simulations were carried out to study the behaviour of the robust estimates in the presence of outliers, and these estimates were also compared to the ordinary classical estimates. To avoid the computational problems involving high-dimensional integrals, the author proposed a robust Monte Carlo Newton—Raphson (RMCNR) algorithm for fitting GLMM's. Lele, Nadeem and Schmuland (2012) used data cloning, a simple computational method that exploits advances in Bayesian computation, in particular the Markov Chain Monte Carlo method, to obtain maximum likelihood estimators of the parameters in these models. This method also led to a simple estimator of the asymptotic variance of the maximum likelihood estimator. A frequenist method was suggested to obtain prediction intervals for random effects. Data cloning in the GLMM context by analyzing the Logistic–Normal model for over-dispersed binary data, and the Poisson–Normal model for repeated and spatial counts data was used. Normal–Normal and Binary–Normal mixture models were considered to show how data cloning can be used to study estimability of various parameter. Stroup (2012) introduced the modern concepts, methods and applications of generalized linear mixed models. Hu *et al.*(2015) presented a predictive algorithm which can analyze the network performance in various network conditions and traffic patterns. The said approach is based on the best predictive generalized linear mixed model (GLMM). The parameters of the best predictive GLMM are estimated by minimizing the mean squared prediction error (MSPE). To expedite the parameter learning with the big data collected through the network, the said algorithm introduced regularization, LASSO, and an innovative bootstrap. The merits of this approach validated

through data and simulation are that (1) the highest prediction accuracy even under a model misspecification; and (2) the least computation time compared to the Estimation-oriented GLMM with LASSO and Stepwise Selection GLMM. A major computational advantage of the proposed method is that, unlike some of the current approaches, this method does not require the EM (Expectation-Maximization algorithm) procedure.

Pitt (2003) presented a Generalized Linear Model of disability income claim termination rates based on the data from 1980 to 1998. The model was simplified and was intended to demonstrate the possibilities of using such an approach in the preparation of "standard" tables. Furthermore, a full scale GLM of incidence rates using the 1995 to 1998 data was presented. All available characteristics present in the data were included.The significance of the individual characteristics showed the variation in the model according to the characteristics. Jiao and Chen (2004)used generalized linear models in production model and sequential population analysis to assess the stock of the Atlantic cod. And it was recommend that generalized linear models should be used to identify the appropriate error structure in modeling fish population dynamics.Decker (2012) proposed an attempt to improve upon the traditional Berquist-Sherman Method by using a generalized linear model of case reserves as the basis for restating case reserves at earlier evaluations rather than using average case reserves as the basis.

Xu(2014) autocorrelation in the repeated-measures data, developed one-level and nested two-level nonlinear mixed effects (NLME) models, constructed on the selected base model; the NLME models incorporated random effects of the tree and plot. The best random-effects combinations for the NLME models were identified by Akaike's information criterion, Bayesian information criterion and 2 logarithm likelihood. Heteroscedasticity was reduced with two residual variance functions, a power function and an exponential function. The autocorrelation was addressed with three residual autocorrelation structures: a first-order autoregressive structure [AR(1)], a combination of first-order autoregressive and

moving average structures [ARMA(1,1)] and a compound symmetry structure (CS). (Archontoulis and Miguez 2015) have proposed steps in fitting nonlinear models as described by a flow diagram and discussed each step separately providing examples and updates on procedures used. The following steps have been considered: (i) choose candidate models, (ii) set starting values, (iii) fit models, (iv) check convergence and parameter estimates, (v) find the "best" model among competing models, (vi) check model assumptions (residual analysis), and (vii) calculate statistical descriptors and confidence intervals. The associated feedback mechanisms are also addressed (i.e., model variance homogeneity). In particular, we emphasize the first step (choose candidate models) by providing an extensive library of nonlinear functions (77 equations with the associated parameter meanings) and examples of typical applications in agriculture.

Ngo (2016) introduced generalized linear models using a systematic approach to adapting linear model methods on non-normal data. It also applied different statistical tests to assess the goodness fit and identify potential problems occurring in the model. Tang*et al*.(2016) have introduced a new Bayesian hierarchical generalized linear models, called spike-and-slab LASSO GLMs, for prognostic prediction and detection of associated genes using large-scale molecular data. The proposed model employs a spike-and-slab mixture double-exponential prior for coefficients that can induce weak shrinkage on large coefficients, and strong shrinkage on irrelevant coefficients. They have developed a fast and stable algorithm to fit large-scale hierarchal GLMs by incorporating expectation-maximization (EM) steps into the fast cyclic coordinate descent algorithm. The proposed approach integrates nice features of two popular methods, i.e., penalized LASSO and Bayesian spike-and-slab variable selection. The performance of the proposed method have been assessed via extensive simulation studies. The results have shown that the proposed approach can provide not only more accurate estimates of the parameters, but also better

prediction. They have demonstrated the proposed procedure on two cancer data sets: a well-known breast cancer data set consisting of 295 tumors, and expression data of 4919 genes; and the ovarian cancer data set from TCGA with 362 tumors, and expression data of 5336 genes. Their analyses showed that the proposed procedure can generate powerful models for predicting outcomes and detecting associated genes. The methods have been implemented in a freely available R package BhGLM.

Adjakossa and Nuel (2017)in their paper have focused onthe selection of fixed effects along with the estimation of fixed effects, random effects and variance components in the linear mixed-effects model. They have introduced a selection procedure based on an adaptive ridge (AR) penalty of the profiled likelihood, where the covariance matrix of the random effects is Cholesky factorized. This selection procedure is intended to both low and high-dimensional settings where the number of fixed effects is allowed to grow exponentially with the total sample size, yielding technical difficulties due to the non-convex optimization problem induced by L_0 penalties. Through extensive simulation studies, the procedure have been compared to the LASSO selection and appears to enjoy the model selection consistency as well as the estimation consistency.Schelldorfer, Meier and Bühlmann (2014) proposed l_1 penalized algorithm for fitting high-dimensional generalized linear mixed models (GLMMs). GLMMs can be viewed as an extension of generalized linear models for clustered observations. Our LASSO-type approach for GLMMs should be mainly used as variable screening method to reduce the number of variables below the sample size. They have suggested a refitting by maximum likelihood based on the selected variables only. This is an effective correction to overcome problems stemming from the variable screening procedure that are more severe with GLMMs than for generalized linear models. They have illustrated the performance of our algorithm on simulated as well as on real data examples. Freund and Schapire (1996)introduced a new “boosting” algorithm called

AdaBoost which, theoretically, can be used to significantly reduce the error of any learning algorithm that consistently generates classifiers whose performance is a little better than random guessing. They also introduced the related notion of a "pseudo-loss" which is a method for forcing a learning algorithm of multi-label concepts to concentrate on the labels that are hardest to discriminate. They further described experiments carried out to assess how well AdaBoost with and without pseudo-loss, performs on real learning problems. They have performed two sets of experiments. The first set compared boosting to Breiman's "bagging" method when used to aggregate various classifiers (including decision trees and single attribute-value tests). They also compared the performance of the two methods on a collection of machine-learning benchmarks. In the second set of experiments, they studied in more detail the performance of boosting using a nearest-neighbor classifier on an OCR problem.

PRELIMINARY SUMMARY OF THE DATA

The graphical and the numerical summary of the data provides a comprehensive study of the data. One can be able to know the distribution of the data. R functions and Excel are used for the accomplishment of meeting this requirement.

2.1 Summary features of the data

The summary of the data mainly consists of two aspects:

2.1.1 Numerical summary

2.1.2 Graphical summary

2.1.1 Numerical summary

Numerical summary consists of summary features of the characters (variables) of the horticultural/floricultural/pathological data. The main features of the numerical summary mainly discussed are minimum, maximum, mean, standard deviation, coefficient of variation, skewness, kurtosis, simple growth rate, compound growth rate. These summary features are obtained for all the variables of the floricultural/horticultural/pathological datasets. The features are calculated from Excell and other by making use of the function R software packages.

The Numerical summary of the floricultural dataset discussed in Section 1.5.1 is as follows:

Table 2.1: Numerical summary of floricultural data set

	BUD	EMER	PHT	SLT	DM	LA	DF	LP
Maximum	119.78	141.28	46.50	34.50	2.50	4.00	38.45	38.45
Minimum	110.45	138.16	34.00	21.50	1.00	1.50	34.00	34.00
Mean	114.17	139.54	41.25	30.04	1.68	3.03	36.28	36.28
SD	3.001	0.836	4.564	4.104	0.433	0.572	1.059	1.059
CV	0.026	0.005	0.110	0.136	0.257	0.189	0.029	0.029
Skewness	0.658	0.587	-0.623	-0.756	0.037	-0.234	0.161	0.161
Kurtosis	-1.23	-0.349	-1.374	-1.147	-0.594	0.271	-0.566	-0.566

The summary features of the variables for the pathological data discussed

in Section 1.5.2 is

Table 2.2: Numerical summary of pathological data set

Mortality					
Maximum	28	SD	8.462579	Kurtosis	-1.29736
Minimum	0	CV	0.674757		
Mean	12.54167	Skewness	0.109122		

The summary features of the variables for the horticultural data discussed in Section 1.5.3 is:

Table2.3: Numerical summary of horticultural data sets

	Yield(kg)	**TCA**
Maximum	144.913	74
Minimum	1.593	4
Mean	65.88527273	39
SD	44.14194232	22.47915701
CV	0.669981932	0.576388641
Skewness	0.040746608	1.47731E-17
Kurtosis	-1.227071701	-1.220817204

Area ('000 hectare)		Production ('000 MT)		Productivity (MT per hectare)	
Maximum	171.00	Maximum	1966.42	Maximum	13.07
Minimum	46.19	Minimum	190.45	Minimum	4.12
Mean	90.57	Mean	8.77	Mean	9.40
SD	36.69	SD	4.30	SD	1.67
CV (%)	4.05	CV(%)	4.91	CV (%)	1.78
Skewness	0.86	Skewness	0.81	Skewness	-0.82
Kurtosis	-0.40	Kurtosis	0.19	Kurtosis	1.64
SGAR (%)	2.50	SGAR(%)	9.32	SGAR (%)	1.94
CGAR(%)	0.03	CGAR(%)	0.06	CGAR (%)	0.02

2.1.2 Graphical summary

A graphical summary of the data provides a visual picture of the data.

`Boxplot()`

`Boxplot()` is meant for comprehensive presentation of data. It shows centre as well as spread of a distribution. Thus, variability can be depicted along with point of centrality. A line box is placed at the median value. The width of the box is equal to inter quartile range *IQR*, which is the difference between the third and first quartile. The width of the box shows the variability present within the character. Wiskers, are lines on both sides of the box that extend the edge of the box to either sides of the extreme value or to a distance of $1.5 \times IQR$ from the median whichever is less, (e.g., Khan and Mir(2005)). Box plot of data is obtained by using `boxplot()` function available in R. Argument to this function is data object whose box plot is required. The general format of the function is

Boxplot(data)

To generate various box plots the commands are:

```
>boxplot(Tulipdata$EMER,data=Tulipdata,ylab=”Days taken to sprouting of tulip bulbs”)

>boxplot(Tulipdata$BUD, data=Tulipdata,ylab=”Time of Budding”)

>boxplot(Tulipdata$PHT,data=Tulipdata,ylab=”Plant height of tulip”)

>boxplot(Tulipdata$SLT,data=Tulipdata,ylab=”Scalp length of tulip”)

>boxplot(venturiadata$mortality,data=venturiadata,ylab=‘‘mortality”)

>boxplot(Applehpd$Yield,data=Applehpd,ylab=”Yield”)

>boxplot(Applehpd$TCA,data=Applehpd,ylab=”TCA”)

>boxplot(appleforecast$Area,data=appleforecast,ylab=”Area”)

>boxplot(appleforecast$production,data=appleforecast,ylab=”Production”)

>boxplot(appleforecast$productivity,data=appleforecast,ylab=”Productivity”)
```

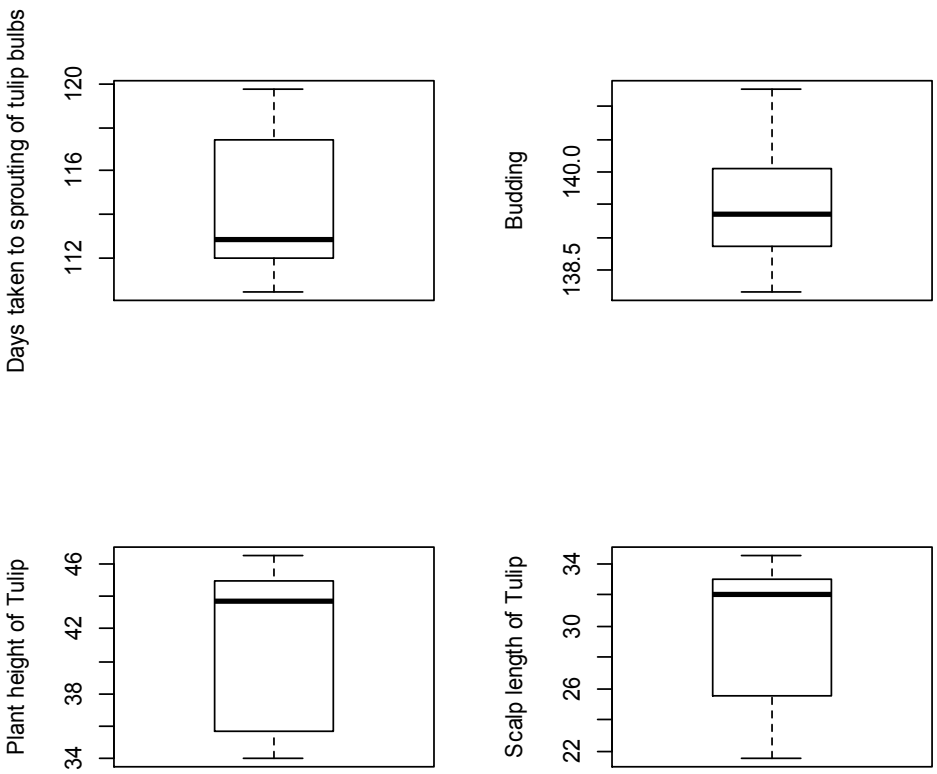


Fig 2.1: Box plot of floricultural data set

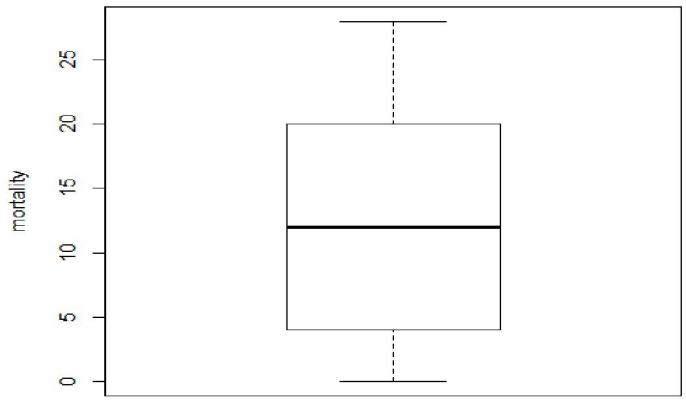


Fig. 2.2: Box plot of pathological data set

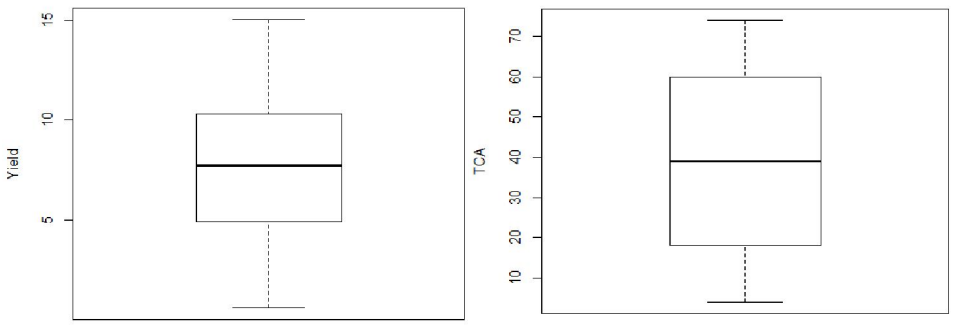


Fig. 2.3: Box plot of horticultural data set

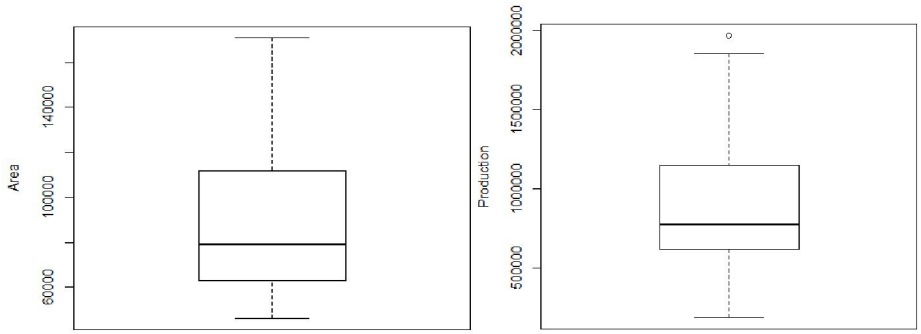


Fig. 2.4: Box plot of area and production of apple

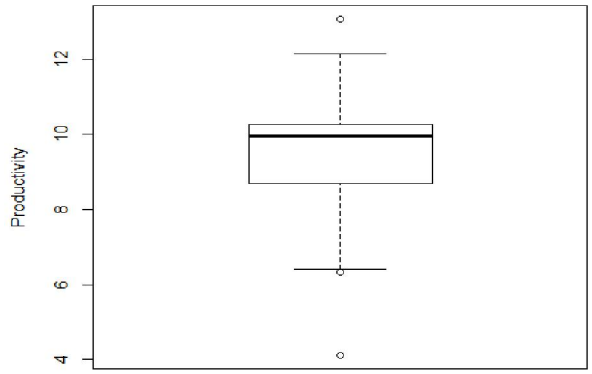


Fig. 2.5: Box plot of productivity of apple

Similarly, to get box plot for the Yield, Trunk Cross-Sectional Area of apple the commands are:

```
>boxplot(Yieldkg~tca,data=Applehpd,ylad="Yieldin
Kg",xlab="Trunk Cross-Sectional Area")
```

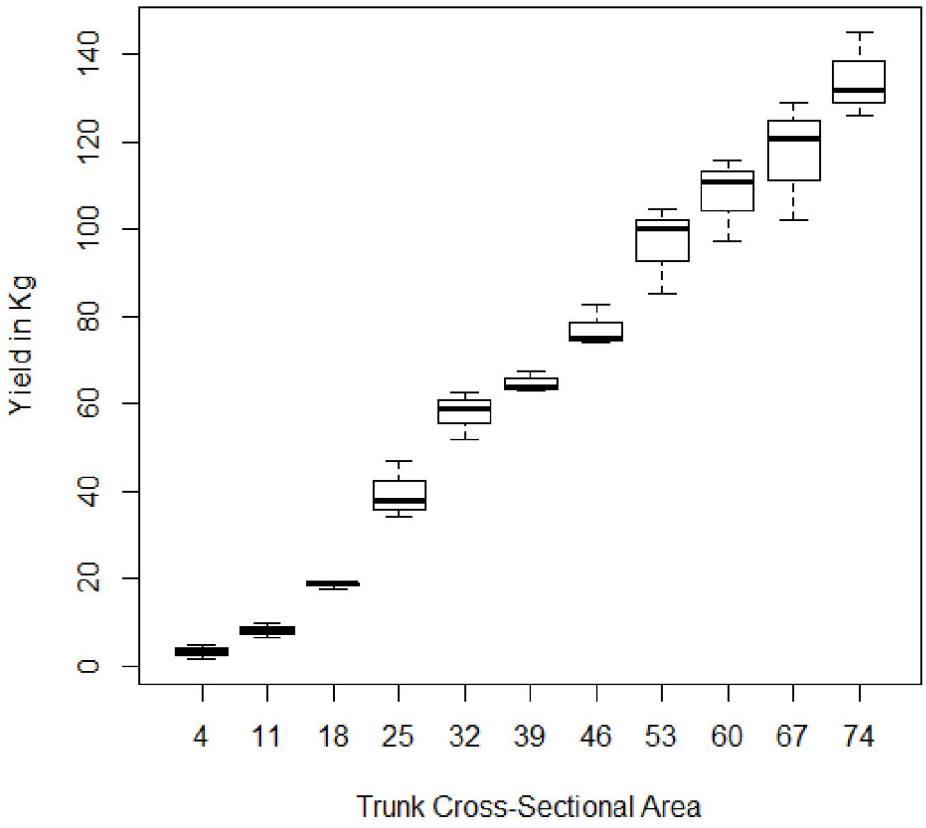


Fig 2.6: Box plot of yield at each trunk cross-sectional area

Quantile-Quantile plot (`QQ-plot`)

The quantile quantile plot is a plot of one set of quantiles against another set of quantile. There are two main forms of `QQ-plot` these are `qqnorm` and `qqline`. The most frequently used form i.e., `qqnorm` checks whether data set comes from a particular normal distribution. In this type of plot one set of quantiles consists of the ordered set of data values and other set of quantiles are from normal distribution. If the points in the plot cluster along the straight line the

data set probably has the normal distribution. The second form i.e., `qqline` fits a line through a normal `qqplot` to check the distribution shape. The general form of these functions is

```
>qqnorm(data)
>qqline(data)
```

To get qqnorm and qqline for different characters of all the datasets we have

```
>qqnorm(venturiadata$mortality)
>qqline(venturiadata$mortality)
>qqnorm(Applehpd$Yield)
>qqline(Applehpd$Yield)
>qqnorm(Applehpd$TCA)
>qqline(Applehpd$TCA)
>qqnorm(appleforecast$Area)
>qqline(appleforecast$Area)
>qqnorm(appleforecast$Production)
>qqline(appleforecast$Production)
>qqnorm(appleforecast$Productivity)
>qqline(appleforecast$Productivity)
```

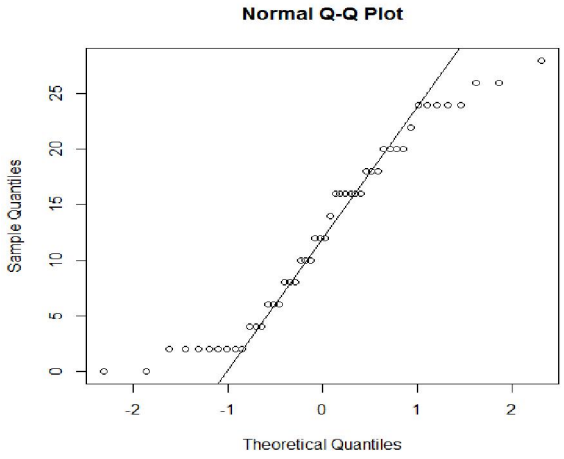


Fig. 2.7: Quantile quantile plot of pathological data set

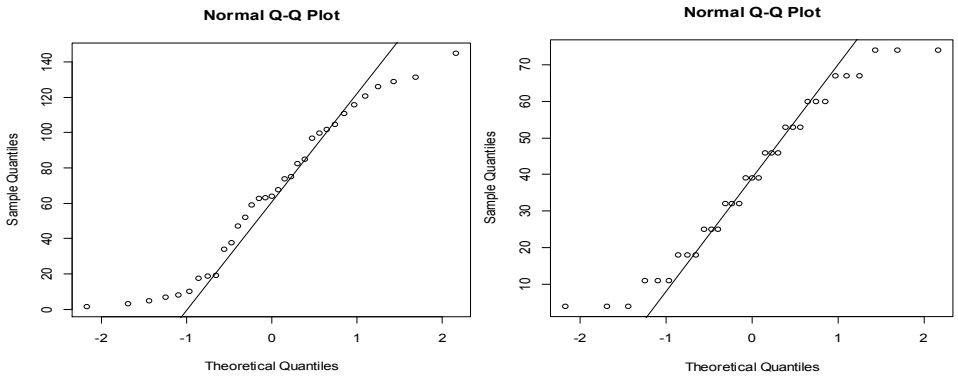


Fig. 2.8: Quantile quantile plot of horticultural data set

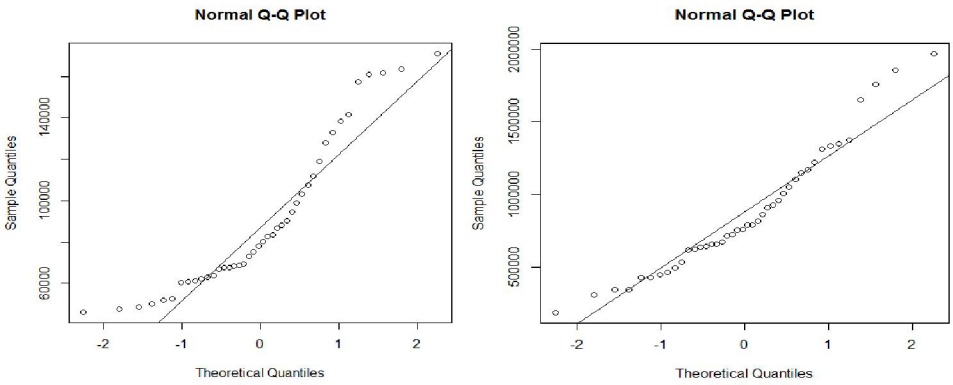


Fig. 2.9: Quantile quantile plot of area and production of apple

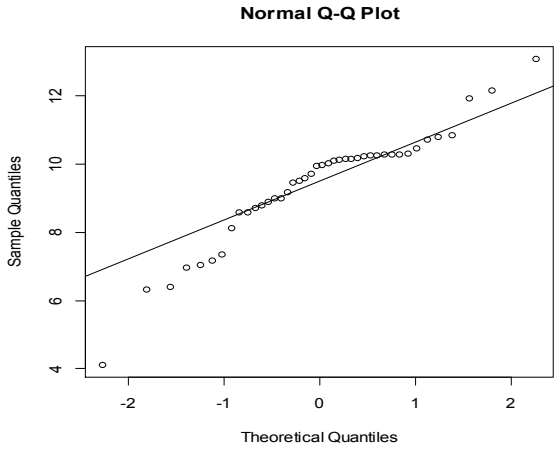


Fig. 2.10: Quantile quantileplot of productivity of apple

Chapter-3

EXTENSION OF LINEAR MODELS ALLOWING THE RESPONSE FROM AN EXPONENTIAL FAMILY OF DISTRIBUTIONS

A GLM is defined by specifying three components. The response should be a member of the exponential family distribution, a linear predictor that involves the regressor variables or covariates and the link function describes how the mean of the response and a linear combination of the predictors are related.

In a GLM the distribution of Y is from the exponential family of distributions which take the general form:

$$f(y \mid \theta, \phi) = \exp\left[\frac{y(\theta) - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{3.1}$$

The θ is called the canonical parameter and represents the location while ϕ is called the dispersion parameter and represents the scale. The most commonly used examples are:

1. Normal or Gaussian:

The density function for the Normal distribution is given by

$$f(y \mid \theta, \phi) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

$$= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]$$

So we can write $\theta = \mu, \phi = \sigma^2, a(\phi) = \phi, b(\theta) = \theta^2/2$ and

$$c(y, \phi) = -(y^2/\phi + \log(2\pi\phi))/2$$

2. Poisson:

The density function for the Poisson distribution is given as

$$f(y \mid \theta, \phi) = e^{-\mu}\mu^{y} / y!$$

$$= \exp(y \log \mu - \mu - \log y!)$$

So we write $\theta = \log(\mu), \phi = 1, a(\phi) = 1, b(\theta) = \exp(\theta)$ and $c(y, \phi) = -\log y!$

3. Binomial:

The density function for the Binomial distribution is given as

$$f(y \mid \theta, \phi) = \binom{n}{y}\mu^{y}(1-\mu)^{n-y}$$

$$= \exp(y \log \mu + (n-y)\log(1-\mu) + \log\binom{n}{y}$$

$$= \exp(y \log \frac{\mu}{1-\mu} + n \log(1-\mu) + \log\binom{n}{y}$$

So we write $\theta = \log \frac{\mu}{1-\mu}, b(\theta) = n \log(1-\mu) = n \log(1 + \exp\theta)$ and

$$c(y, \phi) = \log\binom{n}{y}$$

4)Gamma

The density of the gamma distribution can be given as

$$f(y) = \frac{1}{\Gamma(\nu)}\lambda^{\nu} y^{\nu-1} e^{-\lambda y}; y > 0$$

Where ν describes the shape and λ describes the scale of the distribution. However, for the purposes of a GLM, it is convenient to reparameterize by putting $\lambda = \nu / \mu$ to get:

$$f(y) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-\left(\frac{y\nu}{\mu}\right)}; y > 0$$

Now, $E(Y)=\mu$ and $Var(Y)=\mu^2/v=E(Y)^2/v$. The dispersion parameter is $\phi=\nu^{-1}$. The gamma distribution can arise in various ways. The sum of ν independent and identically distributed exponential random variables with rate λ has a gamma distribution. The χ^2 distribution is a special case of the gamma where $\lambda=1/2$ and $v=df/2$. The canonical parameter is $-1/\mu$, so the canonical link is $\eta=-1/\mu$. However, we typically remove the minus (which is fine provided we take account of this in any derivations) and just use the inverse link. We also have $b(\theta)=\log(1/\mu)=-\log(-\theta)$.

The utility of the gamma GLM arises in two different ways. Certainly, if we believe the response to have a gamma distribution, the model is clearly applicable. However, the model can also be useful in other situations where we may be willing to speculate on the relationship between the mean and the variance of the response but are not sure about the distribution. Indeed, it is possible to grasp the mean to variance relationship from graphical displays with relatively small datasets, while assertions about the response distribution would require a lot more data. Similar is the case with the agricultural datasets. The gamma distribution is merely used in the agriculture to model the different agricultural situations.

There are three common choices of link function:

I. The canonical link is $\eta=\mu^{-1}$. Since $-\infty<\eta<\infty$, the link does not guarantee $\mu>0$ which could cause problems and might require restrictions on β or on the range of possible predictor values. On the other hand the reciprocal link has some advantages. The Michaelis-Menten model has:

$$E(Y)=\mu=\frac{\alpha_0 x}{1+\alpha_1 x}$$

which can be represented after some re-expression as:

$$\eta = \frac{\alpha_1}{\alpha_0 + 1} \Big/ (\alpha_0 x) = \mu^{-1}$$

As x increases, $\eta \to \alpha_1 / \alpha_0$, which means that the mean μ will be bounded. The inverse link can be useful in such situations where we know the mean response to be bounded.

II. The log link, $\eta = \log \mu$,should be used when the effect of the predictors is suspected to be multiplicative on the mean. When the variance is small, this approach is similar to a Gaussian model with a logged response.

III. The linear link, $\eta = \mu$, is useful for modelling sums of squares or variance components which are χ^2. This is a special case of the gamma.

5)Inverse Gaussian

The density of an inverse Gaussian random variable, $Y \sim IG(\mu, \lambda)$ is:

$$f(y \mid \mu, \lambda) = (\lambda / 2\pi y^3)^{1/2} \exp[-\lambda(y-\mu)^2 / 2\mu^2 y]; y, \mu, \lambda > 0$$

The mean is μ and the variance is μ^3 / λ . The canonical link is $\eta = 1/\mu^2$ and the variance function is $V(\mu) = \mu^3$.

The inverse Gaussian has been applied in the modeling of lifetime distributions with non-monotone failure rates and in the first passage times of Brownian motions with drift but it is rarely used for the agricultural setups.

The exponential family distributions have mean and variance:

$$E(Y) = \mu = b'(\theta)$$
$$\text{var}(Y) = b''(\theta) a(\phi)$$

The mean is a function of θ only while the variance is a product of functions of the location and the scale. $b''(\theta)$ is called the variance function and describes how the variance relates to the mean.

3.1 Link Function

Let us suppose we may express the effect of the predictors on the response through a linear predictor:

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x^T \beta \qquad (3.1.1)$$

The link function, g, describes how the mean response, $E(Y) = \mu$, is linked to the covariates through the linear predictor:

$$\eta = g(\mu)$$

In principle, any monotone continuous and differentiable function will do, but there are some convenient and common choices for the standard GLMs. In the Gaussian linear model, the identity link, $\eta = \mu$ is the obvious selection, but another choice would give $y = g^{-1}(x^T \beta) + \varepsilon$. This does not correspond directly to a transform on the response: $g(y) = x^T \beta + \varepsilon$. In a GLM, the link function is assumed known whereas in a single index model, g is estimated. For the Poisson GLM, the mean μ must be positive so $\eta = \mu$ will not work conveniently since η can be negative. The standard choice is $\mu = e^{\eta}$ so that $\eta = \log \mu$ which ensures $\mu > 0$. The canonical link has g such that $\eta = g(\mu) = \theta$, the canonical parameter of the exponential family distribution.

3.2 Estimation and Computational Method

The parameters, β, of a GLM can be estimated using maximum likelihood. The likelihood function of (3.1) is given as

$$l(\theta, \phi \mid y) = \log(f(y \mid \theta, \phi))$$

$$= \frac{y(\theta) - b(\theta)}{a(\phi)} + c(y, \phi) \tag{3.2.1}$$

The first derivative of the log likelihood function with respect to the parameter of interest is the score function. The scale parameter, ψ, for the time being is treated as the nuisance parameter. The score function for (3.2.1) thus obtained is

$$l(\theta \mid \phi, y) = \frac{\partial}{\partial \theta} l(\theta \mid \phi, y) = \frac{y - \partial / \partial \theta b(\theta)}{a(\phi)} \tag{3.2.2}$$

Equating $l(\theta \mid \phi, y)$ to zero and solving for the parameters of interest gives the maximum likelihood estimate, $\hat{\theta}$. Sometimes we can maximize this analytically and find an exact solution for the MLE but the Gaussian GLM is the only common case where this is possible. Typically, we must use numerical optimization. By applying the Newton-Raphson method with Fisher scoring, McCullagh and Nelder (1989) show that the optimization is equivalent to iteratively reweighted least squares (IRWLS). All statistical software uses some iterative root-finding method to find maximum likelihood estimates; the advantage of iterative weighted least squares is that it finds these estimates for any generalized linear model specification based on an exponential family (Green 1984). This technique is carried out in three parts i.e., first is the simple root finding, next is the weighted regression and finally the iterating algorithm.

3.2.1 Newton-Raphson and Root Finding

The maximum likelihood estimate is obtained by Newton-Raphson which is the most famous and widely used technique and is based on Newton's method of finding the roots for polynomial equations. Newton's method is based on a Taylor series expansion. Suppose x_1 is a point such that $f(x_1) = 0$. This a root of the function, $f()$, in the sense that it provides a solution to the polynomial

expressed by the function. Since it is difficult to work with the full length of the Taylor series expansion so we use some subset of the terms as

$$0 \cong f(x_0) + (x_1 - x_0) f'(x_0) \quad (3.2.3)$$

This is referred to as the Gauss-Newton method because it is based on the Newton's algorithm but leads to a least square solution in multivariate problems. Newton's method rearranges (3.2.3) to produce at the $(j+1)^{th}$ step,

$$x^{(j+1)} = x^{(j)} - \frac{f(x^{(j)})}{f'(x^{(j)})}$$

So that continuously improved estimates are produced until $f(x^{(j+1)})$ is sufficiently close to zero. The Newton-Raphson algorithm when to mode finding in a statistical setting adapts (3.2.3)in order to find the root of the score function (the first derivative of the log likelihood) (3.2.2). Treating the score function (3.2.2) as the function of analysis from the Taylor expansion, the iterative estimates are produced by

$$\theta^{(j+1)} = \theta^{(j)} - \frac{\partial / \partial \theta l(\theta^{(j)} \mid y)}{\partial^2 / \partial \theta \partial \theta' l(\theta^{(j)} \mid y)} \qquad (3.2.4)$$

Generalizing (3.2.4) by allowing multiple coefficients where the goal is to estimate a k-dimensional $\hat{\theta}$. The updated multivariate likelihood equation is provided by

$$\theta^{(j+1)} = \theta^{(j)} - \frac{\partial}{\partial \theta} l(\theta^{(j)} \mid y) \left(\frac{\partial}{\partial \theta \partial \theta'} l(\theta^{(j)} \mid y) \right)^{-1} \qquad (3.2.5)$$

Sometimes it becomes difficult to calculate the Hessian matrix, $H = \frac{\partial^2}{\partial \theta \partial \theta' l(\theta^{(j)\mid} \mid y)}$ in such situation it is replaced by its expectation with regard to θ, $A = E_\theta(\partial^2 / \partial \theta \partial \theta' l(\theta^{(i)} \mid y))$. This modification is called Fisher scoring (1925). At each step of the Newton-Raphson algorithm, a system of equations

determined by the multivariate normal equations must be solved which are of the form:

$$(\theta^{(j+1)} - \theta^{(j)})A = -\frac{\partial}{\partial \theta^{(j)}} l(\theta^{(j)} \mid y)$$

3.2.2 Weighted Least Squares

The standard technique for compensating for non-constant error variance (heterosedasticity) is to insert a diagonal matrix of weights, Ω, into the calculation of an estimate in the estimation of a linear regression coefficients such that the heterosedasticty is mitigated. The Ω matrix is created by taking the error variance of the i^{th} case (estimated or known), v_i, and assigning the inverse to the i^{th} diagonal: $\Omega_{ii} = 1/v_i$ (large error variances are reduced by multiplication of the reciprocal). The weighted least squares estimator gives the best linear unbiased estimate (BLUE) of the coefficient estimator in presence of heterosedasticity.

3.2.3 Iterative Weighted Least Squares

Sometimes the individual variances used to make the reciprocal diagonal values for Ω are unknown and cannot be estimated easily, but it is known that these individual variances are the function of the mean of the outcome variances: $v_i = f(E[Y_i])$. Thus, if the expected value of the outcome variable, $E[Y_i] = \mu$, and the form of the relation function, $f()$, are known then this is a very straightforward estimation procedure. Unfortunately, even if the variance structures are dependent on the mean function, it is relatively rare to know the exact form of the dependences.Iteratively estimating the weights is the solution to such problems, the estimates are improved on each cycle using the mean function. So the algorithm iteratively estimates these quantities using progressively improving weights. This proceeds as follows:

I. Assign starting values to the weights, generally equal to one $1/v_i^{(1)} = 1$ and construct the diagonal matrix Ω guarding against division by zero.

II. Estimate θ using weighted least squares with the current weights. The j^{th} estimate is $\hat{\theta}^{(j)} = (X'\Omega^{(j)}X)^{-1}X'\Omega^{(j)}Y$

III. Update the weights using the new estimated mean vector $1/v_i^{(j+1)} = VAR(\mu_j)$

IV. Repeat steps 2 and 3 until convergence (i.e., $X\hat{\theta}^{(j)} - X\hat{\theta}^{(j+1)}$ is sufficiently close to zero).

Under very general conditions, satisfied by the exponential family of distributions, the iterative weighted least squares procedure provides the mode of the likelihood function, thus producing the maximum likelihood estimate of the unknown coefficient vector, $\hat{\theta}$. Furthermore, the matrix produced by $\hat{\sigma}^2(X'\Omega X)^{-1}$ converges in probability to the variance matrix of $\hat{\theta}$ as desired.Since we have an explicit link function identified in a generalized linear model, the form of the multivariate normal equation is modified to include this embedded transformation.

$$(\theta^{(j+1)} - \theta^{(j)})A = -\frac{\partial l(\theta^{(j)} \mid y)}{\partial g^{-1}(\theta)} \frac{\partial g^{-1}(\theta)}{\partial(\theta)}$$

When considering the choice of model for some data, we should define the range of possibilities. The null model is the smallest model we will entertain while the full or saturated model is the most complex. The null model represents the situation where there is no relation between the predictors and the response. Usually this means we fit a common mean μ for all y, that is, one parameter only.

3.3 Model Selection

Akaike information criteria is used for the selection of an appropriate distribution.Akaike information criteria is given by

$$AIC = -2\log p(L) + 2p \qquad (3.3.1)$$

The AIC criterion, which is minus twice the maximized likelihood plus twice the number of parameters, has often been used as a way to choose between models. Smaller values are preferred. However, when computing a likelihood function, it is common practice to discard parts that are not functions of the parameters. This has no consequence when models with same distribution for the response are compared since the parts discarded will be equal.For responses with different distributions, it is essential that all parts of the likelihood be retained.

In the saturated model, the data is explained exactly. Typically, we need to use nparameters for n data points. This can often be achieved by fitting a sufficiently high-orderpolynomial or by treating the numerical values of quantitative predictors as codes,thereby changing them into qualitative predictors. If enough interactions are included, themodel will be saturated. This model tells us no more than the data itself and is usuallyuninformative.A statistical model describes how we partition the data into systematic structure andrandom variation. The null model represents one extreme where the data is representedentirely as random variation, while the saturated or full model represents the data asbeing entirely systematic.The full model does give us a measure of how well any model could possibly fit andso we might consider the difference between the log-likelihood for the full model, $l(y, \phi \mid y)$ and that for the model under consideration, $l(\hat{\mu}, \phi \mid y)$ expressed as a likelihoodratio statistic:

$$2(l(y, \phi \mid y) - l(\hat{\mu}, \phi \mid y))$$

Provided that the observations are independent and for an exponential family distribution,when $a_i(\phi) = \phi / w_i$, this simplifies to:

$$\sum_i 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))/\phi \qquad (3.3.2)$$

Where, $\tilde{\theta}$ is the estimate under the full (saturated) model and $\hat{\theta}$ is the estimate underthe model of interest. The above can be written simply as $D(y,\hat{\mu})/\phi$ where $D(y,\hat{\mu})$ iscalled the deviance and $D(y,\hat{\mu})/\phi$ is called the scaled deviance. Deviances for thecommon GLMs are given as:

GLM	Deviance
Gaussian	$\sum_i (y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_i [y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$
Binomial	$2\sum_i [y_i \log(y_i/\hat{\mu}_i) + (m - y_i)\log((m - y_i)/(m - \hat{\mu}_i))]$
Gamma	$2\sum_i [-\log(y/\hat{\mu}) + (y - \hat{\mu})/\hat{\mu}]$
Inverse Gaussian	$\sum_i (y_i - \hat{\mu}_i)^2/(\mu_i^2 y_i)$
Pearson's χ^2 statistic	$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$

Where $V(\hat{\mu}) = Var(\hat{\mu})$ is an alternative measure of discrepancy that is sometimes used in place of the deviance. There are two main types of hypothesis test we shall employ. The goodness of fit test simply asks whether the current model fits the data. The other type of test compares two nested models where the smaller model represents a linear restriction on the parameters of the larger model. The goodness of fit test can be viewed as model comparison test if we identify the smaller model with the model of interest and the larger model with thefull or saturated model. For the goodness of fit test, we use the fact that, under certain conditions, provided the model is correct, the scaled Deviance and the Pearson's χ^2 statistic are both asymptotically χ^2 with degrees of freedom equal to the number of identifiable parameters. For GLMs such as the Gaussian, we usually do not know

the value of the dispersion parameter ϕ, and so this test cannot be used. For comparing a larger model, W, to a smaller nested model, ω ,the difference in the scaled deviances, $D_\omega - D_W$ is asymptotically χ^2 with degrees of freedom equal to the difference in the number of identifiable parameters in the two models. For the Gaussian model and other models where the dispersion ϕ is usually not known, this test cannot be directly used. However, if we insert an estimate of ϕ we may compute an F-statistic of the form:

$$\frac{(D_\omega - D_W)/(df_\omega - df_W)}{\hat{\phi}}$$

Where $\hat{\phi} = \chi^2/(n-p)$ is a good estimate of the dispersion. For the Gaussian model, $\hat{\phi} = RSS_W / df_W$ and the resulting F-statistic has an exact F distribution for the null. For other GLMs with free dispersion parameters, the statistic is only approximately F distributed. For every GLM except the Gaussian, an approximate null distribution must be used whose accuracy may be in doubt particularly for smaller samples. However, the approximation is better when comparing models than for the goodness of fit statistic.

Residuals represent the difference between the data and the model and are essential to explore the adequacy of the model. In the Gaussian case, the residuals are $\hat{\varepsilon} = y - \hat{\mu}$. These are called response residuals for GLMs, but since the variance of the response is not constant for most GLMs, some modification is necessary. We would likeresiduals for GLMs to be defined such that they can be used in a similar way as in the Gaussian linear model.The Pearson residual is comparable to the standardized residuals used for linear models and is defined as:

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}$$

where, $V(\mu) = b''(\theta)$. These are just a rescaling of $y - \hat{\mu}$.Notice that $\sum r_P^2 = \chi^2$ and hence the name. Pearson residuals can be skewed for non-normal responses. The deviance residuals are defined by analogy to Pearson residuals. The Pearson residual was r_P such that $\sum r_P^2 = \chi^2$ so we set the deviance residual as r_D such that $\sum r_D^2 = Deviance = \sum d_i$.

Thus:

$$r_D = sign(y - \hat{\mu})\sqrt{d_i}$$

3.3.1 Quasi likelihood

Computation of $\hat{\beta}$ and standard errors is often not enough and some form of inference is required. To compute a deviance, we need a likelihood and to compute a likelihood we need a distribution. At this point, we need a suitable substitute for a likelihood that can be computed without assuming a distribution.

Let Y_i have a mean μ_i and variance $\phi V(\mu_i)$. We assume that Y_i are independent. We define a score, U_i:

$$U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}$$

Now,

$$E(U_i) = 0$$

$$Var(U_i) = \frac{1}{\phi V(\mu_i)}$$

$$-E\frac{\partial U_i}{\partial \mu_i} = -E\frac{-\phi V(\mu_i) - (Y_i - \mu_i)\phi V'(\mu_i)}{[\phi V(\mu_i)]^2} = \frac{1}{\phi V(\mu_{i)})}$$

These properties are shared by the derivative of the log-likelihood, l'. This suggests that we can use U in place of l'. So we define:

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$$

The intent is that Q should behave like the log-likelihood. We then define the log quasi-likelihood for all n observations as:

$$Q = \sum_{i=1}^{n} Q_i$$

The usual asymptotic properties expected of maximum likelihood estimators also hold for quasi-likelihood based estimators as may be seen in (McCullagh 1983) notice that the quasi-likelihood depends directly only on the variance function and that the choice of distribution also determines only the variance function. So the choice of variance function is associated with the random structure of the model while the link function determines the relationship with the systematic part of the model.

For the variance functions associated with the members of the exponential family distribution, the quasi-likelihood corresponds exactly to the log-likelihood. However, there is an advantage to using the quasi-likelihood approach for models with variance functions corresponding to the binomial and Poisson distribution. The regular GLMs assume $\phi = 1$ whereas the corresponding quasi-binomial and quasi-Poisson GLMs allow the dispersion ϕ to be a free parameter which is useful in modeling over-dispersion. One curious possibility is that some choices of $V(\mu)$ may not correspond to a known, or even any, distribution.

$\hat{\beta}$ is obtained by maximizing Q. Everything proceeds as in the standard GLMs except for the estimation of ϕ since the likelihood approach is not reliable here. We recommend:

$$\hat{\phi} = \frac{X^2}{n - p}$$

Although quasi-likelihood estimators are attractive because they require fewer assumptions, they are generally less efficient than the corresponding regular likelihood based estimator. So if you have information about the distribution, you are advised to use it.

The inferential procedures are similar to those for standard GLMs. Recall that the regular deviance for a model is formed from the difference in log-likelihoods for the model and the saturated model:

$$D(y,\hat{\mu}) = -2\phi\sum_i (l(\hat{\mu}_i \mid y_i) - l(y_i \mid y_i))$$

So by analogy the quasi-deviance is $-2\phi Q$ because the contribution from the saturated model is zero. The ϕ cancels, so the quasi-deviance is just:

$$Q = -2\sum_i \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)}\, dt$$

3.4 Model diagnostics

We may divide diagnostic methods into two types. Some methods are designed to detect single cases or small groups of cases that do not fit the pattern of the rest of the data. Outlier detection is an example of this. Other methods are designed to check the assumptions of the model. These methods can be subdivided into those that check the structural form of the model, such as the choice and transformation of the predictors, and those that check the stochastic part of the model, such as the nature of the variance about the mean response. Here, we focus on methods for checking the assumptions of the model. For linear models, the plot of residuals against fitted values is probably the single most valuable graphic. For GLMs, we must decide on the appropriate scale for the fitted values. Usually, it is better to plot the linear predictors $\hat{\eta}$ rather than the predicted responses $\hat{\mu}$.

The binomial, Gaussian and Poisson GLMs are by far the most commonly used, but there are a number of less popular GLMs which are useful for particular types of data. The gamma and inverse Gaussian are intended for continuous, skewed responses. We will study these distributions and show how these distributions can be fitted on the agricultural data.

When dealing with the agricultural data it is common practice to fit the data sets by the linear regression models or by simply applying some design of experiment. The data is seldom checked for the normality. We have collected one such data set in which the data was not normal in its variables. We have applied generalized linear model based on Gamma and Inverse Gaussian distributions, on suchagricultural dataset. We have checked Akaike Information Criteria (AIC) for four different distributions i.e., Gamma, Inverse Gaussian, Normal and Multinomial distributions to evaluate which distribution fits our data set best. AIC provided the evidence that the data was not normal and instead of correcting the data for normality we proceed as such and fit the data by generalized linear model approach. Further, since these models are not nested so one may point out that AIC is not an appropriate approach for this reason we have used the deviance approach for comparison purpose.

3.5 Numerical illustration

The Floricultural data set consists of three varieties of Tulip (*Tulipa* sp.) viz *Hollandia, Carribean Parrot* and *Red Beauty* on which four treatments *Trichoderma. harzianum, T. viride, carbendazim* and *Gliocladium virens*were applied for the management of *Fusarium oxysporum* Schlecht f.sp.*tulipae*Apt each treatment was replicated three times. Eight different parameters were taken into consideration and these parameters were EMER (days taken to sprouting of tulip bulbs), BUD, PHT (plant height of tulip), SLT (scalp length of tulip), DM (Diameter of Flower), LA (leaf area), DF (Duration of Flower), LP (leaves per plant).

Table 3.1:AIC value of different distributions for various variables

Distribution	Link	Bud	Emer	PHT	SLT	DM	LA	DF	LP
Gamma	Log	**106.56**	**105.58**	268.32	204.49	106.31	114.74	**157.12**	170.53
Inverse Gaussian	Identity	108.6	107.18	**264.16**	**197.25**	109.47	114.34	156.31	**169.1**
Normal	Identity	109.62	110.15	271.44	199.92	**98.49**	**106.93**	157.69	170.15
Multinomial	Cumulative Logit	1368	1560	622.16	648.04	186	269.35	1224	1128

Table 3.1 gives the AIC values of different distributions with different link functions. AIC value for Gamma and Inverse Gaussian is minimum for most of the Variables.However, when computing a likelihood, it is common practice to discard parts that are not functions of the parameters. This has no consequence when models with same distribution for the response are compared since the parts discarded will be equal.For responses with different distributions, it is essential that all parts of the likelihood be retained.Thus, we have obtained the null deviance and residual deviance of normal, gamma and inverse Gaussian distributions. And the distributions are compared on the basis of the deviance values.

Table 3.2: AIC and deviance for Gamma and Inverse Gaussian distributions

	EMER			**BUD**		
	Null deviance	**Residual deviance**	**AIC**	**Null deviance**	**Residual deviance**	**AIC**
Gamma	0.0239	0.00018	105.58	0.00125	0.00016	106.56
Normal	315.311	2.3956	110.15	24.4707	3.1195	109.62
	PHT			**SLT**		
	Null deviance	**Residual deviance**	**AIC**	**Null deviance**	**Residual deviance**	**AIC**
Inverse Gaussian	0.0113	0.0003	264.16	0.0248	0.0003	197.25
Normal	729.250	20.306	271.44	589.6875	6.7917	199.92

Table 3.2 gives the null deviance and the residual deviance of the Gamma and Inverse Gaussian distributions in comparison with the Normal distribution. We note that the null deviance and Residual Deviance are minimum for Gamma and Inverse Gaussian as compared to the Normal distribution. Comparison of fitted Gamma and Inverse Gaussian generalized linear models for some of the

variables with Normal generalized linear models and are given in Table 3.3,3.4,3.5,3.6.

A comparison of the generalized linear model based on Gamma distribution and Normal distribution for the Emergence variable is given in the Table 3.3.Further, we obtain the dispersion parameter for Normal GLM taken to be 0.1996361 and for Gamma GLM taken to be 1.534627e-05.

Table 3.4 gives the comparison of the generalized linear model based on the Gamma and Normal distributions for the Bud variable. Further, the Dispersion parameter for Normal GLM is taken to be 0.2599542 and for Gamma GLM taken to be 1.331048e-05.

Table 3.3:Comparison of fitted Gamma GLM and Inverse Gaussian GLM for emergence

EMER				
Coefficients	Gamma		Normal	
	Estimate (Std.Error)	**t value Pr(>|t|)**	**Estimate (Std.Error)**	**t value Pr(>|t|)**
(Intercept)	4.715413 ±(0.0032)	1474.225 < 2e-16 ***	111.6497 ±(0.3648)	306.044 < 2e-16 ***
VARIETYV2	0.049096 ±(0.003917)	12.533 2.98e-08 ***	5.6192 ±(0.4468)	12.576 2.86e-08 ***
VARIETYV3	0.008582 ±(0.003917)	2.191 0.0489*	0.9617 ±(0.4468)	2.152 0.0524.
TREATT2	-0.01055 ±(0.004129)	-2.555 0.0252*	-1.1722 ±(0.471)	-2.489 0.0285 *
TREATT3	-0.00165 ±(0.004129)	-0.4 0.696	-0.1733 ±(0.471)	-0.368 0.7193
TREATT4	0.002933 ±(0.004129)	0.71 0.4911	0.3367 ±(0.471)	0.715 0.4884
REPLICATR2	-0.00208 ±(0.003917)	-0.53 0.6058	-0.2183 ±(0.4468)	-0.489 0.6339
REPLICATR3	0.008166 ±(0.003917)	2.085 0.0591.	0.9192 ±(0.4468)	2.057 0.0621.
VARIETYV2:TREATT2	0.007829 ±(0.004523)	1.731 0.1091	0.8667 ±(0.5159)	1.68 0.1188
VARIETYV3:TREATT2	0.005495 ±(0.004523)	1.215 0.2478	0.61 ±(0.5159)	1.182 0.26
VARIETYV2:TREATT3	0.008519 ±(0.004523)	1.883 0.0841.	0.9833 ±(0.5159)	1.906 0.0809.
VARIETYV3:TREATT3	0.002759 ±(0.004523)	0.61 0.5532	0.3067 ±(0.5159)	0.594 0.5633
VARIETYV2:TREATT4	-0.00476 ±(0.004523)	-1.052 0.3133	-0.5467 ±(0.5159)	-1.06 0.3102
VARIETYV3:TREATT4	-0.00104 ±(0.004523)	-0.229 0.8224	-0.1133 ±(0.5159)	-0.22 0.8298

*** highly significant, * significant

Table 3.4: Comparison of fitted Gamma GLM and Normal GLM for bud

Coefficients	BUD			
	Gamma		Normal	
	Estimate (Std.Error)	**t value Pr(>|t|)**	**Estimate (Std.Error)**	**t value Pr(>|t|)**
(Intercept)	4.9292792 ±(0.0030)	1654.749 < 2e-16 ***	138.275 ±(0.4163)	332.157 <2e-16***
VARIETYV2	0.0143561 ±(0.0036)	3.935 0.00198 **	2.00333 ±(0.5098)	3.929 0.002**
VARIETYV3	0.0032348 ±(0.0036)	0.887 0.39269	0.44917 ±(0.5098)	0.881 0.3956
TREATT2	0.0066581 ±(0.0038)	1.731 0.109	0.92889 ±(0.5374)	1.728 0.1095
TREATT3	0.001152 ±(0.0038)	0.3 0.76965	0.16222 ±(0.5374)	0.302 0.7679
TREATT4	0.00820 ±(0.00384)	2.133 0.05425.	1.14556 ±(0.5374)	2.132 0.0544.
REPLICATR2	0.0024652 ±(0.0036)	0.676 0.51205	0.34417 ±(0.5098)	0.675 0.5125
REPLICATR3	0.003326 ±(0.0036)	0.912 0.37991	0.46833 ±(0.5098)	0.919 0.3764
VARIETYV2:TREATT2	-0.0051033 ±(0.0042)	-1.211 0.24906	-0.71 ±(0.5887)	-1.206 0.2511
VARIETYV3:TREATT2	-0.0046534 ±(0.0042)	-1.105 0.29099	-0.64667 ±(0.5887)	-1.098 0.2936
VARIETYV2:TREATT3	0.0020241 ±(0.0042)	0.48 0.63953	0.28667 ±(0.5887)	0.487 0.6351
VARIETYV3:TREATT3	0.0004093 ±(0.0042)	0.097 0.9242	0.05667 ±(0.5887)	0.096 0.9249
VARIETYV2:TREATT4	-0.0030368 ±(0.0042)	-0.721 0.48481	-0.42 ±(0.5887)	-0.713 0.4892
VARIETYV3:TREATT4	-0.0017025 ±(0.0042)	-0.404 0.69322	-0.23667 ±(0.5887)	-0.402 0.6948

*** highly significant, ** moderately significant

Similarly, we can obtain for the rest of the variables for which the Gamma distribution comes out to be the best fit.

Table 3.5: Comparison of fitted Inverse Gaussian GLM and Normal GLM for plant height of tulip

PHT				
	Inverse Gaussian		Normal	
Coefficients	**Estimate (Std.Error)**	**t value Pr(>|t|)**	**Estimate (Std.Error)**	**t value Pr(>|t|)**
(Intercept)	43.396002 ±(1.06671)	37.845 7.45e-14***	43.6 ±(1.062)	41.048 2.83e-14***
VARIETYV2	-8.800894 ±(1.00394)	-6.963 1.51e-05***	-8.75 ±(1.301)	-6.727 2.11e-05***
VARIETYV3	0.512883 ±(1.077021)	0.347 0.1344	0.4583 ±(1.301)	0.352 0.7307
TREATT2	1.577131 ±(1.003133)	1.071 0.0054**	1.278 ±(1.371)	0.932 0.3698
TREATT3	1.335991 ±(1.016243)	0.911 0.3801	0.7778 ±(1.371)	0.567 0.581
TREATT4	1.077929 ±(1.232142)	0.737 0.0352*	1.056 ±(1.371)	0.77 0.4563
REPLICATR2	0.799654 ±(1.014294)	0.582 0.5714	0.7083 ±(1.301)	0.545 0.5961
REPLICATR3	-0.004129 ±(1.025796)	-0.003 0.9976	-0.5 ±(1.301)	-0.384 0.7074
VARIETYV2:TREATT2	-0.466416 ±(1.067413)	-0.318 0.7561	-0.5 ±(1.502)	-0.333 0.745
VARIETYV3:TREATT2	-0.886513 ±(1.020643)	-0.515 0.6158	-0.8333 ±(1.502)	-0.555 0.5892
VARIETYV2:TREATT3	1.038911 ±(1.054181)	0.714 0.0586.	1.167 ±(1.502)	0.777 0.4524
VARIETYV3:TREATT3	-0.150103 ±(1.001288)	-0.088 0.1311	-1.345e-14 ±(1.502)	0.000 1.000
VARIETYV2:TREATT4	-0.636558 ±(1.053066)	-0.438 0.4691	-0.6667 ±(1.502)	-0.444 0.6651
VARIETYV3:TREATT4	0.437391 ±(1.027096)	0.253 0.8044	0.5 ±(1.502)	0.333 0.745

*** highly significant, ** moderately significant, * significant

Table 3.5 gives the fitted generalized linear model based on the Inverse Gaussian and Normal distributions for the PHT variable. We also obtain the dispersion parameter for Normal GLM taken to be 1.69213 and for Inverse Gaussian GLM taken to be 2.526654e-05. Since SLT variable also had the minimum null deviance and minimum residual deviance for Inverse Gaussian distribution so we have fitted the generalized linear model based on the Inverse Gaussian distribution for the SLT variable as well.

Table 3.6: Comparison of fitted Inverse Gaussian GLM and Normal GLM for scalp length of tulip

SLT				
	Inverse Gaussian		Normal	
Coefficients	**Estimate (Std.Error)**	**t value Pr(>|t|)**	**Estimate (Std.Error)**	**t value Pr(>|t|)**
(Intercept)	32.0758 ±(0.0046)	47.549 4.90e-15***	32.2361 ±(0.614)	52.48 1.51e-15***
VARIETYV2	-6.2064 ±(0.0362)	-8.431 2.19e-06***	-6.25 ±(0.752)	-8.308 2.55e-06***
VARIETYV3	1.5276 ±(0.0799)	1.736 0.10813	1.5417 ±(0.752)	2.049 0.06296.
TREATT2	0.2548 ±(0.0531)	0.299 0.02031*	0.1667 ±(0.793)	0.21 0.83706
TREATT3	-0.3596 ±(0.0554)	-0.42 0.0816	-0.4444 ±(0.793)	-0.56 0.58548
TREATT4	0.3192 ±(0.0505)	0.375 0.01402*	-0.1667 ±(0.793)	-0.21 0.83706
REPLICATR2	0.563 ±(0.0804)	0.7 0.49713	0.25 ±(0.7523)	0.332 0.74539
REPLICATR3	-0.2425 ±(0.0986)	-0.304 0.76663	-0.4583 ±(0.7523)	-0.609 0.55373
VARIETYV2:TREATT2	-1.49 ±(0.0412)	-1.771 0.10188	-1.5 ±(0.8687)	-1.727 0.10984
VARIETYV3:TREATT2	-0.5051 ±(1.0104)	-0.5 0.62615	-0.5 ±(0.8687)	-0.576 0.57554
VARIETYV2:TREATT3	-0.339 ±(0.0574)	-0.395 0.69952	-0.3333 ±(0.8687)	-0.384 0.7079
VARIETYV3:TREATT3	-0.3084 ±(1.0202)	-0.302 0.0676	-0.3333 ±(1.8687)	-0.384 0.7079
VARIETYV2:TREATT4	-2.7189 ±(0.0362)	-3.251 0.00694**	-2.6667 ±(0.8687)	-3.07 0.00972**
VARIETYV3:TREATT4	-0.3057 ±(1.017)	-0.301 0.06885	-0.3333 ±(1.8687)	-0.384 0.7079

*** highly significant, ** moderately significant, * significant

Table 3.6 gives the comparison of the generalized linear model based on the Inverse Gaussian distribution and Normal distribution. And the dispersion parameter for Normal GLM taken to be 0.5659722 and for Inverse Gaussian GLM taken to be 2.187822e-05. Similarly, we can obtain it for rest of the variables for which the Inverse Gaussian distribution comes out to be the best fit.

We see that the maximum values for Gamma GLM and Inverse Gaussian GLM are statistically significant. On the basis of being statistically significant we may infer that the models are the best fits. But these models are not nested and have different distributions for the response, which makes direct comparison problematic.So we will take a look at their dispersion parameters for all the fitted models we see that every time the dispersion parameter for normal GLM has higher value then the gamma GLM and Inverse Gaussian GLM which will result in higher kurtosis of normal GLM in comparison to the Gamma GLM and Inverse Gaussian GLM i.e, the models with non-normal response should be fitted by the Generalized linear models. Further, when our data is non-normal correcting the data may result in the loss of information. Thus generalized linear model approach is the most suited approach for such datasets.

Chapter- 4

EXTENSION OF LINEAR MIXED MODELS ALLOWING NON-NORMAL RESPONSE

4.1 Generalized Linear Mixed Models

Generalized linear models that contain random effects in addition to the fixed effects are known as generalized linear mixed models (GLMMs). The generalized linear mixed model (GLMM) relates the conditional mean for the j^{th} cluster to the fixed and random effects as follows:

$$E(y_{n_j} \mid \delta_j) = g^{-1}(X_j\beta + Z_j\delta_j) \tag{4.1.1}$$

Where y_{n_j} is the vector of responses at the j^{th} cluster, g is a differentiable monotonic link function, η_j is the linear predictor given by $\eta_j = X_j\beta + Z_j\delta_j, X_j$ is the $(n_j \times p)$ matrix of fixed effects model terms associated with the j^{th} cluster, β is the corresponding $(p \times 1)$ vector of fixed effects regression coefficients, δ_j is the $(q \times 1)$ vector of random factor levels associated with the j^{th} cluster. There are n_j observations in the j^{th} cluster for a total of $n = \sum_{j=1}^{m} n_j$ observations. The conditional response, $y \mid \delta$, is assumed to have an exponential family member distribution. Each of the q random effects are assumed normally distributed with mean zero. The variance-covariance matrix of the vector of random effects in the j^{th} cluster is denoted D_j. The D_j is typically taken to be the same for each cluster. The conditional mean given by

$$E(y \mid \delta) = g^{-1}(\eta) = g^{-1}(X\beta + Z\delta) \tag{4.1.2}$$

The variance-covariance matrix of the vector of conditional responses is $\text{var}(y \mid \delta) = S$, where

$$S = A^{1/2}(\eta)RA^{1/2}(\eta) \tag{4.1.3}$$

With $A(\eta)$ being the diagonal matrix that contains the variance functions associated with the assumed probability distribution of the response. The variance functions are evaluated at the linear predictor, η. R is the user specified correlation matrix, which is common to all clusters.

The major difference between generalized linear mixed models (GLMMs)and generalized linear models (GLMs) is the presence of the random effects in the GLMMs. When we use GLMMs, the data consists of clusters (e.g., patients, whole plots) and the levels of the random effects correspond to the clusters. We can use GLMMs to model both the conditional and marginal means. GLMMs have the added advantage of explicitly modeling variance components. In GLMs the data are not structured into clusters and the observations are assumed independent of one another. In GLMMs, we assume that the conditional response (i.e., the response conditioned on a fixed setting of the random effects) follows a probability distribution that is part of the exponential family. For GLMs, however, weassume that the unconditional response,y, follows an exponential familymember distribution.

4.2 Estimation and Computational Method

When the random effects and the data are normal, and we have an identity link, we have the standard linear mixed model

$$y = X\beta + Z\delta + \varepsilon \tag{4.2.1}$$

The marginal distribution of y is easily obtained by taking the expectation and variance operators through the model, producing $E(y) = X\beta$ and $\text{var}(y) = ZDZ' + S$. We know that $y \sim N(X\beta, ZDZ' + S)$ since a linear combination of normal variables is also normal. For GLMMs, however, we assume a non-normalconditional response, $y \mid \delta$, and normal random effects. In these cases, obtaining the marginal distribution of y is a more challenging task.

Maximum likelihood Estimation (MLE) is widely used to obtain parameter estimates in GLMM. Using this method, maximum likelihood estimators are obtained by maximizing the likelihood function. Maximum likelihood estimation requires exact information about data distribution. In other words, the maximum likelihood estimate must be based on a full likelihood function. However, this method of estimation often involves high-dimensional integrals in which analytical solutions for these integrals are often hard to be obtained, particularly if the response variable is not normally distributed. To solve this problem, numerical approximation methods are inevitable.In general, the approximation method for parameter estimation in GLMM consists of: (a) simplification of problem analytically (using Laplace approximation to integrate the likelihood function), such as Penalized Quasi Likelihood (Breslow and Clayton ,1993) and Hierarchical Generalized Linear models or HGLM (Lee and Nelder 1996)(b) computer intensive methods, such as Monte Carlo EM algorithm (Booth*et al*. 1999), Markov Chain Monte Carlo or MCMC (Zeger*et al*. 1991)and Gauss-Hermite quadrature or GHQ(Pan and Thompson 2003). The Laplace approximation approximates the integrand, PQL approximates the data and AGQ approximates theintegral.The penalized quasi-likelihood method (PQL) of Breslow and Clayton (Breslow and Clayton 1993) estimates generalized linear mixed models using the Laplace approximation via ignoring one term of the approximation. Though arrived at through different expansions, the PQL and the pseudo-likelihood methods produce identical parameter estimates since the objective functions minimized by the two methods differ only by a constant (Capanu*et al.,*2013).

4.2.1 Penalized quasi-likelihood Parameter Estimation

Maximum likelihood estimation requires exact information about data distribution. In other words, the maximum likelihood estimate must be based on a full likelihood function. If we only have knowledge about means or the relationship between means and variances then we employ quasi likelihood

methods in the estimation. Since the requirement (input) for this method is weaker than the full likelihood method then the quasi likelihood method is robust toward misspecification models. In GLMM, the estimation of variance components is an interest. If the exact distribution of data is unknown, then estimation of variance component is carried out using quasi-likelihood methods in which a penalty term on random effect is imposed. This is known as the Penalized Quasi-Likelihood (PQL) method. The purpose of adding penalty is to avoid some arbitrary values of random effect and to force the random effect approximate to zero (McCullogh and Searle 2008). The Laplace approximation to the data is usually implemented in PQL. In GLMM, the mean for $y_i, i = 1,2,...,m$ conditioned on random effects $b = (b_1, b_2, ..., b_m)'$ is $E(y_i \mid b) = h(x_i'\beta + z_i'b)$ Approximation for vector response data y_i is given by

$$y_i = \mu_i + \varepsilon_i = E(y_i \mid b) + \varepsilon_i = h(x_i'\beta + z_i'b) + \varepsilon_i \tag{4.2.2}$$

Where: β is a fixed vector parameter, x_i and z_i are known vector, $h(\cdot)$ is an inverse of link function $g(\cdot)$, and $\eta_i = g(\mu_i) = x_i'\beta + z_i'b$.

A Taylor expansion for the data y_i (4.2.2) about current $\hat{\beta}$ and $\hat{b}$ is

$$\begin{aligned} y_i &\approx h(x_i'\hat{\beta} + z_i'\hat{b}) + h'(x_i'\hat{\beta} + z_i'\hat{b})x_i'(\beta - \hat{\beta}) + h'(x_i'\hat{\beta} + z_i'\hat{b})z_i'(\beta - \hat{\beta}) + \varepsilon_i \\ &= \mu_i' + V(\hat{\mu}_i x_i'(\beta - \hat{\beta}) + V(\hat{\mu}_i)z_i'(\beta - \hat{\beta}) + \varepsilon_i \end{aligned} \tag{4.2.3}$$

Equation (4.2.3) can also be written as

$$\begin{aligned} y_i^* &\equiv \hat{V}_i^{-1}(y_i - \hat{\mu}_i) + x_i\hat{\beta} + z_i\hat{b}_i \\ &\approx x_i\beta + z_i b_i + \varepsilon_i^* \end{aligned} \tag{4.2.4}$$

Where y_i^* is known as a pseudo-response. The model fitting is done iteratively between updating pseudo-response and fitting linear mixed model(LMM). In other words, model fitting GLMM by using PQL methods is in fact fitting LMM to pseudo-response y_i^*.

The iterative algorithm for model fitting is as follows: (i) Set values for β and variance component θ. Calculate empirical Bayes estimator for b_i and pseudo-response Y_i^*, (ii) Based on pseudo-response Y_i^* fit the model and update β and θ accordingly. Repeat these two steps until convergence is achieved.

Conditional variance for y_i given the random effects b is

$$\text{var}(y_i \mid b) = a_i'(\phi)v(\mu_i) \tag{4.2.5}$$

Where: ϕ is a dispersion parameter, $a_i(\cdot)$ is a known function which is often specified equal to $\phi / w_i; w_i$ is a weight which its value is specified and $v(\cdot)$ is a known variance function.

The parameter estimation using PQL method also can be explained through derivation of the quasi-likelihood function. For a response variable $y = (y_1, y_2, ..., y_n)'$ in which $y_1, y_2, ...y_n$ are independent and $b \sim N(0, G)$, then the quasi-likelihood function is defined by

$$L_Q \propto |G|^{-1/2} \int \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} d_i - \frac{1}{2} b' G^{-1} b \right\} db \tag{4.2.6}$$

Where $d_i = -2\int_{y_i}^{\mu_i} \frac{y_i - u}{a_i(\phi)v(u)} du; d_i$ is also called (quasi) deviance.

By using Laplace approximation, logarithm of L_Q, which is denoted by l_Q, is given by:

$$l_Q \approx c - \frac{1}{2}\log|G| - \frac{1}{2}\log|q''|(\tilde{b}) - q(\tilde{b}) \tag{4.2.7}$$

Where c does not depend on parameter, $q(b) = \frac{1}{2}(\sum_{i=1}^{n} d_i + b' G^{-1} b)$ and $\tilde{b}$ is a value which minimize $q(b)$ Specifically, b is a solution for $q(b) = 0$. It means :

$$G^{-1}b-\sum_{i=1}^{n}\frac{y_i-\mu_i}{a_i(\phi)v(\mu_i)g^{'}(\mu_i)}z_i=0 \tag{4.2.8}$$

Where $\mu_i = x_i^{'}\beta + z_i^{'}b$. The second derivative for $q(b)$ is

$$q^{"}(b)=G^{-1}\sum_{i=1}^{n}\frac{z_i z_i^{'}}{a_i(\phi)v(\mu_i)\{g^{'}(\mu_i)\}^2}+r \tag{4.2.9}$$

Where r is a residual with its expected value is zero. If the denominator at formula () is denoted by w_i^{-1} and the residual r is ignored, then the approximation for $q^{"}(b)$ is given by

$$q^{"}(b)\approx Z^{'}WZ+G^{-1} \tag{4.2.10}$$

Where Z is a matrix which has element $z_i^{'}$ at i^{th} row and $W=diag(w_1,...,w_n)$, where w_i is an iteration weight GLM. By combining (4.2.9) and (4.2.10), the approximation for logarithm of quasi-likelihood function is given by:

$$l_Q = c-\frac{1}{2}(\log|I+Z^{'}WZG|+\sum_{i=1}^{n}\tilde{d}_i+\tilde{b}^{'}G^{-1}\tilde{b}) \tag{4.2.11}$$

Where $\tilde{d}_i$ is d_i with α is replaced with $\tilde{b}$.

A further approximation for logarithm of quasi-likelihood function could be obtained by assuming the GLM weights are not constant but they are functions of mean. The first term of the bracket (4.2.11) depends on β only through W, so that the term could be ignored. Therefore, the approximation for logarithm of quasi-likelihood function is given by

$$l_{PQ}\approx c-\frac{1}{2}\left(\sum_{i=1}^{n}\tilde{d}_i+\tilde{b}^{'}G^{-1}\tilde{b}\right) \tag{4.2.12}$$

Formula (4.2.12) is a penalized quasi-likelihood (PQL) where $\tilde{b}$ is a value which minimized $q(b)$. It means, if β is given, then $\tilde{b}$ will maximize l_{PQ}.

Because $\tilde{b}$ depends on β, then it could be written as $\tilde{b} = \tilde{b}(\beta)$. For a fixed θ, consider $\hat{\beta}$ as a value that maximize l_{PQ}. As a result, $\hat{\beta}$ and $\tilde{b}$ maximize

$$l_{PQ}(\beta, b) = -\frac{1}{2}\left(\sum_{i-1}^{n} d_i + b' G^{-1} b\right) \tag{4.2.13}$$

Where $l_{PQ}(\beta, b) = \frac{1}{2}\left(\sum_{i=1}^{n} d_i + b' G^{-1} b\right)$ is a function of β and b with $\tilde{b} = \tilde{b}(\hat{\beta})$.

The standard method for maximizing (4.2.13) requires solving a nonlinear equation system, $\partial l_{PQ} / \partial \beta = 0$ and $\partial l_{PQ} / \partial b = 0$, or it is equivalent to the solution for formula (4.2.14) and (4.2.15) as follows:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_i}{a_i(\phi) v(\mu_i) g'(\mu_i)} G^{-1} b = 0 \tag{4.2.14}$$

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_i}{a_i(\phi) v(\mu_i) g'(\mu_i)} - G^{-1} b = 0 \tag{4.2.15}$$

In practice, random effects involved in GLMM are often large enough, so that solutions of (4.2.14) and (4.2.15) will be of high-dimensional space. In other words, it is necessary to determine solutions simultaneously from a large number of nonlinear systems. Standard procedure to solve this problem, such as Newton-Raphson, is inefficient and its convergence is very slow when the dimension of the solution is high. Jiang developed a nonlinear Gauss-Seidel algorithm to solve (4.2.14) and (4.2.15). Jiang showed that the algorithm convergence globally in virtually all typical situations of GLMM. On the other hand, Breslow and Clayton proposed an iterative procedure to solve (4.2.14) and (4.2.15) by modifying Fisher scoring algorithm which was developed by Green.

An interesting feature of Breslow and Clayton method is that it exploits a close correspondence with the mixed model equation developed by Henderson *et al.*(1959). First,defineaworkingvector $\tilde{y} = (\tilde{y}_i); i = 1,2,...,n$,where

$\tilde{y}_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$ and where η_i and μ_i are evaluated at the current estimators of β and b. Then, the solution for (4.2.14) and (4.2.15) based on Fisher scoring can be found as the iterative solution to the equation system as follows,

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & G^{-1} + Z'WZ \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'W \\ Z'W \end{pmatrix} \tilde{y} \qquad (4.2.16)$$

Because W depends on β and b, we must update W for each iteration step. Equivalently, solution (4.2.16) could also be given as follows:

$$\begin{aligned} \beta &= X'V^{-1}X)^{-1}X'V^{-1}\tilde{y} \\ b &= GZ'V^{-1}(\tilde{y} - X\beta) \end{aligned} \qquad (4.2.17)$$

Where: $V = W^{-1} + ZGZ'$, assuming that W^{-1} exist. For a large data sets, inverse of V may be not simple computationally, unless V has a certain special structure, such as block diagonal.

According to Jiang, PQL estimators are inconsistent. In other words, the bias due to approximation will not be zero. If Laplace approximation is used, the reduction bias will be larder as the order of approximation getting large. The PQL method has computation advantages since it is usually easy to carry out except in the case higher order approximation. Moreover the PQL results usually work well when the variance components are small. This is because the Laplace approximation becomes accurate when the variance components are close to zero.

4.2.2 Laplace approximation

In GLMM, the conditional distribution of response variable Y for j^{th} unit within i^{th} cluster with the random effect b_i is assumed independent and follows exponential family distributions:

$$f(y_i \mid b_i) = \exp\left\{\phi^{-1}\left[y_{ij}\eta_{ij} - \Psi(\eta_{ij})\right] + c(y_{ij}, \phi)\right\} \qquad (4.2.18)$$

Where: $\mu_{ij} = E(Y_{ij} \mid b_i) = h(\eta_{ij})$ is a conditional mean and $\eta_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}b_i$ is a linear predictor. The probability density function of random effects b_i is denoted by $f(b_i)$ and it is assumed that $(b_i) = 0$ and $\text{var}(b_i) = G; b_i \sim iidN(0, G)$. Therefore, in GLMM it is necessary to specify $f(y_{ij} \mid b_i)$ and $f(b_i)$. Based on the specification of $f(y_{ij} \mid b_i)$ and $f(b_i)$, the marginal density of $Y, f(y_{ij})$ is given by:

$$f(y_{ij}) = \int \prod_{j=1}^{n_i} f(y_{ij} \mid b_i) f(b_i) db_i \tag{4.2.19}$$

Where n_i is a sample size for i^{th} cluster. In Linear Mixed Model (LMM), an analytical or closed form solution for integral (4.2.19) could be obtained since the marginal density $f(y_{ij})$ is a normal density. In general, it is not easy to compute $f(y_{ij})$ because $f(y_{ij} \mid b_i) f(b_i)$ can be a complex function and the integrals are of high dimension. Therefore approximation methods are required to solve the integral.

In GLMM, the construction of likelihood function is based on the marginal density $f(y_{ij})$. Suppose there are m clusters and the number of units for each cluster is $n_i; i = 1,2,...,m$. then, the likelihood function for m clusters are obtained as follows :

$$L(\theta) = \prod_{i=1}^{m} f(y_{ij})$$

$$= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij} \mid b_i) f(b_i) db_i \tag{4.2.20}$$

And the Laplace approximation can be used to approximate the integral by re-written the integral term on (4.2.20) as :

$$\int \prod_{j=1}^{n_i} f(y_{ij} \mid b_i) f(b_i) db_i$$

$$= \int f(y \mid b) f(b) db$$
$$= \int e^{\log f(y|b) f(b)} db = \int e^{g(b)} db \tag{4.2.21}$$

Where $g(b) = \log f(y \mid b) f(b)$

We aim to choose $\hat{b}$ so that $g(b)$ is maximized by fulfilling the necessary and sufficient conditions $g'(b) = 0$ and $g''(\hat{b}) < 0$. The second order Taylor expansion around $\hat{b}$ for $g(b)$ is given by:

$$g(b) \approx \tilde{g}(b) = g(\hat{b}) + (b - \hat{b}) g'(\hat{b}) + \frac{1}{2}(b - \hat{b})^2 g''(\hat{b})$$

$$= g(\hat{b}) - \frac{1}{2}(b - \hat{b})^2 \left(- g''(\hat{b}) \right) \tag{4.2.22}$$

It can be seen that $e^{\tilde{g}(b)}$ is proportional to the normal density (μ_L, σ_L^2);where $\mu_L = \hat{b}$ and $\sigma_L^2 = -\frac{1}{g''(\hat{b})}$.

Then, the Laplace approximation for the likelihood $L(\theta)$ is given by:

$$L(\theta) = \int e^{g(b)} db \approx \int e^{\tilde{g}(b)} db$$

$$= \exp(g(\hat{b}) \int \exp\left(- \frac{1}{2\sigma_L^2} (b - \mu_L)^2 \right) db = \exp\left(g(\hat{b}) \right) \sqrt{2\pi\sigma_L^2}$$

(4.2.23)

It can also be stated that

$$(b \mid y) = \frac{f(y \mid b) f(b)}{f(y)} \alpha \exp\left(g(b) \right) \approx const \exp\left(- \frac{1}{2\sigma_L^2} (b - \mu_L)^2 \right).$$

Here, $B \mid Y = y \approx N(\mu_L, \sigma_L^2)$. This approximation performs better for larger cluster n_i. Moreover, the accuracy of this approximation increases if higher order of

Taylor expansion is used. However, this approximation is less accurate if the variance of random effects is large.

4.2.3 LASSO Method

Let y_{it} denote observation t in cluster $i, i = 1,2,...,n; t = 1,2,...,T_i$, collected in $y_i^T = (y_{i1},...,y_{iT_i})$. Let $x_{it}^T = (1, x_{it1},..., x_{itp})$ be the covariate vector associated with fixed effects and $z_{it}^T = (z_{it1},..., z_{itq})$ be the covariate vector associated with random effects. It is assumed that the observations y_{it} are conditionally independent with mean $\mu_{it} = E(y_{it} \mid b_i, x_{it}, z_{it})$ and variance $\text{var}(y_{it} \mid b_i) = \phi v(\mu_{it})$, where $v(\cdot)$ is a known variance function and ϕ is a scale parameter. The GLMM considered here is of the form

$$g(\mu_{it}) = x_{it}^T \beta + z_{it}^T b_i = \eta_{it}^{par} + \eta_{it}^{rand} \tag{4.2.24}$$

Where, g is a monotonic and continuously differentiable link function, $\eta_{it}^{par} = x_{it}^T \beta$ is a linear parametric term with parameter vector $\beta^T = (\beta_0, \beta_1,..., \beta_p)$ including intercept and $\eta_{it}^{rand} = z_{it}^T b_i$ contains the cluster-specific random effects $b_i \sim N(0, Q)$, with $q \times q$ covariance matrix Q. An alternative form that we also use is

$$\mu_{it} = h(\eta_{it}); \eta_{it} \beta_0 + \eta_{it}^{par} + \eta_{it}^{rand}$$

Where $h = g^{-1}$ is the inverse link function.

Model (4.2.24) can be expressed in the matrix notation by collecting observations within one cluster, and the model takes the form

$$g(\mu_i) = X_i \beta + Z_i b_i,$$

Where, $X_i^T = (x_{i1},..., x_{iTi})$ denotes the design matrix of the i^{th} cluster and $Z_i^T = (z_{i1},..., z_{iTi})$. For all observations one obtains

$$g(\mu) = X\beta + Zb,$$

with $X^T = [X_1^T,...,X_n^T]$ and block diagonal matrix $Z = blockdiag(Z_1,...,Z_n)$. For the random effects vector $b^T = (b_1^T,..,b_n^T)$ one has a normal distribution with block-diagonal covariance matrix $Q_b = diag(Q,...,Q)$.

For GLMM it is assumed that the conditional density of y_{it}, given explanatory variables and the random effect b_i, is of exponential family type

$$f(y_{it} \mid x_{it}, b_i) = \exp\left\{\frac{(y_{it}\theta_{it} - k(\theta_{it}))}{\phi} + c(y_{it}, \phi)\right\}$$

Where $\theta_{it} = \theta(\mu_{it})$ denotes the natural parameter, $k(\theta_{it})$ is a specific function corresponding to the type of exponential family, $c(\cdot)$ the log normalization constant and ϕ the dispersion parameter (Fahrmeir and Tutz 2013).

One of the popular methods discussed above to maximize GLMMs is penalized quasi-likelihood (PQL), which has been suggested by Breslow and Clayton(1993), Lin and Breslow (1996) and Breslow and Lin (1995). Typically the covariance matrix $Q(\ell)$ of the random effects b_i depends on an unknown parameter vector ℓ. In penalization-based concepts the joint likelihood- function is specified by the parameter vector of the covariance structure ℓ together with the dispersion parameter ϕ, which are collected in $\gamma^T = (\phi, \ell^T)$, and parameter vector $\delta^T = (\beta^T, b^T)$. The corresponding log likelihood is

$$l(\delta,\gamma) = \sum_{i=1}^{n} \log\left(\int f(y_i \mid \delta,\gamma)\, p(b_i,\gamma)\, db_i\right) \qquad (4.2.25)$$

Where, $p(b_i,\gamma)$ denotes the density of the random effects. Breslow and Clayton (1993) derived the approximation

$$l^{app}(\delta,\gamma)=\sum_{i=1}^{n}\log(f(y_i\mid\delta,\gamma))-\tfrac{1}{2}b^TQ(\ell)^{-1}b \tag{4.2.26}$$

Where the penalty term $b^TQ(\ell)^{-1}b$ is due to the approximation based on the Laplace method.

PQL usually works within the profile likelihood concept. It is distinguished between the estimation of δ, given the plugged-in estimate $\hat{\gamma}$, resulting in the profile-likelihood $l^{app(\delta,\hat{\gamma})}$,and the estimation of γ.

The log-likelihood (4.2.25) is expanded to include the penalty term $\lambda\sum_{i=1}^{p}|\beta_i|$. Approximation along the lines of Breslow and Clayton (1993) yields the penalized log-likelihood

$$l^{pen}(\beta,b,\gamma)=l^{pen}(\delta,\gamma)=l^{app}(\delta,\gamma)-\lambda\sum_{i=1}^{p}|\beta_i| \tag{4.2.27}$$

For given $\hat{\gamma}$ the optimization problem reduces to

$$\hat{\delta}=\arg\max_{\delta}l^{pen}(\delta,\hat{\gamma})=\arg\max_{\delta}\left[l^{app}(\delta,\hat{\gamma})-\lambda\sum_{i=1}^{p}|\beta_i|\right] \tag{4.2.28}$$

We will use a full gradient algorithm that is based on the algorithm of Geoman given by Goeman (2010). Here the penalty term from the optimization problem of equation (4.2.25) is replaced by $\sum_{i=1}^{p}\lambda|\beta_i|$, where $\lambda_i=0$ is chosen for unpenalized parameters.The penalty used in (4.2.27) and (4.2.28) can be seen as a partially penalized approach if the whole parameter vector $\delta^T=(\beta^T,b^T)$ is considered.

So, for the LASSO approach an algorithm is obtained for maximizing the penalized log-likelihood $l^{pen}(\delta,\gamma)$ from equation (4.2.28). In contrast to the approaches of Shevade and Keerthi(Shevade and Keerthi 2003), Kim and Kim (Kim and Kim 2004) and Genkin *et al.*(Genkin, Lewis and Madigan 2007), where

only a single component is updated at a time, it follows the gradient of the likelihood from a given starting value of δ and uses the full gradient at each step. Similar to Goeman (2010) the algorithm can automatically switch to a Fisher scoring procedure when it gets close to the optimum and therefore avoids the tendency to slow convergence which is typical for gradient ascent algorithms. An additional step is needed to estimate the variance-covariance components Q of the random effects. The algorithm is carried out in the following steps as:

1. Initialization

   Compute starting values $\hat{\beta}^{(0)}, \hat{b}^{(0)}, \hat{\gamma}^{(0)}$ and set $\hat{\eta}^{(0)} = X\hat{\beta}^{(0)} + Z\hat{b}^{(0)}$.

2. Iteration

   For $l = 1,2,...$ until convergence:

   a) Calculation of the log-likelihood gradient for given $\hat{\gamma}^{(l-1)}$

      with $s(\delta) = \partial l^{app}(\delta) / \partial \delta$ derive:

      $$s_0^{pen}(\hat{\delta}^{(l-1)}) = s_0(\hat{\delta}^{(l-1)}); s_i^{pen}(\hat{\delta}^{(l-1)}) = s_i(\hat{\delta}^{(l-1)}); i = p+1,...,p+ns$$

      Furthermore, for $i = 1,...,p$ derive:

$$s_i^{pen}(\hat{\delta}^{(l-1)}) = \begin{cases} s_i(\hat{\delta}^{(l-1)}) - \lambda sign(\hat{\beta}_i^{(l-1)}); & if \hat{\beta}_i^{(l-1)} \neq 0 \\ s_i(\hat{\delta}^{(l-1)}) - \lambda sign(s_i(\hat{\delta}^{(l-1)})); & if \hat{\beta}_i^{(l-1)} = 0 and \mid s_i(\hat{\delta}^{(l-1)}) \mid > \lambda \\ 0; & otherwise \end{cases}$$

      Where

$$sign(x) = \begin{cases} 1; if x > 0 \\ 0; if x = 0 \\ -1; if x < 0 \end{cases}$$

   b) Calculation of the directional second derivative

Let $A := [X, Z]$ and $k = daig(0,...,0,Q^{-1},...,Q^{-1})$ be a block-diagonal penalty matrix with a diagonal of $p+1$ zeros corresponding to the fixed effects and then n times the matrix Q^{-1}.

Then the Fisher matrix is given in closed form as $F^{pen}(\delta) = A^T W(\delta) A + K$, with $W(\delta) = D(\delta)\Sigma^{-1}(\delta)D(\delta)^T$ and $D(\delta) = \partial h(\eta)/\partial \eta, \Sigma(\delta) = \mathrm{cov}(y \mid \delta)$. The directional second derivative is given for every δ and every direction vector $v \in \Re^{p+1+ns}$ by:

$$l''_{pen}(\delta; v) = v^T F^{pen}(\delta) v$$

c) Optimum of Taylor approximation

Based on the Taylor approximation used in Goeman (2010), we derive

$$t_{edge}^{(l-1)} = \min_i \left\{ -\frac{\hat{\delta}_i^{(l-1)}}{s_i^{pen}(\hat{\delta}^{(l-1)})} : sign(\hat{\delta}_i^{(l-1)}) = -sign[s_i^{pen}(\hat{\delta}^{(l-1)})] \neq 0 \right\}$$

And

$$t_{opt}^{(l-1)} = \frac{\left\| s^{pen}(\hat{\delta}^{(l-1)}) \right\|_2}{l''_{app}(\hat{\delta}^{(l-1)}, s^{pen}(\hat{\delta}^{(l-1)}))}$$

With $\|\cdot\|_2$ denoting the L_2 norm.

d). update

$$\hat{\delta}^{(l)} = \begin{cases} \hat{\delta}^{(l-1)} + t_{edge}^{(l-1)} s^{pen}(\hat{\delta}^{(l-1)}) & if t_{opt}^{(l-1)} \geq t_{edge}^{(l-1)} \\ \hat{\delta}_{NR}^{(l-1)} & if t_{opt}^{(l-1)} < t_{edge}^{(l-1)} and sign(\hat{\delta}_{NR}^{(l)}) = sign(\hat{\delta}^{(l-1)}) \\ \hat{\delta}^{(l-1)} + t_{opt}^{(l-1)} s^{pen}(\hat{\delta}^{(l-1)}) & otherwise \end{cases}$$

Where $\hat{\delta}_{NR}^{(l)}$ denotes the Fisher scoring estimate.

e) Computational of variance-covariance components

Estimates $\hat{Q}^{(l)}$ are obtained as approximate EM-type estimate or by alternativemethods yielding the update $\ell^{(l)}$. If necessary the whole vector $\hat{\gamma}^{(l)}$ is completed by an estimate of the dispersion parameter.

3. Re-estimation

In a final step a model that includes only the variables corresponding to non-zero parameters of $\hat{\beta}$ is fitted. A simple Fisher scoring, resulting in the final estimates $\hat{\delta}, \hat{Q}$ is used.

4.3 Statistical Inference on Regression Coefficients and Variance Components

An appropriate test on the variance component associated with a particular random effect determines if that effect belongs in the model, which in turn allows the analysts to reduce the covariance matrix. One can compare two covariance structures through a REML pseudo-likelihood ratio test (LRT).In SAS 9.2, the REML pseudo-likelihood ratio test for comparing two covariance structures uses the COVTEST statement in SAS PROC GLIMMIX. This approach fits the full model, which is the model with the most complex covariance structure of interest, and then takes the pseudo-data from the last iteration to calculate the associated REML pseudo-likelihood value. Let l^R_{FULL} denote this restricted log-pseudo likelihood value. The procedure then uses the same pseudo-data to fit the reduced model and computes the corresponding pseudo-likelihood. Let l^R_{RED} denote this restricted log-pseudo-likelihood value. The appropriate likelihoodratio test statistic, $\hat{\lambda}$, is

$$\hat{\lambda} = l^R_{FULL} - l^R_{RED} \tag{4.3.1}$$

Once again, the distribution of this test statistic depends on whether any of the parameters fall on the boundary of the variance parameter space as defined by Self and Liang in 1987(Self and Liang 1987) andVerbeke and Molenberghs in 2007(Molenberghs and Verbeke 2007).The tests used to compare covariance structures for GLMMs are conducted with the full fixed effects vector, β.

4.4 Comparison of the Estimation Methods

The above estimation procedures are compared on the basis of average relative bias, average squared relative bias, average absolute bias and average squared deviation. Suppose act_i denotes the actual value of the *ith* unit, and est_i is any estimate of $act_i; i = 1,..m$ then

- Average relative bias

$$ARB = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{est_i - act_i}{act_i}\right|$$

- Average squared relative bias

$$ASRB = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{est_i - act_i}{act_i}\right)^2$$

- Average absolute bias

$$AAB = \frac{1}{m}\sum_{i=1}^{m}\left|est_i - act_i\right|$$

- Average squared deviation

$$ASD = \frac{1}{m}\sum_{i=1}^{m}\left(est_i - act_i\right)^2$$

We have used Generalized linear mixed models (GLMMs) to model correlated and clustered mortality responses of the *Venturia inequilis* organism responsible for apple scab. Various estimation methods of GLMM have been

proposed ranging from numerical integration techniques (for example (Booth and Hobert, 1999)) over "joint maximization methods" (Breslow and Clayton, 1993; (Schall 1991), in which parameters and random effects are estimated simultaneously, to fully Bayesian approaches (Fahrmeir and Lang 2001). Overviews on thecurrent methods are found in McCulloch and Searle (2008). We have used some of these estimation methods in our study.But it was found that due to the heavy computational problems in GLMMs modeling usually is restricted to few predictor variables. When many predictors are available, estimates become unstable. Therefore, it is important to note that procedures to select the relevant variables are important in modelling. Classical approaches to the selection of predictors are based on test statistics with the usual stability problems of forward-backward algorithms, which are due to the inherent discreteness of the method (for example (Breiman 1996). A more timely approach to variable selection is based on boosting methods, which haveoriginally been developed within the machine learning community as a method to improveclassification. A first breakthrough was the AdaBoost algorithm. Breiman considered the AdaBoost algorithm as a gradient descent optimization technique(Breiman 1998) and Friedman extended boosting methods to include regressionproblems(Friedman 2001). B¨uhlmann and Yu showed how to fit smoothing splines by boosting baselearners and introduced the concept of componentwise boosting, which may be exploited toselect predictors(Bühlmann and Yu 2003). For linear mixed models the incorporationof random effects has been considered (Tutz and Reithinger 2007), first attempts to fitunivariate GLMMs were proposed by Tutz and Groll (Tutz and Groll 2010).An alternative approach to variable selection that has received much attention is based onpenalized regression techniques. The LASSO proposed by Tibshirani has become a verypopular approach to regression that uses an L_1-penalty on the regression coefficients. This hasthe effect that all coefficients are shrunken towards zero and some are set exactly to zero(Tibshirani 1996).

4.5 Numerical Illustration

Data generated on the mortality of *Venturia inequilis* organism responsible for apple scab as discussed in chapter-1 Section 1.5.2. The data is fitted by different estimation procedures to evaluate which fits the data best. The fitted generalized linear mixed model by the maximum likelihood estimation method is shown in Table 4.1.

Table 4.1: Generalized Linear mixed model by maximum likelihood estimation method

Random effects:

Groups	Name	Variance	Std.Dev.
Location	(Intercept)	4.546e-18	2.132e-09
Residual		1.363e+01	3.691e+00

Number of obs: 48, groups:location, 3

Fixed effects:

	Estimate	Std.Error	t-value	p value
(Intercept)	4.2917	1.4097	3.044	0.034
concC2	12.1667	1.5070	8.073	<2e-16 ***
concC3	17.3333	1.5070	11.502	<2e-16 ***
concC4	0.6667	1.5070	0.442	0.680
treatT2	1.6667	1.5070	1.106	0.534
treatT3	-0.6667	1.5070	-0.442	0.786
treatT4	1.8333	1.5070	1.217	0.309

*** highly significant

Table 4.1gives the fixed and random effects of the variables when we fit the linear mixed model on the given data where the response is non-normal. It can be seen from the fixed effects table that most of the parameters are insignificant and also the estimates have larger estimated values. This indicates that the estimation method used could poorly estimate the effects. In order to improve the estimate values we have used another method of estimation that is the Laplace Approximation method. The fitted generalized linear mixed model by Laplace Approximation method is shown in table 4.2.

Table 4.2:Generalized Linear mixed model by Laplace Approximation
Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.55130	0.14848	10.448	<2e-16 ***
concC2	1.23353	0.14670	8.409	<2e-16 ***
concC3	1.49664	0.14282	10.479	<2e-16 ***
concC4	0.12516	0.17712	0.707	0.480
treatT2	0.13177	0.11496	1.146	0.252
treatT3	-0.05799	0.12044	-0.481	0.630
treatT4	0.14404	0.11463	1.257	0.209

*** highly significant

Table 4.2 gives the values of random effects, fixed effects and correlation values of the variables when the fit generalized linear model by Laplace Approximation method. The p-value for fixed effects indicates that by the application of this method the concentration parameter has improved. The standard error has also decreased in comparison to the values when we had fitted the generalized linear mixed model by the maximum likelihood estimation method. We further look for better estimation procedure which can given more improved results then the above two methods. For this reason we have fitted the

generalized linearmixed model by another estimation procedure i.e., the penalized quasi likelihood method. Table 4.3 shows the fitted generalized linear mixed model by the penalized quasi likelihood estimation method.

Table 4.3: Generalized linear mixed model by Penalized quasi likelihood

Random effects:

	(Intercept)	Residual
StdDev:	7.431224e-06	1.349411

Fixed effects

	Value	Std.Error	t-value	p-value
(Intercept)	1.5513019	0.0167980	7.155516	0.0000
concC2	1.2335316	0.0141926	5.758984	0.0000
concC3	1.4966424	0.1085291	7.177139	0.0000
concC4	0.1251631	0.0586114	0.483982	0.6311
treatT2	0.1317693	0.1678448	0.785066	0.4372
treatT3	-0.057987	0.1758453	-0.329763	0.7433
treatT4	0.1440394	0.1673661	0.860625	0.3947

Table 4.3gives the values of the generalized linear mixed model when fitted by the penalized quasi likelihood estimation method. From the fixed effects table as can be seen clearly that there is larger difference in the values of the parameters from those fitted by the Laplace Approximation. Now, we have fitted the generalized linear mixed model by the LASSO method of estimation on the agricultural data in order to see whether it makes any difference to the fixed and the random effects fitted by the other methods of estimation as shown above. Table 4.4 shows the fitted generalized linear mixed model by the LASSO method of estimation.

Table 4.4: Generalized linear mixed model by LASSO method

Random Effect

Groups	Name	Std.Dev.
location	(Intercept)	0.046917

Fixed Effects:

	Estimate	StdErr	z.value	p.value
(Intercept)	1.551112	0.056953	27.2349	<2e-16 ***
as.factor(conc)C2	1.233532	0.093567	13.1834	<2e-16 ***
as.factor(conc)C3	1.496642	0.124445	12.0265	<2e-16 ***
as.factor(conc)C4	0.125163	0.198492	0.6306	0.0583.
as.factor(treat)T2	0.131769	0.091989	3.4324	0.0220*
as.factor(treat)T3	-0.05798	0.107680	-0.5385	0.5402
as.factor(treat)T4	0.144039	0.143584	4.0032	0.0158*

*** highly significant; * significant

Table 4.3 gives the random effects, fixed effects and correlation values when the data is fitted by the LASSO method of estimation of generalized linear mixed model. We can clearly see that the parameter values show drastic changes when fitted by this method of estimation. It is clear that LASSO method of estimation is the best estimation procedure for fitting the agricultural data which includes both the fixed and random effects with non-normal response. In supportof the above study we further compare the above methods on the basis of the values of Average Relative Bias, Average Squared Relative Bias, Average Absolute Bias and Average Squared Deviation. The values of Average Relative Bias, Average Squared Relative Bias, Average Absolute Bias and Average Squared Deviation are based on the estimated values that we obtained for the above parameters by the different estimation methods.

Table 4.5: Comparison of estimation methods using four different criteria

Estimation	ARB	ASRB	AAB	ASD
MLE	0.589	0.198	13.67	20.32
LA	0.376	0.092	5.34	12.56
PQL	0.478	0.182	7.45	9.34
LM	**0.132**	**0.081**	**4.90**	**3.90**

MLE=Maximum Likelihood Estimation, LA=Laplace Approximation, PQL=Penalized Quasi Likelihood, LM=Lasso Method, ARB=Average Relative Bias, ASRB=Average Squared Relative Bias, AAB=Average Absolute Bias, ASD= Average Squared Deviation

Table 4.5gives the values of the different criteria for the four estimation methods and on looking keenly on the table the value of Average Relative Bias, Average Squared Relative Bias, Average Absolute Bias and Average Squared Deviation is minimum for LASSO method and this can be inferred that LASSO method of estimation of generalized linear mixed models is the best estimation method for the such type of the data with non-normal response.

Chapter-5

EXTENSION OF BASIC NONLINEAR MIXED EFFECT MODEL TO INCORPORATE THE HETEROSCEDASTICITY AND WITHIN-GROUP CORRELATED ERRORS

Nonlinear mixed-effects models extend linear mixed-effects models by allowing the regression function to depend nonlinearly on fixed and random effects. Because of its greater flexibility, an NLME model is generally more interpretable and parsimonious than a competitor empirical LME model based, say, on a polynomial function. Also, the predictions obtained from an NLME model extend more reliably outside the observed range of the data. The greater flexibility of NLME models does not come without cost, however. Because the random effects are allowed to enter the model nonlinearly, the marginal likelihood function, obtained by integrating the joint density of the response and the random effects with respect to the random effects, does not have a closed-form expression, as in the LME model. As a consequence, an approximate likelihood function needs to be used for the estimation of parameters, leading to more computationally intensive estimation algorithms and to less reliable inference results.

An important practical difference between NLME and LME models is that the former require starting estimates for the fixed-effects coefficients. Determining reasonable starting estimates for the parameters in a nonlinear model is somewhat of an art, although some general recommendations are available (Bates and Watts 1988). There are far more similarities than differences between LME and NLME models. Both models are used with grouped data and serve the same purpose: to describe a response variable as a function of covariates, taking into account the correlation among observations in the same group. Random effects are used to represent within-group dependence in both LME and NLME models, and the assumptions about the random effects and the within-group errors are identical in the two models.

Nonlinear mixed-effects models are mixed-effects models in which some,

or all, of the fixed and random effects occur nonlinearly in the model function. They can be regarded either as an extension of linear mixed-effects models in which the conditional expectation of the response given the random effects is allowed to be a nonlinear function of the coefficients, or as an extension of nonlinear regression models for independent data (Bates and Watts, 1988) in which random effects are incorporated in the coefficients to allow then to vary by group, thus inducing correlation within the groups.

5.1 Single-Level of Grouping

By far the most common application of NLME models is for repeated measures data—in particular, longitudinal data. The nonlinear mixed-effects model for repeated measures proposed by Lindstrom and Bates can be thought of as a hierarchical model(Lindstrom *et al*., 1990). At one level the jth observation on the ith group is modeled as

$$y_{ij} = f(\phi_{ij}, v_{ij}) + \varepsilon_{ij},\ i = 1,...,M,; j = 1,...,n_i \qquad (5.1.1)$$

Where M is the number of groups, n_i is the number of observations on the ith groups, f is a general, real-valued, differentiable function of a group-specific parameter vector ϕ_{ij} and covariate vector v_{ij}, and ε_{ij} is a normally distributed within-group error term. The function f is nonlinear in at least one component of the group-specific parameter vector ϕ_{ij}, which is modeled as

$$\phi_{ij} = A_{ij}\beta + B_{ij}b_i, \qquad b_i \sim N(0, \Psi), \qquad (5.1.2)$$

Where β is a p-dimensional vector of fixed effects and b_i is a q-dimensional random effects vector associated with the ith group (not varying with j with variance-covariance matrix ψ. The matrices A_{ij} and B_{ij} are of appropriate dimensions and depend on the group and possibly on the values of some covariates at the jth observation. This model is a slight generalization of that described in Lindstrom and Bates (1990) in that A_{ij} and B_{ij} can depend on j.

This generalization allows the incorporation of "time-varying" covariates in the fixed effects or the random effects for the model. It is assumed that observations corresponding to different groups are independent and that the within-group errors ε_{ij} are independently distributed as $N(0,\sigma^2)$ and independent of the b_i. The assumption of independence and homoscedasticity for the within –group errors can be relaxed, as in 3.4.

Because f can be any nonlinear function of ϕ_{ij}, the representation of the group-specific coefficients ϕ_{ij} could be chosen so that A_{ij} and B_{ij} are always simple incidence matrices. However, it is desirable to encapsulate as much modeling of the ϕ_{ij} as possible in this second stage, as this simplifies the calculation of the derivatives of the model function with respect to β and b_i, used in the optimization algorithm.

We can write (5.1.1) and (5.1.2) in matrix form as

$$\begin{aligned} y_i &= f_i(\phi_i, v_i) + \varepsilon_i, \\ \phi_i &= A_i\beta + B_i b_i, \end{aligned} \tag{5.1.3}$$

for $i = 1,...,M$ where

$$y_i = \begin{bmatrix} y_{i1} \\ \cdot \\ \cdot \\ \cdot \\ y_{in_i} \end{bmatrix}, \phi_i = \begin{bmatrix} \phi_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{in_i} \end{bmatrix}, \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{in_i} \end{bmatrix}, f_i(\phi_i, v_i) = \begin{bmatrix} f(\phi_{i1}, v_{i1}) \\ \cdot \\ \cdot \\ \cdot \\ f(\phi_{in_i}, v_{in_i}) \end{bmatrix}, v_i = \begin{bmatrix} v_{i1} \\ \cdot \\ \cdot \\ \cdot \\ v_{in_i} \end{bmatrix},$$

$$A_i = \begin{bmatrix} A_{i1} \\ \cdot \\ \cdot \\ \cdot \\ A_{in_i} \end{bmatrix}, B_i = \begin{bmatrix} B_{i1} \\ \cdot \\ \cdot \\ \cdot \\ B_{in_i} \end{bmatrix} \tag{5.1.4}$$

5.2 Multilevel NLME Models

The single-level NLME model (5.1.1) can be extended to data grouped according to multiple, nested factors by modifying the model for the random effects in (5.1.2). For example, the multilevel version of the Lindstrom and Bates (1990) model for two levels of nesting is written as a two-stage model in which the first stage expresses the response y_{ijk} for the kth observation on the jth second-level group of the ith first-level group as

$$y_{ijk} = f(\phi_{ijk}, v_{ijk}) + \varepsilon_{ijk}; i = 1,...,M,; j = 1,...,M_i,; k = 1,...,n_{ij}, \quad (5.2.1)$$

Where M is the number of first-level groups, M_i is the number of second-level groups within the ith first-level group, n_{ij} is the number of observations on the jth second-level group of the i^{th} first-level group, and ε_{ijk} is a normally distributed within-group error term. As in the single-level model, f is a general, real-valued, differentiable function of a group-specific parameter vector ϕ_{ijk} and a covariate vector v_{ijk}. It is nonlinear in at least one component of ϕ_{ij}. The second stage of the model expresses ϕ_{ij} as

$$\begin{aligned} &\phi_{ijk} = A_{ijk}\beta + B_{ijk}b_i + B_{ijk}b_{ij} \\ &b_i \sim N(0,\psi_1), bij \sim N(0,\psi_2) \end{aligned} \quad (5.2.2)$$

As in the single-level model (5.1.2), β is a p-dimensional vector of fixed effects, with design matrix A_{ijk}, which may incorporate time varying covariates. The first-level random effects b_i are independently distributed q_1 dimensional vectors with variance-covariance matrix Ψ_1. The second-level random effects b_{ij} are q_2-dimensional independently distributed vectors with variance-covariance matrix Ψ_2, assumed to be independent of the first-level random effects. The random

effects design matrices $B_{i,jk}$ and B_{ijk} depend on first and second level groups and possibly on the values of some covariates at the k^{th} observation. The within-group errors ε_{ijk} are independently distributed as $N(0,\sigma^2)$ and are independent of the random effects. The assumption of independence and homoscedasticity for the within-group errors can be relaxed,

We can express (5.2.1) and (5.2.2) in matrix form as

$$\begin{aligned} y_{ij} &= f_{ij}(\phi_{ij}, v_{ij}) + \varepsilon_{ij}, \\ \phi_{ij} &= A_{ij}\beta + B_{i,j}b_i + B_{ij}b_{ij}, \end{aligned} \tag{5.2.3}$$

For $i = 1,2,...,M; j = 1,2,...M_{i,}$ where

$$y_{ij} = \begin{bmatrix} y_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ y_{ijn_{ij}} \end{bmatrix}, \phi_{ij} = \begin{bmatrix} \phi_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{ijn_{ij}} \end{bmatrix}, \varepsilon_{ij} = \begin{bmatrix} \varepsilon_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{ijn_{ij}} \end{bmatrix}, f_{ij}(\phi_{ij}, v_{ij}) = \begin{bmatrix} f(\phi_{ij1}, v_{ij1}) \\ \cdot \\ \cdot \\ \cdot \\ f(\phi_{ijn_{ij}}, v_{ijn_{ij}}) \end{bmatrix}, v_{ij} = \begin{bmatrix} v_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ v_{ijn_{ij}} \end{bmatrix},$$

$$A_{ij} = \begin{bmatrix} A_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ A_{ijn_{ij}} \end{bmatrix}, B_{i,j} = \begin{bmatrix} B_{i,j1} \\ \cdot \\ \cdot \\ \cdot \\ B_{i,jn_{ij}} \end{bmatrix}, B_{ij} = \begin{bmatrix} B_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ B_{ijn_{ij}} \end{bmatrix}.$$

Extensions of the NLME model to more than two levels of nesting are straightforward. For example, with three levels of nesting the second-stage model for the group-specific coefficients is

$$\begin{aligned} &\phi_{ijkl} = A_{ijkl}\beta + B_{i,jkl}b_i + B_{ij,kl}b_{ij} + B_{ijkl}b_{ijk}, \\ &b_i \sim N(0,\psi_1), b_{ij} \sim N(0,\psi_2), b_{ijk} \sim N(0,\psi_3) \end{aligned}$$

In this chapter we extend the basic nonlinear mixed effects model to allow heteroscedastic correlated within group errors. We describe how the `nlme()` function is used to fit the extended nonlinear mixed effects model and illustrate its various capabilities through examples. We also show the estimation of the simple nonlinear mixed effects models can be applied to the extended model and decomposition of variance, covariance structure of within group errors into two independent components: a variance structure and a correlation.

5.3 General formulation of Extended Nonlinear Mixed Effects model

The basic single-level NLME model (5.1.3) assumes that the within-group errors ε_i are independent $N(0,\sigma^2 I)$ random vectors. The extended single-level NLME model relaxes this assumption by allowing heteroscedastic and correlated within-group errors, being expressed for $i = 1,...,M$ as

$$\begin{aligned} &y_i = f_i(\phi_i, v_i) + \varepsilon_i; \\ &\phi_i = A_i\beta + B_i b_{i,} \\ &b_i \sim N(0,\Psi), \varepsilon_i \sim N(0,\sigma^2\Lambda_i) \end{aligned} \qquad (5.3.1)$$

Where, Λ_i are positive-definite matrices parameterized by a fixed, generally small, set of parameters λ. As in the basic NLME model, the within-group errors ε_i are assumed to be independent for different i and to be independent of the random effects b_i. The σ^2 is factored out of the Λ_i for computational reasons (it can then be eliminated from the profiled likelihood function).

Similarly, the extended two-level NLME model generalizes the basic two-level NLME model (5.2.3) described by letting

$$\varepsilon_{ij} \sim N(0, \sigma^2 \Lambda_{ij});, i = 1,...,M; j = 1,...,M_i,$$

Where Λ_{ij} are positive-definite matrices parameterized by a fixed λ vector. This readily generalizes to a multilevel model with Q levels of random effects.

5.3.1 Estimation and computational methods

The estimation procedure of the nonlinear mixed effects model is described in this section. Because Λ_i is positive-definite, it admits an invertible square-root $\Lambda_i^{1/2}$ (Thisted, 1988), with inverse $\Lambda_i^{-1/2}$, such that

$$\Lambda_i = \Lambda_i^{T/2} \Lambda_i^{1/2} \text{ and}$$

$$\Lambda_i^{-1} = \Lambda_i^{-1/2} \Lambda_i^{-T/2}$$

Letting

$$\begin{aligned} y_i^* &= \Lambda_i^{-T/2} y_i \\ f_i^*(\phi_i, v_i) &= \Lambda_i^{-T/2} f_i(\phi_i, v_i) \\ \varepsilon_i^* &= \Lambda_i^{-T/2} \varepsilon_i \end{aligned} \qquad (5.3.2)$$

And noting that

$$\varepsilon_i^* \sim N\left[\Lambda_i^{-T/2} 0, \sigma^2 \Lambda_i^{-T/2} \Lambda_i \Lambda_i^{-1/2}\right] = N(0, \sigma^2 I),$$

Thus, the model (5.3.1) can be written as

$$\begin{aligned} y_i^* &= f_i^*(\phi_i, v_i) + \varepsilon_i^* \\ \phi_i &= A_i \beta + B_i b_i \\ b_i &\sim N(0, \Psi);, \varepsilon_i^* \sim N(0, \sigma^2 I); i = 1,2,...,M \end{aligned}$$

That is, y_i^* is described by a basic Nonlinear mixed effects model.

Since, the differential of the linear transformation $y_i^* = \Lambda_i^{-T/2} y_i$ is simply $dy_i^* = |\Lambda_i|^{-1/2} dy_i$, the log-likelihood function corresponding to the extended Nonlinear mixed effectsmodel (5.3.1) is expressed as

$$l(\beta, \sigma^2, \Delta, \lambda \mid y) = \sum_{i=1}^{M} \log\, p(y_i \mid \beta, \sigma^2, \Delta, \lambda)$$

$$= \sum_{i=1}^{M} \log p(y_i^* \mid \beta, \sigma^2, \Delta, \lambda) - \frac{1}{2} \sum_{i=1}^{M} \log|\Lambda_i|$$

$$= l(\beta, \sigma^2, \Delta, \lambda \mid y^*) - \frac{1}{2} \sum_{i=1}^{M} \log|\Lambda_i|$$

The log-likelihood function $l(\beta, \sigma^2, \Delta, \lambda \mid y^*)$ corresponds to a basic NLME model with model function f_i^*.

5.3.1.1 Alternating Algorithm

The PNLS step of the alternating algorithm for the extended NLME model consists of minimizing, over β and $b_i, i = 1,2,...,M,$ the penalized nonlinear least-square function

$$\sum_{i=1}^{M} \left[\left\| y_i^* - f_i^*(\beta, b_i) \right\|^2 + \left\| \Delta b_i \right\|^2 \right] = \sum_{i=1}^{M} \left\{ \left\| \Lambda_i^{-T/2} \left[y_i - f_i(\beta, b_i) \right] \right\|^2 + \left\| \Delta b_i \right\|^2 \right\}$$

The derivative matrices and working vector used in the Gauss-Newton algorithm for the PNLS step and also in the LME step are defined as

$$\hat{X}_i^{*(w)} = \frac{\partial f_i^*}{\partial \beta^T}\bigg|_{\hat{\beta}^{(w)}, \hat{b}_i^{(w)}} = \Lambda_i^{-T/2} \hat{X}_i^{(w)}$$

$$\hat{Z}_i^{*(w)} = \frac{\partial f_i^*}{\partial b_i^T}\bigg|_{\hat{\beta}^{(w)}, \hat{b}_i^{(w)}} = \Lambda_i^{-T/2} \hat{Z}_i^{(w)}$$

$$\hat{w}_i^{*(w)} = y_i^* - f_i^*\left(\hat{\beta}^{(w)}, \hat{b}_i^{(w)}\right) + \hat{X}_i^{*(w)} \hat{\beta}^{(w)} + \hat{Z}_i^{*(w)} \hat{b}_i^{(w)} = \Lambda_i^{-T/2} \hat{w}_i^{(w)}$$

Where,

$\hat{X}_i^{*(w)}$ is the response vector

$\hat{Z}_i^{(w)}$ is the fixed effects design matrix

$\hat{w}_i^{*(w)}$ is the random effects design matrix

The LME approximation to the log-likelihood function of the extended single-level NLME model is

$$l^*_{LME}(\beta,\sigma^2,\Delta,\lambda \mid y) = l_{LME}(\beta,\sigma^2,\Delta,\lambda \mid y^*) - \frac{1}{2}\sum_{i=1}^{M}\log|\Delta_i|$$

Which has the same form as the log-likelihood of the extended single-level LME model. The log-restricted-likelihood for the extended NLME model is similarly defined.

5.3.2 Laplacian and Adaptive Gaussian Approximations

For the extended single-level NLME model, the objective function which is minimized to produce the conditional modes $\hat{b}_i$ used in the Laplacian and adaptive Gaussian approximations is

$$g^*(\beta,\Delta,\lambda,y_i,b_i) = \left\|\Lambda_i^{-T/2}\left[y_i - f_i(\beta,b_i)\right]\right\|^2 + \left\|\Delta b_i\right\|^2$$

The corresponding approximation to the second –derivative matrix of g^* with respect b_i evaluated at $\hat{b}_i$ is:

$$\frac{\partial\partial^2 g^*(\beta,\Delta,\lambda,y_i,b_i)}{\partial b_i \partial b_i^T}\bigg|_{\hat{b}_i} \cong G^*(\beta,\Delta,\lambda,y_i) = \frac{\partial f_i(\beta,b_i)}{\partial b_i}\bigg|_{\hat{b}_i}\Lambda_i^{-1}\frac{\partial f_i(\beta,b_i)}{\partial b_i^T}\bigg|_{\hat{b}_i} + \Delta^T\Delta$$

The modified Laplacian approximation to the log-likelihood of the extended single-level NLME model is then given by:

$$l^*_{LA}(\beta,\sigma^2,\Delta,\lambda,\mid y) = -\frac{N}{2}\log(2\pi\sigma^2) + M\log|\Delta|$$
$$-\frac{1}{2}\left\{\sum_{i=1}^{M}\log\left|G^*(\beta,\Delta,\lambda,y_i\right| + \sigma^{-2}\sum_{i=1}^{M}g^*(\beta,\Delta,\lambda,y_i,\hat{b}_i)\right\} - \frac{1}{2}\sum_{i=1}^{M}\log|\Lambda_i|$$

And the adaptive Gaussian approximation is given by

$$l^*_{AGQ}(\beta,\sigma^2,\Delta,\lambda,|\, y) = -\frac{N}{2}\log(2\pi\sigma^2) + M\log|\Delta| - \frac{1}{2}\sum_{i=1}^{M}\log\left|G^*(\beta,\Delta,\lambda,y_i)\right|$$

$$+\sum_{i=1}^{M}\log\left(\sum_{j}^{N_{GQ}}\exp\left\{-g^*\left[\beta,\Delta,\lambda,y_i,\hat{b}_i+\sigma(G^*)^{-\frac{1}{2}}(\beta,\Delta,\lambda,y_i)z_j\right]/2\sigma^2+\left\|z_j\right\|^2/2\right\}\prod_{k=1}^{q}w_{jk}\right)$$

$$-\frac{1}{2}\sum_{i=1}^{M}\log|\Lambda_i|$$

In order to keep the optimization problem feasible, an "iteratively reweighted" scheme is used to approximate the variance function. The fixed and random effects used in the variance function are replaced by their current estimates and held fixed during the log-likelihood optimization. New estimates for the fixed and random effects are then produced and the procedure is repeated until convergence. In the case of alternating algorithm, the estimates for the fixed and random effects obtained in the PNLS step are used to calculate the variance function weights in the LME step. If the variance function does not depend on either the fixed effects or the random effects, then no approximation is necessary.

The R function `nls` uses a Gauss-Newton algorithm to determine the nonlinearleast squares estimates of the parameters in a nonlinear regression model.

5.4 Variance function for modeling heterocedasticity

Variance functions are used to model the variance structure of the within group errors using covariates.They have been studied in detail in the context of nonlinear mixed effects models by Davidian and Giltinan(Davidian and Giltinan 1995). Following Davidian and Giltinan (1995) we define the general variance function model for the within group errors in the extended single level nonlinear mixed effects model (4.1.3) as:

$$\text{var}(e_{ij}\,|\,b_i) = \sigma^2 g^2(\mu_{ij}, v_{ij}, \delta); i = 1,2,...,M; j = 1,2,...,n_i \qquad (5.4.1)$$

Where $\mu_{ij} = E[y_{ij} \mid b_i], v_{ij}$ is a vector of variance covariates, δ is a vector of variance parameters and $g(\cdot)$ is the variance function, assumed continuous in δ. For example, if the within group variability is believed to increase with same power of absolute value of a covariate v_{ij}, we can write the variance model as :

$$\text{var}(e_{ij} \mid b_i) = \sigma^2 (v_{ij})^{2\delta}$$

The variance function in this case is $g(x, y) = \mid x \mid^y$ and the covariate v_{ij} can be the expected value μ_{ij}. The variance function formulation (5.4.1) is very flexible and inductive, because it allows the within group variance to depend on the fixed effects β, and the random effects b_i, through the expected values μ_{ij}. However as discussed in Davidian and Giltinan (1995), it poses some theoretical and computational difficulties as the within group errors and the random effects can no longer assumed to be independent. Under the assumption that $E[e_{ij} \mid b_i] = 0$, it is easy to verify that $\text{var}(e_{ij}) = E(\text{var}(e_{ij} \mid b_i))$, so that the dependence on the unobserved random effects can be avoided by integrating them out of the variance mode. Because the variance function of g is generally non-linear in b_{ij} integrating the random effects out of the variance model (5.4.1) does not lead to a computationally feasible optimization procedure. Instead we proceed as in Davidian and Gillinan (1995) and use an approximate variance model in which the expected value u_{ij} are replaced by their BLUPs

$$\hat{\mu}_{ij} = x_{ij}^T \beta + z_{ij}^T \hat{b}_i$$

Where x_{ij} and z_{ij} denotes, respectively the j^{th} row of X_i and Z_i, thus

$$V(e_{ij}) \approx \sigma^2 g^2 (\hat{\mu}_{ij}, v_{ij}, \delta); i = 1,2,...,M; j = 1,2,...,n_i \qquad (5.4.2)$$

Under this approximation, the within group error are assumed independent of the random effects as in (4.1.3) and the results in section 5.3.1 can still be used.

Note that if the conditional variance model (5.4.1) does not depend on μ_{ij}, (5.4.2) gives the exact marginal variance and no approximation is required.

When the conditional variance model (5.4.1) depends on μ_{ij}, the optimization algorithm follows an "iteratively reweighted" scheme, for given $\beta^{(t)}, \theta^{(t)}, \lambda^{(t)}$, the corresponding BLUP's $\hat{\mu}_{ij}^{(t)}$ can be obtained and held fixed while the objective function is optimized to produce new estimates $\beta^{(t+1)}, \theta^{(t+1)}, \lambda^{(t+1)}$ which in turn give updated BLUP's $\hat{\mu}_{ij}^{(t+1)}$ with the process iterating until convergence. The resulting estimates approximate the (restricted) maximum likelihood estimates. When the variance model does not involve μ_{ij} the likelihood can be directly optimized producing the exact (restricted) maximum likelihood estimates.

5.5 Decomposing the within group variance covariance structure

The Λ_i matrices can always be decomposed into a product of simpler matrices:

$$\Lambda_i = V_i C_i V_i \tag{5.5.1}$$

Where, V_i is diagonal and C_i is a correlation matrix, that is a positive definite matrix with all diagonal elements equal to one. The matrix V_i in (5.5.1) is not uniquely defined, as we multiply any number of its rows by -1 and still get the same decomposition. To ensure uniqueness, we require that all the diagonal elements of V_i be positive. It is easy to verify that:

$\text{var}(e_{ij}) = \sigma^2 [v]_{ij}^2$, and

$cor(e_{ij}, e_{jk}) = [c_i]_{jk}$

So that V_i describes that variance and C_i describes the correlation of the within group errors e_i. This decomposition of Λ_i into variance structure component and a correlation structure component is convenient both theoretically and

computationally. It allows us to model and develop codes for the two structures separately and to combine them into a flexible family of models for the within group variance covariance.

5.6 Fitting of Nonlinear Mixed Effects Model

The `nlme` function of R-software includes several optional arguments. The `model` argument is required and consists of either a two-sided formula specifying the nonlinear model to be fitted, or an `nlsList` object. Any R nonlinear formula can be used, giving the function considerable flexibility. As with the `nls` fits, there is an advantage of encapsulating the model expression in an R function when fitting an `nlme` model, in that it allows analytic derivatives of the model function to be passed to `nlme` and used in the optimization algorithm. The R function `derive` can be used to createmodel functions that return the value of the model and its derivatives as a `gradient` attribute. If the value returned by the model function does not have a `gradient` attribute, numerical derivatives are used in the optimization. The arguments `fixed` and `random` are formulas, or lists of formulas, defining the structures of the fixed and random effects in the model. Each parameter in the model will usually have an associated fixed effect, but it may, or may not, have an associated random effect. Because the `nlme` model assumes that all random effects have expected value zero, the inclusion of a random effect without a corresponding fixed effect would be unusual. Any covariate defined in the `fixed` and `random` formulas can, alternatively, be directly incorporated in the model formula. However, declaring the covariates in `fixed` and `random` allows for more efficient calculation of derivatives and is useful for update methods.

`Data` names a data frame in which any variables used in `model,fixed, random`, and `groups` are to be evaluated. The`groups` argument is a one-sided formula, or an R expression, which when evaluated in `data`, returns a factor with the group label of each observation. The

`start`argument provides a list, or a vector, of starting values for the iterative algorithm. Table 5.1 lists the most important methods for class `nlme`.

Table 5.1: Main`nlme` methods

`ACF`	empirical autocorrelation function of within-group residuals
`Anova`	likelihood ratio or Wald-type tests
`augPred`	predictions augmented with observed values
`coef`	estimated coefficients for different levels of grouping
`fitted`	fitted values for different levels of grouping
`fixef`	fixed-effects estimates
`intervals`	confidence intervals on model parameters
`loglik`	log-likelihood at convergence
`pairs`	scatter-plot matrix of coefficients or random effects
`plot`	diagnostic Trellis plots
`predict`	predictions for different levels of grouping
`print`	brief information about the fit
`qqnorm`	normal probability plots
`ranef`	random-effects estimates
`resid`	residuals for different levels of grouping
`summary`	more detailed information about the fit
`update`	update the `lme` fit
`Variogram`	semivariogram of within-group residuals

We illustrate the use of these methods through the use of the data of Applehpd data as discussed in chapter-1.We can now use the `nlme` method to display the results and to assess the quality of the fit. We first fit a model using the `nlme()` function of R-software where we assume homoscedasticity i.e., we fit a homoscedastic nonlinear mixed effects model. The results obtained are given in Table 5.2.

Table 5.2: Summary of the results of the fixed effects obtained by fitting Homocedastic Nonlinear mixed effects model

	Value	Std.Error	t-value	p-value
Asym	139.14164	9.666241	14.3946000.000	
Xmid	40.79059	2.289960	17.812790	0.000
scal	13.92837	1.288380	10.810760	0.000

Table 5.2 provides the values of different fixed effects along with their standard error, t-value and p-value.From the table it is clear that all the three parameters have a significant role in the yield.

Table 5.3: 95% confidence intervals for fixed effects

	lower	estimate	upper
Asym	124.23299	140.54153	156.85007
xmid	37.36638	41.34033	45.31428
scal	11.67813	14.17846	16.67878

Table 5.4: 95% confidence intervals for random effects

	lower	estimate	upper
sd(Asym)	2.667605	7.347657	20.2384
Within-group standard error	5.373930	**6.921892**	8.915744

Table 5.3 and 5.4 provides the approximate 95% confidence intervals for the fixed and random effects of the homosedastic mixed effects model. It is clear from the above table that the within group standard error has a larger value of 6.92 with an interval of 5.37-8.91.

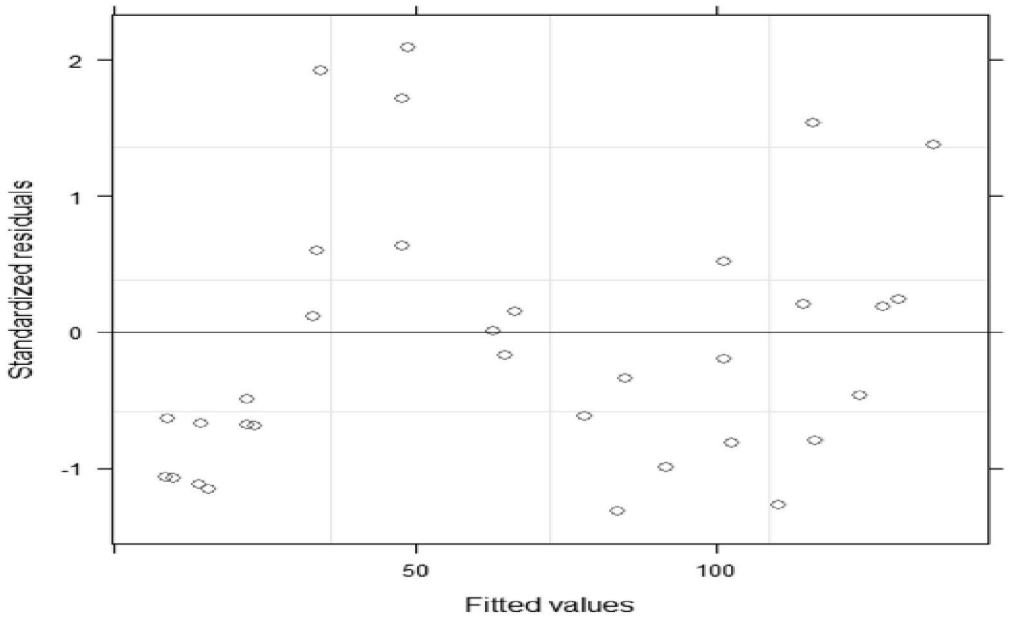


Fig. 5.1: Plot of standardized residuals versus fitted values for the homoscedastic fitted model

From the above Fig. 5.1 there is clear indication of the within-group heterosedasticity because the fitted values are bounded away from zero. So we use the varPower (`varPower()` function of `nlme()` library of R) variance function to model the heterosedasticity.

5.6.1 Variance function in nlme

The `nlme` library of R-software provides a set of classes of variance functions, the `varfunc` classes, that can be used to specify the within group variance models in the extended nonlinear mixed effects model. Table 5.1 lists the standard `varfunc` classes induced in the nlme library. The `varfunc` constructors have the same name as their corresponding class.

Table 5.5: Standard var-Func classes

VarFixed	Fixed variance
VarIdent	Different variance per stratum
VarPower	Power of covariance
VarExp	Exponential of covariates
VarConstPower	Constant plus power of covariates
VarCanb	Combination of variance function

The two main arguments for most of the `varFunc` constructors are value and form. The first specifies the value of the variance parameter δ and the second is a one-sided formula specifying the variance covariate V and optionally a stratification variable for the variance parameters- different parameters are used for each level of stratified variable.

5.6.2 Using variance function with nlme()

The above mentioned variance functions can be used in the `nlme()` functions using the weights argument. By default, `weights=NULL`, corresponding to a homoscedastic variance model for the within group errors. Variance models can be specified in `weights` either as a one-sided formula, in which case it is passed as the single argument to the `varFixed` constructors, or as a `varFunc` object, created using the standard constructors given in 5.5.1. To illustrate the use variance function in nlme we continue with the Applehpd dataset in which the presence of heterosceadsticity can be seen so we next fit the heteroscedastic nonlinear mixed effects model. The fitted heteroscedastic nonlinear mixed effects model is given in Table 5.6.

Table 5.6: Summary of results of the fixed effects obtained by fitting Heterosedastic Nonlinear mixed effects model

	Value	Std.Error	t-value	p-value
Asym	116.97947	5.484656	21.328500	0.000
xmid	33.99852	1.506414	22.569170	0.000
scal	9.26038	0.631472	14.664730	0.000

Table 5.6 provides the values of different fixed effects along with the standard error, t-value and p-value. From the table it is clear that there is significant increase in parameters associated with yield.

Table 5.7: 95% confidence intervals for fixed effects

	lower	estimate	upper
Asym	106.267502	116.979473	127.69144
xmid	31.056378	33.998525	36.94067
scal	8.027059	9.260375	10.49369

Table 5.8: 95% confidence intervals for random effects

	lower	estimate	upper
sd(Asym)	1.2434	3.4566	7.5677
Within-group standard error	1.6758	2.65434	4.5456

As can be seen from the table above that the within group error assumes a smaller value of 2.65 with a 95% interval of 1.67-4.55.

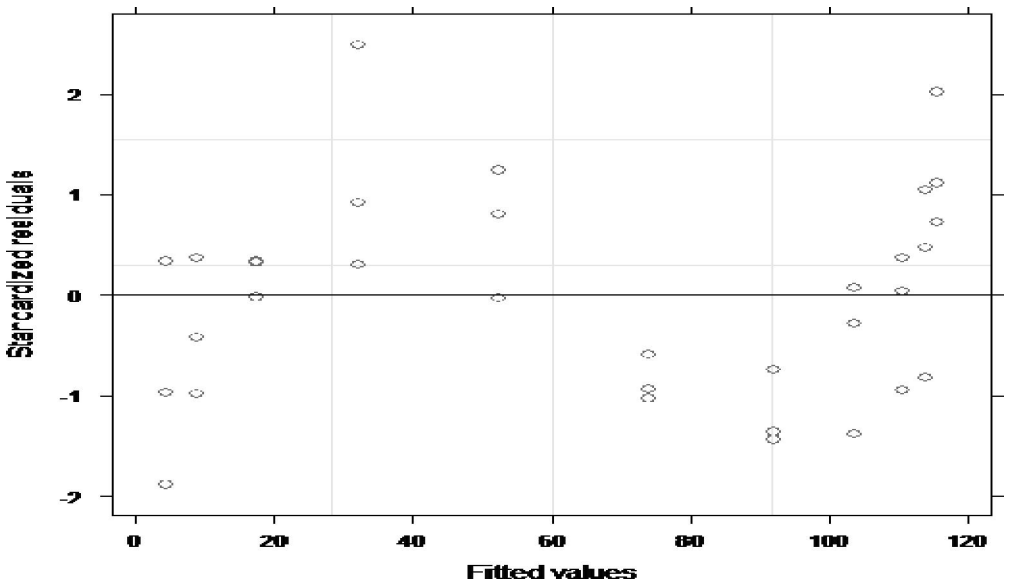


Fig. 5.2: Plot of standardized residuals versus fitted values for the VarPower (heteroscedastic) fitted model

We can test the significance of the variance parameter in the varPower model using the anova method which as expected strongly rejects the assumption of homosedasticity. The results obtained are given in the Table 5.8.

Table 5.9: Empirical comparison of the fitted models i.e Homosedastic nonlinear mixedeffects model and Heterosedastic nonlinear mixed effects model

Model	AIC	BIC	logLik	L.Ratio	p-value
Homocedastic	238.0995	247.0785	113.0497		
Heterosedastic	236.0881	243.5707	110.044	10.011	0.00152

Thus, it can be concluded from the above table that the heterosedastic model is the better fit since AIC/BIC is lowest also the likelihood ratio test is significant which supports our conclusion.

5.6.3 Correlation structure for modeling dependence

Correlation structures are used to model dependence among observations. In case of the Nonlinear mixed effects models, correlation structures are used to model dependence among the within groups errors. Historically correlation structures have been developed for two main classes of data: time series data and spatial data. To establish a general framework for correlation structures, we assume that the within-group errors e_{ij} are associated will position vector ρ_{ij}. The correlation structures used in this chapter are assumed to be isotropic (Cressie 1993); i.e. the correlation between two within group errors e_{ij}, e_{ij}' are assumed to depend on the corresponding position vector ρ_{ij}, ρ_{ij}' onlythrough same distance between them say $d(\rho_{ij}, \rho_{ij}')$ andnot on the particular values they assume. The general within group correlation structure for single level grouping for $i = 1,..., M$ and $j = 1,..., n_i$ can be expressed as

$$Cor(e_{ij}, e_{ij}') = h[d(\rho_{ij}, \rho_{ij}'), \rho] \qquad (5.6.1)$$

Where ρ is vector of correlation parameters and $h(\cdot)$ is a correlation function taking values between -1 and 1 assumed continuous in ρ and such that $h(0, \rho) = 1$, that is if two observations have identical position vectors, they are the same observations and therefore have correlation 1.

5.6.4 Spatial correlation structures

These were originally proposed to model dependence in data indexed by continuous two dimensional position vectors, such as geostatistical data, lattice data and point patterns. Here we consider only isotropic spatial correlation structures which can be given as continuous function of same distance between position vectors that are easily generalized to any finite number of position dimensions. The basic reference for spatial correlation structure used will be the mixed effects models(Diggle*et al*., 1994).

For simplicity of notations we denote by e_x the observations taken at position $x=(x_1,...,x_r)^T$. Any distance metric may be used with isotropic spatial correlation structures, the most common being the Euclidean distance which in our case can be defined as :

$$d_E(e_x, e_y) = \sqrt{\sum_{i=1}^{r}(x_i - y_i)^2}$$

Other popular choices are Manhattan or L1 distance

$$d_{Man}(\varepsilon_x, \varepsilon_y) = \sum_{i=1}^{r} | x_i - y_i |$$

and maximum distance

$$d_{\max}(\varepsilon_x, \varepsilon_y) = \max_{i=1,...r} | x_i - y_i |$$

Spatial correlation structures can generally be represented by their semivariogram, instead of their correlation function (Cressie 1993). The semivariogram of an isotropic spatial correlation structure with a distance function $d(\cdot)$ can be defined as:

$$\lambda(d(e_x, e_y), \lambda) = \frac{1}{2}\text{var}(e_x - e_y) = \frac{1}{2}E(e_x - e_y)^2 \qquad (5.6.2)$$

with the last equality following from $E(e_x) = E(e_y) = 0$. The within group errors can be standardized to have unit variance, without changing their correlation structure. So without loss of generality, we assume that $\text{var}(e_x) = 1 \forall x$. In this case $\gamma(.)$ will depend only on the correlation parameters ρ and it is easy to verify that :

$$\gamma(s, \rho) = 1 - h(s, \rho)$$

It follows from $h(0,\rho)=1$ that $\gamma(0,\rho)=0$. The standardized residuals $r_{ij}=\frac{(y_{ij}-\hat{y}_{ij})}{\hat{\sigma}_{ij}}$ with $\sigma_{ij}^2=\text{var}(e_{ij})$, are the primary quantities used for estimating the semivariogram. The classical estimator of the semivariogram (Matheron 1962)can be expressed as

$$\hat{\gamma}(\text{s})=\frac{1}{2\text{N}(\text{s})}\sum_{\text{i}=1}^{\text{M}}\sum_{\text{d}(\text{p}_{\text{ij}},\text{p}_{\text{ij}})}(\text{rij}-\text{rij}')^2 \qquad (5.6.3)$$

where $N(s)$ denotes the number of residual pairs at a distance s of each other. Because $\hat{\gamma}(\text{s})$ uses the squared differences between residual pairs, it can be quite sensitive to outliers. Furthermore, because each residual r_{ij} appears in n_i-1 squared differences in (5.6.3), a single outlier can affect the estimation of the semivariogram at several distances. A robust estimator of the semivariogram proposed by Cressie and Hawkins(Cressie and Hawkins 1980), uses the square-root differences to reduce the influence of outliers.

$$\hat{\gamma}(s)=\left(\frac{1}{2N(s)}\sum\sum|r_{ij}-r_{ij}^{'}|^{1/2}\right)^4\Big/0.457+0.494/N(s) \qquad (5.6.4)$$

5.6.5 Correlation structures in nlme

The `nlme` library of R-software provides a set of classes of correlation structures, the `corstruct()` classes, which can be used to specify within-group correlation models in the extended nonlinear mixed effects model (Table 5.10) lists the standard `corstruct` classes in the `nlme()` library. The corstructure constructors have the same names as other corresponding classes.

Table 5.10:Standard construct classes

Corcompsymm	Compound summary
Corsymm	General
Cor ARI	Autoregressive of order 1
Cor ARM	Autoregressive moving average
Cor Exp	Exponential
Cor Gaus	Guassian
Cor Lin	Linear
Cor ratio	Rational
Cor sphare	Spherical

The two main arguments to most of the `construct` constructors are `value` and `form`. The first specified the values of the correlation parameters and the second is a one sided formula specifying the position vector and optionally a grouping variable for the data-observations in different groups are assumed independent.The argument fixed available in all construct constructors may be used to specify fixed correlation structures, wherecoefficient are not allowed to change during the numerical optimization in the modeling functions. `If fixed = true` the coefficients in the structure are fixed otherwise not. Default is `fixed = FALSE`. Several methods are available for each `corstruct` class, including initialize which initializes position vector and grouping variables, and `cormatrix`, which extracts the within-group correlation matrix.

5.6.6 Using correlation structure with nlme

Correlation structures can be specified in `nlme` through the correlation argument. By default `correlation =NULL` corresponding to uncorrelated within group errors. Correlation structures can be specified as `corstruct` objects, created using the standard constructor. In this section we will describe the use of correlation models in `nlme()` through the analysis of examples of grouped data with correlated within group errors. While assessing the adequacy of a correlation model, it is often useful to consider diagnose plots of the normalized residuals, if the within group variance covariance model is correct, the normalized

residuals should be approximately distributed as independent N (0, 1) random vectors.

We revisit the Applehpd data set discussed in chapter one to illustrate the use of `corStruct` classes in nlme in combination with variance functions. The corStruct classes representing spatial correlation structures are corExp, corGaus, corLin, corRatio and corSpher. The augment form is one sided formula specifying a position vector and optionally, a grouping variable. The coordinates of the position vector must be numeric variables, but are otherwise unrestricted. The augment metricis a character string specifying a metric to be used for calculating the between pairs distances. Possible values include "Euclidean", "maximum" and "manhattan" corresponding to the metrics.

The Variogram method for the `nlme()` class estimates the sample semivariogram from the residuals of the `nlme()` object. The augments `resType` and `robust control,` respectively, are used to decide what type of the residuals should be used("pearson" or "response") and whether the `robust algorithm ()` or the `classical algorithm` () should be used to estimate the semivariogram. The defaults are `resType="pearson"` and `robust=False,` so that classical estimates of the semivariogram are obtained from the standardized residuals. The augment form is a one-sided formula specifying the position vector to be used for the semivariogram calculations. The results obtained by using the `varCorr()` model are summarized in the Table 5.11.

Table 5.11: Variogram of the data obtained at different distances for different pairs of observations

Variogram	Distance	No. of pairs
0.4921405	7	30
1.1092445	14	27
1.5519567	21	24
1.4869038	28	21
1.0774536	35	18
0.7486675	42	15
0.6162776	49	12
0.6793361	56	9
1.2092667	63	6
2.4403386	70	3

The columns in the table returned by Variogram represent, respectively, the sample semivariogram, the distances, and the number of residual pairs used in the estimation. Because of the imbalance in the tca measurements, the number of the residual pairs used at each distance varies considerably, making some semivariogram estimates more reliable than the others.In general, the number of residual pairs used in the semivariogram estimation decreases with distance, making the values at large distances unreliable. We can control the maximum distance for which semivariogram estimates should be calculated using the argument `maxDist()`. A graphical representation of the sample semivariogram is obtained with the plot method for class Variogram is reported in Fig. 5.3 below.

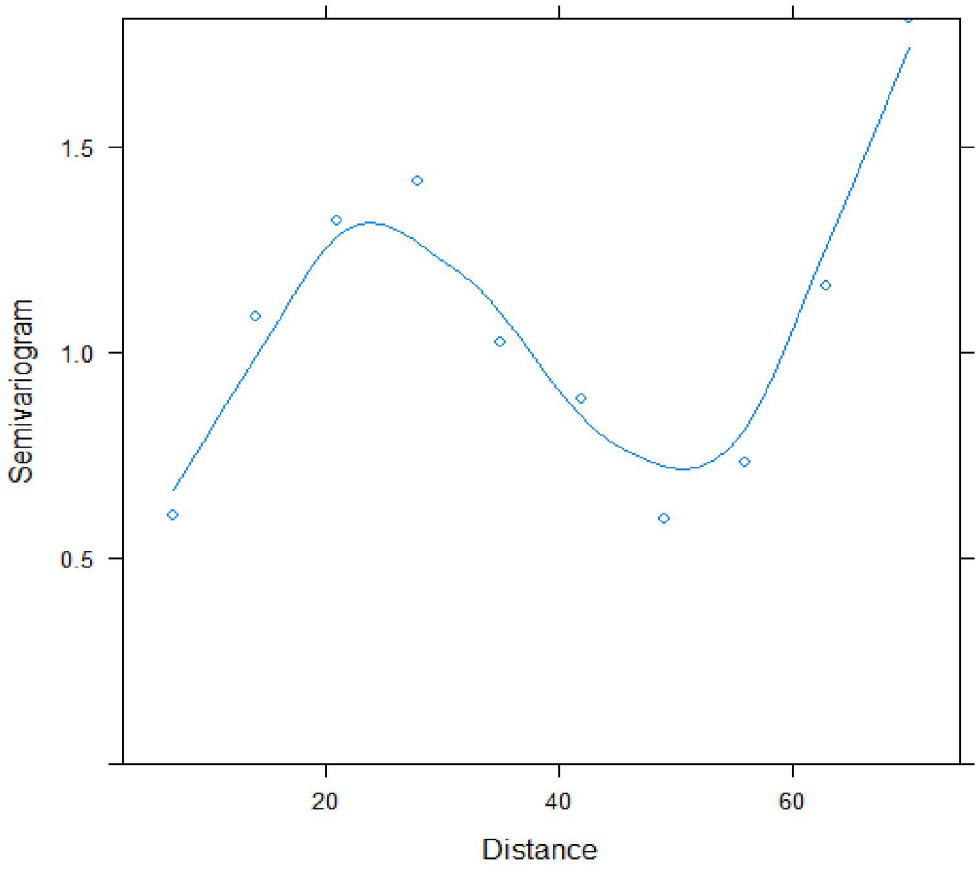


Fig. 5.3: Sample semivariogram estimates corresponding to the standardized residuals of the fitted varPower model

As can be seen from Fig. 5.3, a loess smoother is to enhance the visualization ofsemivariogram patterns. The semivariogram seems to increase with distance up to 25 and then decreases at 50 and then again goes up at 53. The empirical ACF is used to investigate the correlation at different lags. The ACF method can also be used with nlme objects.

5.12: The empirical ACF at different lags

Lag	ACF
0	1.00000000
1	0.24204168
2	-0.28284766
3	-0.42792059
4	-0.46384734
5	-0.05096052
6	0.33300306
7	0.57722063
8	0.13677655
9	-0.22742765
10	-1.02560784

Because they are based on fewer residual pairs, empirical autocorrelations at larger lags are less reliable. We can control the number of lags calculated in ACF using the maxLag argument. We use it in the plot of empirical ACF, displayed in figure 5.4.

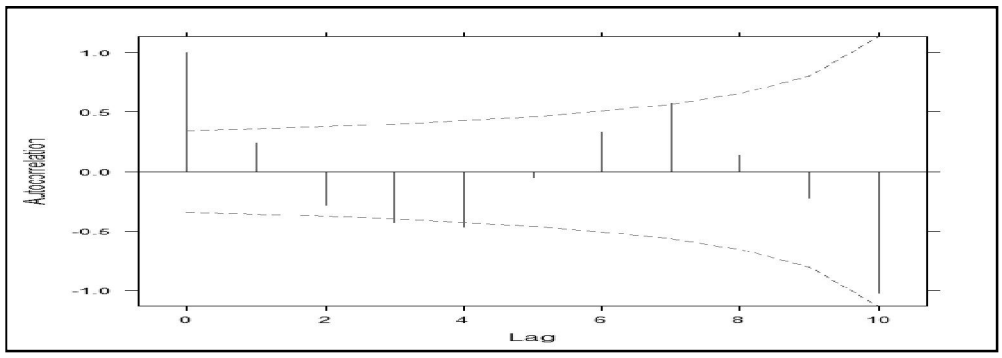


Fig 5.4: Empirical autocorrelation function corresponding to the standardized residuals of the fitted objects

The figure 5.4 shows that only the lag-1 autocorrelation is significant at the 5% level, but the lag-2 is significant. This suggests two different candidate correlation structures for modeling the within-group error covariance structure: $AR(1)$ and $MA(2)$. The two correlation models are not nested, but can be compared using the information criteria provided by the anova method, AIC and BIC. The empirical lag-1 autocorrelation is used as a starting value for the corAR1 coefficient. The results obtained by the anova method is summarized in table 5.13.

Table 5.13: Empirical comparison of the fitted Autoregressive of order 1 model and Autoregressive moving average (0,1) model

Models	AIC	BIC	logLik	L.Ratio test	p-value
Autoregressive of order 1	231.60	242.07	-108.80		
Autoregressive moving Average(0,1)	229.28	240.26	-107.64	2.312	0.1283

The $AR(1)$ model uses one fewer parameter than the $MA(2)$ model to give a larger log-likelihood and hence is the preferred model by both AIC and BIC.

An alternative, "intermediate model between the $AR(1)$ and $MA(2)$ correlation structures is the $ARMA(1,1)$ model. This structure has an exponentially decaying ACF for lags ≥ 2, but allows greater flexibility in the lag-1 autocorrelation. Because the $AR(1)$ model is nested within the $ARMA(1,1)$ model, they can be compared by anova method. The results are summarized in table 5.14.

Table 5.14: Empirical comparison of the fitted Autoregressive of order 1 model and Autoregressive moving average (1,1) model

Models	AIC	BIC	logLik	L.Ratio test	p-value
Autoregressive of order (1)	231.60	242.07	-108.80		
Autoregressive Moving Average(1,1)	224.28	235.26	-105.64	4.201	0.0318

The $ARMA\ (1,1)$ gives a significantly better representation of the within-group correlation, as indicated by the small p-value for the likelihood ratio test. So we use $ARMA\ (1,1)$ model for the within-group errors, and the results obtained are summarized in table 5.15.

Table 5.15: Summary of results of the fixed effects obtained by fitting $ARMA\ (1,1)$ model

	Value	Std.Error	t-value	p-value
Asym	134.93	8.83	15.28	0.000
xmid	39.57	2.59	15.24	0.000
scal	12.35	1.38	8.97	0.000

Table 5.15 provides the values of different fixed effects along with their standard error, t-value and p-value. We assess the variability in $ARMA\ (1,1)$ model with the intervals method. The 95% confidence intervals for the variance components are obtained as:

Table 5.16: 95% confidence intervals for fixed effects

	lower	estimate	upper
Asym	117.68	134.93	152.18
xmid	34.50	39.57	44.64
scal	9.68	12.34	15.04

Table 5.17: 95% confidence intervals for random effects

	lower	estimate	upper
sd(Asym)	2.566	4.877	9.444
Within-group standard error	0.0243	0.044	0.354
Correlation structure range	0.486	0.544	0.947

As can be seen from both the tables, the confidence interval is bounded away from zero, suggesting that the $ARMA\ (1,1)$ model produced a significantly better fit.

When an lme object includes a corStruct object, we can further assess the adequacy of the correlation model with the plot method. In this case, instead of a loess smoother,the fitted semivariogram corresponding to the corStruct object is displayed in the plot.

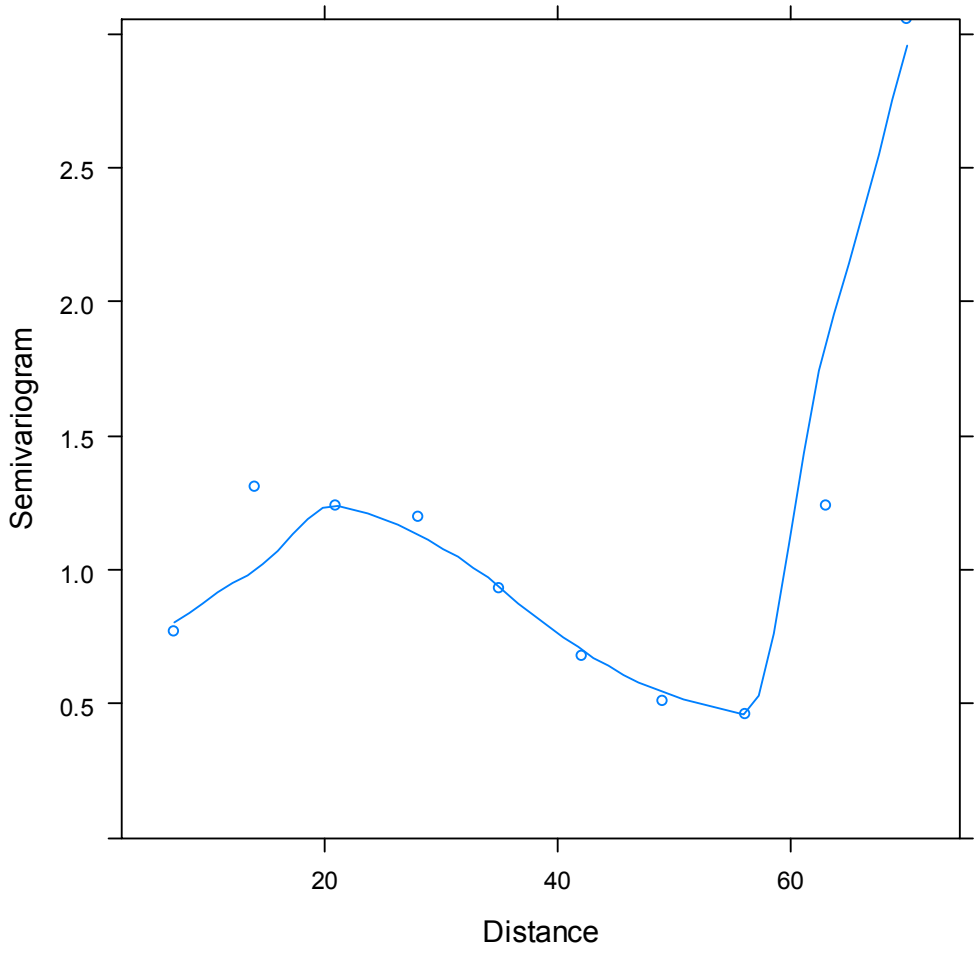


Fig 5.5: Sample semivariogram estimates corresponding to the standardized residuals of the fitted Autoregressive Moving Averages (1,1)

A loess smoother is added to the plot to enhance the visualization of patterns in the semivariogram. The robust semivariogram estimator is used to reduce the influence of an outlying value at distance 1 on the loess smoother. The sample semivariogram estimates in Fig 5.5 appear to vary randomly around y=1 line, suggesting that the normalized residuals are approximately uncorrelated and hence the $ARMA\ (1,1)$ model is adequate model.

Chapter 6

NON-PARAMETRIC APPROACH OF LINEAR REGRESSION MODEL

The classical approach of estimating a regression function is called a parametric regression estimation. In a parametric regression model it is assumed that the structure of the regression function is known and depends only on finitely many parameters, and the data is used to estimate the (unknown) values of these parameters. In linear regression it is assumed that the regression function is a linear combination of the components of $x=(x^{(1)},...,x^{(d)})^T$, *i.e.*,

$$m(x^{(1)},...,x^{(d)})=a_0+\sum_{i=1}^{d}a_i x^{(i)}((x^{(1)},...,x^{(d)})^T\in R^d) \tag{6.1}$$

for some unknown $a_0,...,a_d\in R$. Then data is used to estimate these parameters, e.g., by the principle of least squares method, where the coefficients $a_0,...,a_d$ of the linear function are used such that it best fits the given data:

$$(\hat{a}_0,...,\hat{a}_d)\arg\min_{a_0,...,a_d\in R^d}\left\{\frac{1}{n}\sum_{j=1}^{n}\left|Y_j-a_0-\sum_{i=1}^{d}a_i X_j^{(i)}\right|^2\right\} \tag{6.2}$$

Here $X_j^{(i)}$ denotes the i^{th} component of X_j and $z=\arg\min_{x\in D}f(x)$ is the abbreviation for $z\in D\, f(z)=\min_{x\in D}f(x)$. And finally the estimate is defined by

$$\hat{m}_n(x)=\hat{a}_0+\sum_{i=1}^{d}\hat{a}_i x^{(i)}((x^{(1)},...,x^{(d)})^T\in R^d) \tag{6.3}$$

Parametric estimates usually depend only on a few parameters, thereby making them suitable for small sample sizes n, if the parametric model is appropriately chosen. Furthermore, they can be easily interpreted. In a linear model (when $m(x)$ is a linear function) for instance, the absolute value of the

coefficient $\hat{a}_i$ indicates how much change in the value of i^{th} component of X influences the value of Y , and the sign of $\hat{a}_i$ describes the nature of influence.

However, the parametric estimates have a big drawback. Regardless of the data, a parametric estimate does not approximately fit the regression function better than the best function which has the assumed parametric structure. This inflexibility concerning the structure of the regression function is avoided by so-called nonparametric regression estimates. The nonparametric approach is to choose f from some smooth family of functions. Thus the range of potential fits to the data is much larger than the parametric approach. We do need to make some assumptions about f that it has some degree of smoothness and continuity, but these restrictions are far less limiting than the parametric way.

Smoothing is the term for relatively simple nonparametric regression; when it is apparent that the values of data or the distribution of data are being "ironed out", the procedure is called smoothing. The establishment of the field of smoothing originates with the fact that techniques categorized as smoothing are beneficial data analysis methods when considered in isolation, and the historical circumstance that the evolution of smoothing has become integrated into the development of non-parametric regression, which forms a significant realm within the field of statistical analysis. Smoothing should not, however, be defined as simple nonparametric regression because parametric regression is also used for smoothing. Hence, smoothing is divided into that by parametric regression and that by nonparametric regression in the strict sense.

The relationships among these terms are drawn as a Venn diagram in fig 6.1, which shows that nonparametric regression is involved in both regression equation estimation and probability density function estimation.

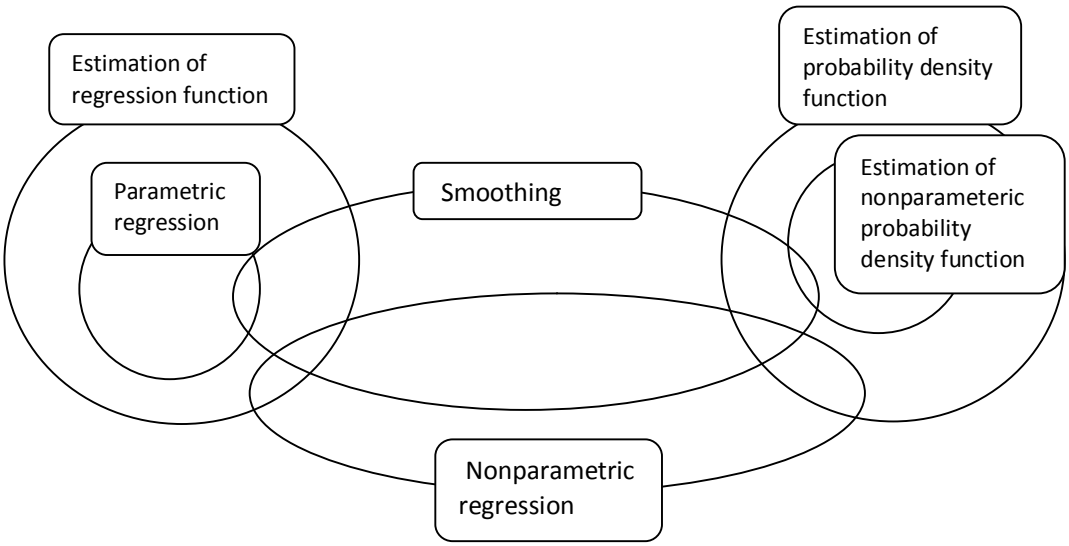


Fig. 6.1: Venn Diagram of the relationship between parametric regression and nonparametric regression

6.1 General formulation of Nonparametric regression model

The nonparametric regression model can be written as

$$Y_i = m(X_i) + \varepsilon_i \tag{6.1.1}$$

Where $\varepsilon_i = Y_i - m(X_i)$ satisfies $E(\varepsilon_i \mid X_i) = 0$. Thus, Y_i can be considered as the sum of the value of the regression function at X_i and some error ε_i, where the expected value of the error is zero. This motivates the construction of the estimates by local averaging, i.e., estimation of $m(x)$ by the average of those Y_i where X_i is "close" to x. Such an estimate can be written as

$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x).Y_i \tag{6.1.2}$$

Where the weights $W_{n,i}(x) = W_{n,i}(x, X_1, ..., X_n) \in \Re$ depend on $X_1, ..., X_n$. Usually the weights are nonnegative and $W_{n,i}(x)$ is "small" if X_i is "far" from x.

An example of such an estimate is the partitioning estimate. Here one chooses a finite or countably infinite partition $\mathrm{P}_n = \{A_{n,1}, A_{n,2}, \ldots\}$ of $\Re^d$ consisting of Borel sets $A_{n,j} \subseteq \Re^d$ and defines, for $x \in A_{n,j}$, the estimate by averaging $Y_i's$ with the corresponding $X_i's$ in $A_{n,j}$, i.e.,

$$m_n(x) = \frac{\sum_{i=1}^{n} I_{\{X_i \in A_n, j\}} Y_i}{\sum_{i=1}^{n} I_{\{X_i \in A_{n,j}\}}} \text{ for } x \in A_{n,j}, \tag{6.1.3}$$

Where I_A denotes the indicator function of set A, so

$$W_{n,i}(x) = \frac{I_{\{X_i \in A_{n,j}\}}}{\sum_{i=1}^{n} I_{\{X_i \in A_{n,j}\}}} \text{ for } x \in A_{n,j}$$

Here and in the following we use the convention $\frac{0}{0} = 0$.

The second example of a local averaging estimate is the Nadaraya-Watson kernel estimate. Let $K : \Re^d \to \Re_+$ be a function called the kernel function, and let $h > 0$ be a bandwidth. The kernel estimate is defined by

$$m_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}, \tag{6.1.4}$$

So

$$W_{n,i}(x) = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)}$$

If one uses the so called naïve kernel (or window kernel) $K(x) = I_{\{\|x\| \le 1\}}$, then

$$m_n(x) = \frac{\sum_{i=1}^{n} I_{\{\|x-X_i\| \le h\}} Y_i}{\sum_{i=1}^{n} I_{\{\|x-X_i\| \le h\}}},$$

i.e., one estimates $m(x)$ by averaging $Y_i's$ such that the distance between X_i and x is not greater than h .

For more general $K : \Re^d \to \Re_+$ one uses a weighted average of the Y_i, where the weight of Y_i (i.e., the influence of Y_i on the value of the estimate at x) depends on the distance between X_i and x (Györfi *et al.*, 2006).

6.2 Estimation and computational method

6.2.1 Kernel estimators

In its simplest form the kernel estimators is just a moving average estimator. The estimate of m called $\hat{m}_h(x)$ is

$$\hat{m}_h(x) = \tfrac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right) Y_j = \tfrac{1}{n} \sum_{j=1}^{n} w_j Y_j \qquad (6.2.1)$$

where , $w_j = K\left(\frac{x - X_j}{h}\right) / h$

K is a kernel where $\int K = 1$. The moving average kernel is rectangular, but smoother kernels can give better results, h is called the bandwidth, window width or smoothing parameter. It controls the smoothness of the fitted curve.

If the $x's$ are spaced very unevenly, then this estimator can give poor results. This problem is somewhat ameliorated by the Nadaraya-Watson estimator

$$m_h(x) = \frac{\sum_{j=1}^{n} w_j Y_j}{\sum_{j=1}^{n} w_j} \qquad (6.2.2)$$

This estimator simply modifies the moving average estimator so that it is a true weighted average where the weights for each Y will sum to one.

It is worth understanding the basic asymptotic of kernel estimators. The optimal choice of h gives:

$$MSE(x) = E(m(x) - \hat{m}_h(x))^2 = O(n^{-4/5})$$

MSE stands for mean squared error and this decreases at a rate proportional to $n^{-4/5}$ with the sample size. Compare this to the typical parametric estimator where $MSE(x) = O(n^{-1})$, but this only holds when the parametric model is correct. So the kernel estimator is less efficient. Indeed, the relative difference between the MSEs becomes substantial as the sample size increases. However, if the parametric model is incorrect, the MSE will be $O(1)$ and the fit will not improve past a certain point even with unlimited data. The advantage of the nonparametric approach is the protection against model specification error. Without assuming much stronger restrictions on m, nonparametric estimators cannot do better than $O(n^{-4/5})$.

The implementation of a kernel estimator requires two choices: the kernel and the smoothing parameter. For the choice of kernel, smoothness and compactness are desirable. We prefer smoothness to ensure that the resulting estimator is smooth. We also prefer a compact kernel because this ensures that only data, local to the point at which m is estimated, is used in the fit. This means that the Gaussian kernel is less desirable, because although it is light in the tails, it is not zero, meaning in principle that the contribution of every point to the fit must be computed. The optimal choice under some standard assumptions is the Epanechnikov kernel.

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2); |x| < 1 \\ 0; otherwise \end{cases} \qquad (6.2.3)$$

This kernel has the advantage of some smoothness, compactness and rapid computation. This latter feature is important for larger datasets, particularly when re-sampling techniques like bootstrap are being used. Even so, any sensible choice of kernel will produce acceptable results, so the choice is not crucially important.

The choice of smoothing parameter h is critical to the performance of the estimator and far more important than the choice of kernel. If the smoothing parameter is too small, the estimator will be too rough; but if it is too large, important features will be smoothed out.

6.2.2 Splines

The model is $Y_i = m(X_i) + \varepsilon_i$, so the spirit of least squares, we might choose $\hat{m}$ to minimize the $MSE = \frac{1}{n}\sum (y_i - m(x_i))^2$. The solution is $\hat{m}(x_i) = y_i$. Suppose we choose $\hat{m}$ to minimize a modified least squares criterion:

$$\frac{1}{n}\sum (Y_i - f(x_i))^2 + h\int [m''(x)]^2 dx \qquad (6.2.4)$$

Where $h > 0$ is the smoothing parameter and $\int [m''(x)]^2 dx$ is a roughness penalty. WhenM is rough, the penalty is large, but when m is smooth, the penalty is small. Thus the two parts of the criterion balance fit against smoothness. This is the smoothing spline fit.

For this choice of roughness penalty, the solution is of a particular form: $\hat{m}$ is a cubic spline. This means that $\hat{m}$ is a piecewise cubic polynomial in each interval $(x_i, x_i + 1)$(assuming that the $x_i's$ are stored). It has the property that $\hat{m}, \hat{m}'$ and $\hat{m}''$are continuous. Given that we know the form of the solution, the estimation is reduced to the parametric problem of estimating the coefficients of the polynomials. This can be done in a numerically efficient way.

Other choices of roughness penalty can be considered, where penalties on higher-order derivatives lead to fitswith more continuous derivatives. We can

also use weights by inserting them in the sum of squares part of the criterion. This feature is useful when smoothing splines are means to an end for some larger procedure that requires weighing. A robust version can be developed by modifying the sum of squares criterion to:

$$\sum \rho(y_i = f(x_i)) + h\int [m''(x)]^2 dx \tag{6.2.5}$$

Where $\rho(x) = |x|$ is one possible choice.

The other class of splines is the regression splines. Regression splines differ from smoothing splines in the following way: For regression splines, the knots of the B-splines used for the basis are typically much smaller in number than the sample size. The number of knots chosen controls the amount of smoothing. For smoothing splines, the observed unique x values are the knots and h is used to control the smoothing. It is arguable whether the regression spline method is parametric or nonparametric, because once the knots are chosen, a parametric family has been specified with a finite number of parameters. It is the freedom to choose the number of knots that make the method nonparametric. One of the desirable characteristics of a nonparametric regression estimator is that it should be consistent for smooth functions. This can be achieved for regression splines if the number of knots is allowed to increase at an appropriate rate with the sample size.

6.2.3 Lowess

The Lowess procedure implements a nonparametric method for estimating regression surfaces pioneered by Cleveland, Devlin and Grosse (Cleveland, Devlin and Grosse 1988), Cleveland and Grosse (Cleveland and Grosse 1991), and Cleveland, Grosse and Shyu (Cleveland, Grosse and Shyu 1992). The Lowess procedure allows great flexibility because no assumptions about the parametric form of the regression surface are needed. The Lowess procedure is suitable when there are outliers in the data and a robust fitting method is necessary. Lowess is an acronym for locally weighted scatterplot smoother.

Assume that for $i = 1,..., n$, the i^{th} measurement y_i of the response y and the corresponding measurement x_i of the vector x of p predictorsare related by

$$y_i = g(x_i) + \varepsilon_i \tag{6.2.6}$$

Where g is the regression function and ε_i is a random error. The idea of local regression is that near $x = x_0$, the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class. Such a local approximation is obtained by fitting a regression surface to the data points within a chosen neighborhood of the point x_0.

In the Lowess method, weighted least squares is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the smoothing parameter, in each local neighborhood controls the smoothness of the estimated surface. Data points in a given local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood. The computational method is explained as:

Step 1: Defining the window width

The first step is to define the window width m, that encloses the closest neighbours to each data observation (the window half-width is labelled *h)*. We call this focal X.

Step 2: Weighting the data

We then choose a weight function to give greatest weight to observations that are closest to the focal X observation. In practice, the tricube weight function is usually used.Let $Z_i = (X_i - X_0)/h$, which is the scaled distance between the predictor value for the i^{th} observation and the focal X .

$$W_T(z) = \begin{cases} (1-|z|^3)^3 \; for \; |z|<1 \\ 0 \; for \; |z| \geq 1 \end{cases}$$

Here h_i is the half-width of the window centered on X_i.Notice that observations more than h (the half-window or bandwidth of the local regression) away from the focal X receive a weight of 0.

Step 3: Locally weighted least squares

A polynomial regression using weighted least squares (using the tricube weights) is then applied to the focal X observation, using only the nearestneighbour observations to minimize the weighted residual sum of squares. Typically a local linear regression or a local quadratic regression is used, but higher order polynomials are also possible.

$$Y_i = A + B_1(x_i - x_0) + B_2(x_i - x_0)^2 + ... + B_p(x_i - x_0)^p + \varepsilon_i$$

From this regression, we then calculate the fitted value for the focal X value and plot it on the scatterplot.

Step 4: The Nonparametric Curve

Steps 1-3 are carried out for each observation in the data. There is a separate local regression for each value of X .

A fitted value from these regressions for each focal X is plotted on the scatterplot.The fitted values are connected, producing the local polynomial nonparametric regression curve.

The growth rates of agricultural crops are mostly estimated by the linear regression models. However, it might be the case that these models may not fit the data well. Under such conditions it becomes essential to apply nonparametric regression, which is based on fewer assumptions. In last few years, nonparametric regression technique for functional estimation has become increasingly popular as a tool for data analysis.To study the nonparametric regression of trends and growth rates, long term data for last 42 years pertaining to the area, production

and productivity of apple is taken. To find out the path of the production process different parametric trend models are also fitted.Among the fitted models, the best model is selected on the basis of their goodness of fit (R^2) value and significance of the coefficients. Further, the values of Finite Sample Corrected AIC (AIC_C), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Max Absolute Percentage Error (MaxAPE), Max Absolute Error (MaxAE) are calculated for the area, production and productivity trends under nonparametric regression and parametric regression approaches.

- Finite Sample Corrected AIC

$$AIC_C(p,q) = -2\log[likelihood\,(p,q)] + 2(p+q+1)\frac{n}{n-p-q-2}$$

- Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(est_i - act_i)^2}{n}}$$

- Mean Absolute Percentage Error

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{est_i - act_i}{act_i}\right|}{n} \times 100$$

- Mean Absolute Error

$$MAE = \sum_{i=1}^{n}\left|\frac{(est_i - act_i)}{n}\right|$$

- MaxAPE

$$MaxAPE = Max\left|\frac{est_i - act_i}{act_i}\right| \times 100$$

- MaxAE

$$MaxAE = Max\sum_{i=1}^{n}\left|est_i - act_i\right|$$

Then on the basis of these values the parametric and nonparametric regression approaches are compared and the approach with minimum values of

AIC_c, RMSE, MAPE, MAE, MaxAPE and MaxAE is selected as the best fit for the long term trends.

6.3 Numerical illustration

The parametric models fitted here are the quadratic model or second degree polynomial model $Y_t = b_0 + b_1 t + b_2 t^2$ and the cubic model or the third degree polynomial model $Y_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$. The dependent variable Y is area, production and productivity and independent variable X is the time points (years). The fitted models are shown in the table 6.1.

Table 6.1: Trends in area, production and productivity of apple in Jammu and Kashmir

	R^2	Constant b_0	b_1	b_2	b_3	RMSE	MAPE	MAE	MaxAPE	MaxAE
Area	0.92	345.9	3.092	-0.024	0.041	11.09	2.89	4.09	5.83	32.56
Production	0.91	37.32	-1.980	0.436	0.002	17.45	4.90	19.45	6.21	51.34
Productivity	0.90	645.33	14.98	-0.432		76.90	6.34	34.78	17.98	91.34

Area in '000 hectares, Production in '000 metric tons, Productivity in metric ton per hectare

The value of b_2 for area is negative which indicates that area under apple cultivation decreased in the middle part of the cultivation period and the value of b_1 and b_2 being positive clearly indicates that there was an increase in the cultivation area. Further, the negative value of b_1 for production is an indication of the decrease in the production during the initial period of the study and the positive values of b_2 & b_3 indicates an increase in the production.

The predicted values for Area, production and Productivity obtained by nonparametric regression models and are shown in the Table 6.2

Table 6.2: Predicted values of Area, production and Productivity of Apple in Jammu and Kashmir

Lowess	Spline

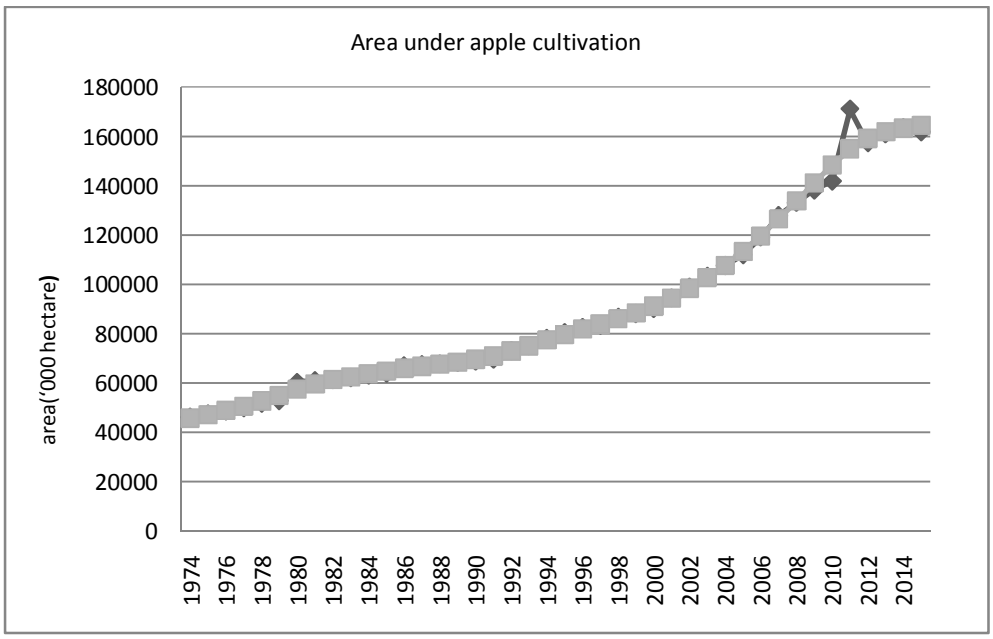
Area ('000hec)	Production ('000MT)	Productivity (MT per hec)	Area ('000hec)	Production ('000MT)	Productivity (MT per hec)
45.568	308.116	6.33712	45.68266	287.2517	6.273
47.197	334.191	6.64209	47.16725	318.7057	6.63
48.901	360.266	6.94705	48.74917	349.9367	6.9818
50.68	385.505	7.23333	50.54234	380.7536	7.3242
52.56	410.744	7.51962	52.63166	410.9015	7.6521
55.041	435.982	7.8059	55.00611	440.0733	7.9601
57.404	460.421	8.07419	57.45991	468.0222	8.2448
59.559	484.859	8.34249	59.56775	494.5811	8.5048
61.251	509.297	8.61079	61.21571	519.7028	8.7401
62.356	533.735	8.84956	62.53253	543.3313	8.9495
63.453	558.173	9.08833	63.68769	565.2982	9.1294
64.679	580.164	9.19719	64.7942	585.5523	9.2772
65.842	602.156	9.30606	65.85267	604.2626	9.394
66.888	624.147	9.41492	66.76057	621.9593	9.4861
67.737	646.139	9.48595	67.54911	639.358	9.5623
68.459	668.13	9.55697	68.37756	656.9153	9.6264
69.48	688.002	9.58398	69.4504	675.0344	9.6811
70.968	707.874	9.61099	70.93599	694.0998	9.7289
72.921	727.745	9.638	72.86528	714.4357	9.7717
75.258	747.617	9.65621	75.07868	736.2788	9.8106
77.689	767.488	9.67441	77.39878	759.9386	9.8477
79.946	795.199	9.73411	79.66308	785.6379	9.884
82.121	822.91	9.79381	81.80116	813.5134	9.9191
84.187	850.62	9.85352	83.86907	843.7533	9.9529
86.246	888.566	9.90056	86.02191	876.6189	9.9863
88.664	926.512	9.94761	88.39281	912.3287	10.0198
91.609	964.458	9.99466	91.16757	951.1022	10.0545
95.012	1012.729	10.04879	94.48858	992.9133	10.0894
98.902	1061	10.10293	98.35757	1037.592	10.1228
103.277	1109.271	10.15707	102.7509	1084.883	10.1533
108.226	1157.542	10.1852	107.7095	1134.453	10.1796
113.837	1205.813	10.21333	113.3316	1185.896	10.2008
119.873	1255.705	10.19789	119.6613	1238.73	10.2158
126.023	1305.596	10.18245	126.542	1292.397	10.2241
132.965	1355.488	10.16701	133.7363	1346.308	10.2251
141.107	1405.38	10.1621	141.1204	1399.893	10.2184
148.241	1455.272	10.1572	148.443	1452.555	10.2028
154.045	1506.627	10.14936	154.8748	1503.812	10.1782
157.469	1557.982	10.14153	159.171	1553.919	10.149
160.438	1609.337	10.13369	161.7802	1603.617	10.1204
163.016	1660.692	10.11562	163.3824	1653.437	10.095
165.222	1712.048	10.09755	164.4919	1703.698	10.0737

The fitted nonparametric models for Area under cultivation is shown in Table 6.3.

Table 6.3: Trends in area of Apple in Jammu and Kashmir using non-parametric regression

	Loess	Splines
Bandwidth	0.23	0.74
R^2	0.987	0.996
AIC_c	1.77	1.02
RMSE	3.03	2.21
MAPE	1.33	1.22
MAE	1.35	1.14
MaxAPE	9.91	8.43
MaxAE	16.95	14.13

Trend analysis of area using nonparametric regression is presented in the table6.3. In Table 6.3 the value of R^2 is 0.987 for Loess. The AIC_c, RMSE, MAPE, MAE, MaxAPE and MaxAE values comes out be small for nonparametric regression for the area under apple cultivation in Jammu and Kashmir.According to the state's horticulture department, around 1.5 million tonnes of apples areproduced in Kashmir annually. The production of apples in the state is growing every year as a result the percentageshare of Jammu & Kashmir in national production has also been increasing steadily; it has increased from, 63.5% inFY2006 to 77.2% in FY2010. The apple production in the year 2004-05 was 10933.33 MT and in year it reached to1852.41 in the year 2010-11 (Sheikh and Tripathi, 2013). The area under the apple cultivation has increased over the years of study and is shown in Fig. 6.2.



obs_area=observed area, pred_area=predicted area

Fig. 6.2: Observed and expected trends of area under apple cultivation using spline in Jammu and Kashmir

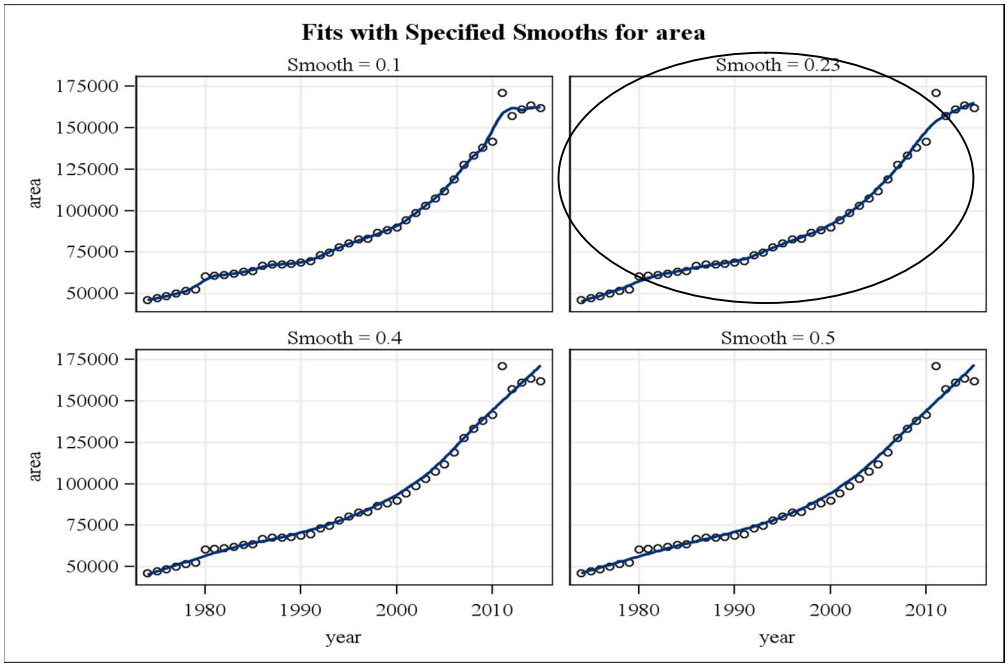


Fig. 6.3: Fits with specified smooths for area of apple production

The values of area are initially fitted at the smoothing parameters in order to obtain the best fit of the data points we obtain the graph of the data points in the neighborhood of the smoothing parameters and look for the curve which covers all the points of the data. The one which covers maximum points is the best fit of the data points. In fig 6.3 the smooth curve fits are obtained for area in the neighborhood of smoothing parameters i.e., at 0.10, 0.23, 0.40 and 0.50. It is observed that the best fit is obtained at smooth=0.23.

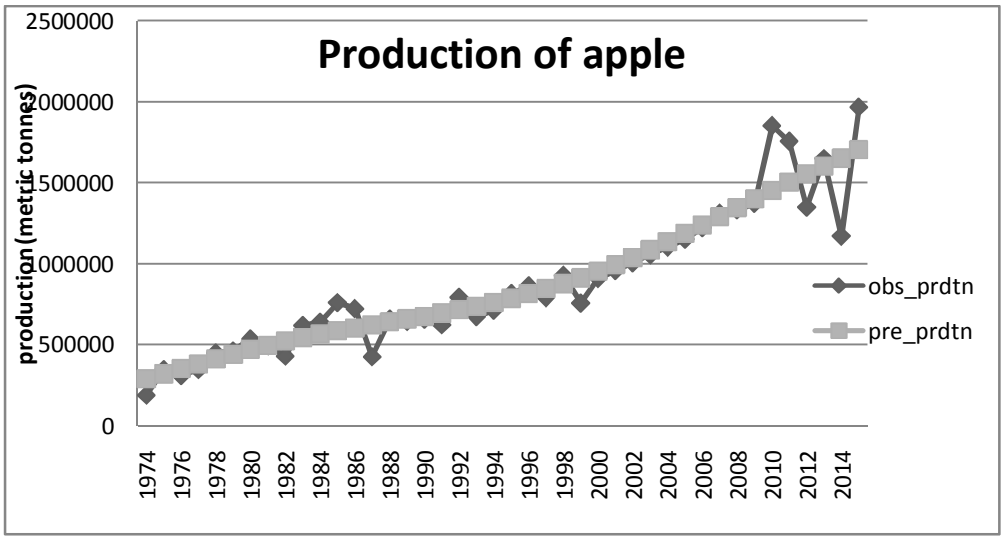
The fitted nonparametric regression models for production are shown in Table 6.4.

Table 6.4: Trends in production of Apple in Jammu and Kashmir using non-parametric regression

	Loess	**Splines**
Bandwidth	0.66	0.165
R^2	0.985	0.998
AIC_c	0.25	0.10
RMSE	1.35	1.04
MAPE	11.73	10.89
MAE	9.15	8.63
MaxAPE	6.18	5.08
MaxAE	4.80	4.03

In Table 6.4 the values for AIC_c,RMSE, MAPE, MAE, MaxAPE and MaxAE for production of apple in Jammu and Kashmir is minimum obtained by using the nonparametric regression models as compared to the fitted parametric model as shown in the Table 6.1.

According to the state's horticulture department, around 1.5 million tonnes of apples areproduced in Kashmir annually. The production of apples in the state is growing every year as a result the percentageshare of Jammu & Kashmir in national production has also been increasing steadily; it has increased from, 63.5% inFY2006 to 77.2% in FY2010. The apple production in the year 2004-05 was 10933.33 MT and in year it reached to1852.41 in the year 2010-11 (Sheikh and Tripathi,2013). The increasing trend in the production over the years of study is shown in the Fig. 6.4.



obs_prdtn=observed production, pre_prdtn=predicted production

Fig. 6.4: Observed and expected trends of production of apple using splines

It can be observed that upto 2014-15 there is sharp increase in production and productivity. However, a decline in production and productivity can also be observed during the year 2015-16 is observed which is due to the floods that occurred during the said year (Islam and Shrivastava, 2017).

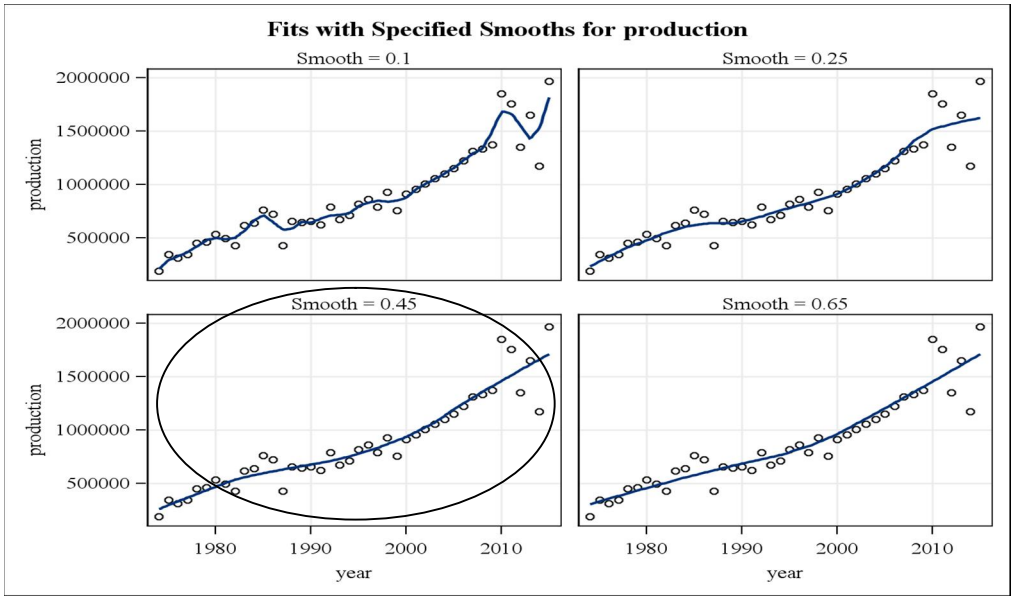


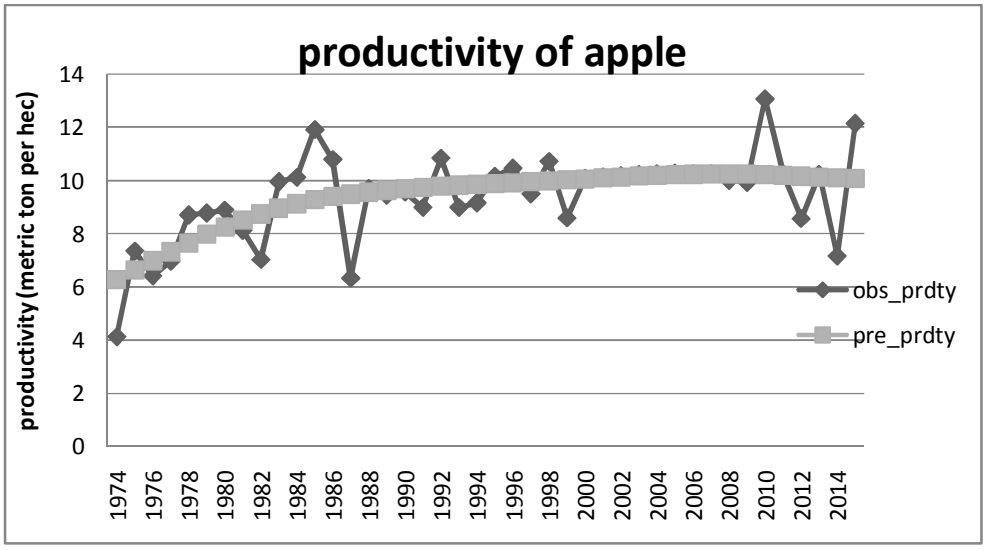
Fig. 6.5: Fits with specified smooths for production of apple

In Fig. 6.5 smooth fits for production are plotted in the neighborhood of the smoothing parameter at 0.10, 0.25, 0.45 and 0.65 and it is observed that the best fit obtained for smooth=0.45.

Table 6.5: Trends in productivity of Apple in Jammu and Kashmir using non-parametric regression

	Loess	**Splines**
Bandwidth	0.53	0.49
R^2	0.988	0.997
AIC_c	1.69	1.12
RMSE	1.22	0.99
MAPE	10.12	9.11
MAE	2.84	0.84
MaxAPE	5.37	2.13
MaxAE	6.08	3.16

Even values of RMSE, MAE, MAPE, MaxAE and MaxAPE for productivity of Jammu and Kashmir for non-parametric regression has observed lower values than the parametric regression. This is clear indication of the superiority of these techniques over the parametric models.These models perform very well in visualizing the past trends where the parametric models fails to. The increasing trend in the productivity over the years of study is shown in the fig 6.6.



obs_prdty=observed productivity, pre_prdty=predicted productivity

Fig. 6.6: Observed and expected trends of productivity of apple using spline

Fig. 6.6 provides the fits for productivity in the neighborhood of the smoothing parameters i.e., at smooths equal to 0.15, 0.25, 0.40 and 0.54. The best fit is observed to be at the smooth=0.54.

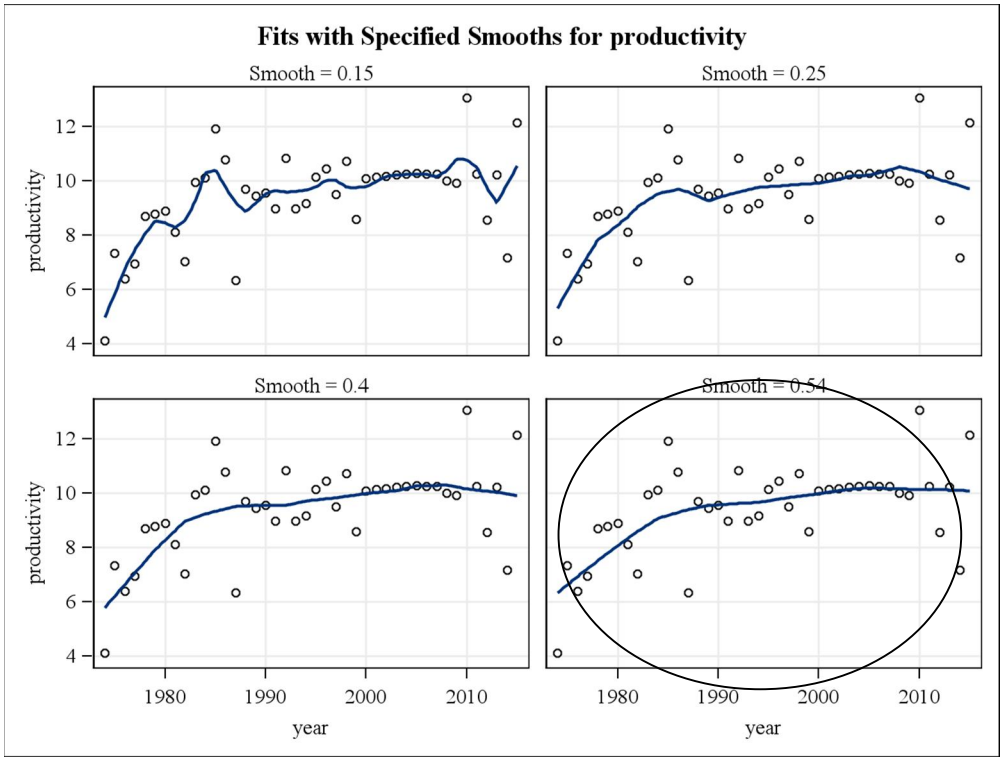


Fig. 6.7: Fits with specified smooths for productivity of apple

It is observed that there is dramatic increase in the area under apple cultivation and in the production as well as productivity. In order to maintain the trend more and more land is to be brought under the apple cultivation. Parametric regression usually utilized in studying the trend seems not to perform better than the nonparametric regression. And thus nonparametric regression is the best fit for the trend analysis of the apple production of Jammu and Kashmir.

Chapter-7

SUMMARY AND CONCLUSION

Often there are situations in which the distribution of the agricultural data is not normal and comes from an exponential family of distribution or it is sometimes the case that the distribution is unknown. The data sets are corrected for normality and are analysed by the classical linear regression model approaches or by applying any of the designs from the design of experiments. There are approaches or techniques available which are seldom or are not at all used in the agricultural setups. We have worked out some of the techniques for such datasets.

The whole thesis is divided into seven chapters. The chapter-1 gives the overview of the exponential family of distributions, generalized linear models, generalized linear mixed models, link functions associated with them, nonlinear mixed models, heteroscedasticity and within group correlation in the nonlinear mixed effects modeland nonparametric regression. This chapter also provides the codes of the functions developed in SAS/R software that were used in the study for the ease of implementation, brief resume of the literature is also provided in this chapter.

The Chapter-2 includes the preliminary summary of the data used in the overall study. This chapter further includes the numerical and the graphical summary of the data used. This chapter gives an idea of the distribution of the data which has been used in the overall thesis. This chapter also includes the codes of the functions that have been developed.

The Chapter-3 contains the overall information of the generalized linear models and the deviance approach of model selection.AIC selection criterion is used to check the distribution of a real floricultural data set. Further the deviance of the data is also obtained for the dataset. The generalized linear model is fitted on the non-normal horticultural data whose response follows an exponential family of distribution. In particular response which follows Gamma and Inverse

Gaussian distributions are studied.Further the fitted generalized linear models are compared with the linear regression (with the response following the normal distribution). It is concluded that the generalized linear models fits the horticultural data best when the data comes from an exponential family of distributions.

In Chapter-4 the generalized linear mixed model is fitted by four different estimation methods viz. Maximum Likelihood Method, Penalized Quasi Likelihood, Laplace Approximation and LASSO Method. This is shown by the numerical illustration of the horticultural data. The four estimation methods are compared by average relative bias, average squared relative bias, average absolute bias and average squared deviation. And the estimation method with minimum values is preferred.It has been found that average relative bias, average squared relative bias, average absolute bias and average squared deviation is considered to be the best fit for the generalized linear mixed models.It was observed that the LASSO method of estimation had minimum values for average relative bias, average squared relative bias, average absolute bias and average squared deviation. Thus, the LASSO method of estimation is considered to be better fit than the other estimation methods for fitting the generalized linear mixed models.

In Chapter-5 the nonlinear mixed effects model is extended to include heteroscedastic, correlated within group errors. The estimation and the computational method can be extended when the heteroscedastic and correlated within group errors situation exists in the nonlinear mixed effects models. The chapter includes several classes of variance functions to characterise heteroscedasticity and several classes of correlation structures to represent serial and spatial correlation are introduced and it is shown that how the latter two can be combined to flexibly model the within-group variance-covariance structure. The nlme() function R software is used to fit the extended nonlinear mixed effects model and describe a suite of S classes and methods to implement variance functions (varFunc) and correlation structures (corStruct). Any of these classes or

others defined by users can be used in nlme to fit the extended nonlinear mixed effects model.

In the Chapter-6 of the thesis nonparametric regression approach is introduced to be used for the long term trend analysis.The nonparametric regression approach is applied on the long term horticultural data set. The data is fitted by the parametric as well as the nonparametric regression approaches. And both the approaches are compared on the basis of Finite Sample Corrected AIC (AIC_C), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Max Absolute Percentage Error (MaxAPE), Max Absolute Error (MaxAE). The regression approach with minimum values for Finite Sample Corrected AIC (AIC_C), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Max Absolute Percentage Error (MaxAPE), Max Absolute Error (MaxAE) is selected as the best fit for the long term trend analysis. It is found that the nonparametric approach is the best fit for the long term trend analysis or when some past information is provided.

LITERATURE CITED

Adjakossa, E. and Nuel, G. 2017. Fixed effects selection in the linear mixed-effects model using adaptive ridge procedure for l0 penalty performance.*arXiv preprint arXiv* **1705**: 01308.

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician***46**: 175-185.

Archontoulis, S. V. and Miguez, F. E. 2015, Nonlinear regression models and applications in agricultural research. *Agronomy Journal***107**: 786-798.

Bates, D. M. and Watts, D. G. 1980. Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society*. Series B (Methodological) pp. 1-25.

Bates, D. and Watts, D. 1988. Nonlinear regression analysis and its applications Wiley, New York.

Bates, D. M. W., Bates, D. G. D. M. and Watts, D. G. 1988. Nonlinear Regression Analysis and Lts Applications,

Beitler, P. J. and Landis, J. R. 1985. A mixed-effects model for categorical data. *Biometrics* pp. 991-1000.

Booth, J. G. and Hobert, J. P. 1998. Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association***93**: 262-272.

Booth, J. G. and Hobert, J. P. 1999. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society:* Series B (Statistical Methodology)**61**: 265-285.

Breiman, L. 1996. Heuristics of instability and stabilization in model selection. *The Annals of Statistics***24**: 2350-2383.

Breiman, L. 1998. Arcing Classifier (with Discussion and a Rejoinder by the Author). *The Annals of Statistics***26**: 801-849.

Breslow, N. E. and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association***88**: 9-25.

Breslow, N. E. and Lin, X. 1995. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika***82**: 81-91.

Bühlmann, P. and Yu, B. 2003. Boosting with the L 2 Loss: Regression and classification. *Journal of the American Statistical Association***98**: 324-339.

Cai, Z. and Tsai, C. L. 1999. Diagnostics for nonlinearity in generalized linear models. *Computational Statistics and Data Analysis***29**: 445-469.

Capanu, M., Gönen, M. and Begg, C. B. 2013. An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in Medicine***32**: 4550-4566.

Charytanowicz, M., Czachor, H., Dobrzański, B. and Niewczas, J. 2015. Nonparametric regression approach: applications in agricultural science. *Czasopismo Techniczne*.

Chen, J., Zhang, D. and Davidian, M. 2002. A monte carlo em algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics***3**: 347-360.

Cleveland, W. S., Devlin, S. J. and Grosse, E. 1988. Regression by local fitting: methods, properties and computational algorithms. *Journal of Econometrics***37**: 87-114.

Cleveland, W. S. and Grosse, E. 1991. Computational methods for local regression. *Statistics and Computing***1**: 47-62.

Cleveland, W., Grosse, E. and Shyu, W. 1992. Statistical models in S, chapter chapter 8: local regression models. *Wadsworth and Brooks, Cole*.

Cressie, N. and Hawkins, D. M. 1980. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology***12**: 115-125.

Cressie, N. 1993. Statistics for spatial data (Revised Ed) Wiley. New York.

Davidian, M. and Gallant, A. R. 1992. Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to Quinidine. *Journal of Pharmacokinetics and Biopharmaceutics***20**: 529-556.

Davidian, M. and Giltinan, D. 1995. Nonlinear Models for Repeated Measurement Data Chapman and Hall London Google Scholar.

Decker, L. 2012. A Glm-based approach to adjusting for changes in case reserve adequacy. In Casualty Actuarial Society E-Forum, Summer.

Dey, D. K., Ghosh, S. K. and Mallick, B. K. 2000. Generalized Linear Models: A Bayesian Perspective, CRC Press.

Diggle, P., Liang, K. and Zeger, S. 1994. Analysis of Longitudinal Data: Oxford Statistical Science Series.

Fahrmeir, L. and Lang, S. 2001. Bayesian inference for generalized additive mixed models based on markov random field priors, *Journal of the Royal Statistical Society:* Series C (Applied Statistics)**50**: 201-220.

Fahrmeir, L. and Tutz, G. 2013. Multivariate statistical modelling based on generalized linear models, Springer Science & Business Media.

Fisher, R.A (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.

Fisher, R. A. 1934. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London.* Series A**144**: 285-307.

Fox, J. 2005. Introduction to nonparametric regression. Lecture Notes. http: //socserv. mcmaster. ca/jfox/Courses/Oxford.

Freund, Y. and Schapire, R. E. 1996. Experiments with a new boosting algorithm. In *Icml*, Bari, Italy pp. 148-156.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* pp. 1189-1232.

Genkin, A., Lewis, D. D. and Madigan, D. 2007. Large-scale bayesian logistic regression for text categorization. *Technometrics***49**: 291-304.

Gill, J. 2000. Generalized Linear Models: A Unified Approach (Vol. 134), Sage Publications.

Gilmour, A., Anderson, R. and Rae, A. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika***72**: 593-599.

Goeman, J. J. 2010. L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal***52**: 70-84.

Green, P. J. 1984. Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *Journal of the Royal Statistical Society.* Series B (Methodological) pp. 149-192.

Groll, A. and Tutz, G. 2014. Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing***24**: 137-154.

Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. 2006. A distribution-free theory of nonparametric regression, Springer Science & Business Media.

Härdle, W. and Linton, O. 1994. Applied nonparametric methods. *Handbook of Econometrics***4**: 2295-2339.

Harville, D. A. and Mee, R. W. 1984. A mixed-model procedure for analyzing ordered categorical data. *Biometrics* pp. 393-408.

Hastie, T. J. and Tibshirani, R. J. 1990. Generalized additive models, Volume 43 of monographs on statistics and applied probability.

Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. (1959), "The Estimation of Environmental and Genetic Trends from Records Subject to Culling," *Biometrics*, 15, 192-218.

Hu, K., Choi, J., Sim, A. and Jiang, J. 2015. Best predictive generalized linear mixed model with predictive lasso for high-speed network data analysis. *International Journal of Statistics and Probability***4**: 132.

Islam, R. T. and Shrivastava, S. 2017. A study on area, production and productivity of apples in J&K from 2006-07 to 2015-16. *International Journal of Scientific Research and Management***5**: 6513-6519.

Jiao, Y. and Chen, Y. 2004. An application of generalized linear models in production model and sequential population analysis. *Fisheries Research***70**: 367-376.

Jørgensen, B. 1983. Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika***70**: 19-28.

Khan, A.A. and Mir, A.H. 2005. Applications of R-software in agricultural data analysis . *SKUAST Journal of Research* 7(1): 36-64.

Kim, Y. and Kim, J. 2004. Gradient lasso for feature selection. **In**: *Proceedings of the twenty-first international conference on Machine learning*, ACM pp. 60.

Laird, N. M. and Ware, J. H. 1982. Random effects models for longitudinal data. *Biometrics* pp. 963-974.

Lane, P. W. and Nelder, J. A. 1982. Analysis of covariance and standardization as instances of prediction. *Biometrics* pp. 613-621.

Lee, Y. and Nelder, J. A. 1996. Hierarchical generalized linear models. *Journal of the Royal Statistical Society.* Series B (Methodological) pp. 619-678.

Lele, S. R., Nadeem, K. and Schmuland, B. 2012. Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*.

Leonard, T., Hsu, J. S. and Tsui, K. W. 1989. Bayesian marginal inference. *Journal of the American Statistical Association***84**: 1051-1058.

Liang, K. Y. and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika***73**: 13-22.

Lindsey, J. 1974. Construction and comparison of statistical models. *Journal of the Royal Statistical Society.* Series B (Methodological) pp. 418-425.

Lindstrom, M. J. and Bates, D. M. 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics* pp. 673-687.

Lin, X. and Breslow, N. E. 1996. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association***91**(435): 1007-1016.

Livanis, G. T., Salois, M. and Moss, C. 2009. A nonparametric kernel representation of the agricultural production function: Implications for economic measures of technology. **In**: *83rd Annual Conference of the Agricultural Economics Society,* Dublin pp. 22.

Matheron, G. 1962. Traité De Géostatistique Appliquée, Tome I: Mémoires Du Bureau De Recherches Géologiques Et Minières. *Editions Technip, Paris* pp. 14.

McCullagh, P. 1983. Quasi likelihood functions. *The Annals of Statistics* pp. 59-67.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models (Vol. 37), CRC Press.

McCulloch, C. E. 1997. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association***92**: 162-170.

McCullogh,C.E., Searle, S.R. 2008. Generalized, Linear and Mixed Models. 2nd Edition John Wiley & Sons,Canada.

McLean, R. A., Sanders, W. L. and Stroup, W. W. 1991. A unified approach to mixed linear models. *The American Statistician***45**: 54-64.

Nelder, J.A. and Wedderburn, R.W.M. 1972. Generalized linear models. *Journal of Royal Statistical Society Series* A pp. 370-384.

Ngo, T.H.D. and Puente, L., C.A. 2016. Generalized linear models for non normal data. *Biometrika*, paper 8380.

Pan, J. and Thompson, R. 2003. Gauss hermite quadrature approximation for estimation in generalised linear mixed models. *Computational Statistics***18**: 57-78.

Pierce, D. A. and Schafer, D. W. 1986. Residuals in generalized linear models. *Journal of the American Statistical Association***81**: 977-986.

Pinheiro JCBates, D. (2000), "Mixed Effects Models in S and S-Plus: Springer."

Pitt, D. 2003. A new mathematical model of Australian disability experience.

Pregibon, D. 1981. Logistic regression diagnostics. *The Annals of Statistics* pp. 705-724.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Statistics Journal***2**: 1-21.

Radhakrishna-Rao, C. and Toutenburg, H. 1999. Linear models: least squares and alternatives, Springer.

Schabenberger, O. and Gregoire, T. 1996. Population averaged and subject specific approaches for clustered categorical data. *Journal of Statistical Computation and Simulation***54**: 231-253.

Schall, R. 1991. Estimation in generalized linear models with random effects. *Biometrika***78**: 719-727.

Schelldorfer, J., Meier, L. and Bühlmann, P. 2014. Glmmlasso: An algorithm for

high-dimensional generalized linear mixed models using L1-Penalization. *Journal of Computational and Graphical Statistics***23**: 460-477.

Self, S. G. and Liang, K. Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association***82**: 605-610.

Sheikh, S. H. and Tripathi, A. 2013. Socio-economic conditions of apple growers of Kashmir valley: A case study of district Anantnag. *International Journal of Educational Research and Technology***4**: 30-39.

Sheiner, L. B. and Beal, S. L. 1980. Evaluation of methods for estimating population pharmacokinetic parameters. I. michaelis menten model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics***8**: 553-571.

Shenoy, S., Gorinevsky, D. and Boyd, S. 2015. Non parametric regression modeling for stochastic optimization of power grid load forecast. **In**: *American Control Conference (ACC)* IEEE pp. 1010-1015.

Shevade, S. K. and Keerthi, S. S. 2003. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics***19**: 2246-2253.

Sinha, S. K. 2004. Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association***99**: 451-460.

Stigler, S. M. 1981. Gauss and the invention of least squares. *The Annals of Statistics* pp. 465-474.

Stroup, W. W. 2012. Generalized linear mixed models: modern concepts, methods and applications, CRC Press.

Stroup, W. W. and Kachman, S. D. 1994. Generalized linear mixed models-an overview.

Tang, Z., Shen, Y., Zhang, X. and Yi, N. 2016. The spike and slab lasso generalized linear models for prediction and associated genes detection. *Genetics***116**: 192-195.

Thisted, R. A. (1988) *Elements of Statistical Computing. Numerical Computation*. NewYork: Chapman & Hall.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* Series B (Methodological) pp. 267-288.

Tierney, L. and Kadane, J. B. 1986. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association***81**: 82-86.

Tutz, G. and Reithinger, F. 2007. A boosting approach to flexible semiparametric mixed models. *Statistics in Medicine***26**: 2872-2900.

Tutz, G. and Groll, A. 2010. Generalized linear mixed models based on boosting. **In**: *Statistical Modelling and Regression Structures*, Springer pp. 197-215.

Verbeke, G and Molenberghs, G. 2007. Likelihood ratio, score and wald tests in a constrained parameter space. *The American Statistician***61**: 22-27.

Vieu, P. 1994. Choice of regressors in nonparametric estimation, Computational *Statistics and Data Analysis***17**: 575-594.

Vonesh, E. F. and Carter, R. L. 1992. Mixed effects nonlinear regression for unbalanced repeated measures. *Biometrics* pp. 1-17.

Williams, D. 1987. Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* pp. 181-191.

Wolfinger, R. and O'connell, M. 1993. Generalized linear mixed models a pseudo likelihood approach. *Journal of statistical Computation and Simulation***48**: 233-243.

Xu, H. 2014. Nonlinear mixed effects (Nlme) diameter growth models for individual china-fir (*Cunninghamia lanceolata*) trees in southeast China. *PloS One* 9: e104012.

Zeger, S. L. and Karim, M. R. 1991. Generalized linear models with random zeffects; A gibbs sampling approach. *Journal of the American Statistical Association***86**: 79-86.

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Faculty of Horticulture, Division of Agricultural Statistics,
Shalimar Campus, Srinagar – 190 025
-::o::-

CERTIFICATE

Certified that all the corrections/amendments as suggested by External Examiner Prof. Inderjeet Singh Grewal during Viva-Voce examination held on 20-09-2018 have been incorporated in the manuscript entitled **"Study on Varied Aspects of Linear/Nonlinear Models and Their Application in Agriculture"** submittedby**Ms.Yasmeena Ismail(Regd. No. 2015-614-D).**

(Prof.S. A. Mir)
Chairman
Advisory Committee