2019

Analysis of Longitudinal Data and Model Selection

Md Abdulla Al Mamun
University of Windsor

Analysis of Longitudinal Data and Model Selection

by

Md Abdulla Al Mamun

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada

Analysis of Longitudinal Data and Model Selection

by

Md Abdulla Al Mamun

APPROVED BY:

S. B. Provost, External Examiner
Western University

Y. Aneja
Odette School of Business

M. Hlynka
Department of Mathematics and Statistics

M. Belalia
Department of Mathematics and Statistics

S. R. Paul, Advisor
Department of Mathematics and Statistics

May 17, 2019

Declaration of Originality

I hereby certify that I am the sole author of this dissertation and that no part of this dissertation has been published or submitted for publication.

I certify that, to the best of my knowledge, my dissertation does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my dissertation and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies Office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

Abstract

An important issue in regression analysis of longitudinal data is model parsimony, that is, finding a model with as few regression variables as possible while retaining good properties of the parameter estimates. In this vein, joint modelling of mean and variance taking into account the intra subject correlation has been standard in recent literature (Pourahmadi, 1999, 2000; Ye and Pan, 2006; and Leng, Zhang, and Pan, 2010). Zhang, Leng, and Tang (2015) propose joint parametric modelling of the means, variances and correlations by decomposing the correlation matrix via hyperspherical co-ordinates and show that this results in unconstrained parameterization, fast computation, easy interpretation of the parameters, and model parsimony. We investigate the properties of the estimates of the regression parameters through semiparametric modelling of the means and variances and study the impact of this to model parsimony. An extensive simulation study is conducted. Three datasets, namely, a biomedical dataset, an environmental dataset and a cattle dataset are analysed.

In longitudinal studies, researchers frequently encounter covariates that are varying over time (see for example Huang, Wu, and Zhou, 2002). We consider a generalized partially linear varying coefficient model for such data and propose a regression spline based approach to estimate the mean and covariance parameters jointly where the correlation matrix is decomposed via hyperspherical co-ordinates. A simulation study is conducted to investigate the properties of the estimates of the regression parameters in terms of bias and standard error and to analyse a real data set taken from a multi-

center AIDS cohort study.

The problem of model selection in regression analysis through the use of forward selection, backward elimination and stepwise selection has been well developed in the literature. The main assumption in this, of course, is that the data are normally distributed and the main tool used here is either a t test or an F test. However, properties of these model selection procedures in the framework of generalized linear models are not well-known. We study here the properties of these procedures in generalized linear models, of which the normal linear regression model is a special case. The main tools that is being used are the score test, the F-test, other large sample tests, such as, the likelihood ratio test and the Wald test; the AIC and the BIC are included in the comparison. A systematic study, through simulations, of the properties of this procedure is conducted, in terms of level and power, for normal, Poisson and binomial regression models. Extensions for over-dispersed Poisson and over-dispersed binomial regression models are also given and evaluated. The methods are applied to analyse three data sets.

In practice, it often occurs that an abundance of zero counts arise in data where a discrete generalized linear model may fail to fit but a zero-inflated generalized linear model can be the ideal choice. Researchers often encounter a large number of covariates in such model and need to decide which are potentially important. To find a parsimonious model we develop a model selection procedure using the score test, the Wald test and the likelihood ratio test; also the AIC and the BIC are included in the comparison. Simulation studies are carried out to investigate the performance of these procedures, in terms of level and power, for zero-inflated Poisson and zero-inflated binomial regression models. The methodology is illustrated through two real examples.

Dedication

Dedicated to my beloved parents and family.

Acknowledgements

First of all I am very grateful to my creator Almighty for His countless blessings over me and my family in every part of our life. I would like to take this opportunity to express my utmost gratitude to Dr. Sudhir Paul, my supervisor, for his supervision, encouragement, financial support and many invaluable suggestions and comments for this research. I really appreciate that he lead me to grow in this wonderful research area. His guidance has not only trained my knowledge and expertise, but also cultivated me with open-mindedness, ability of critical thinking, and skills of effective communication which are essential for my future career. I have also learned from him the spirit of hard working, persistence, patience and creativity, which will benefited me for the rest of my life.

I also would like to thank the committee members, Dr. Yash Aneja, Dr. Myron Hlynka and Dr. Mohamed Belalia for their constructive comments and valuable suggestions. I would like to express my sincere gratitude to the external examiner of my dissertation Dr. Serge Provost, Department of Statistical and Actuarial Sciences, Western University, London, Ontario, for his critical review, valuable suggestions and constructive criticisms.

Thanks to the Department of Mathematics and Statistics for providing me the financial support in the form of graduate assistantships, Ontario Graduate Scholarship (OGS). Thanks are also due to Dr. Paul for his financial support from his NSERC grant throughout my study period. Many thanks to my fellow graduate students.

Big achievement depends on sacrifice. It is my family who has sacrificed the most. Regardless of many limitations, stress and hardship during my study, the love, emotion and the continuous encouragement from my wife Mahmuda Karim kept me on track that helped me to reach where I am today. Huge love and affection to my two little angels, my son Ehan and daughter Ereena for providing me joyous company during my graduate studies.

Md Abdulla Al Mamun

May 17, 2019

Windsor, Ontario, Canada

Contents

List of Tables

Chapter 1

Introduction

In this dissertation, two important aspects are discussed: (i) joint estimation of mean and covariance parameters in longitudinal data when the covariates are time invariant and time variant, and (ii) model selection in generalized linear models and zero-inflated generalized linear models.

Longitudinal data arise in many subject-matter areas such as biostatistics, medical and public health sciences, environmental studies and social sciences. Longitudinal studies are characterized by observing the same subjects repeatedly over a period of time. For example, HIV patients may be followed over time and monthly measures such as CD4 counts (CD4 cells are white blood cells that fight infection), or viral load are collected to characterize immune status and disease burden respectively. Usually the subjects are assumed to be independent, while repeated measures data of the same subject are correlated and thus require special statistical techniques for valid analysis and inference.

In longitudinal data analysis, it has been shown that the choice of covariance model

affects standard error estimates and failure to adopt an appropriate covariance structure can lead to a loss of efficiency in estimating the regression parameters (Liang and Zeger, 1986; Diggle, Heagerty, Liang, and Zeger, 2002; Lin and Carroll, 2006). Covariance model choice also affects predictions and imputations. Taylor and Law (1998) show that individual predictions of future CD4 counts are noticeably affected by assumptions about the covariance structure. Moreover inference on the covariance structure itself is of interest when researchers try to understand biological processes and to answer questions like how observations are correlated to each other. Thus, it is important to get the covariance/correlation model correct.

To estimate mean and variance parameters in longitudinal data, several authors have used different approaches (see for example, Liang and Zeger, 1986; Lin and Carroll, 2006; Fan, Huang, and Li, 2007; Qu, Lindsay, and Li, 2000; Diggle *et al.*, 2002; and Fan and Wu, 2008). However, such approaches do not apply the correlation structure directly and cannot flexibly incorporate covariates that may help to explain the covariations. Joint modelling for the mean and covariance becomes a popular approach to overcome this limitations for longitudinal data analysis; see for example, Zhang, Leng, and Tang (2015), Leng, Zhang, and Pan (2010) and Pourahmadi (1999, 2000).

Zhang *et al.* (2015) propose a parametric method to estimate the mean and the covariance parameters jointly in longitudinal data. However, in some applications parametric regression model is too restrictive and cannot faithfully capture the true underlying relationship between the response and covariate. In such cases nonparametric or semiparametric models are appealing because they allow the data to speak for themselves in determining the form of the relationship between the response and covariates. In this dissertation we develop an estimation procedure for the mean and

covariance parameters semiparametrically in longitudinal data.

Time varying coefficient models are the natural extension of classical parametric models that provide a very important tool to explore the dynamic pattern in many scientific areas, namely health science, epidemiology, economics and so on. In longitudinal studies, researchers frequently encounter covariates that are varying over time. For instance, to analyze the Multi-Center AIDS Cohort study, Huang, Wu, and Zhou (2002) find that the baseline function varies over time. Also whether the effect of PreCD4 is constant over time is unclear which motivate us to consider time varying coefficient model. We develop an estimation procedure of the mean and covariance parameters simultaneously for generalized partially linear varying coefficient model in this dissertation.

The salience of model selection in regression analysis for normally distributed response variable is very familiar and is extensively applied in many areas, for instance, engineering, natural sciences, and social sciences. For model selection the use of forward selection, backward elimination and stepwise selection has been well developed in the literature. However, properties of these model selection procedures in generalized linear models are not well-known. Moreover, in practice, it often occurs that a particular count (for example zero) may arise in the data more than the expected number. Examples include, the number of cigarettes smoked by students in a university, the number of insurance claims for a certain type of risk, the number of earthquakes by geographical location etc. A discrete generalized linear model may result in inconsistent estimates of such data, so a zero-inflated generalized linear model can be the ideal choice. Considerable attention has been given to the problem of estimating the parameters involved in the model. However, in practice, scientist encounter a large number of covariates from where need to select potentially important covariates for

the model. We investigate the properties of model selection in both generalized linear models and zero-inflated generalized linear models in this dissertation.

In Chapter 2, we discuss some preliminaries and review joint mean-covariance regression models, varying coefficient models in longitudinal data, $C(\alpha)$ test (Neyman, 1959), Wald test (Wald, 1943) and likelihood ratio test (Neyman and Pearson, 1928). We also review generalized linear models and hyperspherical coordinates. Furthermore, we review the B-spline and the penalized spline smoothing technique.

In Chapter 3, we develop procedures for estimating mean and covariance parameters semiparametrically in longitudinal data. To estimate non-parametric functions we use regression splines. We further apply a penalized spline in order to investigate whether it produces improvement in estimation of the non-parametric functions. A simulation study is conducted to investigate the performance of the estimators of the regression parameters in terms of bias and efficiency, the effect of fixing the number of knots and the effect of misspecifying the error distribution (robustness study). Three real data sets are analyzed.

In Chapter 4, we deal with generalized partially linear varying coefficient models in longitudinal data and develop estimation procedures of mean and covariance parameters jointly. A regression spline is used to estimate the coefficient of time variant covariates. A simulation study is performed to examine the performance of the estimated regression parameters in terms of bias and standard errors. The method is applied to a real data set.

In Chapter 5, we develop model selection procedures for generalized linear models using the score test. Other large sample tests, namely, the likelihood ratio test and the Wald test are included in the comparison. A systematic study, through simulations,

of the properties of this procedure is conducted, in terms of level and power, for normal, Poisson and binomial regression models. Extensions of the model selection procedure for over-dispersed Poisson and over-dispersed binomial regression models are also developed. The methods are applied to analyse three real data sets.

In Chapter 6, we derive model selection procedures for zero-inflated generalized linear models using the score test, Wald test and likelihood ratio test. We further consider model selection through AIC and BIC for comparison. A simulation study is performed to investigate the properties of these procedures in terms of level and power, for zero-inflated Poisson and zero-inflated binomial regression models. The methodology is illustrated through two real examples.

Finally, conclusions of the thesis with the summary of findings and a plan for future study are presented in Chapter 7.

It is noted that few topics are repeated in the chapters because the chapters are intended for submission as distinct papers.

Chapter 2

Some Preliminaries and Review of Current Literature

2.1 Joint Mean-Covariance Regression Model in Longitudinal Data

Suppose longitudinal measurements $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})'$, $(i = 1, \ldots, n)$ are collected from n subjects at times $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{im_i})'$. Assume that $\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i)$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})'$ and Σ_i are an $m_i \times 1$ vector and an $m_i \times m_i$ positive definite matrix, respectively. The mean μ_{ij} of y_{ij} can usually be modelled by a linear regression, $\mu_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}$, where $\mathbf{x}_{ij}$ denotes the $p \times 1$ covariates associated with the j-th observation of the i-th subject and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression parameters.

Due to the positive definiteness of the covariance matrix Σ_i, it can be decomposed by Cholesky decomposition. Then there exists a unique lower triangular matrix T_i

with 1's as diagonal entries and a unique diagonal matrix D_i with positive diagonal entries such that

$$T_i \Sigma_i T_i' = D_i \quad \text{or} \quad \Sigma_i^{-1} = T_i' D_i^{-1} T_i,$$

where the below-diagonal entries of T_i are the negatives of the coefficients of

$$\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}),$$

where $\hat{y}_{ij}$ is the linear least squares predictor of y_{ij} based on its predecessors $y_{i(j-1)}, \ldots, y_{i1}$ and the diagonal entries of D_i are the prediction error variances (innovation variances) $\sigma_{ij}^2 = \text{var}(y_{ij} - \hat{y}_{ij})$ for $1 \leq j \leq m_i$ and $1 \leq i \leq n$. Note that ϕ_{ijk} and $\log \sigma_{ij}^2$ are unconstrained. Thus they may model in terms of covariates as $\log \sigma_{ij}^2 = \mathbf{z}_{ij}' \boldsymbol{\lambda}$ and $\phi_{ijk} = \mathbf{w}_{ijk}' \boldsymbol{\gamma}$ respectively, where $\mathbf{z}_{ij}$ and $\mathbf{w}_{ijk}$ are $q \times 1$ and $d \times 1$ vectors of known covariates, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ are parameters for the variances and correlations of $\mathbf{y}_i$, respectively.

Pourahmadi (1999) proposed the following joint mean-covariance model by combining all three aforementioned models

$$\mu_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta}, \quad \log \sigma_{ij}^2 = \mathbf{z}_{ij}' \boldsymbol{\lambda}, \quad \text{and} \quad \phi_{ijk} = \mathbf{w}_{ijk}' \boldsymbol{\gamma}. \tag{2.1}$$

The maximum likelihood estimating equations for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are

$$\sum_{i=1}^{n} \mathbf{X}_i' \Sigma_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)) = 0,$$
$$\sum_{i=1}^{n} \mathbf{Z}_i^{*\prime} D_i^{-1} (r_i - \mathbf{Z}_i^{*} \gamma) = 0,$$
and
$$\frac{1}{2} \sum_{i=1}^{n} H_i' (D_i^{-1} e_i - \mathbf{1}_{m_i}) = 0,$$

where the matrix $\mathbf{Z}_i^*$, of order $m_i \times (q+1)$, has typical row $\mathbf{z}_{ij}^{*\prime} = \sum_{k=1}^{j-1} r_{ik} \mathbf{w}_{ijk}'$. Also $H_i = (h_{i1}', \ldots, h_{im_i}')'$, $e_i = (e_{i1}, \ldots, e_{im_i})'$, with $e_{ij} = (r_{ij} - \hat{r}_{ij})^2$ and $\hat{r}_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$ are the $m_i \times (d+1)$ matrix of covariates and the $m_i \times 1$ vector of squared fitted residuals, respectively. The parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are obtained by solving estimating equations using an iteratively re-weighted least squares algorithm.

It is noted that joint mean-covariance model (2.1) proposed by Pourahmadi (1999, 2000) is valid only for balanced longitudinal data. Pan and MacKenzie (2003) generalized the model for unbalanced longitudinal data and Ye and Pan (2006) extended further within the framework of generalized estimating equations. Although Pourahmadi's (1999, 2000) and Pan and MacKenzie's (2003) methods require the normal distributional assumption, the Ye and Pan (2006) approach depends on the existence of the first four moments of responses only.

However all aforementioned models do not reveal the correlation structure between longitudinal measurements directly and may encounter difficulty in interpreting the covariation structure. To overcome this, most recently Zhang, Leng, and Tang (2015) propose a joint mean-variance correlation modelling approach that targets directly the variances and correlations in the longitudinal data. They decompose the correla-

tion matrix via hyperspherical co-ordinates using angles and trigonometric functions. Their proposed joint regression model for the means, variances and correlations is as follows

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}, \quad \log\sigma^2_{ij} = \mathbf{z}'_{ij}\boldsymbol{\lambda}, \quad \text{and} \quad \phi_{ijk} = \mathbf{w}'_{ijk}\boldsymbol{\gamma},$$

where $\mathbf{x}_{ij}$ are the usual known covariates as mentioned earlier, $\mathbf{z}_{ij}$ and $\mathbf{w}_{ijk}$ may contain baseline covariates, as well as polynomials in time (time related to longitudinal data) and their interactions.

2.2 Semiparametric Mean-Covariance Regression Model in Longitudinal Data

Model misspecification may produce biased estimation, which is even more severe than misspecification of the covariance. It is natural to relax the parametric assumption and an attractive approach is the semiparametric regression model which retains the flexibility of the nonparametric model but avoids the need to use a fully nonparametric model. Leng, Zhang, and Pan (2010) propose the following semiparametric models for the mean and the covariance structure for longitudinal data

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(t_{ij}), \quad \log\sigma^2_{ij} = \mathbf{z}'_{ij}\boldsymbol{\lambda} + f_2(t_{ij}), \quad \text{and} \quad \phi_{ijk} = \mathbf{w}'_{ijk}\boldsymbol{\gamma},$$

where two nonparametric functions $f_1(\cdot)$ and $f_2(\cdot)$ are unknown smooth functions which are parametrized by regression splines, and all of the regression parameters are solved iteratively by fixing the other parameters. Here the term nonparametric

is not meant that completely lack of parameters but that the number and nature of the parameters are flexible and not fixed in advance.

2.3 Varying Coefficient Model in Longitudinal Data

To study the association between the covariates and the response variable in longitudinal data, the following linear model is used in practice:

$$y(t) = \beta_0 + \mathbf{x}'(t)\boldsymbol{\beta} + \epsilon(t), \tag{2.2}$$

where both covariates and response variable are collected at time t. See for example Diggle *et al.* (2002).

However model (2.2) does not allow the association to vary over time, even though the covariates and the response variable change over time. To overcome this, Zeger and Diggle (1994) propose a semiparametric model as follows

$$y(t) = \beta_0(t) + \mathbf{x}'(t)\boldsymbol{\beta} + \epsilon(t), \tag{2.3}$$

where $\beta_0(t)$ in model (2.3) can be estimated by the kernel, polynomial and smoothing spline methods (Brumback and Rice, 1998; Fan and Li, 2004).

The model (2.3) allows the intercept to vary over time only, not the coefficients of the other covariates. However researchers sometime encounter a situation where covariates are time variant, for instance in a Multi-Center AIDS Cohort study (Kaslow

et al., 1987); whether or not PreCD4 has a constant effect over time is unclear. Taking this into account Fan, Huang, and Li (2007) introduce a semiparametric varying-coefficient partially linear model, in which the covariate effects are constant over time for some covariates and time-varying for others, as follows

$$y(t) = \mathbf{x}'(t)\boldsymbol{\alpha}(t) + \mathbf{z}'(t)\boldsymbol{\beta} + \epsilon(t), \tag{2.4}$$

where $y(t)$ is the response variable, $\mathbf{x}(t)$ and $\mathbf{z}(t)$ are the covariate vectors at time t, $\boldsymbol{\alpha}(t)$ comprises p unknown smooth functions, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter vector and $E\left[\epsilon(t)|x(t), z(t)\right] = 0$. They model the variance function nonparametrically and correlation structure parametrically, but mainly focus on the improvement in the estimation of the mean regression function using a possibly misspecified covariance structure.

To model jointly the mean and the covariance, Qin, Mao, and Zhu (2015) propose a general semiparametric model using the modified Cholesky decomposition. They consider a generalized partially linear varying coefficient model

$$\begin{aligned} g(\mu_{ij}) &= \mathbf{x}'_{ij}\boldsymbol{\alpha}(t_{ij}) + \mathbf{z}'_{ij}\boldsymbol{\beta}, \qquad \phi_{ijk} = \mathbf{w}'_{ijk}\boldsymbol{\gamma}, \\ \log(\sigma^2_{ij}) &= \mathbf{u}'_{ij}\mathbf{f}(t_{ij}) + \mathbf{v}'_{ij}\boldsymbol{\lambda}, \end{aligned} \tag{2.5}$$

where $\mathbf{x}_{ij} \in \mathbb{R}^p$ and $\mathbf{z}_{ij} \in \mathbb{R}^q$ are covariate vectors for the time varying coefficients and constant coefficients at time t_{ij}, respectively; $\mathbf{w}_{ijk}, \mathbf{u}_{ij}$ and $\mathbf{v}_{ij}$ are $(d \times 1), (h \times 1)$ and $(m \times 1)$ vectors of covariates, respectively; $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are the regression coefficients; $\boldsymbol{\alpha}(t_{ij})$ and $\mathbf{f}(t_{ij})$ comprises p and h unknown smooth functions. A regression spline-based approach within the framework of generalized estimating equations is proposed

to estimate the parameters in the mean and the covariance.

2.4 B-spline

Splines are used to describe complicated curves when the true functional form is not known. A B-spline of order ℓ is a piecewise polynomial function of degree $\ell - 1$ in a variable x. Any smooth function (say $f(x)$) can be approximated by B-spline basis functions. Let $a = a_0 < a_1 < \cdots < a_{k_n} < a_{k_n+1} = b$ be a partition of the interval $[a, b]$. Using $\{a_i\}$ as the internal knots, the B-spline basis of order ℓ, $B_i^{(\ell)}(x)$, is defined recursively as

$$B_i^{(\ell)}(x) = \frac{x - a_i}{a_{i+\ell-1} - a_i} B_i^{(\ell-1)}(x) + \frac{a_{i+\ell} - x}{a_{i+\ell} - a_{i+1}} B_{i+1}^{(\ell-1)}(x),$$

with

$$B_i^{(1)}(x) = \begin{cases} 1, & a_i \leq x < a_{i+1} \\ 0, & \text{otherwise.} \end{cases}$$

Then $K = k_n + \ell$ B-spline basis functions of order ℓ form a basis for the linear spline space. Thus $f(x)$ can be approximated by $f(x) = \sum_{i=1}^{K} \alpha_i B_i^{(\ell)}(x)$.

2.5 Penalized Spline

Although B-splines are a good option for estimation in goodness of fit, they may under-fit or over-fit the data due to wrong number and location of knots selection. Increasing the number of knots does not always provide a better fit (Griggs, 2013). Moreover, too many knots may result in over fitting the data, and modelling noise

instead of the signal. There is an optimal number of knots which can be found by trial and error. However this can be time consuming, especially for large complicated datasets.

To avoid this issue and optimize the fit, penalized spline (P-spline) imposes a penalization upon the coefficients of the B-spline basis functions to optimize the fit (Eilers and Marx, 1996), thus imposing a penalization upon the parameters $\alpha_1, \alpha_2, \ldots, \alpha_K$, so that they are constrained such that

$$\sum_{i=1}^{K} \alpha_i^2 \leq C.$$

Then minimize log-likelihood l subject to $\boldsymbol{\beta}' D \boldsymbol{\beta} \leq C$, where $D = \begin{bmatrix} 0_{p\times p} & 0_{p\times K} \\ 0_{K\times p} & I_{K\times K} \end{bmatrix}$.

Using Lagrange multipliers we can solve the equation to find the optimal $\hat{\boldsymbol{\beta}}$ value for a given λ value.

2.6 Hyperspherical Coordinate System

There is a unique correspondence between 3-dimensional Cartesian coordinate systems and 3-dimensional spherical coordinate systems. Every point $(x_1, x_2, x_3) \in \mathbb{R}^3$ is represented by the 3-dimensional spherical coordinates (r, ϕ_1, ϕ_2) as follows

$$\begin{aligned} x_1 &= r\cos(\phi_1), \\ x_2 &= r\sin(\phi_1)\cos(\phi_2), \\ x_3 &= r\sin(\phi_1)\sin(\phi_2), \end{aligned} \tag{2.6}$$

where $0 \leq r < \infty$, $0 \leq \phi_1 \leq \pi$, and $0 \leq \phi_2 < 2\pi$.

The n-dimensional hyperspherical coordinates, a generalization of the 3-dimensional spherical coordinates, consist of a radial coordinate r and $(n-1)$ angular coordinates $\phi_1, \ldots, \phi_{n-1}$, where the angles $\phi_1, \ldots, \phi_{n-2}$ range over $[0, \pi]$ radians and ϕ_{n-1} ranges over $[0, 2\pi)$ radians. Similar to (2.6) it is possible to construct a unique correspondence between n-dimensional Cartesian coordinate systems and n-dimensional hyperspherical coordinate systems.

If $(x_1, \ldots, x_n)$ are the Cartesian coordinates, then $x_i (i = 1, \ldots, n)$ can be expressed in terms of r, $\phi_1, \ldots, \phi_{n-2}$, and ϕ_{n-1} as follows

$$
\begin{aligned}
x_1 &= r\cos(\phi_1), \\
x_2 &= r\sin(\phi_1)\cos(\phi_2), \\
x_3 &= r\sin(\phi_1)\sin(\phi_2)\cos(\phi_3), \\
&\vdots \\
x_{n-1} &= r\sin(\phi_1)\ldots\sin(\phi_{n-2})\cos(\phi_{n-1}), \\
x_n &= r\sin(\phi_1)\ldots\sin(\phi_{n-2})\sin(\phi_{n-1}).
\end{aligned}
$$

2.7 Test Statistics

Let $y_1, y_2, \ldots, y_n$ be a random sample of size n from a distribution that has a probability density function $f(y_1, y_2, \ldots, y_n; \boldsymbol{\theta}, \boldsymbol{\phi})$. For the given data, the likelihood function over the entire parameter space is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\phi} | y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f(y_i | \boldsymbol{\theta}, \boldsymbol{\phi}).$$

Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ are parameters of interest, then $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_s)'$ are treated as nuisance parameters.

In this section the $C(\alpha)$ test, score test, Wald test and likelihood ratio test statistics are described below to test the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0k})'$. Note that under the null hypothesis the estimator of $\boldsymbol{\phi}$ is denoted by $\hat{\boldsymbol{\phi}}$. Further, under the alternative hypothesis the estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are denoted by $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\phi}}$ respectively.

2.7.1 $C(\alpha)$ Test and Score Test

The $C(\alpha)$ test is evaluated by using the partial derivatives of the log-likelihood function with respect to the nuisance parameters and the parameters of interest calculated at the null hypothesis. Let $L(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{y})$ be the likelihood function and l be the log-likelihood function of the data. Define the partial derivatives of the log-likelihood which are evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0k})'$ as

$$\boldsymbol{\psi} = \frac{\partial l}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \left[\frac{\partial l}{\partial \theta_1}, \frac{\partial l}{\partial \theta_2}, \cdots, \frac{\partial l}{\partial \theta_k}\right]'\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

and

$$\boldsymbol{\gamma} = \frac{\partial l}{\partial \boldsymbol{\phi}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \left[\frac{\partial l}{\partial \phi_1}, \frac{\partial l}{\partial \phi_2}, \cdots, \frac{\partial l}{\partial \phi_s}\right]'\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Under the null hypothesis and mild regularity conditions, $\left(\frac{\partial l}{\partial \boldsymbol{\theta}}, \frac{\partial l}{\partial \boldsymbol{\phi}}\right)$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix

$I^{-1}(\boldsymbol{\theta}, \boldsymbol{\phi})$ (Cramer, 1946), where

$$I(\boldsymbol{\theta}, \boldsymbol{\phi}) = \begin{bmatrix} I_{\boldsymbol{\theta\theta}} & I_{\boldsymbol{\theta\phi}} \\ I'_{\boldsymbol{\theta\phi}} & I_{\boldsymbol{\phi\phi}} \end{bmatrix}$$

is the Fisher information matrix with elements

$I_{\boldsymbol{\theta\theta}} = E\left(-\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)$, $I_{\boldsymbol{\theta\phi}} = E\left(-\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\phi}'}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)$, and $I_{\boldsymbol{\phi\phi}} = E\left(-\frac{\partial^2 l}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)$
which are $(k \times k)$, $(k \times s)$ and $(s \times s)$ matrices respectively.

The $C(\alpha)$ test is based on the adjusted score $S = \frac{\partial l}{\partial \boldsymbol{\theta}} - B\frac{\partial l}{\partial \boldsymbol{\phi}}$, where B is the matrix of partial regression coefficients that is obtained by regressing $\frac{\partial l}{\partial \boldsymbol{\theta}}$ on $\frac{\partial l}{\partial \boldsymbol{\phi}}$. According to Bartlett (1953), B and the variance-covariance matrix of S can be written as $I_{\boldsymbol{\theta\phi}} I^{-1}_{\boldsymbol{\phi\phi}}$ and $I_{\boldsymbol{\theta\theta}.\boldsymbol{\phi}} = I_{\boldsymbol{\theta\theta}} - I_{\boldsymbol{\theta\phi}} I^{-1}_{\boldsymbol{\phi\phi}} I'_{\boldsymbol{\theta\phi}}$ respectively. Thus the distribution of the adjusted score $S \sim MN(0,\ I_{\boldsymbol{\theta\theta}.\boldsymbol{\phi}})$ and hence the distribution of $S' I^{-1}_{\boldsymbol{\theta\theta}.\boldsymbol{\phi}} S \sim \chi^2_{(k)}$ (Neyman, 1959).

As the nuisance parameters $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_s)'$ are involved in this statistic, following Moran (1970), replacing $\sqrt{n}$ consistent estimators of the nuisance parameters by $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_s)'$ that are evaluated from the data.

Thus the test statistic becomes

$$\chi^2_{C(\alpha)} = \tilde{S}' \tilde{I}^{-1}_{\boldsymbol{\theta\theta}.\boldsymbol{\phi}} \tilde{S},$$

which is asymptotically distributed as chi-squared with k degrees of freedom (Neyman, 1959).

Note that if the nuisance parameter $\boldsymbol{\phi}$ is replaced by its maximum likelihood estimator $\hat{\boldsymbol{\phi}}$, then the adjusted score function S reduces to $\boldsymbol{\psi}$. The $C\,(\alpha)$ statistic then

becomes

$$S_1 = \hat{\boldsymbol{\psi}}' \hat{I}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}.\boldsymbol{\phi}} \hat{\boldsymbol{\psi}},$$

which is a score test (Rao, 1948).

2.7.2 Wald Test

The Wald test (Wald, 1943) statistic is given by

$$W = \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left[\text{cov}\left(\tilde{\boldsymbol{\theta}}\right)\right]^{-1} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right),$$

where $\text{cov}\left(\boldsymbol{\theta}\right)$ is the inverse of Fisher information matrix whose (j,k)-th element is $E\left[-\frac{\partial^2 l}{\partial\theta_j \partial\theta_k}\right]$.

2.7.3 Likelihood Ratio Test

The likelihood ratio test (Neyman and Pearson, 1928) is the ratio of the maximized likelihood function under the null hypothesis to that under the alternative hypothesis and is defined as

$$\Lambda = \frac{L\left(y_1, y_2, \ldots, y_n; \boldsymbol{\theta}_0, \hat{\boldsymbol{\phi}}\right)}{L\left(y_1, y_2, \ldots, y_n; \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\phi}}\right)}.$$

Taking ln on both sides, the likelihood ratio test statistic can be written as

$$LR = -2\ln\Lambda = 2(\tilde{l}_1 - \hat{l}_0),$$

where $\hat{l}_0$ and $\tilde{l}_1$ are the maximized log-likelihood under the null hypothesis and under the alternative hypothesis respectively.

2.7.4 Comparison of Three Test Statistics

From large sample probability theory, we know that under certain regularity conditions, if the null hypothesis H_0 holds, the score, Wald and LR test statistics asymptotically follow a central Chi-square distribution with k degrees of freedom (see for example Agresti, 2007). Thus, for a fixed significance level $\alpha > 0$, the null hypothesis is rejected if the value of a test statistic is greater than $\chi^2_\alpha(k)$ value.

The score test is particularly attractive to many researchers because one needs to study the distribution of the test statistic under the null hypothesis only. It often maintains, at least approximately, a preassigned level of significance and often produces a statistic that is simple to calculate.

On the contrary, the other two asymptotically equivalent tests (Wald test and LR test) require estimates of the parameters under the alternative hypothesis and often show liberal or conservative behaviour in small samples. In general these two tests are considered for large samples. For more about the comparison of three tests, the reader may look at Rao (2005).

2.8 Generalized Linear Models

A generalized linear model (GLM) is the generalization of ordinary linear regression models that allows for response variables that have error distribution models other

than a normal distribution. It is comprised of three components, namely, a random component, a systematic component, and a link function and can be described as follows

i) Suppose that the joint probability function of the response variable Y with mean μ can be written in the form

$$f(y;\theta) = \exp\left[\phi^{-1}\left\{y\theta - b(\theta)\right\} + C(y,\phi)\right].$$

for some known functions $b(.)$, $C(.)$, canonical or natural parameter θ, and constant ϕ which may be known or a parameter to be estimated. This is said to be in canonical form.

ii) The systematic component connects a set of covariates with a linear predictor in the form

$$\eta = \sum_{j=1}^{p} X_j \beta_j.$$

iii) The link function is a monotone differentiable function of the mean that connects the random and the systematic components. The model links the mean μ to the linear predictor η by $\eta = g(\mu)$, where the link function $g(.)$ is a monotone differentiable function. The mean and the variance of Y are $E(Y) = b'(\theta)$ and $\text{var}(Y) = \phi b''(\theta)$ respectively (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989).

Chapter 3

Joint Estimation of Mean and Covariance Parameters in Longitudinal Data

3.1 Introduction

In the statistical literature, methods to understand the relationship of explanatory variables on each individual outcome variable are well developed and widely applied. However, in most health-related studies, given the technological advancement and sophisticated methods of obtaining and storing data, a need to perform joint analysis of mean and covariance parameters simultaneously and accounting for the correlations is in high demand since a good covariance modelling approach improves statistical inference of the mean of interest (Liang and Zeger, 1986; Diggle *et al.*, 2002; Lin and Carroll, 2006). Furthermore, the covariance structure itself may be of scientific

interest (Fan and Wu, 2008).

Suppose longitudinal measurements $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})'$ and covariate vectors $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i})'$ $(i = 1, \ldots, n)$, with $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ for $j = 1, \ldots, m_i$, collected from n subjects, are observed at times $\mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})'$. In longitudinal data analysis, in order to avoid biased estimation, it is important that statistical analysis takes into account that the repeated observations y_{ij}, $j = 1, \ldots, m_i$ are correlated (Liang and Zeger, 1986; Diggle *et al.*, 2002; Lin and Carroll, 2006). Accordingly we assume that $\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i)$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})'$, $\Sigma_i = D_i R_i D_i$, $D_i = \text{diag}(\sigma_{i1}, \ldots, \sigma_{im_i})$, and $R_i = (\rho_{ijk})_{j,k=1}^{m_i}$ is the correlation matrix of $\mathbf{y}_i$ with $\rho_{ijk} = \text{corr}(y_{ij}, y_{ik})$ being the correlation between the j^{th} and k^{th} measurements of the i^{th} subject. The main purpose in such longitudinal studies is to estimate the parameters involved in the means, the variances and the correlation matrices. This can be done by maximizing the log-likelihood or by solving the maximum likelihood estimating equations. However, the constraints involved in the correlation parameters create a challenge. This can be overcome by decomposing the correlation matrix by Cholesky decomposition.

Zhang, Leng, and Tang (2015) proposed to parametrize the correlation matrix R_i for subject i (we suppress i) via hyperspherical co-ordinates by the Cholesky decomposition $R = TT'$, where $T = (T_{jk})$ is a lower triangular matrix given by

$$T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{21} & s_{21} & 0 & \cdots & 0 \\ c_{31} & c_{32}s_{31} & s_{32}s_{31} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ c_{m1} & c_{m2}s_{m1} & c_{m3}s_{m2}s_{m1} & \cdots & \prod_{l=1}^{m-1} s_{ml} \end{pmatrix},$$

where $c_{jk} = \cos(\phi_{jk})$ and $s_{jk} = \sin(\phi_{jk})$ are trigonometric functions of angles ϕ_{jk}.

For subject i, the total number of angles ϕ_{ijk} $(1 \leq k < j \leq m_i)$ is $m_i(m_i - 1)/2$, which is the same as that of the free parameters in the correlation matrix. The decomposition of R has several advantages (i) diagonal elements of TT' are 1, and all other elements fall between -1 and 1, (ii) TT' is always non-negative definite, satisfying the requirements of a correlation matrix, and (iii) the angles ϕ_{jk} of T as parameters are unconstrained in the range $[0, \pi)$. It also establishes a hierarchical connection between the correlations and the angles (for further discussion on this see Zhang *et al.*, 2015). They then propose a joint regression model for the means, the variances and the correlations as

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}, \quad \log \sigma^2_{ij} = \mathbf{z}'_{ij}\boldsymbol{\lambda}, \quad \phi_{ijk} = \mathbf{w}'_{ijk}\boldsymbol{\gamma},$$

where $\mathbf{x}_{ij}$ are the usual known covariates as mentioned earlier, $\mathbf{z}_{ij}$ and $\mathbf{w}_{ijk}$ may contain baseline covariates, as well as polynomials in time (time related to longitudinal data) and their interactions. The unknown regression parameters $\boldsymbol{\beta}, \boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are of dimensions $p \times 1$, $d \times 1$ and $q \times 1$ respectively. In practice we may choose $\mathbf{w}_{ijk}$ as a polynomial of time lag $(t_{ik} - t_{ij})$. Zhang *et al.* (2015) estimate the parameters $\boldsymbol{\beta}, \boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ via the quasi-Fisher scoring algorithm.

Zhang *et al.* (2015, pp 237) suggest future research on modelling the means, variances and correlations nonparametrically and semiparametrically. In this Chapter we investigate the properties of the estimates of the regression parameters through semiparametric modelling of the means and variances and study the impact of this to model parsimony. For the purpose of comparison we consider three models.
Model 1: the parametric model given above,

Model 2: a model in which only the means are modelled semiparametrically, which is,

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(t_{ij}), \quad \log(\sigma^2_{ij}) = \mathbf{z}'_{ij}\boldsymbol{\lambda}, \quad \phi_{ijk} = \mathbf{w}'_{ijk}\boldsymbol{\gamma},$$

Model 3: a model in which the means and the variances are modelled semiparametrically, which is,

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(t_{ij}), \quad \log(\sigma^2_{ij}) = \mathbf{z}'_{ij}\boldsymbol{\lambda} + f_2(t_{ij}), \quad \phi_{ijk} = \mathbf{w}'_{ijk}\boldsymbol{\gamma}.$$

In Model 2 and Model 3, $f_1(\cdot)$ and $f_2(\cdot)$ are smooth functions parametrized by regression splines.

As in Zhang *et al.* (2015) we decompose the correlation matrix via hyperspherical co-ordinates and as in Lang, Zhang, and Pan (2010) we use B-spline to estimate the unknown functions $f_1(\cdot)$ in Model 2 and $f_1(\cdot)$ and $f_2(\cdot)$ in Model 3. Four further investigations were conducted. The first of these is to see the performance of the estimators of the regression parameters in terms of bias and efficiency. The second is to see the effect of fixing the knots in spline smoothing. The third is a robustness study where the normality assumption of the error distribution is replaced by a mixture of normal distributions and the fourth is to see whether use of a penalized spline results in improved estimation of the non-parametric functions in comparison to using B-spline.

Section 3.2 deals with the method of estimation of the parameters of Model 3 in which the correlation matrix is decomposed via hyperspherical co-ordinates and the unknown functions $f_1(\cdot)$ and $f_2(\cdot)$ are estimated using B-spline basis functions. Esti-

mation in Model 2 is discussed as a special case of Model 3. An extensive simulation study is conducted and results are summarized in Section 3.3. Three real data sets, a CD4 cell data (Kaslow *et al.*, 1987) set, a set of cattle data (Kenward, 1987), and Progesterone hormone data (Brumback and Rice, 1998) are analyzed in Section 3.4. A discussion follows in Section 3.5.

3.2 Estimation in Joint Semiparametric Models

3.2.1 Review of the Estimation of Parameters of Model 1

In longitudinal measurement each subject $\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i)$. Then the joint density function of $\mathbf{y}_i (i = 1, \ldots, n)$ is

$$f(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^{m_i}|\Sigma_i|}} \exp\left(-\frac{1}{2}\mathbf{r}_i'\Sigma_i^{-1}\mathbf{r}_i\right),$$

where $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$, Thus without the constant, the minus twice log-likelihood of the parametric model (3.1) is

$$-2l = \sum_{i=1}^{n} \left[\log|D_i R_i D_i| + \mathbf{r}_i' D_i^{-1} R_i^{-1} D_i^{-1} \mathbf{r}_i\right].$$

The parameters $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are obtained by solving the following estimating equa-

tions via the quasi-Fisher scoring algorithm

$$\begin{aligned}
\boldsymbol{T}_1(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\mathbf{X}_i'\Delta_i\Sigma_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)=0,\\
\boldsymbol{T}_2(\boldsymbol{\lambda};\boldsymbol{\beta},\boldsymbol{\gamma}) &= \frac{1}{2}\sum_{i=1}^{n}\mathbf{Z}_i'(\mathbf{q}_i-\mathbf{1}_{m_i})=0,\\
&\text{and}\\
\boldsymbol{T}_3(\boldsymbol{\gamma};\boldsymbol{\beta},\boldsymbol{\lambda}) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}(\epsilon_{ij}^2-1)+\epsilon_{ij}\sum_{k=1}^{j-1}b_{ijk}\epsilon_{ik}\right]=0
\end{aligned}$$

respectively, where $\Delta_i = \Delta_i(\mathbf{X}_i\boldsymbol{\beta}) = \text{diag}\{\dot{g}^{-1}(\boldsymbol{x}_{i1}'\boldsymbol{\beta}),\ldots,\dot{g}^{-1}(\boldsymbol{x}_{im_i}'\boldsymbol{\beta})\}$, $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse link function $g^{-1}(\cdot)$, $\mu(\cdot) = g^{-1}(\cdot)$, $\mathbf{q}_i = \text{diag}(R_i^{-1}D_i^{-1}\mathbf{r}_i\mathbf{r}_i'D_i^{-1})$, and $b_{ijk} = \sum_{l=k}^{j}\frac{\partial T_{ilk}}{\partial\gamma}a_{ijl}$ with a_{ijl} being the (j,l) element of T_i^{-1}.

3.2.2 Estimation in Model 3 based on B-spline

As discussed in Section 3.1, following Leng *et al.* (2010) we propose joint regression semiparametric modelling of the means and the variances as

$$g(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} + f_1(t_{ij}), \quad \log(\sigma_{ij}^2) = \mathbf{z}_{ij}'\boldsymbol{\lambda} + f_2(t_{ij}), \quad \text{and} \quad \phi_{ijk} = \mathbf{w}_{ijk}'\boldsymbol{\gamma},$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are unknown smooth functions which are parametrized by regression splines. As in Zhang *et al.* (2015), we parametrize R_i via hyperspherical co-ordinates. For simplicity, we assume that f_1 and f_2 have the same smoothness property. Without loss of generality, we assume that the domain of t_{ij} is in the interval $[0,1]$ with partitions $0 = a_0 < a_1 < \cdots < a_{k_n} < a_{k_n+1} = 1$. Using the a_i's as knots, we have $K = k_n + \ell$ normalized B-spline basis functions of order ℓ that form a

basis for the linear spline space. The B-spline basis of order ℓ, $B_i^{(\ell)}(t)$, is defined as

$$B_i^{(\ell)}(t) = \frac{t - a_i}{a_{i+\ell-1} - a_i} B_i^{(\ell-1)}(t) + \frac{a_{i+\ell} - t}{a_{i+\ell} - a_{i+1}} B_{i+1}^{(\ell-1)}(t),$$

and

$$B_i^{(1)}(t) = \begin{cases} 1, & a_i \leq t < a_{i+1} \\ 0, & \text{otherwise.} \end{cases}$$

Note that $B_i^{(\ell)}(t)$ is a polynomial function of degree $\ell-1$. More details on the construction of a B-spline basis can be found in Schumaker (1981). Thus $f_1(t)$ and $f_2(t)$ are approximated by $\boldsymbol{\pi}'(t)\boldsymbol{\alpha}$ and $\boldsymbol{\pi}'(t)\tilde{\boldsymbol{\alpha}}$, respectively, where $\boldsymbol{\pi}(t) = (B_1^{(\ell)}(t), \ldots, B_K^{(\ell)}(t))'$ is the vector of basis functions and $\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}} \in \mathbb{R}^K$ are the spline coefficient vector. Let $\boldsymbol{\pi}_{ij} = \boldsymbol{\pi}(t_{ij})$. With this notation, the nonlinear regression models can be linearized as in what follows

$$\begin{aligned} g(\mu_{ij}) &= \mathbf{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{\pi}'(t_{ij})\boldsymbol{\alpha} = \Pi_{ij}'\boldsymbol{\theta}, \\ \log(\sigma_{ij}^2) &= \mathbf{z}_{ij}'\boldsymbol{\lambda} + \boldsymbol{\pi}'(t_{ij})\tilde{\boldsymbol{\alpha}} = \Upsilon_{ij}'\boldsymbol{\rho}, \text{ and} \\ \phi_{ijk} &= \mathbf{w}_{ijk}'\boldsymbol{\gamma}, \end{aligned}$$

where $\Pi_{ij}' = (\mathbf{x}_{ij}', \boldsymbol{\pi}_{ij}')$, $\Upsilon_{ij}' = (\mathbf{z}_{ij}', \boldsymbol{\pi}_{ij}')$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ and $\boldsymbol{\rho} = (\boldsymbol{\lambda}', \tilde{\boldsymbol{\alpha}}')'$. Suppose $\boldsymbol{\Pi}_i = (\Pi_{i1}', \Pi_{i2}', \ldots, \Pi_{im_i}')'$, $\boldsymbol{\Upsilon}_i = \left(\Upsilon_{i1}', \Upsilon_{i2}', \ldots, \Upsilon_{im_i}'\right)'$. Thus, now the parameters of interest are $\boldsymbol{\theta}$, $\boldsymbol{\rho}$, and $\boldsymbol{\gamma}$. Let $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$. Then, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})' = T_i^{-1} D_i^{-1} \mathbf{r}_i \sim N(\mathbf{0}, I_{m_i})$.

Denoting l to be the log-likelihood apart from a constant it can be shown that

$$\begin{aligned} -2l &= -2\sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \log|\Sigma_i| + \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{n} \log|D_i R_i D_i| + \sum_{i=1}^{n} \mathbf{r}_i' D_i^{-1} R_i^{-1} D_i^{-1} \mathbf{r}_i \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} (\log\sigma_{ij}^2 + \log T_{ijj}^2 + \epsilon_{ij}^2) \end{aligned}$$

Define $\Delta_i = \Delta_i(\boldsymbol{\Pi}_i\boldsymbol{\theta}) = \text{diag}\{\dot{g}^{-1}(\Pi_{i1}'\boldsymbol{\theta}), \ldots, \dot{g}^{-1}(\Pi_{im_i}'\boldsymbol{\theta})\}$ where $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse link function $g^{-1}(\cdot)$ and note that $\mu(\cdot) = g^{-1}(\cdot)$. Then

$$\begin{aligned} \frac{\partial l_i}{\partial \boldsymbol{\theta}} &= \frac{\partial \mu_i'}{\partial \boldsymbol{\theta}} \Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\Pi}_i\boldsymbol{\theta})) \\ &= \boldsymbol{\Pi}_i' \Delta_i \Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\Pi}_i\boldsymbol{\theta})) \end{aligned}$$

$$\begin{aligned} \frac{\partial l_i}{\partial \boldsymbol{\gamma}} &= -\sum_{j=1}^{m_i} \left[\frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}} + \frac{\partial \epsilon_{ij}}{\partial \boldsymbol{\gamma}} \epsilon_{ij} \right] \\ &= \sum_{j=1}^{m_i} \left[-\frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}} - \left(-\frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}} \epsilon_{ij} - \sum_{k=1}^{j-1} b_{ijk}\epsilon_{ik} \right) \epsilon_{ij} \right] \\ &= \sum_{j=1}^{m_i} \left[\frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}} (\epsilon_{ij}^2 - 1) + \epsilon_{ij} \sum_{k=1}^{j-1} b_{ijk}\epsilon_{ik} \right], \end{aligned}$$

where $b_{ijk} = \sum_{l=k}^{j} \frac{\partial T_{ilk}}{\partial \gamma} a_{ijl}$ with a_{ijl} being the (j,l) element of T_i^{-1} and

$$\frac{\partial T_{ijk}}{\partial \boldsymbol{\gamma}} = \begin{cases} T_{ijk}\{-\mathbf{w}_{ijk}\tan(\phi_{ijk}) + \sum_{l=1}^{k-1} \frac{\mathbf{w}_{ijl}}{\tan(\phi_{ijl})}\}, & k < j \\ T_{ijk} \sum_{l=1}^{k-1} \frac{\mathbf{w}_{ijl}}{\tan(\phi_{ijl})}, & k = j \end{cases}$$

$$\begin{aligned}
\frac{\partial l_i}{\partial \boldsymbol{\rho}} &= -\frac{1}{2}\sum_{j=1}^{m_i}\left[\Upsilon_{ij} + 2\frac{\partial \epsilon_{ij}}{\partial \boldsymbol{\rho}}\epsilon_{ij}\right] \\
&= \frac{1}{2}\left[-\sum_{j=1}^{m_i}\Upsilon_{ij} + \sum_{k=1}^{m_i}\Upsilon_{ik}\left(\sum_{j=k}^{m_i} a_{ijk}\frac{r_{ik}}{\sigma_{ik}}\epsilon_{ij}\right)\right] \\
&= \frac{1}{2}\left[-\sum_{j=1}^{m_i}\Upsilon_{ij} + \sum_{k=1}^{m_i}\Upsilon_{ik}\left(\sum_{j=k}^{m_i} a_{ijk}\frac{r_{ik}}{\sigma_{ik}}\left(\sum_{l=1}^{j} a_{ijl}\frac{r_{il}}{\sigma_{il}}\right)\right)\right] \\
&= \frac{1}{2}\boldsymbol{\Upsilon}_i'\left(\mathbf{q}_i - \mathbf{1}_{m_i}\right),
\end{aligned}$$

where $\mathbf{q}_i = \text{diag}\left(R_i^{-1}D_i^{-1}\mathbf{r}_i\mathbf{r}_i'D_i^{-1}\right)$.

Thus the estimating equations for $\boldsymbol{\theta}, \boldsymbol{\rho}$ and $\boldsymbol{\gamma}$ are

$$\begin{aligned}
\boldsymbol{U}_1 &= \sum_{i=1}^{n}\boldsymbol{\Pi}_i'\Delta_i\Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\Pi}_i\boldsymbol{\theta})) = 0, \\
\boldsymbol{U}_2 &= \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\Upsilon}_i'(\mathbf{q}_i - \mathbf{1}_{m_i}) = 0, \\
&\text{and} \\
\boldsymbol{U}_3 &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}}(\epsilon_{ij}^2 - 1) + \epsilon_{ij}\sum_{k=1}^{j-1} b_{ijk}\epsilon_{ik}\right] = 0
\end{aligned}$$

respectively, where $\Delta_i = \Delta_i(\boldsymbol{\Pi}_i\boldsymbol{\theta}) = \text{diag}\{\dot{g}^{-1}(\Pi_{i1}'\boldsymbol{\theta}), \ldots, \dot{g}^{-1}(\Pi_{im_i}'\boldsymbol{\theta})\}$, $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse link function $g^{-1}(\cdot)$, $\mu(\cdot) = g^{-1}(\cdot)$, $\mathbf{q}_i = \text{diag}(R_i^{-1}D_i^{-1}\mathbf{r}_i\mathbf{r}_i'D_i^{-1})$, and $b_{ijk} = \sum_{l=k}^{j}\frac{\partial T_{ilk}}{\partial \gamma}a_{ijl}$ with a_{ijl} being the (j,l) element of T_i^{-1}. As in Zhang *et al.* (2015) these equations are solved by the quasi-Fisher scoring algorithm which is described in Appendix A.

3.2.3 Estimation in Model 2 based on B-spline

If we consider a semiparametric model with only the mean having semi-parametric term as $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(t_{ij})$, $\log(\sigma^2_{ij}) = \mathbf{z}'_{ij}\boldsymbol{\lambda}$, and $\phi_{ijk} = \mathbf{w}'_{ijk}\boldsymbol{\gamma}$, then we need to estimate $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\gamma}$ which are obtained by solving

$$\boldsymbol{V}_1 = \sum_{i=1}^{n} \boldsymbol{\Pi}'_i \Delta_i \Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\Pi}_i\boldsymbol{\theta})) = 0,$$
$$\boldsymbol{V}_2 = \tfrac{1}{2}\sum_{i=1}^{n} \mathbf{Z}'_i(\mathbf{q}_i - \mathbf{1}_{m_i}) = 0,$$
and
$$\boldsymbol{V}_3 = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}}(\epsilon^2_{ij} - 1) + \epsilon_{ij}\sum_{k=1}^{j-1} b_{ijk}\epsilon_{ik}\right] = 0.$$

These equations can also be solved using the algorithm in Appendix A. At convergence the variance-covariance of $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{\boldsymbol{\gamma}}$ are obtained by inverting the Fisher information matrix given in Appendix A.

3.2.4 Penalized Spline

The B-spline methodology, in some applications, produces overfitting of the data (Carroll, Maca, and Ruppert, 1999). In such cases penalized spline (P-spline) has been used to overcome this (Eilers and Marx, 1996). So, here, we further use the penalized spline in Model 2 instead of the B-spline to see whether it produces improvement in estimation of the non-parametric functions. As in Section 3.2.3, $f_1(\cdot)$ can be approximated by $\boldsymbol{\pi}'(t)\boldsymbol{\alpha}$. Now, we impose a penalization upon the parameters $\alpha_1, \alpha_2, \ldots, \alpha_K$, so that they are constrained such that $\sum_{i=1}^{K} \alpha_i^2 \leq C$.

With this constraint the log-likelihood apart from a constant can be written as

$$-2l = \sum_{i=1}^{n} \left[\log|\Sigma_i| + (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\Pi}_i\boldsymbol{\theta}))'\Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\Pi}_i\boldsymbol{\theta}))\right] - \tau\boldsymbol{\theta}' D\boldsymbol{\theta},$$

where $\tau > 0$ is a constant and $D = \begin{bmatrix} 0_{p\times p} & 0_{p\times K} \\ 0_{K\times p} & I_{K\times K} \end{bmatrix}$.

Using Lagrange multipliers, the score equations for $\boldsymbol{\theta}, \boldsymbol{\lambda}$, and $\boldsymbol{\gamma}$ can be written as

$$\begin{aligned}
\boldsymbol{W}_1 &= -\sum_{i=1}^{n} \boldsymbol{\Pi}_i'\Delta_i\Sigma_i^{-1}\Delta_i(\mathbf{y}_i - \boldsymbol{\mu}_i) + \tau D\boldsymbol{\theta} = 0, \\
\boldsymbol{W}_2 &= \tfrac{1}{2}\sum_{i=1}^{n} \mathbf{Z}_i'(\mathbf{q}_i - \mathbf{1}_{m_i}) = 0, \\
&\text{and} \\
\boldsymbol{W}_3 &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\tfrac{\partial \log T_{ijj}}{\partial\boldsymbol{\gamma}}(\epsilon_{ij}^2 - 1) + \epsilon_{ij}\sum_{k=1}^{j-1} b_{ijk}\epsilon_{ik}\right] = 0.
\end{aligned}$$

All these equations can then be solved using the same algorithm as we used in Section 3.2.2. All of the block components of the Fisher information matrix remain the same as Model (3.2.3) except I_{11} which in this case is

$$I_{11} = -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \sum_{i=1}^{n} \boldsymbol{\Pi}_i'\Delta_i\Sigma_i^{-1}\Delta_i\boldsymbol{\Pi}_i + \tau D.$$

3.2.5 Knot Selection

The importance of knot selection in spline smoothing work has been well described in two pioneering papers by He *et al.* (2005) and Leng *et al.* (2010). These authors found that knot selection is less critical for the estimation of $\boldsymbol{\beta}$ than for the estimation of the nonparametric functions involved in model 2 and model 3 discussed in Section 3.1.

However, in most situations, they found it appropriate to use the sample quantiles of $\{t_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i\}$ as knots. We follow their suggestion and note that the number of knots is not prespecified, rather, it depends on the total sample size $N = \sum_{i=1}^{n} m_i$. Through a detailed asymptotic theoretical study, Leng *et al.* (2010) show that the number of internal knots to be used is the integer part of $N^{1/5}$.

3.3 Simulation Study

As indicated in Section 3.1, an extensive simulation investigation is conducted in this section. Our purpose in this simulation, in addition to the study of the performance of the estimators of the regression parameters in terms of bias and efficiency, is to study the effect of fixing the number of knots and the effect of misspecifying the error distribution (robustness study). These simulations are performed in Sections 3.3.1, 3.3.2, and 3.3.3. A further study, in Section 3.3.4, is conducted to compare performance of the estimation methods using B-spline and penalized spline.

3.3.1 Study 1: Properties of the regression parameters

For this purpose we generate response data from each of the 3 models (Model 1, Model 2, and Model 3)

$$\begin{aligned} y_{ij} &= x_{ij1}\beta_1 + x_{ij2}\beta_2 + e_{ij}, \quad \log(\sigma_{ij}^2) = z_{ij1}\lambda_1 + z_{ij2}\lambda_2; \\ y_{ij} &= x_{ij1}\beta_1 + x_{ij2}\beta_2 + f_1(t_{ij}) + e_{ij}, \quad \log(\sigma_{ij}^2) = z_{ij1}\lambda_1 + z_{ij2}\lambda_2; \end{aligned}$$

and

$$y_{ij} = x_{ij1}\beta_1 + x_{ij2}\beta_2 + f_1(t_{ij}) + e_{ij}, \quad \log(\sigma_{ij}^2) = z_{ij1}\lambda_1 + z_{ij2}\lambda_2 + f_2(t_{ij}).$$

For each model, values of parameters considered were $(\beta_1, \beta_2) = (1, 0.5)$, $(\gamma_1, \gamma_2) = (0.35, 0.5)$ and $(\lambda_1, \lambda_2) = (-0.5, 0.2)$. Following Leng *et al.* (2010) we generate the observation times as in what follows.

For each individual we consider a set of scheduled time points $\{0, 1, 2, \ldots, 12\}$. At each scheduled time, except time 0, each individual has a 20% probability of missing a fixed time point. To make it irregular and unequal time distances for different individuals a uniform $[0, 1]$ random variable is added to a non skipped scheduled time. This results in different observed time points t_{ij} per subject. However, t_{ij} is transformed onto $[0, 1]$ while analysis.

For covariates, we take $x_{ij1} = t_{ij} + \delta_{ij}$, where δ_{ij} follows the standard normal distribution and x_{ij2} is generated from a Bernoulli(0.5) distribution. The nonparametric functions are taken as $f_1(t) = \cos(\pi t)$, and $f_2(t) = \sin(\pi t)$. The error $(e_{i1}, \ldots, e_{im_i})$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\Sigma_i = D_i R_i D_i$, where $R_i = T_i T_i'$ with $\mathbf{w}_{ijk} = (1, t_{ij} - t_{ik})'$, $\mathbf{z}_{ij} = \mathbf{x}_{ij}$. The expected sample size (for the calculation of the number of knots) is about 1040 (=100×13×0.8). The number of knots is taken to be $4 \approx 1040^{1/5}$ (He *et al.*, 2005). Here, as can be seen, the number of knots is not prespecified.

Bias of the estimates of the parameters of all three models along with their standard errors, and MSE of the non-parametric functions f_1 and f_2, based on 1000 replications, are given in Table 3.1. Figure 3.1 displays the true and the fitted curves for the non parametric functions f_1 and f_2 of Model 3. Figure 3.2 displays the true and the fitted curve for the non parametric function f_1 of Model 2.

Table 3.1 shows that our semiparametric methods yield similar bias property of the estimates as compared to that for the parametric model. Both the functions yield

Table 3.1: Bias and standard error of the estimated parameters based on 1000 replications

Parameter	True value	Parametric (Model 1)		Semiparametric with mean (Model 2)		Semiparametric with mean and var (Model 3)	
		Bias	SE	Bias	SE	Bias	SE
β_1	1.0	0.0000	0.0000	0.0000	0.0002	0.0001	0.0003
β_2	0.5	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001
γ_1	0.35	0.0003	0.0013	0.0066	0.0014	0.0031	0.0023
γ_2	0.5	0.0004	0.0023	0.0066	0.0025	0.0056	0.0041
λ_1	-0.5	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001
λ_2	0.2	0.0000	0.0002	0.0000	0.0002	0.0000	0.0002
MSE($\hat{f}_1$)					0.0119		0.0109
MSE($\hat{f}_2$)							0.1129

small MSE, which along with Figure 3.1 and Figure 3.2, show that both the true and the fitted curves are very close in both models.

Figure 3.1: Nonparametric functions (green) and their fitted curves (blue) of Model 3

3.3.2 Study 2: Properties of the regression parameters and the functions f_1 and f_2 when number of knots are fixed

We carry out a similar simulation study as in Study 1 using Model 3. However in this study, the number of knots is prespecified. We redo the simulations of Section 3.1 by fixing the number of knots as $k_n = 10$ and $k_n = 20$. Table 3.2 summarizes the

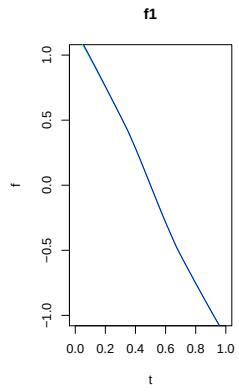

Figure 3.2: Nonparametric function (green) and the fitted curve (blue) of Model 2

results which show that as the number of knots increase, the MSE of the estimated functions f_1 and f_2, bias and standard error of the estimates of the parameters also increase.

Table 3.2: Bias and standard error of the estimated parameters based on 1000 replications when knot is prespecified

Parameter	True value	Semiparametric with mean and var when knot=10		Semiparametric with mean and var when knot=20	
		Bias	SE	Bias	SE
β_1	1.0	0.000	0.0002	0.000	0.0002
β_2	0.5	0.000	0.0001	0.000	0.0001
γ_1	0.35	-0.0215	0.0024	-0.0419	0.0034
γ_2	0.5	-0.0036	0.0041	-0.0057	0.0050
λ_1	-0.5	0.000	0.0001	0.000	0.0001
λ_2	0.2	0.000	0.0001	0.000	0.0001
MSE($\hat{f}_1$)			0.4966		0.5164
MSE($\hat{f}_2$)			0.3740		0.6423

3.3.3 Study 3: A robustness study. Properties of the regression parameters and the functions f_1 and f_2 when error follows mixture of normal distributions

Again, another study, similar to what was done in study 1, has been conducted, in which, for generating the response data of Model 2 and Model 3, we consider a mixture of two multivariate normal distributions with error distributions $N_{m_i}(\mathbf{0}, \Sigma_i)$ and $N_{m_i}(\mathbf{0}, 0.25^2\Sigma_i)$ with equal probability. The results for the simulation study are displayed in Table 3.3 which show that the bias and the standard errors of the estimates, and the MSE of f_1 remain almost the same as those in study 1. However, MSE of the fitted function f_2 in Model 3 increases significantly.

In summary, the mixed modelling affects only the MSE of the estimate of f_2 in the semiparametric Model 3.

Table 3.3: Simulation results for Study 3 in Model 2 and Model 3 over 1000 replications when error terms follow mixture of normal distribution; n=100

Parameter	True value	Semiparametric (Model 2) (without mixed)		Semiparametric (Model 2) (with mixed)		Semiparametric (Model 3) (without mixed)		Semiparametric (Model 3) (with mixed)	
		Bias	SE	Bias	SE				
β_1	1.0	0.0000	0.0002	0.0000	0.0002	0.0001	0.0003	0.0001	0.0002
β_2	0.5	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001
γ_1	0.35	0.0066	0.0049	-0.0032	0.0063	0.0031	0.0023	0.0088	0.0024
γ_2	0.5	0.0066	0.0050	-0.0086	0.0082	0.0056	0.0041	0.0135	0.0044
λ_1	-0.5	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001	0.000	0.0001
λ_2	0.2	0.0000	0.0002	0.0000	0.0001	0.0000	0.0002	0.000	0.0002
MSE($\hat{f}_1$)			0.0119		0.0119		0.0109		0.0110
MSE($\hat{f}_2$)							0.1129		0.6935

3.3.4 Study 4: Comparison of using B-spline and penalized spline in Model 2

A further study is conducted to compare the B-spline and the penalized spline to estimate the nonparametric function f_1 in Model 2. The estimation procedure to estimate all parameters are discussed in Section 3.2.4. To generate the response variable we consider the same mean and variance models as those in study 1 and the results are presented in Table 3.4. Further, Figure 3.3 displays the fitted curve of function f_1 under both the B-spline and the penalized spline. Results in Table 3.4 and Figure 3.3 show no advantage of using the penalized spline over the B-spline.

Table 3.4: Bias ans standard error of the estimated parameters in Model 2 based on 1000 replications using B-spline and penalized spline

Parameter	True value	Semiparametric Model 2			
		B-spline		Penalized spline	
		Bias	SE	Bias	SE
β_1	1.0	0.0000	0.0002	0.0000	0.0002
β_2	0.5	0.0000	0.0001	0.0000	0.0001
γ_1	0.35	0.0066	0.0014	0.0066	0.0014
γ_2	0.5	0.0066	0.0025	0.0065	0.0025
λ_1	-0.5	0.0000	0.0001	0.0000	0.0001
λ_2	0.2	0.0000	0.0002	0.0000	0.0002
MSE($\hat{f}_1$)			0.0119		0.0119

Figure 3.3: Nonparametric functions (green) and their fitted curves (blue)

3.4 Analysis of Three Real Data Sets

In this Section we analyze three real data sets, namely a CD4 cell data set, a set of cattle data, and a Progesterone hormone data set. The first two data sets are available in jmcm package in R and third data set is available at the web site https://content.sph.harvard.edu/fitzmaur/ala2e/. All of these data sets are analysed by using the semiparametric Model 3.

3.4.1 Analysis of CD4 cell data

The data comprise CD4 cell counts of 369 HIV-infected men. In total there are 2376 observations with multiple repeated measurements taken for each individual at different times, covering a period of approximately eight and a half years. The number of measurements for each individual varies from 1 to 12 taken at unequally spaced time points.

This data set has been analysed by others in the past (for example, Zeger and Diggle, 1994 and Ye and Pan, 2006). Most recently Zhang *et al.* (2015), in order to jointly model the mean, correlation and variance structures, fit polynomial regressions of time

$$\begin{aligned} g(\mu_{ij}) &= \beta_0 + x_{ij}\beta_1 + \cdots + x_{ij}^p\beta_p, \\ \phi_{ijk} &= \gamma_0 + w_{ij}\gamma_1 + \cdots + w_{ij}^q\gamma_q, \\ \log(\sigma_{ij}^2) &= \lambda_0 + z_{ij}\lambda_1 + \cdots + z_{ij}^d\lambda_d, \end{aligned}$$

where $w_{ijk} = t_{ij} - t_{ik}$.

They found in terms of Bayesian Information Criterion

$$\text{BIC}(p,q,d) = -2\hat{l}_{\max}/n + (p+q+d+3)\log(n)/n,$$

where $\hat{l}_{\max}$ is the maximum of the log-likelihood, that the model with $(p,q,d) = (8,1,1)$ is the most parsimonious having 13 parameters including the intercept terms β_0, γ_0 and λ_0 with $\hat{l}_{max} = -4892.72$ and BIC =26.72.

We now analyze these data using our method of semiparametric modelling developed in Section 3.2.2. The results, in terms of $\hat{l}_{max}$ and BIC of the parametric and the semiparametric models, are given in Table 3.5. Note that here we provide all relevant information of the 6 most parsimonious models of which the most parsimonious model has $(p,q,d) = (4,1,1)$ with $\hat{l}_{max} = -4877.41$ and BIC =26.58. This model has 9 parameters as opposed to 13 parameters of the model obtained by Zhang *et al.* (2015).

Table 3.5: CD4 cell data: A comparison of various models using parametric (Zhang *et al.*, 2015) and semiparametric approaches

(p,q,d)	No. of par.	Parametric		Semiparametric	
		$\hat{l}_{\max}$	BIC	$\hat{l}_{\max}$	BIC
(4,1,1)	9	-4910.87	26.76	-4877.41	26.58
(3,1,1)	8	-4926.88	26.83	-4882.11	26.59
(8,1,1)	13	-4892.72	26.72	-4874.40	26.63
(8,3,1)	15	-4890.44	26.75	-4871.66	26.64
(3,3,3)	12	-4919.52	26.85	-4879.37	26.64
(8,3,3)	17	-4886.36	26.76	-4872.74	26.68

3.4.2 Analysis of cattle data

This data set consists of 30 cows' 11 bi-weekly weight measurements over a 133-day period to study the effect of treatments on intestinal parasites. Here measurement times were common across animals and no observations were missing. Several authors

analyzed these data; see for example, Pourahmadi (1999, 2000), Pan and MacKenzie (2003). Recently Zhang *et al.* (2015) also analyzed these data using the polynomials given in Section 3.4.1 and obtained the most parsimonious model having $(p, q, d) = (8, 2, 2)$ with $\hat{l}_{max} = -754.02$ and BIC = 51.97. We analyze these data by our proposed method Model 3 and find that the most parsimonious model has $(p, q, d) = (5, 2, 1)$ with $\hat{l}_{max} = -749.53$ and BIC = 51.22 (see Table 5).

Table 3.6: Cattle data: A comparison of various models using parametric (Zhang *et al.*, 2015) and semiparametric approaches

(p,q,d)	No. of par.	Parametric		Semiparametric	
		$\hat{l}_{\max}$	BIC	$\hat{l}_{\max}$	BIC
(5,2,1)	11	-762.81	52.10	-749.53	51.22
(5,2,2)	12	-760.61	52.07	-749.56	51.33
(8,2,2)	15	-755.00	52.03	-754.02	51.97
(9,3,1)	16	-756.92	52.28	-755.83	52.20
(9,3,4)	19	-752.82	52.34	-752.03	52.29
(7,2,2)	14	-761.67	52.37	-760.37	52.28
(8,4,7)	22	-749.90	52.49	-749.19	52.44

3.4.3 Analysis of Progesterone hormone data

These data consist of repeated progesterone metabolite (pregnanediol-3-glucuronide, PdG) measures from day -8 to day 15 in the menstrual cycle (day 0 denotes ovulation day) on a sample of 22 conceptive cycles from 22 women and 29 non-conceptive cycles from another 29 women to study of early pregnancy loss. Altogether 1130 observations were obtained from 51 women, with each woman contributing 9 to 24 observations over time.

As in Brumback and Rice (1998), we take a log transformation of these data to make the normality assumption reasonable. Analysis of these data show that the most parsimonious model, using the semiparametric Model 3, has $(p, q, d) = (2, 2, 1)$ with $\hat{l}_{max} = 31.34$ and BIC = -0.61 (see Table 3.7). This compares favourably, compared

to the parametric Model 1, which has $(p, q, d) = (7, 3, 7)$ with $\hat{l}_{max} = 27.31$ and BIC = 0.47 (see Table 3.7).

Table 3.7: Progesterone hormone data: A comparison of various models using parametric (Zhang *et al.* [1]) and semiparametric approaches

(p, q, d)	No. of par.	Parametric		Semiparametric	
		$\hat{l}_{\max}$	BIC	$\hat{l}_{\max}$	BIC
(2,2,1)	8	-71.66	3.43	31.34	-0.61
(2,2,2)	9	-66.59	3.31	31.34	-0.54
(2,1,1)	7	-129.58	5.62	16.01	-0.09
(2,1,4)	10	-66.98	3.40	16.38	0.12
(3,1,4)	11	-28.97	1.98	16.38	0.21
(3,1,5)	12	-28.20	2.03	19.58	0.16
(2,1,6)	12	-65.84	3.51	23.73	-0.005
(3,2,3)	11	-32.85	2.14	31.34	-0.38
(7,3,7)	20	27.31	0.47	42.06	-0.11

3.5 Discussion

We develop a joint estimation procedure for the mean (regression) and the variance parameters in longitudinal data using semiparametric modelling of the mean and the variance, regression spline, and by decomposing the correlation matrix via hyperspherical co-ordinates. Through an extensive simulation study we compare our method with the parametric method by Zhang *et al.* (2015). Further, the effect of the misspecification of the error distribution and of the number of knots used in the estimation of the nonparametric functions, and whether the penalized spline procedure improves the estimation of the nonparametric functions over the B-spline are investigated. Furthermore three real data sets are analysed.

The main findings of the simulation study are: (a) the parametric modelling and the semiparametric modelling produce similar bias and efficiency property of the regression parameters, (b) increasing the number of knots in the spline procedure decreases

the efficiency of the estimates of the nonparametric functions, and (c) use of the penalized spline does not improve the efficiency of the estimates of the nonparametric functions.

The main advantage of the semiparametric modelling, however, is shown in the analysis of three real data sets, the results of which are summarized in Table 3.8, which produces significant model parsimony. For example, in the Cattle data the most parsimonious model given by Zhang *et al.* (2015) shows $(p, q, d) = (8, 2, 2)$, where p = degree of polynomial in mean, q = degree of polynomial in correlation in time lag and d = degree of polynomial in variance. Whereas, our semiparametric approach shows the most parsimonious model to have $(p, q, d) = (5, 2, 1)$. Similar parsimony advantages are seen in the analysis of the CD4 cell data and the Progesterone hormone data.

Table 3.8: Most parsimonious model with number of parameters

Data	Parametric		Semiparametric	
	(p, q, d)	Number of parameters	(p, q, d)	Number of parameters
CD4	(8,1,1)	13	(4,1,1)	9
Cattle	(8,2,2)	15	(5,2,1)	11
Hormone	(7,3,7)	20	(2,2,1)	8

Chapter 4

Joint Estimation of Mean and Covariance Parameters in Generalized Partially Linear Varying Coefficient Models for Longitudinal Data

4.1 Introduction

In longitudinal data regression modelling, in some instances, time variant regression coefficients play an important role. For example, consider a subset of data from the Multi-Center AIDS Cohort study (Kaslow *et al.*, 1987), analyzed by Qin, Mao, and Zhu (2015), that includes repeated measurements of physical examinations, labora-

tory results and CD4 cell percentage along some covariates of 283 homosexual men who became HIV-positive between 1984 and 1991.

The response variable $y(t)$ is the CD4 cell percentage of a subject at distinct time points after HIV infection. Huang, Wu, and Zhou (2002) took three covariates: smoking status, age and PreCD4. They analyze this data set to describe the trend of mean CD4 percentage depletion over time and to evaluate the effects of cigarette smoking, pre-HIV infection CD4 percentage and age at HIV infection on the mean CD4 percentage after the infection. They consider the following model

$$y(t) = \beta_0(t) + \beta_1(t)\text{smoking} + \beta_2(t)\text{age} + \beta_3(t)\text{preCD4} + \epsilon,$$

where $\beta_0(t)$ can be interpreted as the baseline function.

The results of the hypothesis testing of Huang *et al.* (2002) show that the baseline function varies over time; neither smoking nor age has a significant impact on the mean CD4 percentage, and whether or not PreCD4 has a constant effect over time is unclear which motivate us to consider a time varying coefficient model.

In a recent paper Zhang, Leng, and Tang (2015) consider that covariate effects are time invariant, and we extend their model to include time variant covariates effects. Qin, Mao, and Zhu (2015) proposed a general semiparametric model for the mean and the covariance simultaneously using the modified Cholesky decomposition. We consider a generalized partially linear varying coefficient model (GPLVCM) and propose a regression spline based approach model to estimate mean and variance jointly by decomposing the correlation matrix via hyperspherical coordinates.

In Section 4.2 we describe how to decompose the correlation matrix by using hyper-

spherical coordinates. Furthermore, approximation of time variant terms by B-spline basis is given in Section 4.3. The proposed model and estimation procedure are presented in Section 4.4. A simulation study and real data analysis are given in Sections 4.5 and 4.6, respectively.

4.2 Decomposing Correlation Matrix via Hyperspherical Coordinates

A correlation matrix can be decomposed by hyperspherical coordinates. The following Lemma (Pourahmadi and Wang, 2015) shows the relationship between Cholesky decomposition of order $m \times m$ correlation matrix $R = TT'$ and the hyperspherical parameterization of $T = (T_{jk})$.

Lemma: (a) A positive definite correlation matrix $R = (\rho_{jk})$ can be decomposed as $R = TT'$ where T is a lower triangular matrix given by

$$T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{21} & s_{21} & 0 & \cdots & 0 \\ c_{31} & c_{32}s_{31} & s_{32}s_{31} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2}s_{m1} & c_{m3}s_{m2}s_{m1} & \cdots & \prod_{l=1}^{m-1} s_{ml} \end{pmatrix},$$

where c_{jk} and s_{jk} represent $\cos(\phi_{jk})$ and $\sin(\phi_{jk})$ respectively,

(b) The matrix T is unique if its diagonal entries are positive or equivalently if the angles are restricted to the range $(0, \pi)$.

Note that there are $m(m-1)/2$ unique angles ϕ_{jk} which vary freely in the range $(0, \pi)$. For more details readers are referred to Pinheiro and Bates (1996), Rebonato and Jackel (2000), and Rapisarda *et al.* (2007). Thus the elements of R are of the form

$$\rho_{jk} = c_{j1}c_{k1} + \sum_{l=2}^{j-1} c_{jl}c_{kl} \prod_{p=1}^{l-1} s_{jp}s_{kp} + c_{kj} \prod_{p=1}^{j-1} s_{jp}s_{kp}, \quad 1 \leq j < k \leq m.$$

The advantage of this decomposition is that we can directly parametrize the correlation matrix unconstrainedly using the hyperspherical coordinates which avoids the need to go through the concepts of partial correlations.

4.3 B-spline

The unknown function $f(t)$ can be approximated by a regression spline. Note that a regression spline is a piecewise polynomial function smoothly connected at its knots. Without loss of generality, we assume that the domain of t_{ij} lies in the interval $[0, 1]$. Let $0 = a_0 < a_1 < \cdots < a_{k_n} < a_{k_{n+1}} = 1$ be a partition of the interval $[0, 1]$. Using $\{a_i\}$ as the internal knots, we have $K = k_n + \ell$ normalized B-spline basis functions of order ℓ as $\left\{B_1^{(\ell)}(t), \ldots, B_K^{(\ell)}(t)\right\}$ that form a basis for the linear spline space where $B_i^{(\ell)}(t)$ is defined as

$$B_i^{(\ell)}(t) = \frac{t - a_i}{a_{i+\ell-1} - a_i} B_i^{(\ell-1)}(t) + \frac{a_{i+\ell} - t}{a_{i+\ell} - a_{i+1}} B_{i+1}^{(\ell-1)}(t),$$

and

$$B_i^{(1)}(t) = \begin{cases} 1, & a_i \leq t < a_{i+1} \\ 0, & \text{otherwise.} \end{cases}$$

Thus $f(t)$ is approximated by $\pi(t)'\alpha$ where $\pi(t) = \left(B_1^{(\ell)}(t), \ldots, B_K^{(\ell)}(t)\right)'$ and $\alpha \in \mathbb{R}^K$ is the vector of spline coefficients.

The advantage of regression spline is that it linearizes the nonparametric function so any algorithm designed for linear models can be directly applied to partially linear models. Moreover the B-spline can provide a good approximation with a small number of knots (He *et al.*, 2002).

We consider a cubic spline of order 4 as it is usually smooth enough to fit usual smooth functions which is extensively used in practice (Wolberg and Alfy, 2002). Similar to He *et al.* (2005), the number of the internal knots k_n is taken to be the integer part of $N^{1/5}$ where N is the total number of distinct values of $\{t_{ij}\}$.

4.4 Generalized Partially Linear Varying Coefficient Models

Suppose for subject $i (i = 1, \ldots, n)$, $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})'$ is collected at time $\mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})'$. Each observation consists of a response variable y_{ij} and covariate vectors $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$, which are taken from the i-th subject at time t_{ij}. We denote the conditional mean and variance of y_{ij} by μ_{ij} and σ_{ij}^2 respectively given the covariates $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ at time t_{ij}.

We assume that $\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})'$, $\Sigma_i = D_i R_i D_i$, $D_i = \text{diag}\,(\sigma_{i1}, \ldots, \sigma_{im_i})$, and $R_i = (\rho_{ijk})_{j,k=1}^{m_i}$ is the correlation matrix of $\mathbf{y}_i$ where $\rho_{ijk} = \text{corr}(y_{ij}, y_{ik})$ is the correlation between the j-th and k-th measurements of the i-th subject.

Similar to Zhang *et al.* (2015), we decompose the correlation matrix R_i via hyper-spherical coordinates to decompose $R_i = T_i T_i'$ as discussed in Section 4.2. Since μ_{ij}, ϕ_{ijk} and $\log \sigma_{ij}^2$ are unconstrained, motivated by Qin *et al.* (2015), we propose the following generalized partially linear varying coefficient model (GPLVCM):

$$\begin{aligned} g(\mu_{ij}) &= \mathbf{x}_{ij}'\boldsymbol{\alpha}(t_{ij}) + \mathbf{z}_{ij}'\boldsymbol{\beta}, \\ \phi_{ijk} &= \mathbf{w}_{ijk}'\boldsymbol{\gamma}, \\ &\text{and} \\ \log(\sigma_{ij}^2) &= \mathbf{u}_{ij}'\mathbf{f}(t_{ij}) + \mathbf{v}_{ij}'\boldsymbol{\lambda}, \end{aligned}$$

where $\mathbf{x}_{ij} \in \mathbb{R}^p$ and $\mathbf{z}_{ij} \in \mathbb{R}^q$ are covariate vectors for the time varying coefficients and constant coefficients at time t_{ij}, respectively; $\mathbf{w}_{ijk}, \mathbf{u}_{ij}$ and $\mathbf{v}_{ij}$ are $(d \times 1), (h \times 1)$ and $(m \times 1)$ vectors of covariates, respectively; $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are the regression coefficients; $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_p(t))'$ and $\mathbf{f}(t) = (f_1(t), \ldots, f_h(t))'$ comprises p and h unknown smooth functions respectively. Zhang *et al.* (2015) method is a special case of this model.

For simplicity, we assume that all $\alpha_l(t)$ and $f_s(t)$ have the same smoothness property for $1 \leq l \leq p$ and $1 \leq s \leq h$. Using the procedure of regression spline as discussed in Section 4.3, all unknown smooth functions $\alpha_l(t)$ and $f_s(t)$ are approximated by regression spline $\pi'(t)\psi_l$ and $\pi'(t)\varphi_s$, respectively, where $\pi(t) = (B_1(t), \ldots, B_K(t))'$ is the vector of basis functions, and $\psi_l = (\psi_{l,1}, \ldots, \psi_{l,K})$ and $\varphi_l = (\varphi_{l,1}, \ldots, \varphi_{l,K})$ are the spline coefficient vectors. Then, the nonlinear regression models can be linearized

as in what follows

$$g\left(\mu_{ij}\right) = \Pi_{ij}'\Psi + \mathbf{z}_{ij}'\boldsymbol{\beta} = \tilde{\Pi}_{ij}'\Theta, \quad \phi_{ijk} = \mathbf{w}_{ijk}'\boldsymbol{\gamma},$$

and

$$\log\left(\sigma_{ij}^2\right) = \Upsilon_{ij}'\tilde{\Psi} + \mathbf{v}_{ij}'\boldsymbol{\lambda} = \tilde{\Upsilon}_{ij}'\Lambda,$$

where $\Pi_{ij} = (x_{ij,1}\pi'(t_{ij}), \ldots, x_{ij,p}\pi'(t_{ij}))', \tilde{\Pi}_{ij} = (\Pi_{ij}', \mathbf{z}_{ij}')', \Theta = (\Psi', \boldsymbol{\beta}')', \Psi = (\psi_1', \ldots, \psi_p')' \in \mathbb{R}^{pK}$, $\Upsilon_{ij} = (u_{ij,1}\pi'(t_{ij}), \ldots, u_{ij,h}\pi'(t_{ij}))', \tilde{\Upsilon}_{ij} = (\Upsilon_{ij}', \mathbf{v}_{ij}')', \Lambda = (\tilde{\Psi}', \boldsymbol{\lambda}')', \tilde{\Psi} = (\varphi_1', \ldots, \varphi_h')' \in \mathbb{R}^{hK}$. Define $\tilde{\boldsymbol{\Pi}}_i = \left(\tilde{\Pi}_{i1}', \ldots, \tilde{\Pi}_{im_i}'\right)'$ and $\tilde{\boldsymbol{\Upsilon}}_i = \left(\tilde{\Upsilon}_{i1}', \ldots, \tilde{\Upsilon}_{im_i}'\right)'$ for $i = 1, \ldots, n$.

Let $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$. Then, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})' = T_i^{-1}D_i^{-1}\mathbf{r}_i \sim N(\mathbf{0}, I_{m_i})$. Thus the minus twice log-likelihood, up to a constant, can be written as

$$-2l = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\log\sigma_{ij}^2 + \log T_{ijj}^2 + \epsilon_{ij}^2).$$

Thus, now the parameters of interest are Θ, $\boldsymbol{\gamma}$ and Λ. Omitting details, by usual derivations, the maximum likelihood estimating equations for Θ, $\boldsymbol{\gamma}$ and Λ are

$$\boldsymbol{U}_1 = \sum_{i=1}^{n}\tilde{\boldsymbol{\Pi}}_i'\Delta_i\Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

$$\boldsymbol{U}_2 = -\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}(\epsilon_{ij}^2 - 1) + \epsilon_{ij}\sum_{k=1}^{j-1}b_{ijk}\epsilon_{ik}\right] = 0,$$

and

$$\boldsymbol{U}_3 = \frac{1}{2}\sum_{i=1}^{n}\tilde{\boldsymbol{\Upsilon}}_i'(\mathbf{q}_i - \mathbf{1}_{m_i}) = 0,$$

where $b_{ijk} = \sum_{l=k}^{j}\frac{\partial T_{ilk}}{\partial\gamma}a_{ijl}$ with a_{ijl} being the (j,l) element of T_i^{-1}, $\mathbf{q}_i = \mathrm{diag}(R_i^{-1}D_i^{-1}\mathbf{r}_i\mathbf{r}_i'D_i^{-1})$,

$\Delta_i = \Delta_i(\boldsymbol{\mu}_i(\Theta)) = \text{diag}\{\dot{g}^{-1}(\tilde{\boldsymbol{\Pi}}'_{i1}\Theta), \ldots, \dot{g}^{-1}(\tilde{\boldsymbol{\Pi}}'_{im_i}\Theta)\}$ where $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse link function $g^{-1}(\cdot)$ with $\mu(\cdot) = g^{-1}(\cdot)$. Note that the notation $\sum_{k=1}^{0}$ represents zero throughout the Chapter when $j = 1$.

To solve aforementioned score equations we apply the following quasi-Fisher scoring algorithm where the parameters of interest Θ, Λ and $\boldsymbol{\gamma}$ are solved sequentially one by one by keeping fixed other parameters in optimization:

Step 1 : Select initial values of the parameters as $\Theta^{(0)}, \Lambda^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$. Set $k = 0$

Step 2 : Evaluate Σ_i by using $\Lambda^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$. Update Θ as

$$\Theta^{(k+1)} = \Theta^{(k)} + I_{11}^{-1}\boldsymbol{U}_1|_{\Theta=\Theta^{(k)}}$$

Step 3 : Given $\Theta = \Theta^{(k+1)}$, update Λ and $\boldsymbol{\gamma}$ by using

$$\begin{pmatrix} \boldsymbol{\gamma}^{(k+1)} \\ \Lambda^{(k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}^{(k)} \\ \Lambda^{(k} \end{pmatrix} + \left[\begin{pmatrix} I_{22} & I_{23} \\ I_{32} & I_{33} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{U}_2 \\ \boldsymbol{U}_3 \end{pmatrix} \right] \Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(k)},\ \Lambda=\Lambda^{(k)}}$$

Step 4 : Set $k \leftarrow k + 1$ and repeat steps 2 and 3 until a preferred convergence criteria is satisfied.

It is noted that block components of Fisher information matrix I are I_{11}, I_{12}, I_{13},

I_{22}, I_{23} and I_{33} which are of the form

$$
\begin{aligned}
I_{11} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}'}\right] = \sum_{i=1}^{n}\tilde{\boldsymbol{\Pi}}_i'\Delta_i\Sigma_i^{-1}\Delta_i\tilde{\boldsymbol{\Pi}}_i,\\
I_{12} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\gamma}'}\right] = -\sum_{i=1}^{n}\left[\tilde{\boldsymbol{\Pi}}_i'\Delta_i\frac{\partial\Sigma_i^{-1}}{\partial\boldsymbol{\gamma}'}(E(\mathbf{y}_i)-\boldsymbol{\mu}_i)\right] = 0,\\
I_{13} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\Lambda}'}\right] = 0,\\
I_{22} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[2\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}'}+\sum_{k=1}^{j-1}b_{ijk}b_{ijk}'\right],\\
I_{23} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\Lambda}'}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}\tilde{\Upsilon}_{ij}'+\frac{1}{2}\sum_{k=1}^{j-1}b_{ijk}\sum_{l=k}^{j}a_{ijl}T_{ilk}\tilde{\Upsilon}_{il}'\right],\\
I_{33} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\Lambda}\partial\boldsymbol{\Lambda}'}\right] = \frac{1}{4}\sum_{i=1}^{n}\tilde{\boldsymbol{\Upsilon}}_i'\left[I_{m_i}+R_i^{-1}\circ R_i\right]\tilde{\boldsymbol{\Upsilon}}_i,
\end{aligned}
$$

where '$\circ$' represents the Hadamard product.

4.5 Simulation

A simulation study is conducted in this section to investigate the performance of the estimators of the regression parameters in terms of bias and efficiency. We generate response data from the model

$$y_{ij} = x_{ij1}\alpha_1(t_{ij}) + x_{ij2}\alpha_2(t_{ij}) + z_{ij1}\beta_1 + z_{ij2}\beta_2 + e_{ij},$$

$$\phi_{ijk} = \gamma_1 w_{ijk1} + \gamma_2 w_{ijk2},$$

and

$$\log\left(\sigma_{ij}^2\right) = u_{ij1}f_1(t_{ij}) + u_{ij2}f_2(t_{ij}) + v_{ij1}\lambda_1 + v_{ij2}\lambda_2,$$

for $i = 1, \ldots, 100$, $j, k = 1, \ldots, m_i$, $\alpha_1(t) = \sqrt{t}, \alpha_2(t) = \sin(2\pi t)$, $f_1(t) = 0.5\sin(\pi t)$ and $f_2(t) = 0.5\cos(\pi t)$.

Observation times are in general scheduled but may be randomly missed in practice. Similar to Leng *et al.* (2010) we generate the observation times in the following way. We take a set of scheduled time points $\{0, 1, 2, \ldots, 12\}$ for each subject. At each scheduled time, except time 0, each subject has a 20% probability of missing that fixed time point. To make irregular and unequal time distances for different subjects, a uniform $[0, 1]$ random variable is added to a non skipped scheduled time which produces different observed time points t_{ij} per subject. Note that t_{ij} is converted onto $[0, 1]$ in simulation.

We further take $x_{ij1} = 1$ to include an intercept term. Other covariates are chosen as follows: for a given t_{ij}, $(x_{ij2}, \delta_{ij})'$ is generated from a bivariate normal distribution with mean $\mathbf{0}$, marginal variance 1 and correlation 0.5, and z_{ij2} follows a Bernoulli(0.5) distribution and is independent of x_{ij2} and z_{ij1}. We take $z_{ij1} = \delta_{ij} + t_{ij}$. The error $(e_{i1}, \ldots, e_{im_i})$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\Sigma_i = D_i R_i D_i$, where D_i and R_i are described in Section 4.2. Take $\mathbf{w}_{ijk} = (1, t_{ij} - t_{ik})', \mathbf{u}_{ij} = \mathbf{x}_{ij}, \mathbf{v}_{ij} = \mathbf{z}_{ij}$. For values of regression parameters we take $(\beta_1, \beta_2) = (1, 2), (\gamma_1, \gamma_2) = (0.5, 0.6)$ and $(\lambda_1, \lambda_2) = (-0.5, 0.2)$. The expected sample size is about $1040 = 100 \times 13 \times (1 - 0.2)$. The number of knots is taken to be $4 \approx 1040^{1/5}$ (He *et al.*, 2005).

We generate 1000 data sets to calculate the bias of the estimates of the parameters of our proposed method with standard errors which is presented in Table 4.1 and MSE of α_1, α_2, f_1 and f_2 are presented in Table 4.2.

Table 4.1 and Table 4.2 show that proposed method provides efficient estimates

Table 4.1: Bias and Standard error of the estimated parameters based on 1000 replications

Parameter	True value	Sample mean	Bias	Standard error
β_1	1.0	1.0000	0.0000	0.0214
β_2	2.0	2.0001	0.0001	0.0256
γ_1	0.5	0.4994	-0.0006	0.0071
γ_2	0.6	0.5986	-0.0014	0.0071
λ_1	-0.5	-0.5000	0.0000	0.0286
λ_2	0.2	0.2001	0.0001	0.0496

Table 4.2: MSE of time varying functions

Function	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{f}_1$	$\hat{f}_2$
MSE	0.0211	0.0142	0.0749	0.0135

for the mean regression models and gives consistent estimates for the covariance components.

4.6 Real Data Analysis: A Multi-Center AIDS Cohort Study

We consider a subset from the Multi-Center AIDS Cohort Study as discussed in Section 4.1, where the data include repeated measurements of physical examinations, laboratory results and CD4 cell counts and percentages of 283 homosexual men who became HIV-positive between 1984 and 1991. Each patient was supposed to have measurements taken every 6 months, but it often happened that patients missed or rescheduled their appointments. Therefore, each patient had a different number of repeated measurements and the true observation times were not equally spaced. Each patient has minimum 1 and maximum 14 measurements for these data. Further

details about the design, methods and medical implications of the study can be found in Kaslow *et al.* (1987).

Let x_1 be PreCD4, z_1 be smoking status (1 for a smoker and 0 for a non-smoker) and z_2 be age for each i-th subject. Huang *et al.* (2002) showed that the baseline function varies over time and whether or not x_1 has a constant effect over time is unclear so we consider time varying coefficients for both. Note that x_1 and z_2 are standardized variables with mean 0 and standard deviation 1. We consider the response y as the log-transformed CD4 cell percentage of a subject at distinct time points after HIV infection. To model jointly the mean and covariance structures for the data, we use the following model for mean and variances:

$$\mu_{ij} = \alpha_1(t_{ij}) + \alpha_2(t_{ij})x_{1,ij} + \beta_1 z_{1,ij} + \beta_2 z_{2,ij}, \quad \text{and}$$
$$\log(\sigma_{ij}^2) = f_1(t_{ij}) + f_2(t_{ij})x_{1,ij} + \lambda_1 z_{1,ij} + \lambda_2 z_{2,ij},$$

where $\alpha_1(t)$, $\alpha_2(t)$, $f_1(t)$ and $f_2(t)$ are varying coefficient terms.

In addition for modelling the correlation matrix, similar to Qin *et al.* (2015), we choose $w_{ijk} = (1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2)'$. We estimate all regression parameters as well as standard error (SE) and compare with Qin *et al.* (2015) method in Table 4.3, which indicates that our method has smaller standard error.

4.7 Discussion

We develop a joint estimation procedure for the mean and variance parameters in a generalized partially linear varying coefficient model for longitudinal data by decom-

Table 4.3: Estimates of regression coefficients and their standard error

Parameter	Proposed Method		Qin *et al.* Method	
	Estimate	SE	Estimate	SE
$\hat{\beta}_1$	-0.004	0.029	0.056	0.032
$\hat{\beta}_2$	0.011	0.012	0.014	0.015
$\hat{\gamma}_1$	1.261	0.027	0.846	0.048
$\hat{\gamma}_2$	-0.102	0.024	-0.559	0.048
$\hat{\gamma}_3$	0.024	0.005	0.084	0.010
$\hat{\lambda}_1$	0.268	0.073	-0.445	0.202
$\hat{\lambda}_2$	-0.009	0.033	-0.140	0.123

posing the correlation matrix via hyperspherical coordinates. A simulation study as well as real data analysis indicates that our method fits better. By our method we can include a time variant covariates term, which is more flexible than Zhang *et al.* (2015).

Chapter 5

Model Selection in Generalized Linear Models

5.1 Introduction

The importance of model selection in regression analysis for a normally distributed response variable is well known and is widely used in many fields of study, such as, engineering, biomedical sciences, and social sciences. Consider a normal linear regression model with response variables $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and p covariates $X = (\mathbf{1}_n, \boldsymbol{x_1}, \ldots, \boldsymbol{x_p})$ with $\boldsymbol{x_i} = (x_{1i}, \ldots, x_{ni})'$ and regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ that follow $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ are i.i.d. and $\epsilon_i \sim N(0, \sigma^2)$ distribution. In this setting, the interest is to obtain a regression model with as few regression parameters as possible (parsimonious). The popular method in use, in practice, is one of: forward selection, backward elimination, and stepwise selection procedures through a test of significance of a single regression coefficient, for example, test of $H_0 : \beta_j = 0$, using the

F test (Beale, 1970; Kutner, Nachtsheim, Neter, and Li, 2013). Other model selection procedures, such as, the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978) are also available. However, properties of these model selection procedures are not well-known.

The purpose of this Chapter is to study the properties of these procedures in generalized linear models of which the normal, Poisson, and binomial regression models are special cases.

There are two aspects to the model selection procedure: (a) finding a suitable test statistic for testing the significance of a single regression coefficient, for example, to test $H_0 : \beta_j = 0$, which performs best in holding an appropriate level of significance, say 5% and has overall best power property and (b) finding a model selection procedure using this suitable test statistic, which, again, has the best property with respect to level and power.

For (a) we develop three large sample test statistics, namely, the score test, the likelihood ratio test (LR), and the Wald test. These three tests along with the usual F test are compared using a simulation study.

The score test (Rao 1947) is a special case of the $C(\alpha)$ test (Neyman, 1959) where the nuisance parameters are replaced by maximum likelihood estimates that are $\sqrt{n}$-consistent; here, n denotes the number of observations used in estimating the parameters. The score test is particularly appealing as we only have to study the distribution of the test statistic under the null hypothesis which is that of the basic model. It often maintains, at least approximately, a preassigned level of significance and often produces a statistic that is simple to calculate. On the contrary, two other asymptotically equivalent tests (LR test and Wald test) require estimates of the parameters

under the alternative hypothesis and often show liberal or conservative behaviour in small samples. For further discussion on this, see Rao (2005).

For (b) an extensive simulation study is conducted to compare the properties of the forward selection procedure using the best statistic found in (a), the AIC and the BIC. More discussion of these is given in section 5.3.

In Section 5.2 we develop the three large test statistics, which are then specialized for data from the normal, the Poisson and the binomial distributions. The F statistic used in model selection for data from normal is also discussed. Results of an extensive simulation study are reported in Section 5.2.4 and Section 5.3. Extensions for over-dispersed Poisson and over-dispersed binomial regression models are given and evaluated in Section 5.4. Some examples are given in Section 6.4 and a discussion follows in Section 5.6.

5.2 Generalized Linear Models and the Test Statistics

5.2.1 Generalized Linear Models

Let Y_i be the independent response variables with mean μ_i and variance ϕV_i, where V_i is the variance that $Y_i, (i = 1, \ldots, n)$ is assumed to have under the generalized linear model. Further, suppose that $\mu_i = h(\eta_i)$ and $\eta = X\boldsymbol{\beta}$, where $h(.)$ is the inverse link function, $X = [x_{ir}]$ is the $n \times p$ matrix and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of regression parameters. Further we assume that $x_{i0} = 1$ so that β_0 is the intercept parameter.

5.2.2 The Test Statistics

Our interest is to develop a test statistic for testing the hypothesis that one of the β parameters is zero. As such we consider the null hypothesis $H_0 : \beta_j = 0$ against $H_a : \beta_j \neq 0$ for $j = 0, 1, \ldots, p$.

In the GLM framework, the probability density function of y for $\phi = 1$ is given by

$$f(y;\theta) = \exp\left[a(\theta)y - g(\theta) + c(y)\right], \tag{5.1}$$

where θ is related to η and when a natural link function is used $\theta = \eta = X\boldsymbol{\beta}$. Then the log-likelihood can be written as is

$$l = \sum_{i=1}^{n} \left[a(\theta_i)y_i - g(\theta_i) + c(y_i)\right]$$

from which the score function is obtained as

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij},$$

where $\mu_i = \mathrm{E}(y_i) = \dfrac{g'(X_i\boldsymbol{\beta})}{a'(X_i\boldsymbol{\beta})}$, $\dfrac{\partial \mu_i}{\partial \eta_i} = h'(X_i\boldsymbol{\beta})$ and
$V_i = \mathrm{var}(y_i) = \dfrac{g''(X_i\boldsymbol{\beta})a'(X_i\boldsymbol{\beta}) - a''(X_i\boldsymbol{\beta})g'(X_i\boldsymbol{\beta})}{(a'(X_i\boldsymbol{\beta}))^3}$, where $'$ denotes differentiation with respect to θ. To estimate the parameters $\beta_k, k = 0, 1, \ldots, p$, we need to solve $\dfrac{\partial l}{\partial \beta_k} = 0$ which are non-linear in nature in β_k, so must be solved iteratively (McCullagh and Nelder, 1989).

Note that under the null hypothesis we estimate β_k for $k = 0, 1, \ldots, j-1, j+1, \ldots, p$. Denote these estimates by $\hat{\beta}_k$. Further, under the alternative hypothesis we estimate

β_k, for $k = 0, 1, \ldots, p$. Denote these estimates by $\tilde{\beta}_k$.

The Likelihood Ratio Test

Let $\hat{l}$ and $\tilde{l}$ be the maximized log-likelihood under the null and the alternative hypothesis, respectively. Then, the likelihood ratio statistic is $LR_j = 2\left(\tilde{l} - \hat{l}\right)$.

The Wald Test

The Wald test statistic is given by $W_j = \tilde{\beta}_j / \sqrt{\text{var}(\tilde{\beta}_j)}$, where $\text{var}(\tilde{\beta}_j)$ is obtained from the Hessian matrix at the end of the iterative process.

The Score Test

The score test is a special case of the $C(\alpha)$ test which is based on the partial derivatives of the log-likelihood function with respect to the nuisance parameters and the parameters of interest evaluated at the null hypothesis. Suppose $\delta = \beta_j$ and $\theta = (\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)'$. Define the partial derivatives of the log-likelihood which are evaluated at $\delta = 0$ as

$$\psi = \frac{\partial l}{\partial \delta}\Big|_{\delta=0} = \left[\frac{\partial l}{\partial \beta_j}\right]\Bigg|_{\delta=0} \text{ and}$$
$$\gamma = \frac{\partial l}{\partial \theta}\Big|_{\delta=0} = \left[\frac{\partial l}{\partial \beta_0}, \frac{\partial l}{\partial \beta_1}, \cdots, \frac{\partial l}{\partial \beta_{j-1}}, \frac{\partial l}{\partial \beta_{j+1}}, \cdots, \frac{\partial l}{\partial \beta_p}\right]'\Bigg|_{\delta=0}.$$

The $C(\alpha)$ test is based on the adjusted score $S = \dfrac{\partial l}{\partial \delta} - B\dfrac{\partial l}{\partial \theta}$, where B is the matrix of partial regression coefficients that is obtained by regressing $\dfrac{\partial l}{\partial \delta}$ on $\dfrac{\partial l}{\partial \theta}$. The variance-

covariance of S is $D - AB^{-1}A'$, where $D = E\left[-\frac{\partial^2 l}{\partial \beta_j^2}\right]\Big|_{\delta=0}$, $A = E\left[-\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]\Big|_{\delta=0}$ $(k \neq j)$, which is a $1 \times p$ vector and $B = E\left[-\frac{\partial^2 l}{\partial \beta_k \partial \beta_t}\right]\Big|_{\delta=0}$ $(k, t \neq j)$ which is a $p \times p$ matrix. After replacing θ in S, A, B and D by $\hat{\theta}$, the $C(\alpha)$ statistic takes the form

$$S_j = S'(D - AB^{-1}A')^{-1}S,$$

which is approximately distributed as chi-squared with 1 degree of freedom.

Now, define

$$\begin{aligned} w_i &= \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 V_i^{-1}, \\ W &= \text{diag}(w_1, \ldots, w_n)|_{\beta_j=0}, \quad \boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})', \quad \text{and} \\ X_j &= \begin{bmatrix} 1 & x_{11} & \cdots & x_{1(j-1)} & x_{1(j+1)} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2(j-1)} & x_{2(j+1)} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n(j-1)} & x_{n(j+1)} & \cdots & x_{np} \end{bmatrix}. \end{aligned}$$

Then

$$\begin{aligned} &S = \frac{\partial l}{\partial \beta_j}\Big|_{\beta_j=0} = \sum_{i=1}^{n}\left[w_i(y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}x_{ij}\right]\Big|_{\beta_j=0}, \qquad D = \sum_{i=1}^{n} w_i x_{ij}^2\Big|_{\beta_j=0} = \boldsymbol{x}_j' W \boldsymbol{x}_j, \\ &A = \sum_{i=1}^{n} x_{ij} w_i \left(1, x_{i1}, \ldots, x_{i(j-1)}, x_{i(j+1)}, \ldots, x_{ip}\right)\Big|_{\beta_j=0} = \boldsymbol{x}_j' W X_j, \\ &B = X_j' W X_j, \quad \text{and} \qquad D - AB^{-1}A' = \boldsymbol{x}_j' W\left[I_n - X_j\left(X_j' W X_j\right)^{-1} X_j' W\right]\boldsymbol{x}_j. \end{aligned}$$

Replace $\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p$ by their MLE's under the null hypothesis. Then

the score test statistic is

$$S_j = \frac{\hat{S}^2}{\boldsymbol{x}_j'\hat{W}\left(I_n - X_j\left(X_j'\hat{W}X_j\right)^{-1}X_j'\hat{W}\right)\boldsymbol{x}_j}. \tag{5.2}$$

The above score test can also be obtained from Pregibon (1982). For testing a subset q of the regression parameters equal to zero, Pregibon (1982) obtains a score test given by

$$PS_q = s'X_q\left(X_q'W^{\frac{1}{2}}M_pW^{\frac{1}{2}}X_q\right)^{-1}X_q's,$$

where $M_p = I - W^{\frac{1}{2}}X_p(X_p'WX_p)^{-1}X_p'W^{\frac{1}{2}}$, $W = W^{\frac{1}{2}}W^{\frac{1}{2}}$ and $S = sX$.

Using $q = 1$ in the above, the score test becomes $PS_1 = S'\left(\boldsymbol{x}_j'W^{\frac{1}{2}}M_pW^{\frac{1}{2}}\boldsymbol{x}_j\right)^{-1}S$.

Now

$$\begin{aligned}
\boldsymbol{x}_j'W^{\frac{1}{2}}M_pW^{\frac{1}{2}}\boldsymbol{x}_j &= \boldsymbol{x}_j'W^{\frac{1}{2}}\left(I - W^{\frac{1}{2}}X_p(X_p'WX_p)^{-1}X_p'W^{\frac{1}{2}}\right)W^{\frac{1}{2}}\boldsymbol{x}_j \\
&= \boldsymbol{x}_j'W^{\frac{1}{2}}W^{\frac{1}{2}}\boldsymbol{x}_j - \boldsymbol{x}_j'W^{\frac{1}{2}}W^{\frac{1}{2}}X_p(X_p'WX_p)^{-1}X_p'W^{\frac{1}{2}}W^{\frac{1}{2}}\boldsymbol{x}_j \\
&= \boldsymbol{x}_j'W\boldsymbol{x}_j - \boldsymbol{x}_j'WX_p\left(X_p'WX_p\right)^{-1}X_p'W\boldsymbol{x}_j \\
&= \boldsymbol{x}_j'W\left[I_n - X_p\left(X_p'WX_p\right)^{-1}X_p'W\right]\boldsymbol{x}_j.
\end{aligned}$$

Therefore

$$\begin{aligned}
PS_1 &= S'\left(\boldsymbol{x}_j'W^{\frac{1}{2}}M_pW^{\frac{1}{2}}\boldsymbol{x}_j\right)^{-1}S \\
&= \frac{S^2}{\boldsymbol{x}_j'W\left[I_n - X_p\left(X_p'WX_p\right)^{-1}X_p'W\right]\boldsymbol{x}_j},
\end{aligned}$$

which is identical to S_j.

Asymptotically (for large n), the distribution of each of the test statistics LR_j, W_j^2, and S_j converges to $\chi^2(1)$ (Neyman, 1959). Therefore, for a fixed significance level $\alpha > 0$, we reject the null hypothesis if the value of a test statistic is greater than $\chi^2_\alpha(1)$.

The F Test

The F statistic used in model selection for data from normal is

$$NF = \frac{\text{SSR}(\boldsymbol{x}_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p)/\text{df1}}{\text{SSE}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)/\text{df2}},$$

where $\text{SSR}(\boldsymbol{x}_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p) = \text{SSE}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p) - \text{SSE}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\text{df1} = 1$ and $\text{df2} = n - p - 1$. Here SSE means error sum of squares and $NF \sim F(1, n-p-1)$ if H_0 holds (Kutner *et al.* 2013; pp-267).

5.2.3 Special Cases

We now give the expressions for the three test statistics LR_j, W_j, and S_j for the special cases for which the data distribution is normal, Poisson, and binomial respectively. In each situation, however, for robustness study, we include the same normal theory test statistic given in Section 5.2.2.

(i) For the $N(\mu, \sigma^2)$ distribution with link function $\eta_i = \mu_i$ these statistics are

$$\begin{aligned} LRN_j &= \frac{-1}{\sigma^2}\left[(y_i - \tilde{\mu}_i)^2 - (y_i - \hat{\mu}_i)^2\right], \\ WN_j &= \frac{\tilde{\beta}_j}{\sqrt{\frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n} x_{ij}^2}}, \text{and} \\ SN_j &= \frac{\left(\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{\hat{\sigma}^2}\right) x_{ij}\right)^2}{\mathbf{x}_j' W\left(I - X_j\left(X_j' W X_j\right)^{-1} X_j' W\right)\mathbf{x}_j}, \end{aligned}$$

where $\tilde{\mu}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_p x_{ip}$, $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip}$, $W = \text{diag}(1/\hat{\sigma}^2, \ldots, 1/\hat{\sigma}^2)$ and $\hat{\sigma}^2 = \left(\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\right)/n$.

(ii) For the Poisson(λ) distribution, the link function is $\eta_i = \log(\lambda_i)$. After derivation and simplification we obtain the corresponding test statistics for Poisson distributed data as

$$\begin{aligned} LRP_j &= 2\left[(y_i \log\tilde{\lambda}_i - \tilde{\lambda}_i) - (y_i \log\hat{\lambda}_i - \hat{\lambda}_i)\right], \\ WP_j &= \frac{\tilde{\beta}_j}{\sqrt{\sum_{i=1}^{n}\tilde{\lambda}_i x_{ij}^2}}, \text{and} \\ SP_j &= \frac{\left(\sum_{i=1}^{n}\left(y_i - \hat{\lambda}_i\right) x_{ij}\right)^2}{\mathbf{x}_j' W\left(I - X_j\left(X_j' W X_j\right)^{-1} X_j' W\right)\mathbf{x}_j}, \end{aligned}$$

where $\tilde{\lambda}_i = \exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_p x_{ip})$, $\hat{\lambda}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip})$ and $W = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$.

(iii) Finally, for the Binomial(m, p) distribution with link function $\eta_i = \log(\frac{p_i}{1-p_i})$

the corresponding statistics are

$$
\begin{aligned}
LRB_j &= 2\left[\left(y_i \log\frac{\tilde{p}_i}{1-\tilde{p}_i} + m_i \log(1-\tilde{p}_i)\right) - \left(y_i \log\frac{\hat{p}_i}{1-\hat{p}_i} + m_i \log(1-\hat{p}_i)\right)\right], \\
WB_j &= \frac{\tilde{\beta}_j}{\sqrt{\sum_{i=1}^{n} m_i \tilde{p}_i (1-\tilde{p}_i) x_{ij}^2}}, \text{and} \\
SB_j &= \frac{\left(\sum_{i=1}^{n} (y_i - m_i \hat{p}_i)\, x_{ij}\right)^2}{\mathbf{x}_j' W \left(I - X_j \left(X_j' W X_j\right)^{-1} X_j' W\right) \mathbf{x}_j},
\end{aligned}
$$

where $\dfrac{\tilde{p}_i}{1-\tilde{p}_i} = \exp\left(\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_p x_{ip}\right)$,
$\dfrac{\hat{p}_i}{1-\hat{p}_i} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip}\right)$ and
$W = \text{diag}\left(m_1 \hat{p}_1 (1-\hat{p}_1), \ldots, m_n \hat{p}_n (1-\hat{p}_n)\right)$.

5.2.4 Simulation

A simulation study is now conducted to compare the behaviour of the four test statistics, namely, the score, LR, Wald, and F, in terms of empirical level and power, for testing the significance of a single regression coefficient. We consider a two-variable regression model with link function $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $\lambda = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, and $\dfrac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ for $\text{N}(\mu, \sigma^2)$, $\text{Poisson}(\lambda)$, and $\text{Bin}(m, p)$ distributions respectively.

Suppose our interest is to test $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$ in each case. For empirical levels we take $\beta_0 = 1$, $\beta_1 = -1$, and $\beta_2 = 0$. For power we take $\beta_0 = 1$, and $\beta_1 = -1$ and different values of β_2 as given in Table 5.1 for normal and Poisson distributed data, and Table 5.2 for binomially distributed data. For data

from the binomial distribution, the level and power results may be affected by the binomial index m. To check this we conduct simulations for $m = 10$, $m = 20$, and $m = 40$. For both level and power we consider sample sizes $n = 10, 20, 30$ and 50 for all distributions. Each simulation experiment is based on 10,000 replicated samples. The level and power results are presented in Table 5.1 for normal and Poisson distribution and in Table 5.2 for binomial distribution.

Table 5.1: Empirical level (EL) and power (in %) of the four test statistics; $\alpha = 0.05$

Distr.	Size (n)	Test	EL	Empirical Power β_2										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Normal	10	Score	7.80	7.85	7.90	8.17	8.66	9.06	9.76	10.54	11.19	12.09	13.34
		Wald	9.29	9.37	9.64	9.87	10.31	10.87	11.62	12.50	12.99	14.36	15.45
		LR	11.57	11.63	11.89	12.19	12.92	13.26	14.38	14.98	15.49	16.87	18.61
		F	5.15	5.28	5.45	5.55	5.70	6.12	6.66	7.46	7.86	8.53	9.64
	20	Score	6.57	6.58	6.80	7.47	8.01	9.14	10.38	13.23	14.31	16.39	19.35
		Wald	7.17	7.11	7.45	8.11	8.68	9.96	11.19	14.36	15.34	17.58	20.46
		LR	7.92	7.99	8.30	8.94	9.57	11.08	12.22	15.56	16.81	19.03	22.14
		F	5.32	5.52	5.64	6.26	6.82	7.82	8.82	11.25	12.38	14.31	17.23
	30	Score	5.96	6.01	6.45	7.73	8.67	10.77	12.85	16.16	19.26	22.98	25.66
		Wald	6.42	6.35	6.75	8.15	9.17	11.43	13.38	16.78	20.08	23.76	26.58
		LR	6.83	6.78	7.42	8.71	9.84	12.17	14.18	17.78	21.15	24.95	27.76
		F	5.33	5.45	5.89	6.91	7.83	9.67	11.61	15.12	17.94	21.10	23.72
	50	Score	5.61	6.14	6.60	8.20	11.25	14.64	17.61	22.10	28.57	34.03	40.51
		Wald	5.77	6.42	6.79	8.43	11.62	15.18	18.05	22.56	29.11	34.70	41.09
		LR	6.09	6.75	7.09	8.80	12.05	15.71	18.48	23.25	29.88	35.45	41.96
		F	5.28	5.60	6.19	7.77	10.46	13.84	16.70	21.12	27.56	32.80	39.32
Poisson	10	Score	5.09	5.96	7.80	11.02	17.16	22.56	30.25	37.84	47.75	54.68	62.69
		Wald	4.59	5.37	7.22	10.22	16.24	21.53	29.00	36.56	46.37	53.54	61.65
		LR	5.42	6.33	8.04	11.39	16.41	23.16	30.59	38.42	48.64	55.56	63.54
		F	0.09	0.14	0.22	0.28	0.32	0.40	0.53	0.73	1.05	1.13	1.41
	20	Score	4.74	6.00	11.56	19.42	31.01	44.85	58.27	70.13	79.45	87.37	91.98
		Wald	4.61	5.84	11.41	19.11	30.64	44.53	57.90	69.80	79.14	87.11	91.86
		LR	4.80	6.14	11.66	19.58	31.19	45.12	58.73	70.39	79.58	87.53	92.25
		F	0.02	0.00	0.01	0.04	0.04	0.14	0.22	0.57	0.97	1.24	1.98
	30	Score	4.83	8.50	15.38	27.95	45.02	62.09	76.17	86.93	93.34	96.69	98.59
		Wald	4.79	8.44	15.30	27.70	44.89	61.93	76.06	86.85	93.28	96.65	98.55
		LR	4.82	8.59	15.37	28.00	45.24	62.10	76.31	87.08	93.45	96.74	98.58
		F	0.01	0.01	0.01	0.01	0.12	0.23	0.31	0.80	1.66	2.20	3.70
	50	Score	4.85	9.42	22.10	44.88	66.33	83.94	93.96	97.85	99.51	99.89	99.98
		Wald	4.84	9.43	22.04	44.78	66.27	83.91	93.93	97.83	99.50	99.89	99.98
		LR	4.81	9.45	22.12	44.94	66.40	84.08	93.98	97.87	99.51	99.88	99.98
		F	0.00	0.00	0.01	0.02	0.11	0.30	0.96	2.45	5.05	9.43	13.82

Results in Table 5.1 show that for normally distributed data, the score test and the F test maintain level reasonably well, although, the score test shows some inflated level. As a result, it shows some inflated power. The other two statistics (Wald and LR), show liberal behaviour. Because of this, these two statistics show higher power

Table 5.2: Empirical level (EL) and power (in %) of the four test statistics in binomial distribution; $\alpha = 0.05$

Size (m,n)	Test	EL	Empirical Power β_2										
		0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
	Score	4.87	5.58	7.03	9.23	12.38	16.21	21.11	27.23	33.16	40.14	46.17
	Wald	4.44	5.02	6.32	8.44	11.47	15.10	19.56	25.73	31.54	38.25	44.46
(10,10)	LR	5.20	6.05	7.39	9.66	12.95	16.83	21.88	28.03	33.95	40.80	46.80
	F	0.55	0.80	0.96	1.01	1.47	2.16	2.80	3.75	5.08	6.72	8.14
	Score	4.99	6.19	8.82	13.08	19.81	27.72	36.89	47.03	55.77	63.77	70.34
	Wald	4.75	5.92	8.45	12.65	19.19	26.92	36.08	46.11	54.93	62.92	69.71
(20,10)	LR	5.19	6.31	9.03	13.30	19.97	28.09	37.39	47.58	56.06	64.15	70.71
	F	0.00	0.01	0.00	0.02	0.05	0.06	0.17	0.18	0.42	0.59	0.88
	Score	4.96	6.79	12.59	21.55	33.67	47.59	60.52	71.35	79.71	85.84	89.87
	Wald	4.83	6.57	12.38	21.26	33.33	47.18	60.11	71.01	79.42	85.69	89.78
(40,10)	LR	5.03	6.89	12.68	21.69	33.86	47.66	60.71	71.50	79.89	85.88	89.99
	F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.01
	Score	5.37	6.12	9.71	14.26	21.50	30.95	41.20	51.76	61.75	70.84	78.90
	Wald	5.04	5.89	9.27	13.74	20.94	30.15	40.49	51.07	60.94	70.00	78.30
(10,20)	LR	5.47	6.27	9.90	14.58	21.78	31.22	41.64	52.40	62.17	71.10	79.27
	F	0.19	0.23	0.45	0.79	1.55	2.85	4.70	7.35	11.37	16.63	22.80
	Score	5.30	7.10	13.43	23.86	37.02	53.01	67.69	78.72	86.70	92.63	95.83
	Wald	5.18	6.98	13.21	23.46	36.68	52.61	67.22	78.45	86.55	92.53	95.81
(20,20)	LR	5.30	7.20	13.49	24.03	37.20	53.19	67.97	78.72	86.84	92.72	95.90
	F	0.00	0.00	0.01	0.00	0.01	0.02	0.09	0.12	0.46	0.89	1.43
	Score	5.37	9.14	22.10	42.02	62.80	79.42	90.62	95.66	98.24	99.40	99.71
	Wald	5.31	9.10	21.96	41.84	62.70	79.31	90.54	95.64	98.24	99.39	99.71
(40,20)	LR	5.37	9.23	22.16	42.16	62.92	79.48	90.68	95.69	98.26	99.41	99.71
	F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
	Score	5.35	7.04	11.01	19.80	31.08	44.40	58.13	69.15	79.91	88.09	92.52
	Wald	5.23	6.82	10.74	19.33	30.47	43.89	57.64	68.81	79.55	87.88	92.31
(10,30)	LR	5.40	7.18	11.18	19.98	31.33	44.61	58.45	69.51	80.09	88.27	92.59
	F	0.08	0.19	0.36	1.05	2.26	4.53	8.67	13.76	22.16	32.39	42.40
	Score	5.49	8.79	17.90	33.99	54.11	71.67	85.21	92.22	96.74	98.79	99.45
	Wald	5.43	8.64	17.73	33.70	53.79	71.50	85.03	92.14	96.69	98.78	99.45
(20,30)	LR	5.46	8.83	17.91	34.25	54.22	71.88	85.26	92.28	96.79	98.80	99.48
	F	0.00	0.00	0.00	0.01	0.00	0.03	0.14	0.35	0.88	2.36	4.38
	Score	5.29	11.95	31.16	57.93	81.60	93.48	98.17	99.49	99.88	99.98	99.98
	Wald	5.22	11.90	30.99	57.81	81.56	93.45	98.16	99.49	99.88	99.98	99.98
(40,30)	LR	5.30	11.97	31.24	57.94	81.68	93.52	98.18	99.49	99.88	99.98	99.98
	F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01
	Score	5.24	7.77	16.28	30.41	46.97	65.41	79.71	89.45	95.40	98.27	99.32
	Wald	5.12	7.66	16.15	30.08	46.73	65.09	79.53	89.30	95.30	98.24	99.31
(10,50)	LR	5.26	7.79	16.43	30.59	47.16	65.53	79.80	89.54	95.48	98.28	99.33
	F	0.03	0.05	0.52	1.89	4.12	10.02	19.47	33.85	49.03	64.54	78.18
	Score	5.12	10.22	27.59	52.74	75.87	90.60	97.08	99.13	99.84	99.94	100
	Wald	5.08	10.11	27.49	52.54	75.74	90.53	97.08	99.12	99.84	99.94	100
(20,50)	LR	5.19	10.30	27.73	52.84	75.87	90.66	97.08	99.15	99.84	99.94	100
	F	0.00	0.00	0.00	0.01	0.01	0.06	0.36	1.39	4.98	12.47	24.67
	Score	4.92	15.75	47.95	81.18	95.42	99.37	99.96	99.99	100	100	100
	Wald	4.92	15.72	47.87	81.16	95.42	99.37	99.96	99.99	100	100	100
(40,50)	LR	4.97	15.78	47.91	81.20	95.44	99.37	99.96	99.99	100	100	100
	F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.06

than the other two tests.

Results in Table 5.1 and Table 5.2 show that for data from the Poisson and binomial distributions, the F test performs very badly. The other 3 statistics hold level very

well and their power performances are also similar. Further, results in Table 5.2 show that the size of the binomial index m does not have any effect on size and power of the tests. So, in subsequent sections we choose $m = 40$ as the binomial index.

That the F test does well for data from the normal distribution is reassuring. So, in Section 5.3, we use this test in the study of the performance of the model selection procedures. For data from the Poisson and binomial distributions we use the score test as it has a very simple form, does not need estimates of the regression parameters under the alternative hypothesis, and its level and power properties are, at least, as good as that of the LR and the Wald tests.

5.3 Model Selection

5.3.1 Empirical Level and Power

Following the findings in Section 5.2, our model selection criterion for normally distributed data is based on testing the significance of a single regression coefficient β_j using the F test given in Section 5.2.2. Also as discussed in Section 5.2.4, for data from the Poisson and the binomial distributions we use the score test statistic SP_j and SB_j, respectively, given in Section 5.2.3. Our purpose here is to make a comparative study of performance of forward selection, AIC and BIC with respect to level and power. The other two procedures, backward elimination and stepwise selection are not included in our study, as in practice, these produce a similar final model as that obtained by the forward selection procedure.

Although these model selection procedures are well known, to be helpful to the

readers, we give a brief description of these below.

Forward Selection Procedure: The forward selection starts with only one variable in the model. So, if the model has p regression variables, apart from the intercept, we fit p regression models and calculate the value of the score test statistic for each model. If the score test statistic for testing $H_0 : \beta_j = 0$, for example, is found to be the largest and significant at a specified level of significance, we then keep this variable in the model. We then continue this process by adding one more variable, each time, until no more variables can be included in the model. At the end the final model will have $q \leq p$ variables.

AIC and BIC Criteria: For p covariates in the regression model, we first construct all possible $(2^p - 1)$ models and choose the model having smallest value of $AIC = -2l + 2p$. The process is similar for $BIC = -2l + \ln(n)p$.

As mentioned earlier our purpose is to find the most parsimonious model. Here we illustrate a method of calculating the empirical level using a p variable Poisson regression model with $\ln(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. For given values of the regression parameters and simulated values of the regression variables, we obtain a sample of size n from the Poisson(λ) distribution. We then use the score test statistic for testing $H_0 : \beta_j = 0$ and a model selection procedure, for example, the forward selection procedure and find a model of a subset of the regression variables. We repeat this process 10,000 times and find 10,000 models. If the given value of β_j, is very small, we want to see that the regression variable x_j is in the final model. We then count the number of models in which the variable x_j is included. Let this number be s. Then the empirical level for rejecting $H_0 : \beta_j = 0$ is $s/10,000$. Empirical power is calculated similarly by taking a larger value of β_j during the simulation process.

Simulation Study

Now, we conduct a simulation study to compare the performance of the model selection procedures, the forward selection, the AIC and the BIC with respect to empirical level and power. We consider a 4-variable regression model. Data are drawn from the normal $\mathrm{N}(\mu, \sigma^2)$ regression model, the Poisson(λ) regression model, and the Binomial(m, p) regression model with

$$\begin{aligned} \mu &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \\ \lambda &= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4), \text{and} \\ \frac{p}{1-p} &= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4) \end{aligned}$$

respectively. Suppose we would like to test $H_0 : \beta_1 = 0$. To calculate the empirical level for each distribution we choose $\beta_1 = 0.001$ (a very small value) and for empirical power we take different values of β_1 as given in Table 5.3. The rest of the parameters are set at $\sigma^2 = 2$, $\beta_2 = -0.3$, $\beta_3 = 0.2$, $\beta_4 = 0.3$ for normal and Poisson distributions and $m = 40$, $\beta_2 = 0.2$, $\beta_3 = -0.1$, and $\beta_4 = -0.2$ for the binomial distribution. For each distribution 10,000 replicated samples are taken for sample size $n = 10$, 20, 30 and 50.

For the forward selection procedure we consider $\alpha = 0.05$. Note that for the other two procedures α cannot be fixed.

The level and power results are presented in Table 5.3 which show that the forward selection method using the F test for normally distributed and the score test for Poisson and binomially distributed data always produces a reasonable empirical level (close to the nominal level) irrespective of sample size. The other two procedures, the

Table 5.3: Empirical level (EL) and power (in %) of model selection by the forward selection using the score test (Forward-S), forward selection using F test (Forward-F), the AIC, and the BIC; based on 10,000 replications

Dist.	Size (n)	Method	EL	Empirical Power β_1										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Normal	10	Forw-F	5.09	5.48	5.34	5.82	6.10	6.11	6.77	7.86	7.95	9.11	9.56
		AIC	30.94	31.66	32.31	32.47	33.05	33.12	34.20	35.41	36.08	37.18	40.22
		BIC	27.17	27.83	28.32	28.56	28.86	29.19	30.09	31.34	32.35	33.20	36.25
	20	Forw-F	4.96	5.58	5.67	6.22	7.32	7.93	9.61	10.84	12.85	15.20	16.50
		AIC	21.40	21.53	22.72	23.69	25.72	27.66	29.55	31.41	35.01	39.22	41.26
		BIC	12.32	12.53	13.19	13.77	15.72	17.21	18.82	21.18	23.85	27.16	29.55
	30	Forw-F	4.68	5.25	5.76	6.79	8.10	10.39	11.57	14.40	17.55	21.24	23.58
		AIC	18.87	20.39	20.30	22.35	25.01	28.12	30.94	35.33	39.38	44.07	48.34
		BIC	8.23	8.94	9.59	10.83	12.68	15.27	17.33	20.97	24.54	28.57	31.46
	50	Forw-F	5.18	5.41	6.24	8.36	10.35	13.80	17.54	21.67	26.71	32.20	38.93
		AIC	18.28	18.94	19.99	23.94	27.39	33.15	38.08	43.62	49.93	57.08	63.04
		BIC	5.87	6.19	6.98	9.34	11.46	14.98	18.97	23.36	28.91	34.45	41.03
Poisson	10	Forw-S	6.93	8.53	9.84	12.27	16.58	22.27	27.73	32.98	41.26	47.11	53.98
		AIC	19.08	21.77	23.40	26.61	32.98	39.36	46.25	51.88	59.80	64.21	70.92
		BIC	16.28	18.64	20.73	23.63	29.57	35.66	42.66	48.44	56.68	61.13	68.34
	20	Forw-S	7.01	8.33	12.20	19.26	28.33	38.91	49.33	62.48	71.04	79.04	86.55
		AIC	18.06	19.59	26.85	36.34	48.18	59.23	70.53	80.46	86.63	91.11	94.96
		BIC	10.87	12.16	17.63	25.56	36.63	47.58	59.69	71.45	79.01	85.75	91.56
	30	Forw-S	6.52	8.53	15.21	25.88	40.22	54.97	69.85	80.33	89.06	93.48	96.82
		AIC	17.13	20.23	31.01	44.98	62.01	75.49	85.95	92.23	96.60	98.16	99.20
		BIC	8.08	10.41	18.21	30.06	45.41	60.83	74.85	84.68	92.06	95.46	97.89
	50	Forw-S	5.75	9.46	21.49	41.32	62.66	79.40	90.61	96.69	98.83	99.62	99.87
		AIC	15.71	22.87	40.92	63.18	80.95	91.90	97.02	99.16	99.80	99.9	99.98
		BIC	5.47	9.23	21.42	41.32	62.73	79.56	91.02	96.79	99.02	99.64	99.90
Binomial	10	Forw-S	6.93	9.14	14.03	20.82	30.86	41.57	53.03	64.12	71.59	78.82	84.68
		AIC	18.54	21.34	28.43	38.00	49.14	60.52	70.59	78.97	84.12	88.78	92.23
		BIC	16.03	18.54	25.18	34.54	45.64	57.23	67.65	76.69	82.38	87.15	91.19
	20	Forw-S	6.73	9.99	20.98	37.45	56.55	73.18	84.48	92.07	95.86	98.36	99.06
		AIC	17.38	22.81	39.18	59.48	75.77	88.02	94.06	97.41	98.81	99.70	99.86
		BIC	10.43	14.57	28.20	47.29	66.26	80.88	89.85	95.08	97.84	99.22	99.64
	30	Forw-S	6.20	12.21	29.28	53.98	75.26	90.12	96.24	98.97	99.77	99.90	99.97
		AIC	16.88	26.33	49.26	74.42	89.40	96.81	99.14	99.80	99.95	100	100
		BIC	7.71	14.84	33.73	59.68	80.04	92.56	97.47	99.34	99.83	99.97	99.99
	50	Forw-S	5.47	16.04	46.67	77.80	94.77	99.22	99.82	100	100	100	100
		AIC	16.47	33.71	68.36	90.77	98.64	99.87	99.99	100	100	100	100
		BIC	5.46	15.90	46.31	77.76	94.90	99.30	99.84	100	100	100	100

AIC and BIC produce a highly inflated type I error. The BIC, however, does well for large sample size ($n = 50$) in which case its power performance is also comparable to that of the forward selection procedure using the score test.

Thus, for normal regression models, our recommendation is to use the forward selection procedure using the F test. For Poisson and binomial regression models our recommendation is to use the forward selection procedure using the score test for small to moderate sample sizes and for large n ($n \geq 50$) the BIC should be used

because it has simple form and is easy to compute.

5.4 Over-dispersed Poisson and Over-dispersed Binomial Regression Models

In this section we extend the methods and ideas developed in Sections 5.2 and 5.3 for model selection for Poisson and binomial regression models to over-dispersed Poisson and over-dispersed binomial regression models. Specifically, we deal with model selection procedures in negative binomial regression model and beta binomial regression models. Here also we first develop the score, the LR and the Wald tests for testing the significance of a single regression variable and then for model selection we compare the forward selection, the AIC and the BIC procedures.

5.4.1 Negative Binomial Regression Model

Consider the Negative binomial (NB) distribution with probability mass function

$$f(y;m,c) \quad = \frac{\Gamma(y+c^{-1})}{\Gamma(c^{-1})y!}\left(\frac{cm}{1+cm}\right)^{y}\left(\frac{1}{1+cm}\right)^{c^{-1}}, \tag{5.3}$$

for $y = 0, 1, 2, \ldots,\ m > 0,\ c > -1/m$ with mean $E(y) = m$ and variance $\text{var}(y) = m(1+cm)$ (see Piegorsch, 1990). We denote this distribution as $\text{NB}(m, c)$. In equation (5.3), the term c represents the dispersion parameter which is constant. Clearly, when $c \to 0$, the NB distribution reduces to the Poisson distribution with parameter m.

Let $y_i, i = 1, \ldots, n$, be a random sample from the $NB(m_i, c)$ distribution with

$m_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)$. Then $\dfrac{\partial m_i}{\partial \beta_j} = m_i x_{ij}$. The log-likelihood of the NB regression model then, is

$$l = \sum_{i=1}^{n} \Big\{ y_i \log(m_i) - (y_i + c^{-1}) \log(1 + c m_i) + \sum_{j=1}^{y_i} \log[1 + c(j-1)] \Big\}.$$

The first and second order partial derivatives of the log-likelihood function with respect to the parameters $\boldsymbol{\beta}$ and c and their expected values are given in Appendix B.

Derivation of the test statistics

We follow the same procedure to find the score test for testing $H_0 : \beta_j = 0$ as described in Section 5.2.2. Omitting the details, the score, the Wald and LR statistics are

$$SNB_j = S'(D - A_1 B_{11}^{-1} A_1')^{-1} S = \frac{\left(\sum_{i=1}^{n} \frac{(y_i - \hat{m}_i) x_{ij}}{1 + \hat{c}\hat{m}_i} \right)^2}{\boldsymbol{x}_j' W \left(I_n - X_j \left(X_j' W X_j \right)^{-1} X_j' W \right) \boldsymbol{x}_j},$$

$$WNB_j = \tilde{\theta} \Big/ \sqrt{\operatorname{var}(\tilde{\theta})} = \tilde{\theta} \Big/ \sqrt{\sum_{i=1}^{n} \frac{\tilde{m}_i}{1 + \tilde{c}\tilde{m}_i} x_{ij}^2},$$

$$\begin{aligned} LNB_j = 2 \sum_{i=1}^{n} \Big\{ & y_i \log \frac{\tilde{m}_i}{\hat{m}_i} - (y_i + \tilde{c}^{-1}) \log(1 + \tilde{c}\tilde{m}_i) \\ & + (y_i + \hat{c}^{-1}) \log(1 + \hat{c}\hat{m}_i) + \sum_{l=1}^{y_i} \log \left[\frac{1 + \tilde{c}^{-1}(l-1)}{1 + \hat{c}^{-1}(l-1)} \right] \Big\}, \end{aligned}$$

where $w_i = \dfrac{\hat{m}_i}{1 + \hat{c}\hat{m}_i}$, $W = \operatorname{diag}(w_1, \ldots, w_n)$, $\hat{m}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip})$ and $\tilde{m}_i = \exp(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})$ where $\tilde{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ under the alternative hypothesis.

Simulation

We conduct here two simulation studies; the first is to compare the performance of three test statistics and the other is to compare the performance of model selection by forward selection, AIC and BIC.

Empirical Level and Power of the score, the Wald, and the LR Tests: Data are simulated from the negative binomial regression model NB(m, c) with link function $m = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$. We would like to test the null hypothesis $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$.

For empirical levels: We simulate response data from the negative binomial regression model with $c = 0.03$, $\beta_0 = 2$, $\beta_1 = -0.3$, $\beta_2 = 0$. For power, different values of β_2 are taken as in Table 5.4. The independent variables x_1 and x_2 are generated from the standard normal distribution.

Table 5.4: Empirical level (EL) and power (in %) of the three test statistics in negative binomial distribution; based on 10,000 replications and $\alpha = 0.05$

Size (n)	Test	EL	Empirical Power									
			β_2									
		0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	Score	4.58	5.53	9.24	15.84	23.66	34.23	45.39	55.89	64.35	71.95	79.30
	Wald	6.74	7.87	13.19	20.99	31.70	44.16	55.27	66.22	74.06	80.57	86.99
	LR	6.07	7.04	11.84	19.66	29.60	41.73	52.85	63.78	72.03	78.74	85.78
20	Score	4.85	8.07	17.01	32.94	51.67	68.32	81.37	90.17	95.25	97.66	98.83
	Wald	6.43	10.09	20.59	38.04	56.92	73.49	85.09	92.65	96.72	98.41	99.26
	LR	5.82	9.30	19.09	36.14	55.04	71.87	84.03	91.87	96.33	98.2	99.12
30	Score	4.70	10.18	25.29	47.87	69.99	87.17	94.84	98.37	99.46	99.82	99.93
	Wald	5.80	11.98	28.29	52.41	73.44	89.53	95.97	98.81	99.62	99.89	99.95
	LR	5.35	11.18	26.89	50.50	72.06	88.64	95.46	98.65	99.56	99.87	99.95
50	Score	4.96	12.95	40.10	71.07	91.21	97.96	99.62	99.93	100	100	100
	Wald	5.78	14.40	42.59	73.56	92.28	98.20	99.71	99.94	100	100	100
	LR	5.37	13.73	41.57	72.42	91.80	98.12	99.68	99.94	100	100	100

The level and power results are presented in Table 5.4. The results show that the score test has best level property (empirical level close to the nominal level). The other two statistics show some inflation of the empirical level compared to the nominal level which results in some higher power for the Wald and the LR statistics. So, here

also we use the score test statistic in model selection using the forward selection procedure.

Empirical Level and Power of Model Selection: A simulation study is conducted similar to that in Section 5.3 to compare the property of the model selection procedure through forward selection using the score test with the other two criteria AIC and BIC.

Data are taken from the NB(m, c) distribution with $m = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$, $\beta_0 = 2$, $\beta_1 = 0.001$, $\beta_2 = -0.3$, $\beta_4 = -0.1$ for empirical level and values of β_1 are given in Table 5.5 for power. The value of c is taken $c = 0.03$. Further, as in Section 5.3, sample sizes and nominal level are chosen for $n = 10$, 20, 30 and 50 and $\alpha = 0.05$. The level and power results are presented in Table 5.5.

Table 5.5: Empirical level (EL) and power (in %) of model selection by forward selection using score test, AIC and BIC in negative binomial distribution; based on 10,000 replications

Size (n)	Method	EL	Empirical Power										
			β_1										
		0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	Forward	3.80	4.05	5.04	7.06	9.57	13.23	16.84	22.06	27.63	32.55	38.05
	AIC	23.55	24.18	30.50	37.17	45.99	56.40	64.41	72.19	78.25	83.23	87.64
	BIC	20.53	21.22	27.51	33.70	42.32	52.79	60.92	69.25	75.76	80.95	85.77
20	Forward	4.81	5.49	8.90	13.82	21.25	29.95	40.27	51.03	61.39	70.03	79.00
	AIC	19.08	24.46	38.04	54.20	71.16	83.52	90.73	95.47	97.63	99.03	99.41
	BIC	11.30	15.37	26.39	41.39	59.74	74.27	84.46	91.46	95.12	97.66	98.73
30	Forward	4.86	6.68	11.77	20.47	33.26	46.77	61.55	72.86	83.32	89.71	93.74
	AIC	18.26	26.68	46.99	69.74	85.26	94.46	98.03	99.46	99.77	99.95	100
	BIC	8.59	14.62	30.52	53.73	74.19	87.76	94.80	98.08	99.20	99.76	99.99
50	Forward	4.85	8.15	18.64	34.59	54.58	73.04	86.24	94.13	97.85	99.00	99.78
	AIC	18.26	26.68	46.40	26.61	32.98	39.36	46.25	51.88	59.80	64.21	70.92
	BIC	8.59	14.62	20.73	23.63	29.57	35.66	42.66	48.44	56.68	61.13	68.34

Results in Table 5.5 show similar performance of the forward selection procedure with the score test as the test statistic as in Table 5.4, namely, that its level is close to the nominal level. The other procedures show highly inflated empirical level, even for large n ($n = 50$). For n=10, 20, 30 power of the forward selection procedure is the lowest, because the other two procedures have highly inflated empirical level.

For n=50 power of the forward selection procedure is smaller than that of the other two procedures for smaller values of $\beta_2(= .05, .10)$. However, as β_2 increases power of the forward selection procedure increases dramatically as compared to the other two procedures, even though the forward selection procedure holds level, where as the other procedures are liberal.

5.4.2 Beta Binomial Regression Model

Suppose Y follows a beta binomial distribution with mean μ and dispersion parameter θ, denoted by $Y \sim \text{BB}(k, \mu, \theta)$ if Y has the following probability function

$$P(Y = y) = \begin{pmatrix} k \\ y \end{pmatrix} \frac{\prod_{r=0}^{y-1}(\mu + r\theta) \prod_{r=0}^{k-y-1}(1 - \mu + r\theta)}{\prod_{r=0}^{k-1}(1 + r\theta)},$$

for $y = 0, 1, \ldots, k$, $0 \leq \mu \leq 1$ and $\theta \geq \max[-\mu/(k-1), -(1-\mu)/(k-1)]$ with mean $E(Y) = k\mu$ and variance $\text{var}(Y) = k\mu(1-\mu)[1 + (k-1)\phi]$, where $\phi = \theta/(1+\theta)$ (see Williams, 1975; Paul, 1982).

Note that, as $\theta \to 0$ the $\text{BB}(k, \mu, \theta)$ tends to the Binomial(k, μ) distribution, and for $\theta = 0$, we have $\text{var}(Y) = k\mu(1-\mu)$, and the $\text{BB}(k, \mu, \theta)$ becomes the Binomial(k, μ) distribution.

Let $y_i, i = 1, \ldots, n$ be a random sample from the $\text{BB}(k_i, \mu_i, \theta)$, then the log-likelihood is

$$l = \sum_{i=1}^{n} \left[\sum_{r=0}^{y_i - 1} \log(\mu_i + r\theta) + \sum_{r=0}^{k_i - y_i - 1} \log(1 - \mu_i + r\theta) - \sum_{r=0}^{k_i - 1} \log(1 + r\theta) \right].$$

The mean μ_i is assumed to follow the logistic model $\mu_i(\boldsymbol{x}_i', \boldsymbol{\beta}) = \dfrac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$. Then $\dfrac{\partial \mu_i}{\partial \beta_j} = \mu_i(1-\mu_i)x_{ij}$. The first and second order partial derivatives of l with respect to the parameter $\boldsymbol{\beta}$ and θ and their expected values are given in Appendix B.

Derivation of the test statistics

Using the same procedure as given in Section 5.2, the score test statistics for testing $H_0 : \beta_j = 0$ is

$$SBB_j = \frac{\left(\sum_{i=1}^{n}\left(\sum_{r=0}^{y_i-1}\frac{1}{\hat{\mu}_i+r\hat{\theta}} - \sum_{r=0}^{k_i-y_i-1}\frac{1}{1-\hat{\mu}_i+r\hat{\theta}}\right)\hat{\mu}_i(1-\hat{\mu}_i)x_{ij}\right)^2}{\hat{V}_j},$$

where $\dfrac{\hat{\mu}_i}{1-\hat{\mu}_i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1}x_{i(j-1)} + \hat{\beta}_{j+1}x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip})$ and $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}$ are the maximum likelihood estimates of $\boldsymbol{\beta}$ and θ under the null hypothesis and $\hat{V}_j = V_j(\hat{\mu}, \hat{\theta})$ with

$$V_j = \boldsymbol{x}_j'\left[W - \left(W - \frac{1}{a}UU'\right)X_j V_1^{-1} X_j' W - \left(I - WX_j\left(X_j'WX_j\right)^{-1}X_j'\right)UV_2^{-1}U'\right]\boldsymbol{x}_j,$$

$$W = \mathrm{diag}(w_1, \ldots, w_n), \quad U = (u_1, \ldots, u_n)',$$

$$w_i = (p_{1i} + p_{2i})\mu_i^2(1-\mu_i)^2, \quad u_i = \frac{1}{\theta}[-\mu_i p_{1i} + (1-\mu_i)p_{2i}]\mu_i(1-\mu_i),$$

$$V_1 = X_j'\left(W - \frac{1}{a}UU'\right)X_j, \quad V_2 = a - U'X_j\left(X_j'WX_j\right)^{-1}X_j'U,$$

$$a = \frac{1}{\theta^2}\sum_{i=1}^{n}\left(\mu_i^2 p_{1i} + (1-\mu_i)^2 p_{2i} - p_{3i}\right), \quad p_{1i} = \sum_{r=1}^{k_i}\frac{\Pr(y_i \geq r)}{[\mu_i + (r-1)\theta]^2},$$

$$p_{2i} = \sum_{r=1}^{k_i}\frac{\Pr(y_i \leq k_i - r)}{[1-\mu_i + (r-1)\theta]^2}, \quad \text{and} \quad p_{3i} = \sum_{r=1}^{k_i}\frac{1}{[1+(r-1)\theta]^2}.$$

The Wald test and LR test statistics are as follows

$$WBB_j = \tilde{\theta} \Big/ \sqrt{\text{var}(\tilde{\theta})} = \tilde{\theta} \Big/ \sqrt{\sum_{i=1}^{n} (p_{1i} + p_{2i}) \tilde{\mu}_i^2 (1 - \tilde{\mu}_i)^2 x_{ij}^2},$$

$$LBB_j = 2 \sum_{i=1}^{n} \left[\sum_{r=0}^{y_i - 1} \log \frac{\tilde{\mu}_i + r\tilde{\theta}}{\hat{\mu}_i + r\hat{\theta}} + \sum_{r=0}^{k_i - y_i - 1} \log \frac{1 - \tilde{\mu}_i + r\tilde{\theta}}{1 - \hat{\mu}_i + r\hat{\theta}} - \sum_{r=0}^{k_i - 1} \log \frac{1 + r\tilde{\theta}}{1 + r\hat{\theta}} \right],$$

where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\theta}$ are the maximum likelihood estimates of $\boldsymbol{\beta}$ and θ under the alternative hypothesis with $\dfrac{\tilde{\mu}_i}{1 - \tilde{\mu}_i} = \exp(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}})$.

Simulation Study

Two simulation studies are conducted in this subsection: the first is to compare the performance of three test statistics and other one is to compare the performance of model selection by forward selection through the score test and using AIC and BIC.

Empirical Level and Power of score, Wald, and LR Tests: We take data from the beta binomial regression model $\text{BB}(k, \mu, \theta)$ with link function $\dfrac{\mu}{1 - \mu} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ and would like to test the null hypothesis $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$.

We simulate data from the beta binomial regression model $\text{BB}(k, \mu, \theta)$ with $k = 40$, $\theta = 9$, $\beta_0 = 1$, $\beta_1 = -0.5$, and $\beta_2 = 0$ for empirical level and different values of β_2 given in Table 5.6 for power. The independent variable x_1 and x_2 are generated from the standard normal distribution. The level and power results are presented in Table 5.6.

Table 5.6 clearly shows that score test performs well in terms of level when n is small. However Wald and likelihood ratio test perform well when n is big. Moreover

Table 5.6: Empirical level (EL) and power (in %) of the three test statistics in beta binomial distribution; based on 10,000 replications and $\alpha = 0.05$

Size (n)	Test	EL	Empirical Power										
			β_2										
		0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	Score	8.07	8.85	11.69	13.26	15.86	20.09	22.95	27.64	32.08	36.80	40.88
	Wald	13.93	15.11	17.76	19.56	23.39	27.78	32.58	37.63	42.07	48.50	53.38
	LR	10.47	11.79	14.42	16.20	19.25	23.86	27.99	33.90	38.49	44.55	49.44
20	Score	7.47	8.60	10.95	16.12	21.30	27.54	34.97	40.67	47.27	52.61	56.11
	Wald	8.27	9.69	12.85	18.56	26.55	34.81	43.85	53.47	62.57	69.59	75.50
	LR	7.32	8.64	11.54	17.16	24.94	32.72	42.01	51.20	60.89	67.81	74.05
30	Score	6.96	8.34	12.24	18.16	26.18	35.61	44.60	51.94	43.45	48.16	52.51
	Wald	7.00	8.90	13.82	21.43	32.45	45.17	57.16	68.25	70.52	79.41	84.83
	LR	6.38	8.26	12.93	20.42	31.25	43.81	55.75	67.02	69.66	78.48	84.08
50	Score	6.71	8.92	15.22	25.40	37.17	49.66	59.78	67.86	73.32	75.47	77.70
	Wald	5.96	8.85	16.96	30.61	47.57	63.78	77.83	87.73	93.88	96.92	98.77
	LR	5.79	8.46	16.44	29.99	46.63	63.05	77.15	87.28	93.56	96.68	98.70

empirical power of Wald and likelihood ratio test are close and higher than score test for any sample size.

Empirical Level and Power of Model Selection: We conduct a simulation study similar to Section 5.3 to investigate the behaviour of model selection through the score test using the forward selection procedure in terms of level and power. We further consider model selection using AIC and BIC for comparison.

To calculate empirical level we generate data from the beta binomial regression model $\text{BB}(k, \mu, \theta)$ with $\dfrac{\mu}{1-\mu} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$. We choose $k = 40$, $\beta_0 = 1$, $\beta_1 = 0.001$, $\beta_2 = -0.5$, $\beta_3 = 0.4$ and $\beta_4 = 0.5$ for empirical level. We take different values of β_1 as given in Table 5.7 for empirical power. For each distribution 10,000 replicated samples are taken for sample size $n = 10$, 20, 30 and 50. For the forward selection procedure we consider $\alpha = 0.05$. The level and power results are presented in Table 5.7.

Table 5.7 shows that the level of forward selection using the score test performs better than the other two procedures when n is small ($n = 10$) although the level is not very close to the nominal level. BIC performs better than other two procedures

Table 5.7: Empirical level (EL) and power (in %) of model selection by forward selection using score test, AIC and BIC in beta binomial distribution; based on 10,000 replications

Size (n)	Method	EL	Empirical Power									
			β_1									
		0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	Forward	17.84	18.92	20.69	22.98	25.35	27.07	30.47	32.34	34.98	38.72	40.37
	AIC	33.51	34.87	35.05	36.96	40.18	43.09	46.37	50.91	53.62	57.17	61.85
	BIC	29.78	31.25	31.38	33.29	36.38	39.43	42.82	46.93	50.12	53.31	58.59
20	Forward	25.78	27.43	29.44	33.53	36.98	42.40	47.14	50.83	56.05	59.98	63.42
	AIC	22.89	24.35	28.80	33.22	39.70	46.92	54.29	61.23	68.35	72.86	78.66
	BIC	14.38	15.20	18.69	22.93	28.52	35.12	42.34	49.20	56.89	62.18	68.74
30	Forward	31.41	34.58	37.40	41.52	47.42	52.12	59.21	64.15	68.85	73.76	77.53
	AIC	20.28	22.39	29.12	38.93	49.30	59.18	70.23	78.53	84.62	89.56	93.50
	BIC	9.71	11.21	15.91	24.09	33.12	42.49	54.51	63.57	72.38	79.23	85.35
50	Forward	41.50	43.30	48.17	53.46	58.68	64.90	72.01	78.06	82.12	86.85	89.86
	AIC	18.74	21.58	31.52	45.13	59.25	71.97	82.52	89.80	93.94	97.02	98.53
	BIC	6.46	8.26	15.11	24.86	37.34	51.01	64.62	76.37	83.81	90.37	94.38

when n is large ($n = 50$). Empirical power of model selection using AIC and BIC are close.

5.5 Real Data Analysis

To further illustrate the effectiveness of the model selection procedure, we apply this in three real data sets, namely, normal data, count data and over-dispersed data respectively. The first two data sets are previously analysed in the book Applied Linear Statistical Models by Kutner *et al.* (2013) and both data sets are available on the website
"http://users.stat.ufl.edu/ rrandles/sta4210/Rclassnotes/data/textdatasets/index.html". The third data set is previously analysed by Lawless (1987) and is provided in this Chapter.

Example: 1 We consider a data set about predicting survival time in patients undergoing a particular type of liver operation. From each of 54 patients, the following information is taken under consideration as covariates: x_1: blood clotting score, x_2:

prognostic index, x_3: enzyme function test score, x_4: liver function test score, x_5: age, in years, x_6: indicator variable for gender (0 = male, 1 = female), and x_7 and x_8 represent indicator variables for history of alcohol use as follows

Alcohol Use	x_7	x_8
None	0	0
Moderate	1	0
Severe	0	1

See Kutner *et al.* (2013), page 350.

The response variable is survival time and following Kutner *et al.* (2013) a log transformation is applied to the response variable to make the normality assumption more plausible. In Table 5.8 we provide the variables that entering the model associated with each step of analysis by using forward selection procedure through the score test and find that 4 covariates (blood clotting, prognostic index, enzyme function test score, and history of severe alcohol use) are significant out of 8 covariates, a similar conclusion to Kutner *et al.* (2013) where they apply forward selection using the F test.

Table 5.8: Variable to enter model using forward selection procedure through score test

Step	1st	2nd	3rd	4th
Variable	x_3	x_2	x_8	x_1
Score test	23.09	22.24	18.40	12.61

Example: 2 We consider a count data set which is about the total number of customers who live within a 10-mile radius of Miller Lumber Company stores and visited during a two-week period. A survey is conducted with 110 customers and the following information is taken under consideration as covariates: x_1: number of housing units, x_2: average income, in dollars, x_3: average housing unit age, in years,

x_4: distance to nearest competitor, in miles and x_5: distance to store, in miles. The response variable y is the number of customers who visited the stores.

Using the forward selection procedure through the score test we find that all covariates are significant which agrees with the analysis done in the book where they use the likelihood ratio test. Table 5.9 presents the variables that entering the model associated with each step of analysis by using forward selection through the score test.

Table 5.9: Variable to enter model using forward selection procedure through score test

Step	1st	2nd	3rd	4th	4th
Variable	x_5	x_4	x_2	x_1	x_3
Score test	258.13	36.00	13.83	15.63	4.38

Example: 3 We use the data from Lawless (1987) on Ames salmonella assay given in Margolin, Kaplan, and Zeiger (1981). Margolin *et al.* (1981) reported that quinoline was studied by W. Speck (unpublished) as part of the Environmental Mutagenesis Test Development Program at the National Institute of Environmental Health Sciences. The data in Table 5.10 were obtained from a test with Salmonella strain TA98 and induced rat liver homogenate S9. Dimethyl Sulfoxide was the solvent control, and each was replicated three times. So, the regression variable is the dose level having values $0, 10, 33, 100, 333$ and 1000, and for each dosage there are 3 observations.

Table 5.10: Ames salmonella assay Data

x	0	10	33	100	333	1000
	15	16	16	27	33	20
y	21	18	26	41	38	27
	29	21	33	60	41	42

For finding the appropriate model, we take x_1 as dose level, $x_2 = \log(x_1 + 10)$. Lawless (1987) used an approximation to Morgan *et al.'s* (1981) single hit model,

namely assume the full model satisfies

$$m_i = \exp(\beta_0 + \beta_1 x_i + \beta_2 \log(x_i + 10)).$$

Our purpose is to select as simple a model as possible. Table 5.11 represents values of χ^2-statistics of the variables entering the model through forward selection using three test statistics associated with each step of analysis.

Table 5.11: Test statistic value for variable to enter model using score, Wald and LR test statistics

Step	Variable	Score test χ^2	Wald test χ^2	LR test χ^2
1	x_2	4.03	6.15	4.97
2	x_1	4.71	6.44	5.56

Thus, the final model selected by the forward selection method using all the test statistics is $E(y_i) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$.

5.6 Summary

In this chapter, the score test procedure is developed for testing the significance of any covariate in generalized linear models. Simulation studies show that the score test performs satisfactorily and better than Wald and the likelihood ratio test in terms of empirical level and power under various choices of sample size. Model selection procedure is further developed using the score test through forward selection and is made a comparison with model selection using AIC and BIC. Model selection using the F test is also considered. Simulation studies show that, for small to moderate sample sizes, forward selection using the score test performs better in terms of level and power

than all other selection techniques in all distributions except normally distributed data where the F test performs better. Model selection on negative binomial and beta binomial regression models are also developed and similar conclusions are drawn. Thus, our recommendation is to use the forward selection procedure using the F test for normal regression models. For Poisson, binomial, negative binomial and beta binomial regression models, our recommendation is to use the forward selection procedure using the score test for small to moderate sample sizes and for large n ($n \geq 50$) the BIC should be used as it is computationally much simpler.

Chapter 6

Model Selection in Zero-inflated Generalized Linear Models

6.1 Introduction

Count data or proportional data arise in many biological and epidemiological studies. Discrete generalized linear model (Poisson or binomial) is a very widely used and popular model in analysing of such data. However, in practice, it often occurs that a particular count (for example zero) may arise in the data more than the expected number. For example, consider the number of insurance claims for a certain type of risk. A discrete generalized linear model may not fit such data well, so a zero-inflated generalized linear model (ZIGLM) may be more appropriate (Mullahy, 1986; Lambert, 1992). For more details regarding ZIGLM readers are referred to Cameron and Trivedi (1998, 2005). Considerable attention has been given to the problem of estimating the parameters involved in the model. However, in practice, one is faced

with a large number of covariates from which we need to select potentially important covariates for the model that motivate us to develop the model selection procedure.

In this chapter we develop the model selection procedure in ZIGLM of which zero-inflated Poisson and zero-inflated binomial regression models are considered as special cases. For model selection using the hypothesis testing paradigm, first we require a suitable test statistic for testing the significance of any covariate which performs best in holding an appropriate level of significance and has overall best power. Then we need to find a model selection procedure using that suitable test statistic which has the best property with respect to empirical level and power.

In Section 6.2 we develop score test, Wald test and likelihood ratio test statistics to test the significance of single covariate in ZIGLM, which are then specialized for data from the zero-inflated Poisson and the zero-inflated binomial distribution. An extensive simulation study is conducted to compare the behaviour of the three test statistics, in terms of empirical level and power to test the significance of a single regression coefficient.

The model selection procedure is developed in Section 6.3. An extensive simulation study is performed to compare the properties of the forward selection procedure using the three test statistics. For comparison we further include model selection using the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978). Two examples are presented in Section 6.4 and a discussion follows in Section 6.5.

6.2 Zero-inflated Generalized Linear Models and the Test Statistics

6.2.1 Zero-inflated Generalized Linear Models

Consider the natural exponential family distribution with probability density function $f(y, \theta)$ as

$$f(y; \theta) = \exp[a(\theta)y - g(\theta) + c(y)],$$

where y represents the response variable and θ is an unknown parameter. The zero-inflated exponential family is defined by

$$P(Y = y) = \begin{cases} \omega + (1 - \omega)f(0; \theta), & y = 0 \\ (1 - \omega)f(y; \theta), & y > 0, \end{cases} \tag{6.1}$$

where ω presents the zero inflation parameter. Note that it is possible to take $\omega < 0$, provided that $\omega \geq -f(0, \theta)/(1-f(0, \theta))$. The mean and variance of y in this model are $(1 - \omega)g'(\theta)/a'(\theta)$ and $(1 - \omega)(a'(\theta))^2 \left(g''(\theta) - \dfrac{a''(\theta)g'(\theta)}{a'(\theta)} + \omega(g'(\theta))^2 \right)$ respectively (Deng and Paul, 2000).

Let $y_1, y_2, \ldots, y_n$ be a random sample of size n from the distribution (6.1) with $\theta_i = \theta_i(X_i; \boldsymbol{\beta})$, a function of some $p \times 1$ vector of covariates X_i and a vector of regression parameters $\boldsymbol{\beta}$ for $i = 1, \ldots, n$. Using $\gamma = \dfrac{\omega}{1 - \omega}$, the log-likelihood can be

written as

$$\begin{aligned} l(\gamma,\theta;y) &= \sum_{i=1}^{n} l_i(\gamma,\theta_i;y_i) \\ &= \sum_{i=1}^{n}\left[-\log(1+\gamma)+I_{\{y_i=0\}}\log(\gamma+f(0;\theta_i))+I_{\{y_i>0\}}\log f(y_i;\theta_i)\right] \\ &= \sum_{i=1}^{n}[-\log(1+\gamma)+I_{\{y_i=0\}}\log\left(\gamma+\exp(-g(\theta_i)+c(0))\right)+ \\ &\quad I_{\{y_i>0\}}\left(a(\theta_i)y_i-g(\theta_i)+c(y_i)\right)], \end{aligned}$$

where I is the indicator function. For our convenience, replace $f(0;\theta_i), a(\theta_i)$ and $g(\theta_i)$ with f_0, a_i and g_i respectively. The first and second order partial derivatives of l with respect to the parameters θ_i and γ and their expected values are given in Appendix B.

To test the null hypothesis $H_0 : \beta_j = 0$ against $H_a : \beta_j \neq 0$ involved in $\theta_i = \theta_i(X_i;\boldsymbol{\beta})$, we derive the score test, Wald test and likelihood ratio test statistics in subsequent sections.

6.2.2 The Score Test

Similar to Chapter 5, omitting details, the score test statistic takes the form $S_j = S'(D-AB^{-1}A')^{-1}S$ where the parameter of interest is β_j and the nuisance parameters are $\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p$ and γ with

$$S = \frac{\partial l}{\partial \beta_j}\Big|_{\beta_j=0}, \qquad D = E\left[-\frac{\partial^2 l}{\partial \beta_j^2}\right]\Bigg|_{\beta_j=0},$$

$A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$ and $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}' & B_{22} \end{pmatrix}$, where

$$\begin{aligned} A_1 &= E\left[-\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right]\Bigg|_{\beta_j=0} (k \neq j), \qquad A_2 = E\left[-\frac{\partial^2 l}{\partial\beta_j\partial\gamma}\right]\Bigg|_{\beta_j=0}, \\ B_{11} &= E\left[-\frac{\partial^2 l}{\partial\beta_k\partial\beta_t}\right]\Bigg|_{\beta_j=0} (k,t \neq j), \qquad B_{22} = E\left[-\frac{\partial^2 l}{\partial\gamma^2}\right]\Bigg|_{\beta_j=0}, \\ B_{12} &= E\left[-\frac{\partial^2 l}{\partial\beta_k\partial\gamma}\right]\Bigg|_{\beta_j=0} (k \neq j). \end{aligned}$$

Hence

$$S = \frac{\partial l}{\partial\beta_j}\Big|_{\beta_j=0} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\beta_j}\Big|_{\beta_j=0} = \sum_{i=1}^{n}\left(I_{\{y_i=0\}}\frac{f_0(-g_i')}{\gamma+f_0} + I_{\{y_i>0\}}(a_i'y_i - g_i')\right)\frac{\partial\theta_i}{\partial\beta_j}\Big|_{\beta_j=0}.$$

Using $E\left[I_{\{y_i=0\}}\right] = \dfrac{\gamma+f_0}{1+\gamma}$ and $E\left[I_{\{y_i>0\}}\right] = \dfrac{1-f_0}{1+\gamma}$ and assume W be a $n \times n$ diagonal matrix with i-th diagonal element

$$w_i = E\left[-\frac{\partial^2 l_i}{\partial\theta_i^2}\right] = -\frac{\gamma f_0 (g_i')^2}{(1+\gamma)(\gamma+f_0)} + \frac{g_i''}{1+\gamma} - \frac{1-f_0}{1+\gamma}a_i'' E(y_i),$$

and U be a $p \times 1$ vector with i-th element

$$u_i = E\left[-\frac{\partial^2 l_i}{\partial\theta_i\partial\gamma}\right] = -\frac{f_0 g_i'}{(1+\gamma)(\gamma+f_0)},$$

and

$$a = E\left[-\frac{\partial^2 l}{\partial\gamma^2}\right] = \sum_{i=1}^{n}\frac{1-f_0}{(1+\gamma)^2(\gamma+f_0)}.$$

Let Z be an $n \times p$ matrix with (i,k)-th element $\dfrac{\partial \theta_i}{\partial \beta_k}$, $(i, k \neq j)$ and $\mathbf{z}_j = (\dfrac{\partial \theta_1}{\partial \beta_j}, \cdots, \dfrac{\partial \theta_n}{\partial \beta_j})'$.

Thus

$$\begin{aligned} A_1 &= E\left[-\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]\Big|_{\beta_j=0} (k \neq j) \\ &= \sum_{i=1}^{n} E\left[-\frac{\partial^2 l_i}{\partial \theta_i^2}\frac{\partial \theta_i}{\partial \beta_j}\frac{\partial \theta_i}{\partial \beta_k} - \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}\frac{\partial \theta_i}{\partial \beta_k}\right]\Big|_{\beta_j=0} \\ &= \mathbf{z}_j' W Z, \\ A_2 &= E\left[-\frac{\partial^2 l}{\partial \beta_j \partial \gamma}\right]\Big|_{\beta_j=0} = \sum_{i=1}^{n} E\left[-\frac{\partial^2 l_i}{\partial \theta_i \partial \gamma}\frac{\partial \theta_i}{\partial \beta_j}\right]\Big|_{\beta_j=0} = \mathbf{z}_j' U. \end{aligned}$$

Similarly

$$\begin{aligned} B_{11} &= E\left[-\frac{\partial^2 l}{\partial \beta_k \partial \beta_t}\right]\Big|_{\beta_j=0} (k,t \neq j) = Z'WZ, \qquad & B_{22} &= E\left[-\frac{\partial^2 l}{\partial \gamma^2}\right]\Big|_{\beta_j=0} = a, \\ B_{12} &= E\left[-\frac{\partial^2 l}{\partial \beta_k \partial \gamma}\right]\Big|_{\beta_j=0} (k \neq j) = Z'U, & D &= E\left[-\frac{\partial^2 l}{\partial \beta_j^2}\right]\Big|_{\beta_j=0} = \mathbf{z}_j' W \mathbf{z}_j. \end{aligned}$$

None of the expected value of mixed partial derivatives equal zero, so $AB^{-1}A' = A_1 B_{11}^{-1} A_1' + A_2 B_{21}^{-1} A_1' + A_1 B_{12}^{-1} A_2' + A_2 B_{22}^{-1} A_2'$, where

$$\begin{aligned} A_1 B_{11}^{-1} A_1' &= \mathbf{z}_j' W Z V_1^{-1} Z' W \mathbf{z}_j, \qquad & A_2 B_{21}^{-1} A_1' &= -\frac{1}{a}\mathbf{z}_j' U U' Z V_1^{-1} Z' W \mathbf{z}_j, \\ A_1 B_{12}^{-1} A_2' &= -\mathbf{z}_j' W Z \left(Z'WZ\right)^{-1} Z' U V_2^{-1} U' \mathbf{z}_j, & A_2 B_{22}^{-1} A_2' &= \mathbf{z}_j' U V_2^{-1} U' \mathbf{z}_j, \end{aligned}$$

with $V_1 = Z'\left(W - \frac{1}{a}UU'\right)Z$ and $V_2 = a - U'Z\left(Z'WZ\right)^{-1} Z'U$.

After some simplification

$$AB^{-1}A' = \mathbf{z}_j' \left[\left(W - \frac{1}{a}UU'\right) Z V_1^{-1} Z' W + \left(I - WZ\left(Z'WZ\right)^{-1} Z'\right) U V_2^{-1} U'\right] \mathbf{z}_j.$$

Hence $V_j = D - AB^{-1}A'$ becomes

$$\mathbf{z}_j' \left[W - \left(W - \frac{1}{a} UU' \right) ZV_1^{-1} Z'W - \left(I - WZ \left(Z'WZ \right)^{-1} Z' \right) UV_2^{-1} U' \right] \mathbf{z}_j.$$

Therefore the score statistic is

$$S_j = \frac{\hat{S}^2}{\hat{V}_j},$$

where

$$\begin{aligned} S &= \sum_{i=1}^{n} \left(I_{\{y_i=0\}} \frac{f_0(-g_i')}{\gamma + f_0} + I_{\{y_i>0\}} (a_i' y_i - g_i') \right) \frac{\partial \theta_i}{\partial \beta_j} \Big|_{\beta_j=0}, \quad \text{and} \\ V_j &= \mathbf{z}_j' \left[W - \left(W - \frac{1}{a} UU' \right) ZV_1^{-1} Z'W - \left(I - WZ \left(Z'WZ \right)^{-1} Z' \right) UV_2^{-1} U' \right] \mathbf{z}_j \end{aligned}$$

with $\hat{S}$ and $\hat{V}_j$ replaced by $S(\hat{\theta}, \hat{\gamma})$ where $\hat{\theta}$ and $\hat{\gamma}$ are the maximum likelihood estimates under the null hypothesis.

6.2.3 The Wald Test

The Wald test statistic is given by

$$W_j = \frac{\tilde{\beta}_j}{\sqrt{\text{Var}(\tilde{\beta}_j)}} = \frac{\tilde{\beta}_j}{\sqrt{\sum_{i=1}^{n} \tilde{w}_i \left(\frac{\partial \tilde{\theta}_i}{\partial \beta_j} \right)^2}},$$

where w_i and $\dfrac{\partial \theta_i}{\partial \beta_j}$ are given in Section 6.2.2 and $\tilde{w}_i$ and $\tilde{\theta}_i$ are evaluated under alternative hypothesis.

6.2.4 The Likelihood Ratio Test

Let $\hat{l}$ and $\tilde{l}$ be the maximized log-likelihood under the null and the alternative hypothesis respectively. Then, the likelihood ratio statistic is

$$\begin{aligned} LR_j &= 2\left(\tilde{l}-\hat{l}\right) \\ &= 2\sum_{i=1}^{n}\left[-\log(1+\tilde{\gamma})+I_{\{y_i=0\}}\log(\tilde{\gamma}+\exp(-g(\tilde{\theta}_i)+c(0)))+I_{\{y_i>0\}}\{a(\tilde{\theta}_i)y_i-g(\tilde{\theta}_i)\}\right]- \\ &\quad 2\sum_{i=1}^{n}\left[-\log(1+\hat{\gamma})+I_{\{y_i=0\}}\log(\hat{\gamma}+\exp(-g(\hat{\theta}_i)+c(0)))+I_{\{y_i>0\}}\{a(\hat{\theta}_i)y_i-g(\hat{\theta}_i)\}\right]. \end{aligned}$$

Note that under some regularity conditions, S_j, W_j^2 and LR_j asymptotically converge to $\chi^2(1)$. Therefore, for a fixed significance level $\alpha > 0$, we reject the null hypothesis, if each test statistic is greater than $\chi^2_\alpha(1)$.

6.2.5 Special Cases

We now give the expressions for the three test statistics S_j, W_j, and LR_j for the special cases for which the data distribution is zero-inflated Poisson and zero-inflated binomial respectively.

(i) Zero-inflated Poisson Regression Model

For the zero-inflated Poisson distribution ZIP(λ, γ), Suppose $\theta_i = \log\lambda_i = \mathbf{x}_i'\boldsymbol{\beta}$. Then we have $a(\theta_i) = \theta_i, g(\theta_i) = e^{\theta_i}$. In this case $Z = X$ and $\mathbf{z}_j = \mathbf{x}_j$. Therefore the score

test statistic for testing $H_0 : \beta_j = 0$ is

$$SP_j = \frac{\left(\sum_{i=1}^{n}\left(\frac{-\hat{\lambda}_i e^{-\hat{\lambda}_i}}{\hat{\gamma}+e^{-\hat{\lambda}_i}} I_{\{y_i=0\}} + (y_i - \hat{\lambda}_i) I_{\{y_i>0\}}\right) x_{ij}\right)^2}{\hat{V}_j},$$

where $\hat{\lambda}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip})$ and $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$ are the maximum likelihood estimates of $\boldsymbol{\beta}$ and γ under the null hypothesis and $\hat{V}_j = V_j(\hat{\lambda}, \hat{\gamma})$ with

$$V_j = \mathbf{x}_j' \left[W - \left(W - \frac{1}{a} UU' \right) X V_1^{-1} X' W - \left(I - WX \left(X'WX \right)^{-1} X' \right) U V_2^{-1} U' \right] \mathbf{x}_j,$$

$$w_i = \frac{-\lambda_i e^{-\lambda_i}\gamma + \gamma + e^{-\lambda_i}}{(1+\gamma)(\gamma + e^{-\lambda_i})}\lambda_i, \qquad u_i = \frac{-\lambda_i e^{-\lambda_i}}{(1+\gamma)(\gamma + e^{-\lambda_i})}, \quad \text{and}$$

$$a = \sum_{i=1}^{n} \frac{1 - e^{-\lambda_i}}{(1+\gamma)^2(\gamma + e^{-\lambda_i})}.$$

The Wald test and likelihood ratio test statistics are

$$\begin{aligned} WP_j &= \frac{\tilde{\beta}_j}{\sqrt{\sum_{i=1}^{n} \frac{-\tilde{\lambda}_i e^{-\tilde{\lambda}_i}\tilde{\gamma} + \tilde{\gamma} + e^{-\tilde{\lambda}_i}}{(1+\tilde{\gamma})(\tilde{\gamma}+e^{-\tilde{\lambda}_i})} \tilde{\lambda}_i x_{ij}^2}}, \quad \text{and} \\ LRP_j &= 2[\Big(-\log(1+\tilde{\gamma}) + I_{\{y_i=0\}} \log(\tilde{\gamma} + e^{-\tilde{\lambda}_i}) + I_{\{y_i>0\}}(y_i \log \tilde{\lambda}_i - \tilde{\lambda}_i) \Big) - \\ &\quad \Big(-\log(1+\hat{\gamma}) + I_{\{y_i=0\}} \log(\hat{\gamma} + e^{-\hat{\lambda}_i}) + I_{\{y_i>0\}}(y_i \log \hat{\lambda}_i - \hat{\lambda}_i) \Big)], \end{aligned}$$

where $\tilde{\lambda}_i = \exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_p x_{ip})$ and $\hat{\lambda}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip})$.

The maximum likelihood estimates of β_k $(\forall k \neq j; k = 1, \ldots, p)$ and γ are obtained

by solving the equations

$$\sum_{i=1}^{n}\left[I_{\{y_i=0\}}\frac{-\lambda_i e^{-\lambda_i}}{\gamma+e^{-\lambda_i}}+I_{\{y_i>0\}}(y_i-\lambda_i)\right]x_{ik}=0,\quad \text{and}$$
$$\sum_{i=1}^{n}\left[-\frac{1}{1+\gamma}+I_{\{y_i=0\}}\frac{1}{\gamma+e^{-\lambda_i}}\right]=0.$$

(ii) Zero-inflated Binomial Regression Model

For the zero-inflated binomial distribution ZIBin(m,p,γ), assume $\theta_i=\log\left(\frac{p_i}{1-p_i}\right)=\mathbf{x}_i'\boldsymbol{\beta}$. Then $a(\theta_i)=\theta_i,\ g(\theta_i)=m_i\log(1+e^{\theta_i}),\ f_0=f(0;\theta_i)=(1-p_i)^{m_i}$. Here also $Z=X$ and $\mathbf{z}_j=\mathbf{x}_j$. Thus the score test statistics for testing $H_0:\beta_j=0$ is

$$SB_j=\frac{\left(\sum_{i=1}^{n}\left(\frac{-\hat{f}_0 m_i\hat{p}_i}{\hat{\gamma}+\hat{f}_0}I_{\{y_i=0\}}+(y_i-m_i\hat{p}_i)I_{\{y_i>0\}}\right)x_{ij}\right)^2}{\hat{V}_j},$$

where $\frac{\hat{p}_i}{1-\hat{p}_i}=\exp\left(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_{j-1}x_{i(j-1)}+\hat{\beta}_{j+1}x_{i(j+1)}+\cdots+\hat{\beta}_p x_{ip}\right)$ and $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$ are the maximum likelihood estimates of $\boldsymbol{\beta}$ and γ under the null hypothesis and $\hat{V}_j=V_j(\hat{p},\hat{\gamma})$ with

$$V_j=\mathbf{x}_j'\left[W-\left(W-\frac{1}{a}UU'\right)XV_1^{-1}X'W-\left(I-WX(X'WX)^{-1}X'\right)UV_2^{-1}U'\right]\mathbf{x}_j,$$
$$w_i=-\frac{\gamma f_0 m_i^2p_i^2}{(1+\gamma)(\gamma+f_0)}+\frac{m_ip_i(1-p_i)}{1+\gamma},$$
$$u_i=\frac{-m_ip_if_0}{(1+\gamma)(\gamma+f_0)},\quad \text{and}$$
$$a=\sum_{i=1}^{n}\frac{1-f_0}{(1+\gamma)^2(\gamma+f_0)}.$$

The Wald test and likelihood ratio test statistics are

$$WB_j = \frac{\tilde{\beta}_j}{\sqrt{\sum_{i=1}^{n}\left[-\frac{\tilde{\gamma}m_i^2(1-\tilde{p}_i)^{m_i}\tilde{p}_i^2}{(1+\tilde{\gamma})(\tilde{\gamma}+(1-\tilde{p}_i)^{m_i})}+\frac{m_i\tilde{p}_i(1-\tilde{p}_i)}{1+\tilde{\gamma}}\right]x_{ij}^2}}, \quad \text{and}$$

$$\begin{aligned} LRB_j = 2[&(-\log(1+\tilde{\gamma}) + I_{\{y_i=0\}}\log(\tilde{\gamma}+(1-\tilde{p}_i)^{m_i})+ \\ &I_{\{y_i>0\}}(y_i\log\tfrac{\tilde{p}_i}{1-\tilde{p}_i} + m_i\log(1-\tilde{p}_i)))- \\ &(-\log(1+\hat{\gamma}) + I_{\{y_i=0\}}\log(\hat{\gamma}+(1-\hat{p}_i)^{m_i})+ \\ &I_{\{y_i>0\}}(y_i\log\tfrac{\hat{p}_i}{1-\hat{p}_i} + m_i\log(1-\hat{p}_i)))], \end{aligned}$$

where $\frac{\tilde{p}_i}{1-\tilde{p}_i} = \exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_p x_{ip})$ and $\frac{\hat{p}_i}{1-\hat{p}_i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1}x_{i(j-1)} + \hat{\beta}_{j+1}x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip})$.

Note that the maximum likelihood estimates of β_k $(\forall k \neq j; k = 1, \ldots, p)$ and γ are obtained by solving the equations

$$\sum_{i=1}^{n}\left(-\frac{f_0 m_i p_i}{\gamma+f_0}I_{\{y_i=0\}} + (y_i - m_i p_i)I_{\{y_i>0\}}\right)x_{ik} = 0, \quad \text{and}$$

$$\sum_{i=1}^{n}\left(-\frac{1}{1+\gamma} + \frac{I_{\{y_i=0\}}}{\gamma+f_0}\right) = 0.$$

6.2.6 Simulation

A simulation study is conducted to assess the performance of the three test statistics, in terms of empirical level and power, for testing the significance of a single regression coefficient. We consider a two-variable regression model with link function $\lambda = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, and $\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ for ZIP(λ, γ), and ZIBin(m, p, γ) distributions respectively.

Suppose our interest is to test $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$ in each case. For

empirical level in zero-inflated Poisson distributed data we take $\beta_0 = 1$ and $\beta_1 = -0.3$, and $\beta_2 = 0$. For power we take $\beta_0 = 0.3$ and $\beta_1 = -0.9$ and different values of β_2 as given in Table 6.1.

For zero-inflated binomial distributed data we take $m = 40$, $\beta_0 = 0.2$, $\beta_1 = -0.1$ and $\beta_2 = 0$ to calculate level and for power we take $\beta_0 = 0.2$, $\beta_1 = -0.1$ and different values of β_2 as given in Table 6.1. For both distributions we take $\gamma = 0.25$.

Table 6.1: Empirical level (EL) and power (in %) of the three test statistics; $\alpha = 0.05$

Distr	Size (n)	Test	EL	Empirical Power										
				β_2										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ZIP	10	Score	5.13	5.66	6.25	8.52	10.53	14.37	17.79	23.34	27.32	32.48	36.93
		Wald	12.20	12.69	13.07	15.84	18.81	23.04	27.86	33.75	39.86	46.71	51.62
		LR	7.99	8.43	9.30	11.68	14.54	18.90	23.26	29.50	35.39	42.20	46.88
	20	Score	5.18	6.42	8.96	13.95	21.11	28.55	38.18	47.38	55.71	64.21	71.11
		Wald	8.50	10.39	12.41	18.44	25.68	35.22	45.21	55.56	66.38	74.46	81.33
		LR	6.44	7.83	10.09	15.70	23.27	32.29	42.71	52.59	62.65	71.68	78.52
	30	Score	4.93	6.23	12.12	19.05	30.00	42.82	55.80	66.57	75.94	82.76	88.39
		Wald	6.99	8.32	14.61	22.45	33.86	47.75	61.42	73.62	82.54	89.40	93.90
		LR	5.73	7.02	13.03	20.44	31.77	45.43	59.25	71.03	80.21	87.48	92.36
	40	Score	4.85	7.14	13.88	24.03	39.59	54.72	69.03	80.12	87.48	92.58	95.35
		Wald	6.57	8.89	15.86	26.14	42.71	58.64	74.14	84.71	91.91	95.75	97.78
		LR	5.45	7.87	14.54	24.68	41.22	57.41	72.22	83.20	90.67	94.78	97.29
	50	Score	5.09	8.20	16.27	30.75	47.73	65.84	78.99	88.59	93.76	96.82	98.51
		Wald	6.23	9.21	17.48	32.55	50.86	69.27	82.11	91.42	95.98	98.48	99.40
		LR	5.40	8.34	16.52	31.15	49.12	67.87	81.14	90.33	95.18	97.89	99.19
ZIBin	10	Score	4.24	6.35	11.52	20.26	31.01	42.04	54.07	63.27	70.82	76.97	81.00
		Wald	7.10	9.45	14.76	23.58	35.72	47.88	61.18	70.66	78.34	84.98	88.81
		LR	4.96	7.14	12.35	21.27	32.84	45.06	57.92	67.55	75.05	81.57	86.05
	20	Score	5.00	8.66	21.67	40.10	60.34	76.15	86.04	93.07	96.04	97.72	98.62
		Wald	6.08	9.69	22.76	42.29	63.43	79.85	89.38	95.68	98.05	99.21	99.67
		LR	5.32	8.95	22.20	41.18	62.35	78.54	88.42	94.80	97.70	98.79	99.46
	30	Score	5.04	11.09	29.72	56.89	79.79	90.85	96.38	98.80	99.51	99.81	99.96
		Wald	5.59	11.83	30.90	59.08	81.80	93.04	97.86	99.37	99.79	99.93	100
		LR	5.16	11.32	30.27	58.73	81.28	92.33	97.37	99.18	99.71	99.90	99.99
	40	Score	4.86	13.23	39.62	70.43	89.00	97.07	99.25	99.81	99.99	99.99	100
		Wald	5.33	13.80	41.05	72.20	90.52	98.04	99.60	99.89	100	100	100
		LR	4.99	13.59	40.45	71.68	90.01	97.72	99.53	99.86	100	100	100
	50	Score	4.94	15.76	48.08	79.42	94.95	98.94	99.92	99.99	100	100	100
		Wald	5.38	16.02	49.13	80.93	95.91	99.28	99.97	99.99	100	100	100
		LR	5.12	15.83	48.82	80.40	95.52	99.13	99.94	100	100	100	100

Results in Table 6.1 show that for data from the zero-inflated Poisson and zero-inflated binomial distribution, the score test holds nominal level very well and power performance of all three tests for both distributions is also similar. LR test also performs better for the zero-inflated binomial distribution.

As the score test and likelihood ratio tests perform well, in Section 6.3, we again consider all three test statistics in the study of the performance of the model selection procedures.

6.3 Model Selection

Model selection criteria mainly used here are based on hypothesis testing using the score test, Wald test and likelihood ratio test. In general, our main focus is to select the most significant explanatory variables to simplify the regression model. There are three widely used approaches in stepwise regression: forward selection, backward elimination and bidirectional selection. The main tool here we use is forward selection procedure. The other two procedures are not included in our study, as in most situations these procedures produce the same final model. We further consider model selection using AIC and BIC for comparison. For more details about these procedures readers may see Draper and Smith (1998).

We apply the following procedure for calculating the empirical level using a p variable regression model, for example, zero-inflated Poisson regression model with $\ln(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. We first generate a sample of size n from the zero-inflated Poisson ZIP(λ, γ) distribution for given values of the regression parameters and simulated values of the regression variables. We then use the test statistic for testing $H_0 : \beta_j = 0$ and a model selection procedure, for example, the forward selection procedure and find a model of a subset of the regression variables. We repeat this process 10,000 times and find 10,000 models. If the given value of β_j is very small, we want to see if the regression variable x_j is in the final model. We then count the number of models in which the variable x_j is included. Let this number be c. Then

the empirical level for rejecting $H_0 : \beta_j = 0$ is $c/10,000$. Empirical power is calculated similarly by taking a larger value of β_j during the simulation process.

We intend here to make a comparative study of performance of forward selection using the score test, Wald test and LR test statistics; and model selection using the AIC and BIC with respect to level and power.

6.3.1 Simulation

A simulation study is conducted to examine the performance of model selection through the score test, Wald test and LR test using forward selection in terms of empirical level and power, and the test are compared with one another. We further consider model selection using AIC and BIC.

We consider a 4-variable regression model. Data are drawn from the zero-inflated Poisson ZIP(λ, γ) regression model and the zero-inflated binomial ZIBin(m, p, γ) regression model with

$$\begin{aligned} \lambda &= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4), \quad \text{and} \\ \frac{p}{1-p} &= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4) \end{aligned}$$

respectively. Suppose we would like to test $H_0 : \beta_1 = 0$. To calculate the empirical level for each distribution we choose $\beta_1 = 0.001$ (a very small value) and for empirical power we take different values of β_1 as given in Table 5.3. The rest of the parameters are set at $\gamma = 0.25$, $m = 40$, $\beta_2 = -0.3$, $\beta_3 = 0.2$, $\beta_4 = 0.3$ for zero-inflated Poisson and zero-inflated binomial distribution. For each distribution 10,000 replicated samples are taken for sample size $n = 10$, 20, 30, 40 and 50.

For forward selection procedure we consider $\alpha = 0.05$. Note that for the other two procedures α cannot be fixed. The level and power of model selection in zero-inflated Poisson and zero-inflated binomially distributed data are presented in Table 6.2 and Table 6.3 respectively. In both tables For-S, For-W and For-L represent forward selection using score, Wald and Likelihood ratio test respectively.

Table 6.2: Empirical level (EL) and power (in %) of model selection in zero-inflated Poisson distribution

Dist.	Size (n)	Method	EL	Empirical Power										
				β_1										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ZIP	10	For-S	7.07	7.11	8.12	9.71	11.67	15.09	19.25	22.05	26.32	29.78	33.71
		For-W	13.82	13.81	14.88	16.28	18.72	21.44	26.30	30.26	35.05	39.07	44.62
		For-L	9.37	9.04	10.09	11.60	14.35	17.22	22.09	25.64	30.19	34.72	40.16
		AIC	26.65	26.83	28.80	30.86	33.89	38.44	41.83	46.41	50.95	56.89	59.95
		BIC	22.83	23.31	24.89	26.94	30.21	34.36	39.01	42.76	47.37	54.04	56.81
	20	For-S	7.36	8.19	11.06	14.88	20.29	27.77	35.62	43.30	50.38	57.86	63.78
		For-W	10.93	11.91	14.67	19.21	25.28	32.96	41.09	49.68	57.88	65.77	72.80
		For-L	7.92	9.04	11.62	16.03	21.97	29.42	38.17	46.26	54.82	62.73	69.34
		AIC	20.96	22.11	27.20	33.75	41.22	50.08	58.76	67.43	74.37	80.25	84.48
		BIC	13.20	13.58	17.59	23.55	30.39	39.16	47.32	57.02	65.46	71.97	77.81
	30	For-S	7.26	8.72	13.13	19.99	28.95	38.99	49.53	60.61	69.53	76.90	82.51
		For-W	9.07	10.82	15.42	22.79	32.01	43.00	54.99	65.97	75.74	83.04	88.66
		For-L	7.60	8.94	13.05	20.41	29.88	40.77	52.35	63.71	73.07	80.68	86.65
		AIC	19.67	22.13	28.84	39.29	51.97	62.41	73.21	82.15	87.92	92.76	95.50
		BIC	9.58	11.90	16.76	25.27	36.45	46.65	59.44	70.30	78.27	85.74	90.93
	40	For-S	7.13	9.16	15.33	24.88	37.30	51.52	64.21	75.10	83.03	89.08	92.87
		For-W	8.58	10.38	16.23	26.66	40.01	54.52	67.85	79.45	87.42	92.65	96.08
		For-L	7.08	9.09	14.80	25.06	38.27	52.48	65.81	77.60	85.54	91.54	95.19
		AIC	18.30	21.21	32.34	45.26	60.28	74.29	84.07	91.11	95.16	97.22	98.55
		BIC	7.84	9.32	17.68	27.27	41.34	56.45	69.75	80.17	88.57	92.83	95.96
	50	For-S	6.86	9.41	17.51	30.59	46.42	62.72	75.13	84.77	91.27	94.92	96.97
		For-W	7.60	10.49	18.63	31.53	49.05	65.65	78.73	87.97	93.82	97.04	98.63
		For-L	6.58	9.17	17.20	30.30	47.36	63.68	77.24	86.65	92.79	96.34	98.01
		AIC	17.47	22.45	34.87	51.81	68.33	81.90	90.47	95.55	97.98	99.22	99.67
		BIC	6.48	9.33	17.45	30.96	47.41	64.79	77.71	88.26	93.86	96.83	98.71

Both Tables 6.2 and 6.3 show that the forward selection method using score test always produces a reasonable empirical level (close to the nominal level) and better than the other four procedures irrespective of sample size. Model selection using AIC and BIC produce highly inflated type I error when sample size is small although the BIC does well for sample size $n = 50$. The powers of all procedures are almost similar.

Thus our recommendation is to use forward selection through the score test for

Table 6.3: Empirical level (EL) and power (in %) of model selection in zero-inflated binomial distribution

Dist.	Size (n)	Method	EL	Empirical Power										
				β_1										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ZIBin	10	For-S	9.59	10.61	13.27	18.43	23.96	30.67	38.34	46.13	53.20	59.11	64.95
		For-W	13.43	14.09	17.08	21.38	26.95	33.90	42.32	50.09	57.83	64.36	70.19
		For-L	9.50	10.58	13.36	18.38	24.37	31.52	39.54	47.45	55.32	61.63	67.64
		AIC	21.58	23.21	27.57	34.10	41.98	50.42	58.14	64.78	71.16	76.15	80.03
		BIC	19.06	20.46	24.55	30.99	38.39	46.94	55.09	61.96	68.61	74.02	78.38
	20	For-S	7.80	10.10	17.95	29.98	43.76	58.25	70.34	79.45	85.86	90.04	93.40
		For-W	9.29	11.79	19.80	31.89	46.79	61.52	73.21	82.66	89.04	93.24	95.84
		For-L	7.91	10.22	18.53	30.68	45.15	60.11	72.17	81.45	87.71	92.04	94.92
		AIC	18.09	21.98	33.43	50.21	65.70	77.73	87.08	92.57	95.74	97.67	98.60
		BIC	11.06	13.99	23.30	38.51	54.76	68.52	80.14	88.03	92.48	95.42	97.55
	30	For-S	6.29	10.77	23.73	43.24	62.89	78.04	88.22	93.99	96.83	98.48	99.20
		For-W	7.43	11.82	25.69	45.31	65.62	80.97	90.56	95.58	97.94	99.19	99.64
		For-L	6.57	10.84	24.61	44.32	64.57	80.01	89.75	95.10	97.63	98.98	99.51
		AIC	16.51	24.18	42.79	65.36	82.26	92.35	96.82	98.68	99.62	99.79	99.98
		BIC	7.70	12.42	28.06	48.77	69.66	84.98	92.37	96.70	98.69	99.39	99.84
	40	For-S	5.66	11.36	29.79	55.26	77.37	90.07	96.23	98.58	99.50	99.87	99.94
		For-W	6.23	12.02	31.35	58.38	80.00	91.86	97.40	99.13	99.77	99.96	99.97
		For-L	5.77	11.56	30.60	57.23	78.89	91.29	97.01	98.94	99.68	99.94	99.96
		AIC	15.61	26.69	51.96	76.48	91.63	97.23	99.28	99.86	99.95	99.99	99.99
		BIC	5.99	12.15	33.13	58.88	80.74	92.84	97.64	99.44	99.80	99.95	99.97
	50	For-S	5.26	13.19	36.56	66.93	87.40	96.06	99.01	99.74	99.94	99.98	100
		For-W	5.66	14.01	38.14	68.93	89.35	97.06	99.42	99.90	99.99	100	100
		For-L	5.45	13.50	37.40	68.09	88.66	96.64	99.28	99.85	99.99	100	100
		AIC	15.62	29.29	59.18	83.71	95.72	99.14	99.81	99.97	100	100	100
		BIC	5.17	13.48	37.76	67.56	88.42	96.98	99.32	99.87	99.97	100	100

small to moderate sample sizes as it has a very simple form, and does not need estimates of the regression parameters under the alternative hypothesis. For large sample size ($n \geq 50$), BIC can be used due to its simple form as well as easy of computation.

6.4 Real Data Analysis

To further illustrate the effectiveness of the proposed approach, we apply forward selection procedure in two zero-inflated real data sets. Both data sets are available on http://faculty.econ.ucdavis.edu/faculty/cameron/racd2/RACD2programs.html.

Example: 1 We consider a data set from Gurmu and Trivedi (1996) on 659 obser-

vations about their number of recreational boating trips to Lake Somerville, Texas in 1980. The covariates are x_1= facility's subjective quality ranking on 1 to 5; $x_2 = 1$ if engaged in water-skiing at the lake; x_3= household income of the head of the group; x_4= equal 1 if user's fee paid at Lake Somerville; x_5, x_6 and x_7 are dollar expenditure when visiting Lake Conroe, Lake Somerville and Lake Houston respectively. The number of recreational boating trips is considered as the dependent variable. The data are zero-inflated as more than 65% of the respondents reported taking no trips in the survey period.

We provide score test statistic value of the variables entering the model associated with each step in forward selection procedure in table 6.4.

Table 6.4: Variable to enter model using forward selection procedure through score test

Step	1st	2nd	3rd	4th	5th	6th
Variable	x_4	x_6	x_7	x_2	x_3	x_1
Score test	348.20	115.41	151.98	64.45	20.50	9.20

Table 6.4 show that all covariates are significant except x_5 which is consistent with Cameron and Trivedi (1998, pp-209).

Example: 2 We consider a data set from Deb and Trivedi (1997) on 4406 individuals, aged 66 and over, who are covered by Medicare public insurance program. The data are originally obtained from the US National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. About 15% of the visits are counted as zero counts.

We consider the number of physician office visits as the dependent variable and x_1 = number of hospital stays; x_2 = number of chronic conditions; x_3 = number of years of education; x_4 = family income; and x_5 = age, as covariates.

In table 6.5 we provide values of score test statistics of the corresponding variables that enter the model associated with each step in the forward selection procedure.

Table 6.5: Variable to enter model using forward selection procedure through score test

Step	1st	2nd	3rd	4th	5th
Variable	x_1	x_2	x_3	x_5	x_4
Score test	1551.51	811.90	89.75	17.51	11.30

Table 6.5 shows that all covariates are significant.

6.5 Discussion

In this chapter, we develop model selection procedure in ZIGLM using the score, Wald and LR test statistics. We further consider model selection using the AIC and BIC. Simulation studies show that forward selection using the score test and LR test perform better than the Wald test. Moreover model selection using AIC and BIC produces highly inflated type I error when sample size is small although the BIC does well for sample size $n = 50$. Thus our suggestion is to use the forward selection procedure using the score test for small to moderate sample sizes and for large n ($n \geq 50$) BIC should be used as it is computationally much simpler.

Chapter 7

Summary and Plan for Future Research

7.1 Summary

We develop the procedure of joint estimation of mean and covariance parameters semiparametrically for longitudinal data in Chapter 3. An extensive simulation study is done and overall findings are: both the parametric modelling and the semiparametric modelling produce similar bias and efficiency properties of the regression parameters; increasing the number of knots in the spline procedure decreases the efficiency of the estimates of the nonparametric functions; and use of the penalized spline does not improve the efficiency of the estimates of the nonparametric functions. However, the great benefit of the semiparametric modelling is shown in the analysis of three real data sets to find parsimonious models.

To incorporate time varying covariates in longitudinal data, we develop a joint esti-

mation procedure for the mean and the covariance parameters in generalized partially linear varying coefficient model by decomposing the correlation matrix via hyperspherical co-ordinates in Chapter 4. The simulation study and the real data analysis are used to illustrate the proposed approach.

In Chapter 5, score, Wald and likelihood ratio test statistics are developed to test the significance of a single covariate in generalized linear models. Simulation studies show that the score test maintains nominal level very well among all three test statistics under various choice of sample sizes. Based on the score test, the selection procedure is further developed through forward selection and is made a comparison using F test, AIC and BIC in terms of level and power. Model selection on negative binomial and beta binomial regression models are also developed. According to simulation studies, our suggestion is to use the forward selection procedure through the F test for normal regression models and for Poisson, binomial, negative binomial and beta binomial regression models one should use the forward selection procedure through the score test for small to moderate sample sizes and BIC for large n ($n \geq 50$).

In Chapter 6, a model selection procedure in ZIGLM is developed using the score, the Wald and the LR test statistics and show a comparison is made with model selection using AIC and BIC. Simulation studies show forward selection using the score test and the LR test perform well. Based on simulation studies, our recommendation is to use the forward selection procedure through the score test for small to moderate sample sizes and for large n ($n \geq 50$) BIC should be used as it is easy to compute.

7.2 Future Research

In longitudinal data, missing responses and covariate measurement error are very commonly seen in practice. However ignoring measurement error or omitting missing covariates may produce inconsistent estimators (Little and Rubin, 2002; Carroll *et al.*, 2012; Wang *et al.*, 2008; Yi *et al.*, 2012). The problem of how to handle either missing responses or covariate measurement error under the linear setting is well developed in the literature. However, properties of these in longitudinal settings are not well-known. Most recently Qin, Zang, and Zhu (2016) discuss this issue and propose simultaneous estimation of mean and covariance parameters of partially linear models in generalized estimating equation settings. It will be interesting to incorporate the correlation matrix directly via hyperspherical coordinates and develop a methodology to improve estimation of the parameters.

To estimate mean and covariance parameters simultaneously in longitudinal data most authors consider continuous data. However researchers often encounter longitudinal observations that contain a substantial number of discrete variables (see for example Lynn, 2009; Molenberghs and Verbeke, 2005). Recently Tang, Zhang, and Leng (2017) developed a procedure to estimate mean-correlation regression parameters for a family of discrete responses. Our plan is to extend their model semiparametrically as well as nonparametrically.

Appendix A

A.1 Solution of Estimating Equations of Model 3

We apply the quasi-Fisher scoring algorithm to solve estimating equations of Model 3 where the parameters $\boldsymbol{\theta}$, $\boldsymbol{\rho}$ and $\boldsymbol{\gamma}$ are solved sequentially one by one with other parameter kept fixed in optimization:

Step 1 : Choose initial values of the parameters as $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\rho}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$. Set $k = 0$

Step 2 : Calculate $\boldsymbol{\Sigma}_i$ by using $\boldsymbol{\rho}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$. Update $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + I_{11}^{-1}\boldsymbol{U}_1|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}$$

Step 3 : Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k+1)}$, update $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$ by using

$$\begin{pmatrix} \boldsymbol{\gamma}^{(k+1)} \\ \boldsymbol{\rho}^{(k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}^{(k)} \\ \boldsymbol{\rho}^{(k} \end{pmatrix} + \left[\begin{pmatrix} I_{22} & I_{23} \\ I_{32} & I_{33} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{U}_3 \\ \boldsymbol{U}_2 \end{pmatrix} \right] |_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(k)},\boldsymbol{\rho}=\boldsymbol{\rho}^{(k)}}$$

Step 4 : Set $k \leftarrow k + 1$ and repeat steps 2 and 3 until a desired convergence criteria is

satisfied.

Note that block components of Fisher information matrix I are:

$$
\begin{aligned}
I_{11} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \sum_{i=1}^{n}\boldsymbol{\Pi}_i'\Delta_i\Sigma_i^{-1}\Delta_i\boldsymbol{\Pi}_i,\\
I_{12} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\gamma}'}\right] = -\sum_{i=1}^{n}\left[\boldsymbol{\Pi}_i'\Delta_i\frac{\partial\Sigma_i^{-1}}{\partial\boldsymbol{\gamma}'}(E(\mathbf{y}_i)-\boldsymbol{\mu}_i)\right] = 0,\\
I_{13} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\rho}'}\right] = 0,\\
I_{22} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[2\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}'}+\sum_{k=1}^{j-1}b_{ijk}b_{ijk}'\right],\\
I_{23} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\rho}'}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}\Upsilon_{ij}'+\frac{1}{2}\sum_{k=1}^{j-1}b_{ijk}\sum_{l=k}^{j}a_{ijl}T_{ilk}\Upsilon_{il}'\right],\\
I_{33} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\rho}\partial\boldsymbol{\rho}'}\right] = \frac{1}{4}\sum_{i=1}^{n}\boldsymbol{\Upsilon}_i'\left[I_{m_i}+R_i^{-1}\circ R_i\right]\boldsymbol{\Upsilon}_i,
\end{aligned}
$$

where '$\circ$' represents the Hadamard product and all other necessary symbols are defined in page 27. Note that for two matrices A and B of the same dimension $m\times n$, the Hadamard product $A\circ B$ is a matrix of the same dimension as the operands, with elements given by $[A\circ B]_{ij}=[A]_{ij}[B]_{ij}$ for all $1\le i\le m,\ 1\le j\le n$.

A.2 Block Components of Fisher Information Matrix of the Estimating Equations of Model 2

$$
\begin{aligned}
I_{11} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \sum_{i=1}^{n}\boldsymbol{\Pi}_i'\Delta_i\Sigma_i^{-1}\Delta_i\boldsymbol{\Pi}_i,\\
I_{12} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\gamma}'}\right] = -\sum_{i=1}^{n}\left[\boldsymbol{\Pi}_i'\Delta_i\frac{\partial\Sigma_i^{-1}}{\partial\boldsymbol{\gamma}'}(E(\mathbf{y}_i)-\boldsymbol{\mu}_i)\right] = 0,\\
I_{13} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\lambda}'}\right] = 0,\\
I_{22} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[2\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}'}+\sum_{k=1}^{j-1}b_{ijk}b_{ijk}'\right],\\
I_{23} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\lambda}'}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial\log T_{ijj}}{\partial\boldsymbol{\gamma}}Z_{ij}'+\frac{1}{2}\sum_{k=1}^{j-1}b_{ijk}\sum_{l=k}^{j}a_{ijl}T_{ilk}Z_{il}'\right],\\
I_{33} &= -E\left[\frac{\partial^2 l}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right] = \frac{1}{4}\sum_{i=1}^{n}\mathbf{Z}_i'\left[I_{m_i}+R_i^{-1}\circ R_i\right]\mathbf{Z}_i,
\end{aligned}
$$

where '$\circ$' represents the Hadamard product.

Appendix B

B.1 Expected Values of the Mixed Partial Derivatives in Negative Binomial Regression Model

First and second order partial derivatives of the log-likelihood function of negative binomial regression model with respect to the parameters $\boldsymbol{\beta}$ and c are

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - m_i}{m_i(1+cm_i)} \frac{\partial m_i}{\partial \beta_j}, \\
\frac{\partial l}{\partial c} &= \sum_{i=1}^{n} \left[\frac{log(1+cm_i)}{c^2} - \frac{m_i(y_i + c^{-1})}{1+cm_i} + \sum_{l=1}^{y_i} \frac{l-1}{1+c(l-1)} \right], \\
\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^{n} \left[\frac{y_i - m_i}{m_i(1+cm_i)} \frac{\partial^2 m_i}{\partial \beta_j \partial \beta_k} - \frac{y_i + 2cm_i y_i - cm_i^2}{m_i^2(1+cm_i)^2} \frac{\partial m_i}{\partial \beta_j} \frac{\partial m_i}{\partial \beta_k} \right], \\
\frac{\partial^2 l}{\partial \beta_j \partial c} &= -\sum_{i=1}^{n} \frac{(y_i - m_i)m_i}{(1+cm_i)^2} \frac{\partial m_i}{\partial \beta_j},
\end{aligned}$$

and

$$\frac{\partial^2 l}{\partial c^2} = -\sum_{i=1}^{n} \left[\sum_{l=1}^{y_i} \left(\frac{l-1}{1+c(l-1)} \right)^2 + 2c^{-3} \log(1+cm_i) - \frac{2c^{-2}m_i}{(1+cm_i)} - \frac{(y_i + c^{-1})m_i^2}{(1+cm_i)^2} \right].$$

Assume $m_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)$, then $\dfrac{\partial m_i}{\partial \beta_j} = m_i x_{ij}$, then

$$\begin{aligned} E\left[-\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right] &= \sum_{i=1}^{n}\frac{m_i}{1+cm_i}x_{ij}x_{ik} \quad \text{and} \\ E\left[-\frac{\partial^2 l}{\partial\beta_j\partial c}\right] &= 0. \end{aligned}$$

B.2 Expected Values of the Mixed Partial Derivatives in Beta Binomial Regression Model

First and second order partial derivatives of log-likelihood of beta binomial regression model with respect to the parameters $\boldsymbol{\beta}$ and θ are

$$\begin{aligned} \frac{\partial l}{\partial\beta_j} &= \sum_{i=1}^{n}\left[\sum_{r=0}^{y_i-1}\frac{1}{\mu_i+r\theta} - \sum_{r=0}^{k_i-y_i-1}\frac{1}{1-\mu_i+r\theta}\right]\frac{\partial\mu_i}{\partial\beta_j}, \\ \frac{\partial l}{\partial\theta} &= \sum_{i=1}^{n}\left[\sum_{r=0}^{y_i-1}\frac{r}{\mu_i+r\theta} + \sum_{r=0}^{k_i-y_i-1}\frac{r}{1-\mu_i+r\theta} - \sum_{r=0}^{k_i-1}\frac{r}{1+r\theta}\right], \\ \frac{\partial^2 l}{\partial\beta_j\partial\beta_k} &= -\sum_{i=1}^{n}\left[\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i+r\theta)^2} + \sum_{r=0}^{k_i-y_i-1}\frac{1}{(1-\mu_i+r\theta)^2}\right]\frac{\partial\mu_i}{\partial\beta_j}\frac{\partial\mu_i}{\partial\beta_k} + \\ &\qquad \sum_{i=1}^{n}\left[\sum_{r=0}^{y_i-1}\frac{1}{\mu_i+r\theta} - \sum_{r=0}^{k_i-y_i-1}\frac{1}{1-\mu_i+r\theta}\right]\frac{\partial^2\mu_i}{\partial\beta_j\partial\beta_k}, \\ \frac{\partial^2 l}{\partial\beta_j\partial\theta} &= \sum_{i=1}^{n}\left[-\sum_{r=0}^{y_i-1}\frac{r}{(\mu_i+r\theta)^2} + \sum_{r=0}^{k_i-y_i-1}\frac{r}{(1-\mu_i+r\theta)^2}\right]\frac{\partial\mu_i}{\partial\beta_j}, \text{ and} \\ \frac{\partial^2 l}{\partial\theta^2} &= \sum_{i=1}^{n}\left[-\sum_{r=0}^{y_i-1}\frac{r^2}{(\mu_i+r\theta)^2} - \sum_{r=0}^{k_i-y_i-1}\frac{r^2}{(1-\mu_i+r\theta)^2} + \sum_{r=0}^{k_i-1}\frac{r^2}{(1+r\theta)^2}\right]. \end{aligned}$$

In order to obtain the expected values of the mixed partial derivatives in the beta binomial regression model, we need to evaluate the following terms

$$E\left[\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i+r\theta)^2}\right], \quad E\left[\sum_{r=0}^{k_i-y_i-1}\frac{1}{(1-\mu_i+r\theta)^2}\right],$$
$$E\left[\sum_{r=0}^{y_i-1}\frac{r}{(\mu_i+r\theta)^2}\right], \quad \text{and } E\left[\sum_{r=0}^{k_i-y_i-1}\frac{r}{(1-\mu_i+r\theta)^2}\right].$$

Now,

$$\begin{aligned}
&E\left[\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i+r\theta)^2}\right]=\sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i+r\theta)^2}\right]Pr\left(y_i\right)\\
&=0+\\
&\frac{Pr\left(y_i=1\right)}{(\mu_i+0\cdot\theta)^2}+\\
&\frac{Pr\left(y_i=2\right)}{(\mu_i+0\cdot\theta)^2}+\frac{Pr\left(y_i=2\right)}{(\mu_i+1\theta)^2}+\\
&\frac{Pr\left(y_i=3\right)}{(\mu_i+0\cdot\theta)^2}+\frac{Pr\left(y_i=3\right)}{(\mu_i+1\theta)^2}+\frac{Pr\left(y_i=3\right)}{(\mu_i+2\theta)^2}+\\
&\vdots\\
&+\frac{Pr\left(y_i=k_i\right)}{(\mu_i+0\cdot\theta)^2}+\frac{Pr\left(y_i=k_i\right)}{(\mu_i+1\theta)^2}+\cdots+\frac{Pr\left(y_i=k_i\right)}{(\mu_i+(n_i-1)\theta)^2}\\
&=\frac{Pr\left(y_i\geq 1\right)}{(\mu_i+(1-1)\theta)^2}+\frac{Pr\left(y_i\geq 2\right)}{(\mu_i+(2-1)\theta)^2}+\cdots+\frac{Pr\left(y_i\geq k_i\right)}{(\mu_i+(k_i-1)\theta)^2}.\\
&=\sum_{r=1}^{k_i}\frac{Pr\left(y_i\geq r\right)}{\left[\mu_i+(r-1)\theta\right]^2}.
\end{aligned}$$

$$
\begin{aligned}
&E\left[\sum_{r=0}^{k_i-y_i-1}\frac{1}{(1-\mu_i+r\theta)^2}\right]=\sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{k_i-y_i-1}\frac{1}{(1-\mu_i+r\theta)^2}\right]Pr\left(y_i\right)\\
&=\frac{Pr\left(y_i=0\right)}{(1-\mu_i+0\cdot\theta)^2}+\frac{Pr\left(y_i=0\right)}{(1-\mu_i+1\theta)^2}+\cdots+\frac{Pr\left(y_i=0\right)}{(1-\mu_i+(k_i-1)\theta)^2}+\\
&\quad\frac{Pr\left(y_i=1\right)}{(1-\mu_i+0\cdot\theta)^2}+\frac{Pr\left(y_i=1\right)}{(1-\mu_i+1\theta)^2}+\cdots+\frac{Pr\left(y_i=1\right)}{(1-\mu_i+(k_i-2)\theta)^2}+\\
&\quad\frac{Pr\left(y_i=2\right)}{(1-\mu_i+0\cdot\theta)^2}+\frac{Pr\left(y_i=2\right)}{(1-\mu_i+1\theta)^2}+\cdots+\frac{Pr\left(y_i=2\right)}{(1-\mu_i+(k_i-3)\theta)^2}+\\
&\quad\vdots\\
&\quad+\frac{Pr\left(y_i=k_i-2\right)}{(1-\mu_i+0\cdot\theta)^2}+\frac{Pr\left(y_i=k_i-2\right)}{(1-\mu_i+1\theta)^2}+\\
&\quad+\frac{Pr\left(y_i=k_i-1\right)}{(1-\mu_i+0\cdot\theta)^2}.\\
&=\frac{Pr\left(y_i\leq k_i-1\right)}{(1-\mu_i+(1-1)\theta)^2}+\frac{Pr\left(y_i\leq k_i-2\right)}{(1-\mu_i+(2-1)\theta)^2}+\cdots+\frac{Pr\left(y_i\leq 0\right)}{(1-\mu_i+(k_i-1)\theta)^2}.\\
&=\sum_{r=1}^{k_i}\frac{Pr\left(y_i\leq k_i-r\right)}{\left[1-\mu_i+(r-1)\theta\right]^2}.
\end{aligned}
$$

$$
\begin{aligned}
E\left[\sum_{r=0}^{y_i-1}\frac{r}{(\mu_i+r\theta)^2}\right] &= \sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{y_i-1}\frac{r}{(\mu_i+r\theta)^2}\right]Pr\left(y_i\right)\\
&= \sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{y_i-1}\frac{r\theta}{\theta(\mu_i+r\theta)^2}\right]Pr\left(y_i\right)\\
&= \sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{y_i-1}\frac{\mu_i+r\theta-\mu_i}{\theta(\mu_i+r\theta)^2}\right]Pr\left(y_i\right)\\
&= \frac{1}{\theta}\sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{y_i-1}\frac{Pr\left(y_i\right)}{(\mu_i+r\theta)}\right]-\frac{\mu_i}{\theta}\sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{y_i-1}\frac{Pr\left(y_i\right)}{(\mu_i+r\theta)^2}\right]\\
&= -\frac{\mu_i}{\theta}\sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{y_i-1}\frac{Pr\left(y_i\right)}{(\mu_i+r\theta)^2}\right]\quad\text{since}\quad E\left[\frac{\partial l}{\partial\beta_j}\right]=0\\
&= -\frac{\mu_i}{\theta}\sum_{r=1}^{k_i}\frac{Pr\left(y_i\geq r\right)}{\left[\mu_i+(r-1)\theta\right]^2}.
\end{aligned}
$$

Similarly

$$E\left[\sum_{r=0}^{k_i-y_i-1}\frac{r}{(1-\mu_i+r\theta)^2}\right]=-\frac{1-\mu_i}{\theta}\sum_{r=1}^{k_i}\frac{Pr\left(y_i\leq k_i-r\right)}{\left[1-\mu_i+(r-1)\theta\right]^2}.$$

Suppose μ_i follows the logistic model $\mu_i(\boldsymbol{x}_i',\boldsymbol{\beta})=\dfrac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$. Then $\dfrac{\partial\mu_i}{\partial\beta_j}=\mu_i(1-\mu_i)x_{ij}$. Also assume

$$\begin{aligned}
&\sum_{r=1}^{k_i}\frac{Pr\left(y_i\geq r\right)}{\left[\mu_i+(r-1)\theta\right]^2}=p_{1i},\\
&\sum_{r=1}^{k_i}\frac{Pr\left(y_i\leq k_i-r\right)}{\left[1-\mu_i+(r-1)\theta\right]^2}=p_{2i},\quad\text{and}\\
&\sum_{r=0}^{k_i-1}\frac{1}{(1+r\theta)^2}=\sum_{r=1}^{k_i}\frac{1}{[1+(r-1)\theta]^2}=p_{3i}.
\end{aligned}$$

Thus

$$\begin{aligned}
E\left[-\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right] &= \sum_{i=1}^{n}E\left[\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i+r\theta)^2}+\sum_{r=0}^{k_i-y_i-1}\frac{1}{(1-\mu_i+r\theta)^2}\right]\frac{\partial\mu_i}{\partial\beta_j}\frac{\partial\mu_i}{\partial\beta_k}\\
&= \sum_{i=1}^{n}(p_{1i}+p_{2i})\mu_i^2(1-\mu_i)^2x_{ij}x_{ik}.
\end{aligned}$$

$$\begin{aligned}
E\left[-\frac{\partial^2 l}{\partial\beta_j\partial\theta}\right] &= \sum_{i=1}^{n}E\left[\sum_{r=0}^{y_i-1}\frac{r}{(\mu_i+r\theta)^2}-\sum_{r=0}^{k_i-y_i-1}\frac{r}{(1-\mu_i+r\theta)^2}\right]\frac{\partial\mu_i}{\partial\beta_j}\\
&= \sum_{i=1}^{n}\left[-\mu_ip_{1i}+(1-\mu_i)p_{2i}\right]\frac{\mu_i(1-\mu_i)}{\theta}x_{ij}.
\end{aligned}$$

In order to calculate $E\left[-\frac{\partial^2 l}{\partial\theta^2}\right]$, replace θ^{-1} by ψ. Then the log-likelihood becomes

$$l = \sum_{i=1}^{n}\left[\sum_{r=0}^{y_i-1}\log(\mu_i\psi + r) + \sum_{r=0}^{k_i-y_i-1}\log((1-\mu_i)\psi + r) - \sum_{r=0}^{k_i-1}\log(\psi + r)\right].$$

Differentiating l with respect to ψ twice, we have

$$\frac{\partial^2 l}{\partial\psi^2} = -\sum_{i=1}^{n}\left[\mu_i^2\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i\psi + r)^2} + (1-\mu_i)^2\sum_{r=0}^{k_i-y_i-1}\frac{1}{((1-\mu_i)\psi + r)^2} - \sum_{r=0}^{k_i-1}\frac{1}{(\psi + r)^2}\right].$$

Now

$$\begin{aligned}
E\left[-\frac{\partial^2 l}{\partial\psi^2}\right] &= \sum_{i=1}^{n} E\left[\mu_i^2\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i\psi + r)^2} + (1-\mu_i)^2\sum_{r=0}^{k_i-y_i-1}\frac{1}{((1-\mu_i)\psi + r)^2} - \sum_{r=0}^{k_i-1}\frac{1}{(\psi + r)^2}\right] \\
&= \sum_{i=1}^{n}\sum_{y_i=0}^{k_i}\left[\mu_i^2\sum_{r=0}^{y_i-1}\frac{1}{(\mu_i\psi + r)^2} + (1-\mu_i)^2\sum_{r=0}^{k_i-y_i-1}\frac{1}{((1-\mu_i)\psi + r)^2}\right] Pr\left(y_i\right) - \\
&\quad \sum_{i=1}^{n}\sum_{y_i=0}^{k_i}\left[\sum_{r=0}^{k_i-1}\frac{1}{(\psi + r)^2}\right] Pr\left(y_i\right) \\
&= \sum_{i=1}^{n}\left[\mu_i^2\sum_{r=1}^{k_i}\frac{Pr\left(y_i \geq r\right)}{\left[\mu_i\psi + (r-1)\right]^2} + (1-\mu_i)^2\sum_{r=1}^{k_i}\frac{Pr\left(y_i \leq k_i - r\right)}{((1-\mu_i)\psi + (r-1))^2} - \sum_{r=0}^{k_i-1}\frac{1}{(\psi + r)^2}\right].
\end{aligned}$$

Since $E\left[-\frac{\partial^2 l}{\partial\theta^2}\right] = \theta^{-4}\left[-\frac{\partial^2 l}{\partial\psi^2}\right]$, then

$$\begin{aligned}
E\left[-\frac{\partial^2 l}{\partial\theta^2}\right] &= \sum_{i=1}^{n}\left[\frac{\mu_i^2}{\theta^2}\sum_{r=1}^{k_i}\frac{Pr\left(y_i \geq r\right)}{\left[\mu_i + (r-1)\theta\right]^2} + \frac{(1-\mu_i)^2}{\theta^2}\sum_{r=1}^{k_i}\frac{Pr\left(y_i \leq k_i - r\right)}{(1-\mu_i + (r-1)\theta)^2}\right] - \\
&\quad \frac{1}{\theta^2}\sum_{i=1}^{n}\sum_{r=0}^{k_i-1}\frac{1}{(1+r\theta)^2} \\
&= \frac{1}{\theta^2}\sum_{i=1}^{n}\left[\mu_i^2 p_{1i} + (1-\mu_i)^2 p_{2i} - p_{3i}\right].
\end{aligned}$$

B.3 Expected Values of the Mixed Partial Derivatives in Zero-inflated Generalized Linear Models

Suppose the log-likelihood l of zero-inflated generalized linear model is written in the form $l = \sum_{i=1}^{n} l_i$. Then the first and second order partial derivatives of l_i with respect to θ_i and γ are as follows

$$\begin{aligned}
\frac{\partial l_i}{\partial \theta_i} &= I_{\{y_i=0\}} \frac{f_0(-g_i')}{\gamma + f_0} + I_{\{y_i>0\}}(a_i' y_i - g_i') \\
\frac{\partial l_i}{\partial \gamma} &= -\frac{1}{1+\gamma} + I_{\{y_i=0\}} \frac{1}{\gamma + f_0} \\
\frac{\partial^2 l_i}{\partial \theta_i^2} &= I_{\{y_i=0\}} \left(\frac{f_0(-g_i')^2}{\gamma + f_0} - \frac{f_0^2(-g_i')^2}{(\gamma+f_0)^2} + \frac{f_0(-g_i'')}{\gamma+f_0} \right) + I_{\{y_i>0\}}(a_i'' y_i - g_i'') \\
\frac{\partial^2 l_i}{\partial \theta_i \partial \gamma} &= I_{\{y_i=0\}} \frac{-f_0(-g_i')}{(\gamma+f_0)^2} \\
\frac{\partial^2 l_i}{\partial \gamma^2} &= \frac{1}{(1+\gamma)^2} - I_{\{y_i=0\}} \frac{1}{(\gamma+f_0)^2}
\end{aligned}$$

We have

$$\begin{aligned}
E\left[I_{\{y_i=0\}}\right] &= \Pr(y_i = 0) = \frac{\gamma + f_0}{1+\gamma} \quad \text{and} \\
E\left[I_{\{y_i>0\}}\right] &= \Pr(y_i > 0) = 1 - \Pr(y_i = 0) = \frac{1-f_0}{1+\gamma}.
\end{aligned}$$

Hence expected values of mixed partial derivatives are as follows

$$\begin{aligned}
E\left[-\frac{\partial^2 l_i}{\partial \theta_i^2}\right] &= -\frac{\gamma+f_0}{1+\gamma}\left(\frac{f_0(-g_i')^2}{\gamma+f_0} - \frac{f_0^2(-g_i')^2}{(\gamma+f_0)^2} + \frac{f_0(-g_i'')}{\gamma+f_0} \right) - \frac{1-f_0}{1+\gamma}(a_i'' E[y_i] - g_i'') \\
&= -\frac{\gamma f_0 (g_i')^2}{(1+\gamma)(\gamma+f_0)} + \frac{g_i''}{1+\gamma} - \frac{1-f_0}{1+\gamma} a_i'' E(y_i)
\end{aligned}$$

$$
\begin{aligned}
E\left[-\frac{\partial^2 l_i}{\partial\theta_i\partial\gamma}\right] &= -\frac{\gamma+f_0}{1+\gamma}\frac{-f_0(-g_i')}{(\gamma+f_0)^2} = -\frac{f_0 g_i'}{(1+\gamma)(\gamma+f_0)} \\
E\left[-\frac{\partial^2 l_i}{\partial\gamma^2}\right] &= -\frac{1}{(1+\gamma)^2} + \frac{\gamma+f_0}{1+\gamma}\frac{1}{(\gamma+f_0)^2} = \frac{1-f_0}{(1+\gamma)^2(\gamma+f_0)}
\end{aligned}
$$

Bibliography

Agresti, A. (2007). An Introduction to Categorical Data Analysis. *New Jersey: Wiley.*

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Bartlett, M. S. (1953). Approximate confidence intervals. *Biometrika*, **40**, 12-19.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961-994.

Cameron, A. C. and Trivedi, P. K. (1998). Regression Analysis of Count Data. *Cambridge University Press, Cambridge.*

Cameron, A. C. and Trivedi, P. K. (2005). Microeconometrics: Methods and Applications. *Cambridge University Press, Cambridge.*

Carroll, R. H., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541-554.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2012). Mea-

surement Error in Nonlinear Models: A Modern Perspective. *CRC Press*.

Cramer, H. (1946). Mathematical Methods of Statistics. *Princeton University Press*.

Daniels, M. J. and Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis*, **100**, 2352-2363.

Darper N. and Smith, H. (1998). Applied Regression Analysis. *Wiley Series in Probability and Statistics*.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, **12-3**, 313-336.

Deng, D. and Paul, S. R. (2000). Score tests for zero-inflation in generalized linear models. *The Canadian Journal of Statistics*, **27**, 563-570.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). Analysis of Longitudinal Data. *2nd edition, Oxford University Press*.

Efroymson, M. A. (1960). Multiple Regression Analysis. *Mathematical Methods for Digital Computers. Eds. A. Ralston and H. S. Wilf, New York: John Wiley and Sons*.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Journal of the American Statistical Association*, **11**, 89-121.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710-723.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semi-

parametric estimation of covariance function. *Journal of the American Statistical Association*, **102**, 632-640.

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *Journal of the American Statistical Association*, **103**, 1520-1533.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, **37**, 152-155.

Garside, M. J. (1965). The best sub-set in multiple regression analysis. *Applied Statistics*, **14**, 196-200.

Griggs, W. (2013). Penalized spline regression and its applications. *Whitman College*, 1-51.

Gurmu, S. and Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics*, **14**, 469-477.

Harrell, F. E. (2013). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. *New York: Springer-Verlag New York*.

He, X. M., Zhu, Z. Y., and Fung, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, **89**, 579-590.

He, X. M., Fung, W. K., and Zhu, Z. Y. (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*, **100**, 1176-1184.

Hocking, R. R. and Leslie, R. B. (1967). Selection of the best subset in regression

analysis. *Technometrics*, **9**, 531-554.

Hocking, R. R. (1976). A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1-49.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111-128.

Kaslow, R. A, Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. Jr. (1987). The Multicenter AIDS Cohort Study: Rationale,organization and selected characteristics of the participants. *American Journal of Epidemiology*, **126**, 310-318.

Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR Jr.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**, 296-308.

Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2013). Applied Linear Statistical Models. 5th Edition, *McGraw-Hill*.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15-3**, 209-225.

Leng, C., Zhang, W., and Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *Journal of the American Statistical Association*, **105**, 181-193.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society B*, **68**, 69-88.

Little, R. J. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. *Wiley Series in Probability and Statistics*.

Lynn, P. (2009). Methodology of Longitudinal Surveys. *Wiley*.

Margolin, B. H., Kaplan, N., and Zeiger, E. (1981). Statistical analysis of the Ames salmonella/microsome test. *Proc. Nat. Acad. Sci. U.S.A.*, **76**, 3779-3783.

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

Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data. *Springer-Verlag*.

Moran, P. A. P. (1970). On asymptotically optimal tests of composite hypotheses. *Biometrika*, **57**, 45-75.

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, **33**, 341-365.

Nelder, J. A and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society A*, **135**, 370-384.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika*, **2OA**, 175-240, 264-294.

Neyman, J. (1959). Optimal asymptotic tests for composite hypotheses. *In Probability and Statistics*, Ed. U. Grenander, pp. 213-234. New York: Wiley.

Pan, J. and Mackenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239-244.

Paul, S. R. (1982). Analysis of proportions of affected foetuses in teratological experiments. *Biometrics*, **38**, 361-370.

Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, **46**, 863-867.

Pinheiro, J. D. and Bates, D. M. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, **6**, 289-296.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 677-690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425-435.

Pourahmadi, M. and Wang, X. (2015). Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor. *Statistics and Probability Letters*, **106**, 512.

Pregibon, D. (1982). Score tests in GLIM with applications. *Lecture Notes in Statistics*, **14**, 87-97.

Qin, G., Mao, J., and Zhu, Z. (2015). Joint mean-covariance model in generalized partially linear varying coefficient models for longitudinal data. *Journal of Statistical Computation and Simulation*, 1-17.

Qin, G., Zhang, J., and Zhu, Z. (2016). Simultaneous mean and covariance estimation of partially linear models for longitudinal data with missing responses and covariate measurement error. *Computational Statistics and Data Analysis*, **96**, 24-39.

Qu, A., Lindsay, B. G. and Li, B. (2000). Improving estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-836.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **44**, 50-57.

Rao, C. R. (2005). Score test: Historical review and recent developments, advances in ranking and selection, multiple comparisons, and reliability-methodology and applications, *In: Statistics for Industry and Technology, Balakrishnan N., Kannan N., Nagaraja H.N. (eds.)*, 3-20.

Rapisarda, F., Brigo, D., and Mercurio, F. (2007). Parameterizing correlations: a geometric interpretation. *IMA J. Manag. Math.*, **18**, 55-73.

Rebonato, R. and Jackel, P. (2000). The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. *Journal of Risk*, **2**, 17-27.

Schumaker, L. L. (1981). Spline Functions. *New York: Wiley.* **182, 190**.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Tang, C. Y., Zhang, W., and Leng, C. (2019). Discrete longitudinal data modeling with a mean-correlation regression approach. *Statistica Sinica*, **29**, 853-876.

Taylor, J. M. G. and Law, N. (1998). Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts. *Statistics in Medicine*, **17**, 2381-2394.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 1426-1482.

Wang, C., Huang, Y., Chao, E. C., and Jeffcoat, M. K. (2008). Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics*, **64**, 85-95.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31-4**, 949-952.

Wolberg, G. and Alfy, I. (2002). An energy-minimization framework for monotonic cubic spline interpolation. *Journal of Computational and Applied Mathematics*, **143(2)**, 145-188.

Ye, H. and Pan, J. (2006). Modelling covariance structures in generalized estimating equations for longitudinal data. *Biometrika*, **93**, 927-941.

Yi, G. Y., Ma, Y., and Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, **99**, 151-165.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689-699.

Zhang, W., Leng, C., and Tang, Y. C. (2015). A joint modelling approach for longi-

tudinal studies. *Journal of the Royal Statistical Society B*, **77**, 219-238.

Vita Auctoris

The author was born in Noakhali, Bangladesh in 1979 and grew up in Cumilla, Bangladesh. He obtained his higher secondary degree from Cumilla Victoria Government College, Bangladesh in 1996. From there he went to the Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh where he obtained his B.Sc. (Honours) and M.Sc. in Mathematics in 2003 and 2005 respectively. In 2005 he first joined as a Lecturer in the Department of Computer Science and Engineering, Leading University, Sylhet. After that he moved to the Department of Mathematics, SUST, Sylhet in the following year. He also earned an M.Sc. degree in Mathematics from the Western University, London, Ontario, Canada in 2010. He is currently a candidate for the Doctor of Philosophy in Statistics at the University of Windsor and hopes to graduate in May 2019.