KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

PENALIZED VARIABLE SELECTION AND MODELING: APPLICATION OF H-LIKELIHOOD, JOINT-GLM AND HGLM METHODS TO MODELING CROP YIELD IN THE THREE NORTHERN REGIONS OF GHANA

By

SARPONG, SMART ASOMANING

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN MATHEMATICAL STATISTICS

September 24, 2015

DECLARATION

I hereby declare that this submission is my own work towards the award of the PhD. degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

SARPONG, SMART ASOMANING	.....	.....
Student	Signature	Date

Certified by:

Prof. N.N.N. NSOWAH-NUAMAH	.....	.....
Supervisor (Main)	Signature	Date

Certified by:

Dr. RICHARD K. AVUGLAH	.....	.....
Co-Supervisor	Signature	Date

Certified by:

Prof. S. K. AMPONSAH	.....	.....
Head of Department	Signature	Date

Dedication

I dedicate this dissertation to my God Almighty; my help in ages past and my hope for years to come. Then to my life partner and dear Wife, Family and many other persons who have been and continue to be a blessing to my life. God Bless you all richly... AMEN

Acknowledgement

"For I am the Lord, your God, who takes hold of your right hand and says to you, Do not fear; I will help you (Isaiah 41:13)".

Truly God Almighty holds my hand throughout all these years and helps me. I would therefore like to express my sincere gratitude to Him for seeing me through this PhD programme.

My special thanks also go to my dependable team of supervisors under the joint leadership of Prof N. N. N. Nsowah-Nuamah, Prof Youngjo Lee, Prof S. K. Amponsah and Dr. Richard K. Avuglah. It all started with Prof Louis Munyakazy (Late). After His demise, Prof N. N. N. Nsowah-Nuamah, Prof S. K. Amponsah and Dr. Richard K. Avuglah stepped in and helped me carry on with the then so long journey. May the Blessings of our Lord and Saviour Jesus Christ be your portion for all the help given me throughout this programme.

Then came my visit to the HGLM Laboratory at the Data Science Research Centre of the Seoul National University in South Korea. I am and will forever be grateful to Prof Youngjo Lee, Mr. Sinyuong Ho, Dr. Suhuang Park and all the lovely people of the Statistics Department of the Seoul National University for offering directions, suggestions, and encouragement during the entire duration of my visit and even till today. Miss you all so much and God richly Bless you all.

I am also grateful to all friends and family who in one way or the other contributed towards the successful completion of this PhD. May God Almighty bless you all....AMEN

Abstract

This study proceeded on two paths; to select significant crop yield physical support variables among many potential ones to be included in a model via penalized methods (LASSO, SCAD, H-Likelihood) and to also propose and demonstrate the excellent performance of higher levels and very recent extensions of the Generalized Linear Models (GLM); Joint Generalized Linear Models (JGLM) and Hierarchical Generalized Linear Models (HGLM) in the global quest to developing Statistical Models with highest model accuracy. The analyses is be based on raw data available at the regional Monitoring and Evaluation office of the Linking Farmers to Markets (FtM) project in Tamale - Ghana. Physical support (Fixed effect) variables measured include; crop type, Financial Credit, Training, Study tour, Demonstrative Practicals, Networking Events, Post harvest Equipment, Number of farmers in the FBO and Plot size cultivated. Dependent variable measured is Total Crop Yield whereas the regions and the particular communities were treated as Random variables. After the highly rigorous processes of data analysis the study concluded that, the H-Likelihood method of penalized variable selection performs both selection of significant variables and estimation of their coefficients simultaneously with the least penalize cross-validated errors compared to the SCAD and the LASSO. In modelling the effects of fixed physical support services given to farmer based organizations on crop yield, the GLM with assumed fixed dispersion will not be recommended by this study. The study concludes that the proposed modelling of both mean and dispersion (Joint-GLM) improves the quality of the models significantly. In the case of both fixed and random effects, the, HGLM 2 is highly recommended. This study concludes that the HGLM 2 performs far better, gives a more fitting model and improves the quality of the crop yield models significantly. The study recommends that deliberate effort be put into strengthening the Agricultural support systems as a form of strategy for increasing crop production in Northern Ghana.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of the Study

The rate of food production in many parts of sub-Saharan Africa has not kept pace with the rate of population growth. Whereas the estimates of population growth rate increase at about 3 per cent annually, that of food production increases by only 2 per cent (Rosegrant et al., 2001). The sub-region's per capita deficit in grains and cereals according to Rosegrant et al., (2001) is one of the highest in the world. Way back in 1967, the sub-region's cereal imports was 1.5 million tons. However, just within thirty years down the way, this figure increased to 12 million tons in 1997, and projections have it that the sub-region will require about 27 million tons of cereal imports to satisfy demand by 2020 (Rosegrant et al., 2001). In the long run, importation may not be economically feasible to ameliorate food shortages. Thus, there is a need to increase domestic production to guarantee food security.

Attempts to increase food production in sub-Saharan Africa has been accomplished mostly by expanding the area of land under cultivation. However, because of increasing pressure on farm lands for other domestic and industrial purposes, the scope of agricultural intensification has drastically reduced. Yield increase, rather than farm area expansion is becoming more and more important for increasing food production. This in turn implies that as efficient soil management practices are strictly adhered to, such studies into crop yield physical variables becomes highly imperative to guarantee food security.

In recent years, as a consequence of the general apprehension of international donors and national governments about investments in agricultural research, policy makers and researchers have devoted increasing attention to research efficiency in order to rejuvenate donors' and governments' support and to convince them of the importance of agricultural research. One of the outcomes of this renewed attention to agricultural research efficiency issues has been a series of research activities undertaken by economists to measure the impacts of agricultural research in Africa. A useful product of these research activities is the growing number of studies documenting impacts and rates of return to major food crops' research in Africa (Oehmke and Crawford, 1996; Sanders, 1996).

Strengthening of national agricultural support system has been advocated as a strategy for increasing agricultural production in sub-Saharan Africa by governments in the region and by international development agencies (see, eg., World Bank, 1990; Bindlish and Evenson, 1997). The training and visit system of agricultural extension has been central to this strategy. The world bank supported agricultural extension programmes based on training and visits have been implemented in some 30 sub-Saharan African countries including Ghana. Substantial amount of money and resources have been committed to this system, both by national governments and international development agencies (Bindlish and Evenson, 1997). There is however an emerging controversy as to cost-effectiveness and productivity of a national system of agricultural extension, particularly in sub-Saharan Africa where government's ability to the large recurrent cost that the system entails is limited (see Purcell and Anderson, 1997; Gautam, 1998).

Ghana is still an agriculture-based economy. Agriculture has been the backbone of Ghana's economy in the entire post-independence history (McKay and Aryeetey, 2004). While policy and political failure had caused per capita GDP growth declining until 1980s, the agricultural sector had been less affected than the non-

agricultural sector because it was less intervened by the government than the non-agricultural sector and its growth is primarily led by smallholders for subsistence purpose of production. Agricultural growth in Ghana has been more rapid than growth in the non-agricultural sectors in recent years, expanding by an average annual rate of 5.5 percent, compared to 5.2 percent for the economy as a whole (Bogetic et al., 2007).

Agriculture was about 40 per cent of Ghana's GDP in the late 1990's and was still above 35 per cent until 2007. Only in the recent two years of 2007 and 2008, share of agriculture fell down to 34 per cent and 32 per cent, respectively. Recent decline in the agricultural GDP share in Ghana is the result of faster growth in the services sector, which increased its share in GDP to more than 40 per cent in 2007 and 2008 (World Bank Global Forum on Agriculture, 2010). Thus, it was first time in Ghana's history that agriculture is no more the largest sector in the economy and the service sector has taken this position.

Agricultural growth is at the centre of the Comprehensive African Agriculture Development Programme (CAADP, 2009) agenda because increasing agricultural productivity is necessary to achieve poverty reduction and food output targets, while at the same time reduce production costs and food prices for the poor. Ghana's Medium Term Agriculture Sector Investment Plan (METASIP, 2010) seeks to modernize agriculture which will culminate in a structurally transformed economy evident in food security, employment opportunities and poverty reduction. To this end, as per (CAADP, 2009) directives, the country is to allocate 10 percent of government expenditure to achieve an agricultural gross domestic product (GDP) growth of at least 6 percent annually to achieve the millennium development goal 1 (MDG1) of halving poverty and hunger before the target year of 2015.

Agriculture in Ghana accounts for more than 30 percent of GDP (MoFA, 2011) and three-quarters of export earnings. Yields of most crops in Ghana however are generally low (20-60 percent below their achievable level). For example, the yield of cassava is at 12.4 Mt/ha against a potential yield of 28.0 Mt/ha (MoFA, 2011). The yield of 1.7 Mt/ha for maize is less than a third of the achievable yield of 6.0 Mt/ha.

The agriculture sector makes up over 50 percent of Ghana's total employment and approximately 25 percent of the nation's Gross Domestic Product (GDP). The cocoa industry, in particular, is extremely important for Ghana, contributing around 30 percent of export revenue (MOFA, 2011). The service sector is the fastest growing sector of the economy. As of 2011, this sector accounted for nearly 50 percent of Ghana's GDP, and employed approximately 30 percent of the Ghanaian work force.

Although employment in the industrial sector is less than 20 percent of Ghana's total employment, this sector makes up approximately 25 percent of Ghana's GDP (World Data Bank, 2013). Furthermore, the industrial sector provides the greatest contributions to the country's foreign exchange earnings through exports of oil, gold, bauxite, aluminium, manganese ore, diamonds, natural gas and electricity (World Data Bank, 2013). Such growth patterns in the non-agricultural sector are not consistent with the transformation theory as well as experience of other developing countries in which the role of industry, especially of manufacturing has increased in the development process (Breisinger and Diao, 2008).

Agricultural structure and the regional distribution of agricultural GDP significantly differ across Ghana's agro-ecological zones. These regional differences have important implications for sub-sector-level agricultural growth strategies. The Forest Zone remains the major agricultural producer, accounting for 43 per

cent of agricultural GDP, compared to about 10 per cent in the Coastal Zone, and 26.5 per cent and 20.5 per cent in the Southern and Northern Savannah Zones, respectively (Breisinger et al., 2008). The Northern Savannah zone is the main producer of cereals and livestock. More than 70 per cent of the country's sorghum, maize, millet, cowpeas, groundnuts, beef and soybeans come from the Northern Zone, while the Forest Zone supplies a large share of higher-value products, such as cocoa and livestock (mainly commercial poultry).

The heterogeneous agricultural production structure also indicates differences in the agricultural income structure across regions. The Forest Zone generates about half its agricultural income from two of Ghana's major export goods (cocoa and forestry). Including non-traditional exports and fishery, export agriculture also plays an important role in total agricultural income for the Coast and Southern Savannah Zones. In contrast, 90 per cent of agricultural income in the Northern Zone comes from staple crops and livestock (World Bank Global Forum on Agriculture, 2010).

The analysis presented in this dissertation suggests that a system of support services; Access to credit facility, Training, Study tour, Demonstrative practical, Networking event and Post harvest Equipment, plays an important role in determining crop yield even though their individual and interaction effects on yield is not uniform across farmer based organizations. This research is focused mainly on the production of Maize and Soy beans in the northern parts of Ghana where there is substantial farming activity. Maize and Soy beans are the very much cultivated in these parts of the country due to their vegetation which supports the growth of grains and cereals. Beyond the numbers and descriptive statistics on yield of such crops, this study tries to bring out variables that significantly contribute to yield. We seek to select among access to credit, training, study tour, demonstrative practicals, networking events, post-harvest equipments, size

of plot cultivated and number of farmers; covariates or variables that significantly influence crop yield in the Northern regions of Ghana.

Variable selection is intended to select the "best" subset of predictors. If the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in and also degrees of freedom will be wasted. A major challenge in regression analysis is to decide which predictors among many potential ones are to be included in the model. It is customary to use stepwise selection and subset selection. But these procedures are unstable and ignore the stochastic errors introduced by the selection process.

Several methods, including bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (EN) (Zou and Hastie, 2005), and adaptive lasso (A-LASSO) (Zou, 2006), have been proposed to select variables and estimate their regression coefficients simultaneously. All these methods have common advantages over subset selection procedures; they are computationally simpler, the derived sparse estimators are stable, and they facilitate higher prediction accuracies.

These methods can be cast in the framework of penalized least squares and likelihood. The main advantage of those methods is that they select important variables and estimate the regression coefficients of the covariates, simultaneously. This thesis propose the use of random effect models (H-likelihood) to generate penalty functions for variable selection. It show how the h-likelihood methods overcome such difficulties to allow an oracle variable selection and simultaneously enhance estimation power.

Under the modelling of crop yield, the study propose the JGLM and HGLM approach of Lee and Nelder (2006). With hierarchical generalized linear models (HGLMs) of Lee and Nelder (2006), various scale mixtures can be considered as distributions for β. To be specific, this study suggest the use of a gamma mixture for β and apply it to a very critical area in Ghana's quest for economic growth, food security and expansion in the agricultural sector.

1.2 Problem Statement

Contribution of Agriculture to GDP keeps going down marginally and this give a course to worry. The sector, in 2010, contributed 29.9 per cent to GDP; declined to 25.6 per cent in 2010 and further dipped to 22.7 per cent in 2012 (World Bank Global Forum on Agriculture, 2010). Until date, there has been no concrete study to inform empirically on the influence of some support services continually been provided by the Ministry of Agriculture and other agencies to help boost crop yield and ensure food security. This study seeks to redirect stakeholder's attention to the part played by some of the support variables intended to increase crop yield. Preview to this would help resources to be channelled to the relevant variables which significantly contributes to yield.

Variable selection techniques have been developed to enhance prediction, but their use in decision making has not been well tested. Mostly in literature, these techniques often miss or downplay the importance of certain interaction variables that are key to making decisions. The variable selection techniques being proposed by this study, focuses on finding these important interactions. There are multiple reasons why variable selection might be necessary in a decision making application. One reason is that finding the optimal policy becomes more difficult as the number of spurious variables included in the model increases. Thus, careful variable selection could lead to better policies. Also, due to limited resources, it may only be possible to collect a small number of variables when enacting a

policy in a real world setting.

Researchers are often unsure as to which variables would be most important to collect. Variable selection techniques could help identify these variables. In addition, policies with fewer variables are often easier to understand, so variable selection can improve interpretability. Currently, variable selection for decision making in many fields is predominantly guided by expert opinion. Expert opinion can be a good starting place when there is sufficient domain knowledge and expertise.

In Agricultural and especially crop yield analysis, a combination of predictive variable selection techniques and statistical testing of a small number of interaction variables suggested by expert opinion are most commonly used. Little research has been carried out to evaluate these techniques in decision making, or to suggest how they might be improved. Variable selection is particularly important in the interpretation of Statistical models, especially when the true underlying model has a sparse representation. Identifying null predictors enhances the prediction performances of the fitted model. However, traditional variable selection procedures have two fundamental limitations. First, when the number of predictors p is large, it is computationally infeasible to perform subset selection. Second, subset selection is extremely unreliable because of its inherent discreteness (Breiman, 1996; Fan and Li, 2001).

To overcome these difficulties, several other penalties have been proposed. The L_2-penalty yields ridge regression estimation, but it does not perform variable selection. With the L_1-penalty, specifically, the penalized least squares (PLS) estimator becomes the least absolute shrinkage and selection operator (LASSO), which thresholds predictors with small estimated coefficients (Tibshirani, 1996). LASSO is a popular technique for simultaneous estimation and variable selection,

ensuring high prediction accuracy, and enabling the discovery of relevant predictive variables. Donoho and Johnstone (1994) selected significant wavelet bases by thresholding based on an L_1-penalty. Prediction accuracy can sometimes be improved by shrinking (Efron and Morris, 1975) or setting some coefficients to zero by thresholding (Donoho and Johnston, 1994).

Tibshirani (1996) gave a comprehensive overview of LASSO as a PLS estimation. LASSO has been criticized on the grounds that a single parameter λ is used for both variable selection and shrinkage. It typically ends up selecting a model with too many variables to prevent over shrinkage of the regression coefficients (Radchenko and James, 2008); otherwise, regression coefficients of selected variables are often over shrunken. To overcome this problem, various other penalties have been proposed. Fan and Li (2001) proposed a family of new variable selection methods based on a non-concave penalized likelihood approach called the smoothly clipped absolute deviation (SCAD) penalty for oracle variable selection.

These methods are different from traditional procedures of variable selection in that they delete insignificant variables by estimating their coefficients as 0. As a result, their approaches simultaneously select significant variables and estimate regression coefficients. Recent related studies include (Fan and Li, 2006, Leng et.al, 2006, Potscher and Leeb, 2009, Zou and Li, 2008). More recently, Zou (2006) showed that LASSO does not satisfy Fan and Li's (2001) oracle property, and proposed the adaptive LASSO. This study demonstrates how the h-likelihood method overcome such difficulties to allow an oracle variable selection and simultaneously enhance estimation power.

Modelling of crop yield is another important and integral aspect of this study. By theory, the Generalized Linear Models (GLMs) can be derived from classical normal models by two extensions, one to the random part and another to the

systematic part. Random elements may now come from a one-parameter exponential family, of which the normal distribution is a special case. Distributions in this class include Poisson, binomial, gamma and inverse Gaussian as well as normal. But in practice, even though the GLM is widely noted for its good performance in modelling, some natural discrepancies arise between the data and the fitted values produced. Outliers are observations which have large discrepancies on the y-axis. Discrepancies between the data and the fitted values produced by the model fall into two main classes, isolated or systematic.

Isolated discrepancies appear when a few observations only have large residuals. Such residuals can occur if the observations are simply wrong, for instance where 129 has been recorded as 192. Such errors are understandable if data are hand recorded, but even automatically recorded data are not immune. Robust methods were introduced partly to cope with the possibility of such errors; for a description of robust regression in a likelihood context see, e.g. Pawitan, 2001 (Chapters 6 and 14). Observations with large residuals are systematically down weighted so that the more extreme the value, the smaller the weight it gets. Total rejection of extreme observations (outliers) can be regarded as a special case of robust methods. Robust methods are data driven, and to that extent they may not indicate any causes of the discrepancies.

A useful alternative is to seek to model isolated discrepancies as being caused by variation in the dispersion, and to seek covariates that may account for them. The techniques of joint modelling of mean and dispersion developed and demonstrated in this thesis make such exploration straightforward. Furthermore if a covariate can be found which accounts for the discrepancies this gives a model-based solution which can be checked in the future by policy makers in the field it is applied. Under the modelling of crop yield therefore, this study propose the JGLM and HGLM approach of Lee and Nelder (2006). With hierarchical gener-

alized linear models (HGLMs) of Lee and Nelder (2006), it becomes to consider various scale mixtures of distributions for β and by so doing greatly improving the model. To be specific, this study suggest the use of a gamma mixture for β and apply it to a very critical area in Ghana's quest for economic growth, food security and expansion in the agricultural sector.

1.3 Objectives

The general objective of this study is to select the best physical support covariates that influence crop yield in the three Northern regions of Ghana using recent methods for penalized variable selection and modelling techniques.

1.3.1 Specific objectives

The specific objectives are;

1. To compare the sparsity and number of significant crop yield variable selected by the three penalized methods; LASSO, SCAD, and H-likelihood.

2. To propose the H-likelihood approach to crop yield variable selection compared to other forms of penalized methods ie. LASSO and SCAD based on their estimated penalized cross validated errors.

3. To propose and demonstrate the extended versions of the famous GLM to JGLM and prove its ability to statistically improve fixed effects model quality using crop yield data.

4. To propose and demonstrate the HGLM with gamma mixture as best model approach with the highest model accuracy and prove its ability to statistically improve mixed effects model for crop yield.

1.4 Methodology

Maize and Soybeans are the very much cultivated in the Northern parts of Ghana due to their vegetation which supports the growth of grains and cereals. Beyond the numbers and descriptive statistics on yield of such crops, this study tries to bring out variables that significantly contribute to yield. There are broadly two class of methods; first has to do with the methods of regularized variable selection and second is the Hierarchical Generalized Linear Models.

1.4.1 Profile of study area

The northern region of Ghana is considered the major bread basket of the country, and is also the most susceptible to the vagaries of the weather, especially the lack of rainfall. Unfortunately past agricultural growth and development has been accompanied by increased income inequality, and poverty abatement is lagging in Northern Ghana (Al Hassan and Diao, 2007).

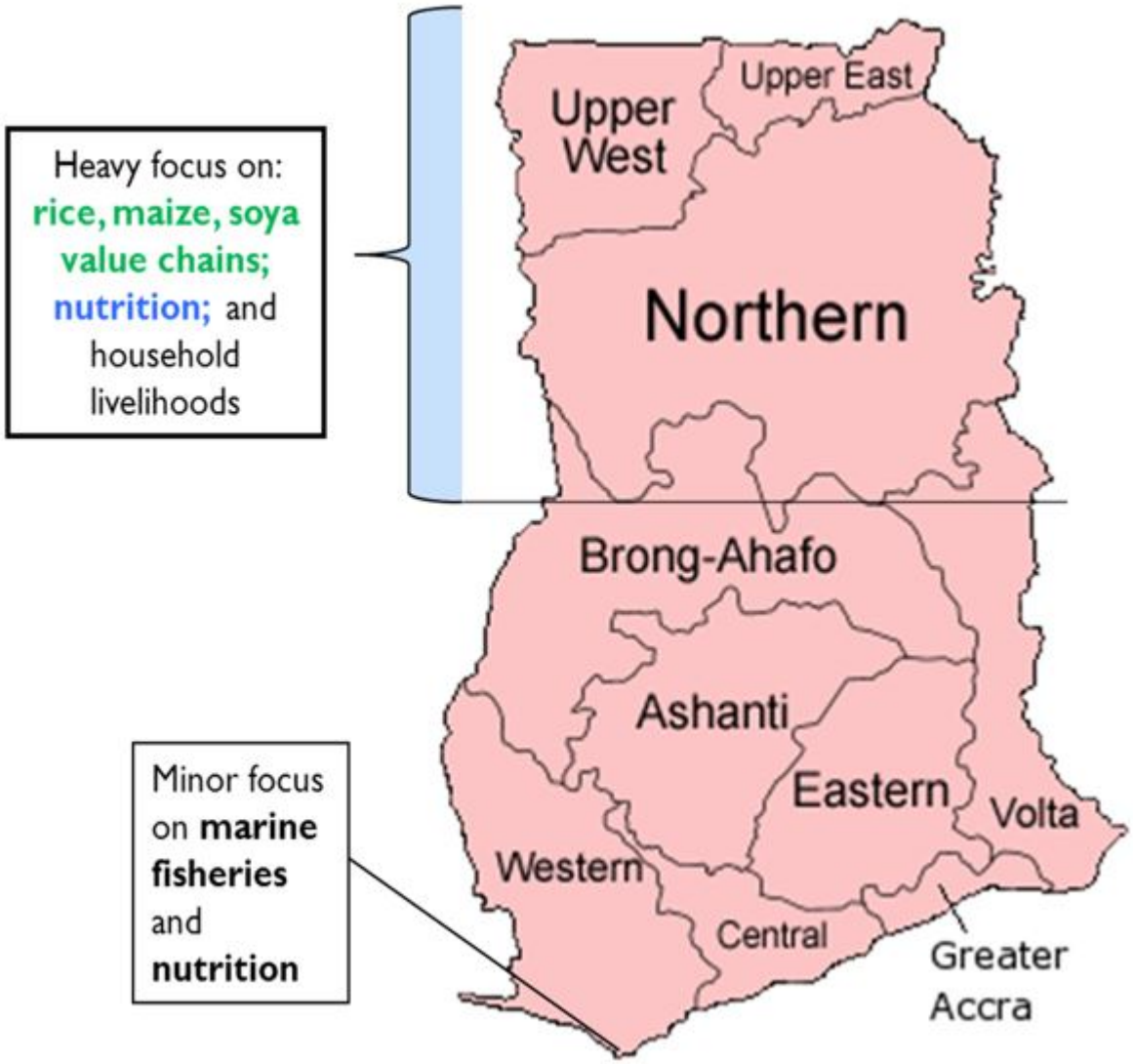

Figure 1.1: Profile map of study area

Over the past three decades, Ghana has experienced significant development and growth, becoming one of Africa's great success stories (IFAD, 2012). In 2011, Ghana's economy experienced the world's greatest economic growth at 13.4 percent (USAID, 2012). According to a 2012 report from the Center for Global Development, Ghana moved from a low-income country to a middle-income country in late 2010- a decade earlier than planned. This dramatic economic growth was fostered by a stable government and relatively conducive investment climate (Moss and Majerowicz, 2012).

Along with impressive economic growth, Ghana has experienced a steady decline in poverty and hunger over the past decade (IFAD, 2012), and is on a fast track to achieving the first Millennium Development Goal of halving 1990 levels of poverty and hunger by 2015 (Wiggins and Leturque, 2011). Unfortunately, this development is not equally distributed throughout the country. In the northern regions, poverty, hunger, poor nutrition and health, and high mortality rates among women and children are persistent, and must be addressed (FAO, 2010). Today, there is a dramatic north-south divide where poverty, as well as food and nutrition insecurity remains widespread in the northern savannah (IFAD 2012, 2010 Ghana Millennium Development Goals Report).

The northern region of Ghana lying between latitude 80^oS and 110^oN and stretching between longitude 30^oW to 10^oE is the largest administrative region in Ghana. About 80 per cent of the estimated 1.8 million population depend on agriculture for their livelihood. The estimated incidence of poverty at 69 per cent (International Fund for Agricultural Development, 2003) makes it one of the poorest regions in Ghana.

The mean annual rainfall is about 1100 mm but rainfall is characterized by high intensity and seasonal and annual variability (Andreini et al., 2000). The monomodal rainfall pattern, typical of the moist savannah agroecological zone of West Africa, results in a growing season lasting for 5-6 months. The soils are classified as Typic Plinthaqualf, Rhodic Paleustalf and Typic Plinthaquept (Soil Survey Staff, 1994), with Alfisols accounting for up to 80 per cent of the land area. The farming system is characterized by very small external inputs as inorganic fertilizers in spite of the inherently low soil nutrient content.

Maize is the most important food crop cultivated by the smallholder farmers, contributing about 20 per cent of calories to the diet. The commonest maize va-

riety is the white, dent-grained, late-maturing (120-day), streak-resistant, open-pollinated variety released by the Crop Research Institute of Ghana in 1989 (Sallah et al., 1993).

The increase in population has amplified pressure on land in northern Ghana. Land tenure and ownership are rooted in the traditional common property system, in which land administration is vested in the village chief who allocates parcels of land according to household needs. However, the gradual commercialization of agriculture has profoundly influenced land tenure and ownership, leading to a general tendency to preferentially allocate land to large-scale commercial farmers. The pressure on available land also manifests in intense competition between farmers and herdsmen on the use of the alluvial plains.

This northern part of Ghana is made up of three main regions; Upper West Region, the Upper East Region and the Northern Region. The largest of these is the Northern Region which incidentally is the largest region in Ghana, covering a land area of about 70,383 square kilometers. However, it has the lowest population density of all ten regions in the country (PPMED, Ghana, 1991) with 80 per cent of its people dependent on farming. The major food crops grown here are yam, millet, rice, maize, sorghum, soybeans, groundnut and cassava.

Tamale is the administrative capital of the Northern Region and the biggest town in Northern Ghana. Although Ghana's have weather cycles consisting of two seasons; rainy seasons and a dry season, the northern region experiences a very short rainy season and an extended dry season (traditionally November-April). During the dry season, there are also Harmattan winds (dry desert winds) which blow from the northeast from December to March, lowering the humidity with hot days and cool nights. However, like most climates, there is some variability, more so in recent years. Annual rainfall is about 1,100 mm (about 43 in) with a range from

about 800 mm to about 1,500 mm. In the Northern region, the Ghana Meteorological Agency (GMA) reported a 10.2 per cent change in the cumulative rainfall between the 30-year average and that for 2009. Those changes for the Upper East and Upper West regions were -3.5 per cent and - 34.5 per cent respectively. All together, the percentage change in rainfall for the northern sector of Ghana was -8.6 per cent.

Average monthly rainfalls over the past 4 decades in the three northern regions has changed. The Upper East Region has a fairly steady rainy season but the Northern and Upper West Regions trended toward a more variable "rainy season" by about one month on average. The regions have a vegetation classified as savannah woodland, with vast areas of grassland, characterized by drought-resistant trees such as the acacia, baobab, shea nut, dawadawa, mango, neem and mahogany.

The soil in this area is mostly silt or loam, thus having the tendency to get water-logged during the rainy season but drying up in the dry season. This, however, works well for the farmers since they grow various types of crops: each with its own soil preference. For example, during the rainy season, rice is a preferred crop since it fares very well on marshy land. Yam, on the other hand, is better cultivated when the land is dried out. Although the type of vegetation supports agricultural production quite well, a major hurdle for farmers is maintaining the soil fertility of the land throughout the various farming cycles.

1.4.2 Data Source and Type

The analyses were based on raw data available at the regional Monitoring and Evaluation office of the Linking Farmers to Markets (FtM) project in Tamale - Ghana. The project is organized by the Alliance for a Green Revolution in Africa (AGRA) with the primary goal of easing the flow of produce from the farm-gate

to the market by linking smallholder farmers to commercial buyers and processors. (FtM Grant Narrative Report, 2011)

Specifically, the project aims at forming alliances with partners to build organizational management, productivity and entrepreneurial skills of smallholder farmers engaged in the production and processing of rice, maize, sorghum, soy beans and cowpeas in the Northern, Upper West and Upper East regions of Ghana. The project is also to link approximately 50,000 smallholder farmers of maize, rice, sorghum, and soy bean in the Northern, Upper West and Upper East regions of Ghana to develop commercial relationships with structural markets such as industrial processors, the Ghana School Feeding Program, the World Food Program's P4P, local entrepreneurs and processors as well as urban consumers in Southern Ghana. (FtM Grant Narrative Report, 2011).

A two (Multi) stage probability sampling technique was used in selecting specific districts and Communities as first and second stages respectively. Three different strata was created for districts in the three regions considered and 7, 3, 3 districts selected proportionately for the Northern, Upper east and upper west regions respectively. At the community selection stage, only communities with recognized Farmer Based Organizations were included for selection. Also at this stage, 7, 3, 3 communities were proportionately selected for Northern, Upper east and Upper west respectively.

In all, data from 800 Maize and Soy bean farmer based organizations (FBOs) were gathered by means of a structured questionnaire. This was later cleaned to 790 distinct observations. The FBOs were randomly selected through a multi-stage random procedure. First, proportional randomizations resulted in selecting three (3) farming communities each from the Upper East and West regions while seven (7) were selected from the Northern Region.

Fixed effect variables measured include; crop type (Maize or Soybean), Financial Credit (Acquired or Not), Training (Acquired or Not), Study tour (Acquired or Not), Demonstrative Practicals (Acquired or Not), Networking Events (Acquired or Not), Post harvest Equipment (Acquired or Not), Number of farmers in the FBO and Plot size cultivated. Beside these 9 fixed effects, 36 two-way interaction terms are also generated as fixed interaction terms. This brings the total number of fixed covariates to 45.

Dependent variable measured is Total Crop Yield. The regions and the particular communities are treated as Random variables.

The main source of knowledge for the successful completion of this study has been the Data Science for Knowledge Creation Research Centre, at the Department of Statistics, Seoul National University - Korea Republic. The Seoul National University Main Library resource centre as well as the E-resource centre of the Kwame Nkrumah University of Science and Technology - Kumasi, Ghana have been extremely helpful sources of knowledge all through this dissertation. However, the internet and other individual Crop yield experts and Statisticians that formed my Team of Supervisors, have and continued to help enrich the progress and outcome of the study.

1.4.3 Methods of analysis and modelling

This study sought to select significant variables among many potential ones to be included in a model via penalized methods and to also propose and demonstrate the excellent performance of higher levels and very recent extensions of the Generalized Linear Models (GLM); Joint Generalized Linear Models (JGLM) and Hierarchical Generalized Linear Models (HGLM) in the development of Statistical Models with highest model accuracy. The researcher proposed the H-Likelihood

method of penalized variable selection as well as the unified JGLM and HGLM with gamma random effects as best methods useful for variable selection and modelling crop yield in the three Northern regions of Ghana respectively.

1.4.4 Variable selection

When we have p variables in a model, the total number of models we can generate is 2^p. As number of variables increase, identifying the optimal model within the large model space can be computationally burdensome. Stepwise regression methods (Miller, 2002) are among the most known subset selection methods, although currently quite out of fashion. Stepwise regression is based on two different strategies, namely Forward Selection (FS) and Backward Elimination (BE). Details of the theory behind the selection procedure of this classical variable selection method is discussed in chapter three.

However, these selection procedures are discrete in the sense that one variable is either added or deleted at a time, and hence provide unstable result (Breiman, 1995). For continuous selection process, penalized likelihood methods have been developed recently, including Negative Garrote (Breiman, 1995), LASSO (Tibshirani, 1996, Fan and Li, 2001, Zou and Hastie, 2005), etc.

Consider the model

$$Y_i = \beta^{\mathbf{T}}\mathbf{X_i} + \varepsilon_{\mathbf{i}},\ \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{n} \tag{1.1}$$

where Y_i is the response variable, $\mathbf{x_i}$ is a p-vector of predictors for the ith subject, β is a p-vector of regression coefficients, and $(\varepsilon_i, \ldots, \varepsilon_n)$ are independent and identically distributed errors. For simplicity, assume that the ε_i s have means 0. Define $l(\beta) = \|y - X\beta\|^2$ where $y = (Y_1, \ldots, Y_n)^T$ and $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})^{\mathbf{T}}$. Then the penalized least squares estimator of β is the minimizer of the objective

function

$$l(\beta) + n \sum_{j=1}^{p} p\lambda(|\beta_j|) \tag{1.2}$$

where $p\lambda(.)$ is a penalty function. Appropriate choices of $p\lambda$ (detailed in chapter 3) yield the aforementioned variable selection procedures. For likelihood-based models, the penalized maximum likelihood estimator is obtained by setting $l(\beta)$ to the minus log-likelihood. This study demonstrates how the h-likelihood methods overcome such difficulties to allow an oracle variable selection and simultaneously enhance estimation power.

1.4.5 Modelling crop yield

The Joint GLM is an extension of the famous GLM. The technique seeks to model isolated discrepancies as being caused by variation in the dispersion, and to seek covariates that may account for them. The techniques of joint modelling of mean and dispersion developed and demonstrated in this thesis gives a model-based solution to finding covariates which account for the discrepancies in our crop yield model.

The thesis also employ the Hierarchical Generalized Linear Models (Lee and Nelder, 1996, 2006). HGLMs is a synthesis of two widely used existing model classes; Generalized Linear Models (GLMs) and Normal Linear Mixed Models. Generalized Linear Mixed models (GLMMs: Breslow and Clayton, 1993), which assumes Gaussian random effects, form a subclass of HGLMs. These models have received increasing attention because of their wide applicability and ease of interpretation. However, likelihood estimation in random effect models is often complicated because of the marginal likelihoods involves an analytically intractable integrals.

To avoid this, various approximations and Bayesian inferential procedures have been proposed. An alternative is to use the hierarchical likelihood (detailed in

chapter 3), which avoids such burdensome numerical integrations. By definition, HGLMs is defined by Lee and Nelder (1996) as;

Conditional on random effects u, the response y follows a GLM family, satisfying $\mathrm{E}(y/u) = \mu$ and $var(y/u) = \phi V(\mu)$,

For which the kernel of the likelihood is given by

$$\sum \frac{y\theta - b(\theta)}{\phi} \tag{1.3}$$

Where $\theta = \theta(\mu)$ is the canonical parameter. The linear predictor takes the form

$$\eta = g(\mu) = X\beta + Zv \tag{1.4}$$

Where $v = v(u)$, for some monotone functions $v(.)$, are the random effects and β are the fixed effects. The random component u follows a distribution conjugate to a GLM family of distributions with parameters λ (detailed in chapter 3).

1.4.6 Statistical Software Used

The R package, a statistical analysis software (R version 3.0.3 (2014-03-06)) was used throughout the analysis.

1.5 Justification of the Study

Over the years, variable selection methods have received much attention and have been applied to various fields. This is because, using uninformative variables will not only waste money and time, but also reduce estimation efficiency or prediction accuracy. Selecting an appropriate set of important variables helps to reduce the variances of parameter estimates. By eliminating some noise variables, precision of the estimates are greatly improved.

Ghana is still an agriculture-based economy. The country's recent development

is characterized by balanced growth at the aggregate economic level, with agriculture continuing to form the backbone of the economy (McKay and Aryeetey, 2004). This study is therefore highly justified as it seeks to rigorously select variables for crop yield with the help of very recent methods of variable selection and parameter estimation. Findings of this thesis would form an empirical basis for Agricultural related Government and Non-governmental stakeholders to focus on the significant aid and supports that would actually maximize yield.

To statisticians and the academia, this thesis seeks to contribute to the ever growing knowledge in the area of Penalized variable selection as a result of the limitations of the classical stepwise variable selection. The study proposes the hierarchical likelihood, the Joint GLM technique and the Hierarchical Generalized Linear Models which are recent extensions of the Fisher likelihood (Fisher, 1935) and the Generalized Linear (Mixed) models (GLMMs: Breslow and Clayton, 1993) respectively. This would among other things stimulate further studies in this area of statistics and its application in other areas of the economy of Ghana as well as elsewhere in the world.

1.6 Limitations and Scope of the Study

The study is an application of statistical variable selection and modelling techniques. The scope of this study is restricted to Maize and Soy bean yield from some randomly selected Farmer based organizations in the Northern, Upper East and Upper West Regions of Ghana. Even though the target crops do very well in the specified areas, many other parts of the country are also noted for their production. Focusing on the three Northern regions only may be a limitation to the study especially if one wants to rely on the findings of this study for country generalization.

The issue of post harvest losses is likely to be a great limitation to the accuracy

of measured crop yield. Also, measurements for plot sizes and number of farmers were based solely on the verbal records of the farmers since the researcher was limited by means of validating the authenticity of all such records for the 800 selected FBOs. This in the researchers view, can introduce some measurement errors and this may be a limitation to the study findings.

The researcher admits that but for the unavailability of data, as frequently the case in many parts of our world, extensive input data on farm management practices, soil condition, climate and other non-physical contributors to yield would have enriched our models. That not withstanding, the study is carefully structured within the confines of the thesis study matter.

1.7 Organization of the Study

This report is organized in five chapters. Chapter 1 is the introductory chapter to the entire study. It takes a critical look at the general background of agricultural contribution to Ghana's economic growth and also looks at the general socio-economic profile of the study areas. The problem statement, research questions and objectives, research methodology, justification of the study as well as scope and limitations of the study are discussed in this chapter. Chapter 2 reviews related literature based on the thesis objectives and preferred models to be used in achieving these objectives. Expected outcome of the study and other comparative results of similar studies are also discussed in this chapter. Chapter 3 describes the theory of model to be used, formulations and methods of solution. Chapter 4 is dedicated to data analysis, results and discussion of study findings. Chapter 5 concludes the entire study by stating specific recommendations to stakeholders based on the major findings made in the study.

Chapter 2

Literature Review

2.1 Introduction

This chapter takes a review into the concept of variable and penalized variable selection methods as well as statistical modelling using the extended versions of the Generalized Linear models (GLM); Joint GLMs and Hierarchical Generalized models (HGLMs). It also discusses Agricultural and crop yield models as found in literature. It presents summary of abstracts and critiquing of various literature with regard to the model being used and the general working title.

2.2 Variable Selection

Variable selection is an important topic in linear regression analysis. In practice, a large number of predictors usually are introduced at the initial stage of modelling to attenuate possible modelling biases. On the other hand, to enhance predictability and to select significant variables, statisticians usually use stepwise deletion and subset selection. Although they are practically useful, these selection procedures ignore stochastic errors inherited in the stages of variable selections. Hence, their theoretical properties are somewhat hard to understand.

2.2.1 Classical variable selection methods

Given a linear model;

$$y = \mathbf{X}\beta + \varepsilon \tag{2.1}$$

Where y is an $N \times 1$ vector of responses, X is an $N \times p$ designed matrix, β is a $p \times 1$ vector of unknown regression coefficients, and ε is an $N \times 1$ vector of random

errors with $\varepsilon \sim N(0, \sigma^2 I_N)$. If we have p variables, there exist 2^p candidate models. As the number of variables increases, the number of computations needed rapidly increases. For efficient and effective computation, many algorithms have been proposed. In this section, the researcher will go over traditional procedures that are in common use.

All possible regressions are in fact to compare 2^p candidate models. However, it requires considerably complex computations. Furnival and Wilson (1974) proposed the leaps and bounds algorithm to perform all possible regression efficiently. They employed the lexicographic algorithm and also performed an exhaustive search. The algorithm is quite useful in linear models with $p < 40$. The idea of the algorithm is to use information obtained from previous steps. As a result, we can reduce the computational burden. Their algorithm offers the best m models of each size, where m is set by the user. They provided the Fortran subroutine which is available in many statistical software. When we find the best subset by the leaps and bound algorithm, Cp, R^2, and R^2adj, are available as a criteria for comparing candidate models. The best subset selection is to choose the best one among all possible subsets. It tends to result in a model with too many variables, and the final model would be very unstable.

For forward selection, the procedure starts with no variables in the model. First, for all variables not included in the model we check which variable has the largest partial F-statistic. If the partial F-statistic is greater than a pre-determined F value, the variable is added to the model. The pre-determined F value is often called 'F-to-enter'. The above procedure is continued until no new variable can be added to the model any more. Roecker (1991) showed that forward selection can provide slightly smaller prediction error and less bias compared to all possible regressions.

Backward elimination is the simplest procedure for variable selection and works in the opposite direction of forward selection. At first, the procedure begins with all variables in the model. The partial F-statistic are then computed for each variable out of the model. If the smallest partial F-statistic is less than a pre-determined F value, the variable is excluded from the model. This pre-determined F value is sometimes called 'F-to-remove'. The backward elimination also stops when the partial F-statistic for variables not belonging to the model are all greater than F-to-remove. Forward selection and backward elimination are more economical than the all possible regressions. Since we start with all variables in the model, backward elimination can be performed only when $p < N$.

The stepwise regression can be thought of as a combination of forward selection and backward elimination. At each step, one variable may be either entered or removed. Therefore, the same variable can be again added to the model after exclusion. Note that this procedure allows the move of only one variable at one step. These methods described above are easy to understand and perform, but the selection results are unstable. Because the selection procedure is discrete; that is, variables are either remained or dropped from the model, even small changes in the data might lead to quite different results for variable selection. This can also result in worse prediction accuracy. In next section, the researcher shall review penalized approaches which are continuous selection procedures.

In general, the method of least squares is used to estimate the regression coefficients from the data. However, in practice, various model selection criteria have been proposed to compare candidate models and to select the best model. Different criteria have different motivations and perform better for some problems in practice. A brief review on those widely used criteria are as follows;

2.2.2 Coefficient of Determination: R^2

The coefficient of determination has been widely used as a measure of the capability of the model to fit the data. It is defined as

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{RSS}{SST} \tag{2.2}$$

Where $\bar{y}$ is the overall mean of y, $SSR = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$ is the regression sum of squares and $SST \sum_{i=1}^{N}(y_i - \bar{y})^2$ is the total sum of squares. This can be viewed as the ratio of the explained variance to the total variance. As the number of parameters used in the model increases, R^2 increases. Therefore, R^2 achieves the maximum when all variables enter in the model. Based on R^2, we select the candidate model having the largest R^2. As a result, the chosen model might be over fitted.

2.2.3 Adjusted R^2

The drawback of R^2 leads to the modification of R^2. The adjusted R^2 is defined as

$$R_a d^2 = 1 - \frac{RSS/(N-p)}{SST/(N-p)} = 1 - \frac{(N-1)MSE}{SST} = 1 - (\frac{N-1}{N-p})(1-R^2) \tag{2.3}$$

The $R_a d^2$ penalizes bigger models. As seen in (1.2), the minimum MSE and the maximum $R_a d^2$ yield the same model selection. That is, comparing models in terms of MSE is identical to that in terms of $R_a d^2$.

2.2.4 Residual Mean Square: MSE

The residual mean square is defined as

$$MSE = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-p} = \frac{RSS}{N-p} \tag{2.4}$$

Where p is the number of variables in the fitted model, RSS is the residual sum of squares and $\hat{y}_i$ is the fitted value of y_i. This is widely used to evaluate how well the model is fitted to the data. We prefer the candidate model with the minimum MSE. For small data sets, the MSE might not work effectively.

2.2.5 Mallows' C_P

The statistic C_p, proposed by Mallows (1973), is defined as

$$C_p = \frac{RSS}{\hat{\sigma}^2} - N + 2p \tag{2.5}$$

where $\hat{\sigma}^2$ is the residual mean squares in the full model. The C_p was motivated as an unbiased estimate of prediction accuracy of the candidate model. If the model with p variables is proper, $E(C_p)$ is approximately equal to p. Therefore, we find points close to the $C_p = P$ line on the plot of C_p versus p. Also, it might be good to select points below the $C_p = P$ line due to random variation. As a result, we prefer choosing the candidate model with small C_p value about equal to p. Generally, many statistical software packages select the model having the smallest C_p. Mallows (1995) studied the property of a C_p plot when p is large and there exist many weak effects. Some modified versions of Mallows' C_p are described with some examples in Miller (2002).

2.2.6 Information Criteria

Akaike Information Criterion (AIC) is originally proposed by Akaike (1973) to consider the number of parameters as a standard comparing the candidate models. His idea is to impose a penalty for model complexity to the log likelihood.
In general, the AIC is defined as

$$AIC = -2log(likelihood) + 2p \tag{2.6}$$

Hurvich and Tsai (1989) showed that AIC brings about over fitting in the small sample, and suggested using $AICc$, a corrected version of AIC,

$$AICc = AIC + \frac{2(p+1)(p+2)}{N-p-2} \tag{2.7}$$

For several variants of AIC, see McQuarrie and Tsai (1998).

Another information criterion is the Bayesian Information Criterion (BIC), proposed by Schwarz (1978),

$$BIC = -2log(likelihood) + plogN \tag{2.8}$$

BIC is motivated in the Bayesian approach to model selection. Schwarz (1978) made an appropriate modification of maximum likelihood using the asymptotic behaviour of Bayes estimators. We desire the model with smaller AIC or BIC. Miller (2002) stated that using AIC tends to choose a little larger models than using Mallows' C_p.

Information criteria and C_p statistic consider the trade-off between σ^2 and p. One cannot say which criterion is better than the others. However, we can consider the behavior of these criteria as follows. When $N > e^2$, BIC penalizes larger models more heavily, and hence it prefers simpler models. Moreover, BIC is asymptotically consistent for model selection. That is, the probability that BIC yields the correct model approaches 1 as $N \to \infty$. Contrary to BIC, AIC tends to select more complex models as $N \to \infty$. BIC also has disadvantages; BIC often chooses too simple models for finite samples.

Hurvich and Tsai (1989) showed that BIC may poorly perform in small samples. They also showed that BIC is consistent when the true model is fixed. If the dimensionality of the true model increases with N, AIC is also consistent (Shibata,

1981).

2.2.7 Prediction Sum of Squares: (PRESS)

Allen (1981) proposed the prediction sum of squares (PRESS) which is defined as

$$PRESS = \sum_{i=1}^{N}(y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{N} e_{(i)}^2 \tag{2.9}$$

where $\hat{y}_{(i)}$ denotes the predicted value of the i^{th} response when the model is fitted without using the i^{th} observation. PRESS provides detailed information about the stability of the candidate models. However, PRESS requires an excessively complex computation. Breiman and Spector (1992) showed that non-resampling estimates including PRESS statistic lead to inaccurate estimates of the mean squared error of prediction. To overcome this problem, they used cross-validation and bootstrap methods.

2.3 Cross Validation

To assess the predictive value of selected significant variables, cross-validation is often recommended (Houwelingen and Cessie, 1990). Before the different CV methods are discussed, two theoretically important definitions for a correct model and the 'true' model are given. I define and consider a correct model as one that contains all important (truly significant) variables. It is assumed that these variables are among the p given variables. If an important variable is missing, the model is termed incorrect (under-fitting). A correct model may contain additional noise variables (over-fitting). The smallest correct model (i.e. the model that contains only the important variables and no more) is termed the true model. It is the one with minimal prediction error. Hence, the general objective of variable selection is to find the true model. If the true model is not among the candidate models, the model closest to the true one is sought.

In most statistical models, frequently employed objective functions to assess the predictive ability in variable selection are n-fold cross-validation (Shao, 1993; Brieman et.al, 1984; Burman, 1989) and leave-one-out cross-validation (LOO-CV) (Allen, 1974; Stone and Roy, 1974; Picard and Cook, 1984).

In an n-fold CV the available training data are split into n disjoint groups of approximately the same size. Then the algorithm is run n times using $(n-1)$ groups as the construction set and one group as validation set. This is done in turn until each group was left out once. Clearly, if $n = m$ (m is sample size) then n-fold CV is LOO-CV, since exactly one object is left out at a time. Unfortunately, n-fold CV and LOO-CV without further constraints are unsuitable objective functions to find the true model. Shao showed that minimizing the LOO-CV estimate for the prediction error does not lead to a statistically consistent choice of the true model in case of multiple linear regression (MLR) (Shao, 1993).

A consistent objective function selects the true model from the candidate models with certainty (probability equal to one) when the sample size is increased to infinity ($m \to \infty$). In this sense, LOO-CV is inconsistent. However, with large sample sizes, LOO-CV identifies the variable subset belonging to the true model (i.e. incorrect models will not be selected), but it also selects additional variables. That means that minimizing the LOO-CV estimate results in over-fitting and thus in a larger prediction error.

The deficiencies of LOO-CV can be overcome by using a leave-multiple-out cross-validation (LMO-CV) for model selection (Shao, 1993). In LMO-CV, the available training data set is split into a validation data set with d objects and a construction data set with the remaining $m-d$ objects. Put differently, d objects are left out for validation. For LMO-CV to be consistent, two requirements need to be fulfilled. First, in LMO-CV the size of the validation data set (d) needs to

be much larger than the size of the construction data set $(m - d)$ (Shao, 1993). Second, the LMO-CV estimate of the prediction error needs to be averaged over a large number of different splits into construction and validation sets. This renders LMO-CV computationally extremely expensive.

In practical applications, the most important parameter of LMO-CV is the choice of the validation data set size (d). For theoretical reasons, Shao recommends the use of

$$d = m - m/(ln(m) - 1) \tag{2.10}$$

A reasonable range for several real data sets was found to be $d \approx 0.4 - 0.6.m$. However, d is generally problem-dependent.

Despite these theoretical findings, LMO-CV could not prevail against LOO-CV and n-fold CV. The latter two CV types were used in many studies applying variable selection (see references cited in (Baumann et.al, 2002) and (Izrailev and Agrafiotis 2002; Jouan-Rimbaud et.al, 1996; Gao et.al, 2002). Yet, in several of these papers, it was recognised that over-fitted models resulted. For example, when augmenting the real data matrix with random variables, it was found that a genetic algorithm (GA) selected some of these augmented random variables (Jouan-Rimbaud et.al, 1996).

In another case, various variable selection procedures were applied to permuted response vectors (randomisation test) and yielded improvements, despite the fact that the response vectors were scrambled (Norinder, 1996). Randomisation tests were used in a different study to reveal and to avoid chance correlation and over-fitting (Leardi and Gonzalez, 1998). A common finding is that selection procedures based on LOO-CV as the objective function improve the internal consistency of the training data sets (decreased internal prediction error), but often do not improve test-set prediction (Norinder, 1996; Golbraikh and Tropsha,

2002). Sometimes, test-set prediction even deteriorates (Norinder, 1996).

In an eye-opening paper, Golbraikh and Tropsha showed that there is little correlation between the internal estimate of the prediction error obtained by LOO-CV and the external estimate of the prediction error obtained by test-set prediction (Golbraikh and Tropsha, 2002). An external estimate of the prediction error (PE) provides an independent assessment of the predictive power of a finally chosen model, since the model to be validated did not see these data before. This independence stems from the fact that these test-set data did not influence the choice of the model at all. An internal estimate of the PE is used to influence the choice of the final model (e.g. cross-validation as objective function in variable selection, or cross-validation as objective function in the selection of the optimal number of latent variables).

Although the validation data are independent of the model building process in a single split of the CV procedure (recall, training data are split into construction data and validation data), the resulting internal estimate of the PE is nonetheless overoptimistic since the same data are repeatedly used to build and to assess the model. As a consequence, the variable selection procedure may learn the training data and their respective splits into construction and validation data by heart. It is intuitively clear that the more similar the construction data to the validation data, the easier it is for a selection procedure to learn the idiosyncrasies of the data. In LOO-CV, only one object is deleted from the training set. Hence, the similarity between the construction data and the validation data is generally largest.

Validation can be made more stringent by leaving out multiple objects in cross-validation (LMO-CV). This has two effects: first, there are less data for constructing the model; and, second, more data are available to assess the model's

quality. Put another way, in LMO-CV, there is less information to build the model but more information to validate the model. Both points force the selection procedure to concentrate on the general patterns in the data and that, in turn, reduces over-fitting.

Clementi and co-workers were the first to recommend the use of LMO-CV for variable selection (Cruciani et.al, 1992). Put precisely, they suggested using a repeated n-fold CV. A repeated n-fold CV consists of B runs of the n-fold CV procedure with different random splits into n disjunct groups. It is a balanced version of LMO-CV, since every object is used exactly B times for assessing the candidate model.

In the aforementioned sequel of papers, Clementi and co-workers introduced the SDEP parameter (standard deviation of prediction error), which is based on a repeated n-fold CV. It is used for the comparison of different models and was found to perform better than LOO-CV. Thus, SDEP became the objective function of the GOLPE (Generating Optimal Linear PLS Estimations) variable selection procedure (Cruciani et.al, 1994). The parameter n of the repeated n-fold CV is usually set to 4 or 5 in GOLPE (Cruciani et.al, 1994). This corresponds to leaving out 25 percent or 20 per cent of the data during cross-validation. The number of repetitions (B) is often set to 100.

Using LMO-CV instead of LOO-CV is one possible way to reduce the amount of over-fitting. Other sensible constraints to avoid over-fitting are restrictions on the maximum number of variables or the maximum number of latent variables (McShane et.al, 1999; Kubinyi, 1996). When using LOO-CV or n-fold CV, restricting the maximum number of latent variables often results in the same model performance with respect to test-set prediction as for LMO-CV (Baumann et.al, 2002). However, this type of constraint requires a priori knowledge of the problem. A

different, innovative approach to avoid over-fitting was followed by Todeschini and co-workers, who computed several indices for a candidate model to detect multicollinearity and chance correlations (Baumann et.al, 2002). If these indices indicate problems with the data, the respective Candidate model is rejected.

Summing up, the susceptibility of LOO-CV to over-fitting was shown theoretically and was recognised by several groups studying practical applications. In this study we also strongly recommend and apply the LMO-CV as justified by Clementi and co-workers for variable selection (Cruciani et.al, 1992).

Precisely in this dissertation, the researcher suggests the use of a repeated 10-fold CV. A repeated 10-fold CV consisting of 100 runs of the 10-fold CV procedure with different random splits into 10 disjoint groups. It is a balanced version of LMO-CV, since every object is used exactly 100 times for assessing the candidate model. The researcher introduces the PCVE (Penalized Cross Validated Errors) which is based on a repeated 10-fold CV. This PCVE is used for the comparison of different penalized methods of variable selection and was found to perform better than LOO-CV. The parameter n of the repeated n-fold CV was set to n=10 in this study. This corresponds to leaving out 10 per cent of the data during cross-validation. The number of repetitions was set to 100.

2.4 Tuning Parameter

Among other variable selection methods, penalized regression models have been popularly used, which penalize the model fitting with various regularization terms to encourage model sparsity, such as the lasso regression (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), the adaptive lasso (Zou, 2006), and the truncated l1-norm regression (Shen et al., 2012). In the penalized regression models, tuning parameters are often employed to balance the trade-off between model fitting and model sparsity, which largely affects the

numerical performance and the asymptotic behaviour of the penalized regression models.

For example, Zhao and Yu (2006) showed that, under the irrepresentable condition, the lasso regression is selection consistent when the tuning parameter converges to 0 at a rate slower than $O(n-1/2)$. Analogous results on the choice of tuning parameters have also been established for the SCAD, the adaptive lasso, and the truncated l_1-norm regression. Therefore, it is of crucial importance to select the appropriate tuning parameters so that the performance of the penalized regression models can be optimized.

In literature, many classical selection criteria have been applied to the penalized regression models, including cross validation (Stone, 1974), generalized cross validation (Craven and Wahba, 1979), Mallows' Cp (Mallows, 1973), AIC (Akaike, 1974), BIC (Schwarz, 1978).

For instances, under certain regularity conditions, Wang et al. (2007) and Wang et al. (2009) established the selection consistency of BIC for the SCAD, and Zhang et al. (2010) also showed the selection consistency of generalized information criterion (GIC) for the SCAD. Most of these criteria follow the route of minimizing the estimated prediction error or maximizing the posterior model probability. To the best of our knowledge, few criteria has been developed directly focusing on the selection of the informative variables.

This thesis proposes a general tuning parameter selection criterion based on a novel concept of variable selection stability. Similar stability measures have been studied in the context of clustering (Ben-Hur et al., 2002; Wang, 2010) and variable selection (Meinshausen and Buhlmann, 2010). The key idea is that if multiple samples are available from the same distribution, a good variable selec-

tion method should yield similar sets of informative variables that do not vary much from one sample to another.

The effectiveness of the proposed selection criterion is demonstrated in a variety of simulated examples and real applications. More importantly, its asymptotic selection consistency is established, showing that the variable selection method with the selected tuning parameter would recover the truly informative variable set with probability tending to 1.

2.5 Penalized Methods for Variable Selection

Consider a linear regression model

$$Y = X'\beta + \varepsilon \tag{2.11}$$

where β is a $p \times 1$ vector of regression coefficients associated with X. We are interested in estimating β when $p \to \infty$ as the sample size $n \to \infty$ and when β is sparse, in the sense that many of its elements are zero. However, the traditional variable selection can introduce severe problems such as biases in estimates of regression parameters and corresponding standard errors, instability of selected variables or an overoptimistic estimate of the predictive value (Chen and George, 1985; Houwelingen and Cassella, 1990; Harrell et.al, 1996; Sauerbrei, 1999). To overcome some of these difficulties several proposals were made during the last few decades.

The best classical subset variable selection suffers from several drawbacks, the most severe of which is its lack of stability as analysed, for instance, by Breiman (1996). In an attempt to automatically and simultaneously select variables, a unified approach via penalized least squares, retaining good features of both subset selection and ridge regression is proposed. The penalty functions have to be

singular at the origin to produce sparse solutions (many estimated coefficients are zero), to satisfy certain conditions to produce continuous models (for stability of model selection), and to be bounded by a constant to produce nearly unbiased estimates for large coefficients.

The bridge regression proposed in Frank and Friedman (1993) and the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996, 1997) are members of the penalized least squares, although their associated L_q penalty functions do not satisfy all of the preceding three required properties. The penalized least squares idea can be extended naturally to likelihood-based models in various statistical contexts.

In many cases, it is reasonable to assume a sparse model, because the number of important covariates is usually relatively small, although the total number of covariates can be large. We use the SCAD method to achieve variable selection and estimation of β simultaneously. The SCAD method is proposed by Fan and Li (2006) in a general parametric framework for variable selection and efficient estimation. This method uses a specially designed penalty function, the smoothly clipped absolute deviation (hence the name SCAD).

Compared to the classical variable selection methods such as subset selection, the SCAD has two advantages. First, the variable selection with SCAD is continuous and hence more stable than the subset selection, which is a discrete and non-continuous process. Second, the SCAD is computationally feasible for high-dimensional data. In contrast, computation in subset selection is combinatorial and not feasible when p is large.

In addition to the SCAD method, several other penalized methods have also been proposed to achieve variable selection and estimation simultaneously. Ex-

amples include the bridge penalty (Frank and Friedman, 1993), LASSO (Tibshirani, 1996), and the Elastic-Net (Enet) penalty (Zou and Hastie, 2005), among others.

Fan and Li (2006) and Fan and Peng (2004) studied asymptotic properties of SCAD penalized likelihood methods. Their results are concerned with local maximizers of the penalized likelihood, but not the maximum penalized estimators. These results do not imply existence of an estimator with the properties of the local maximize without auxiliary information about the true parameter value. Therefore, they are not applicable to the SCAD-penalized maximum likelihood estimators, nor the SCAD-penalized estimator.

Knight and Fu (2000) studied the asymptotic distributions of bridge estimators when the number of covariates is fixed. Huang, Horowitz and Ma (2006) studied the bridge estimators with a divergent number of covariates in a linear regression model. They showed that the bridge estimators have an oracle property under appropriate conditions if the bridge index is strictly between 0 and 1.

Several earlier studies have investigated the properties of regression estimators with a divergent number of covariates. See, for example, (Huber, 1981) and (Portnoy, 1984, 1985). Portnoy proved consistency and asymptotic normality of a class of M-estimators of regression parameters under appropriate conditions. However, he did not consider penalized regression or selection of variables in sparse models.

In this thesis, the researcher studied the asymptotic properties of the SCAD-penalized least squares estimator, abbreviated as LS-SCAD estimator henceforth. It was found that the LS-SCAD estimator can correctly select the non-zero coefficients with probability converging to one and that the estimators of the non-zero coefficients are asymptotically normal with the same means and covariances as

they would have if the zero coefficients were known in advance. Thus, the LS-SCAD estimators have an oracle property in the sense of Fan and Li (2001) and Fan and Peng (2004). In other words, this estimator is asymptotically as efficient as the ideal estimator assisted by an oracle who knows which coefficients are non-zero and which are zero.

Definitions to the LS-SCAD estimator, the consistency and oracle properties as well as the algorithm for computing the LS-SCAD estimator and the criterion for choosing the penalty parameter are detailed in chapter 3 of this thesis.

Fan and Li in 2011 performed Variable Selection via Non-concave Penalized Likelihood and its Oracle Properties. Their proposed methods select variables and estimate coefficients simultaneously. Hence they enable us to construct confidence intervals for estimated parameters. The proposed approaches are distinguished from others in that the penalty functions are symmetric, non-concave on $(0, \infty)$, and have singularities at the origin to produce sparse solutions. Furthermore, the penalty functions should be bounded by a constant to reduce bias and satisfy certain conditions to yield continuous solutions. A new algorithm was proposed for optimizing penalized likelihood functions.

The proposed ideas was widely applicable. They are readily applicable to a variety of parametric models such as generalized linear models and robust regression models. They could also be applied easily to non-parametric modelling by using wavelets and splines. Rates of convergence of their proposed penalized likelihood estimators were established. Furthermore, with proper choice of regularization parameters, they showed that the proposed estimators perform as well as the oracle procedure in variable selection; namely, they work as well as if the correct sub model were known. Their simulation shows that the newly proposed methods compare favourably with other variable selection techniques. Furthermore, stan-

dard error formulas were tested to be accurate enough for practical applications.

The Fan and Li (2011) approach is distinguished from traditional methods (usually quadratic penalty) in that the penalty functions are symmetric, convex on $(0, \infty)$, (rather than concave for the negative quadratic penalty in the penalized likelihood situation), and possess singularities at the origin. A few penalty functions are discussed. They allow statisticians to select a penalty function to enhance the predictive power of a model and engineers to sharpen noisy images. Optimizing a penalized likelihood is challenging, because the target function is a high-dimensional non-concave function with singularities. Their new and generic algorithm yields a unified variable selection procedure. A standard error formula for estimated coefficients is obtained by using a sandwich formula. The formula is tested accurately enough for practical purposes, even when the sample size is very moderate. The proposed procedures are compared with various other variable selection approaches and the results indicated favourable performance of their newly proposed procedures.

For models with main interest in a good predictor, the LASSO by Tibshirani, (1996) has gained some popularity. By minimizing residuals under a constraint, it combines variable selection with shrinkage. It can be regarded, in a wider sense, as a generalization of an approach by Houwelingen and Cessie (1990), who proposed to improve predictors with respect to the average prediction error by multiplying the estimated effect of each covariate with a constant, an estimated shrinkage factor. As the bias caused by variable selection is usually different for individual covariates, Sauerbrei, (1999) extends their idea by proposing parameter-wise shrinkage factors. The latter approach is intended as a post-estimation shrinkage procedure after selection of variables. To estimate shrinkage factors the latter two approaches used cross-validation calibration and also applied them in GLMs and regression models for survival data.

In their article titled "Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited", Houwelingen and Sauerbrei, (2013) proposed the data re-sampling approaches to handle data-dependent model building. In order to assess and compare several strategies, they conducted a simulation study with 15 predictors and a complex correlation structure in the linear regression model. Using sample sizes of 100 and 400 and estimates of the residual variance corresponding to R^2 of 0.50 and 0.71, they considered 4 scenarios with varying amount of information. They also consider two examples with 24 and 13 predictors, respectively. They discussed the value of cross-validation, shrinkage and back-ward elimination (BE) with varying significance level.

They assessed whether 2-step approaches using global or parameter wise shrinkage (PWSF) can improve selected models and compare results to models derived with the LASSO procedure. Besides MSE, they used model sparsity and further criteria for model assessment. The amount of information in the data had an influence on the selected models and the comparison of the procedures. None of the approaches was best in all scenarios. The performance of backward elimination with a suitably chosen significance level was not worse compared to the LASSO and BE models selected were much sparser, an important advantage for interpretation and transportability. Compared to global shrinkage, PWSF had better performance. Provided that the amount of information is not too small, they concluded that BE followed by PWSF is a suitable approach when variable selection is a key part of data analysis.

Two penalization methods, and a hybrid of these, are most commonly used. Ridge regression (Hoerl and Kennard, 1970) uses a penalty on the L_2 norm of the coefficients, which introduces bias in the prediction error in exchange for reduced variance. However, ridge regression keeps all variables in the model and

thus cannot produce a parsimonious model from many variables. LASSO regression (Tibshirani, 1996; 1997) penalizes the L_1 norm, which tends to reduce many coefficients to exactly zero and thus performs variable selection in addition to prediction. However, the LASSO has been noted to be inferior to Ridge regression for prediction in lower dimensional situations, and tends to select only one of a group of collinear variables, which may not always be desirable (Zou and Hastie, 2005).

Zou and Hastie (2005) thus proposed the Elastic Net, penalizing both the L_1 and L_2 norms with individual tuning parameters, as a way to achieve the best of both LASSO and Ridge. These three variants of penalized regression-LASSO, Ridge and Elastic Net have since been applied to a variety of phenotype prediction tasks using genomic data (for example, Sharma, 2008; Shedden, 2008). The elastic net performs simultaneous regularization and variable selection. It is able to perform grouped selection and is appropriate for the $p > n$ problem. It gives a more reliable analytical results on the degree of freedom of the elastic net over LASSO and has Interesting applications in other areas such as sparse PCA and new support kernel machines (Zou and Hastie, 2005)

Several previous simulation studies have investigated properties of the Elastic Net (Zou and Hastie, 2005), the LASSO and Ridge regression (Bovelstad, 2007; Gui and Li, 2005; Yuan and Lin, 2006).

Recently, variable selection methods using a penalized likelihood with penalty functions have been widely studied in various statistical models such as linear models, generalized linear models and Cox's (1972) proportional hazards (PH) models. The main advantage of those methods is that they select important variables and estimate the regression coefficients of the covariates, simultaneously. Such methods, for example, include the least absolute shrinkage and selection

operator (LASSO) by Tibshirani, (1996), smoothly clipped absolute deviation (SCAD) by Fan and Li, (2001, 2002), and adaptive-LASSO (Zou, 2006), etc. but have not compared all these methods with alternative strategies for their application.

In this thesis, the researcher propose a simple but unified penalized h-likelihood method for variable selection of fixed effects in a general class of semi parametric models. Here, the study consider three penalty functions, LASSO, SCAD and h-likelihood (HL; Lee and Oh, 2009). In contrast, the SCAD penalty provides good properties such as oracle property, while the HL penalty is un-bounded at the origin (Lee and Oh, 2009) and gives a very good performance in various low and high dimensional problems (Lee et al., 2010; Lee et al., 2011a,b). Note that the SCAD penalty method leads to an oracle maximum likelihood (ML) estimator, whereas the HL penalty approach gives an oracle shrinkage estimator (Kwon et al., 2013). In other words, an oracle ML estimator is the ML estimator when all covariates with non-zero coefficients are known. Fan and Peng (2004) showed that a local solution of the SCAD penalty is asymptotically equivalent to an oracle ML estimator.

Similarly, an oracle shrinkage estimator is the shrinkage estimator when all covariates with non zero coefficients are known. Kwon et al. (2013) showed that a local solution for the HL penalty is an oracle shrinkage estimator. It is well known that shrinkage estimations would be preferred for prediction (Efron and Morris, 1975; Casella, 1985; Lee and Nelder, 2006). The Simulation results in chapter 4 show that the HL has higher probability of choosing the true model than the LASSO and SCAD methods without losing prediction accuracy.

The study shows that the proposed approach can be easily implemented via a slight modification to the existing h-likelihood estimation procedures (Ha and

Lee, 2003; Ha et al., 2011). It also investigates via crop yield dataset the performances of the three variable-selection methods (LASSO, SCAD and HL) within the framework of the proposed procedure.

The study presents a comprehensive assessment and optimization of these methods via sparse penalized approaches such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and the very recent H-likelihood approach (Lee and Nelder, (2006)) using the crop yield dataset in this dissertation.

The study demonstrates the concepts of Cross-Validation, Shrinkage and Variable Selection by comparing the approaches of each of the above mentioned sparse penalized approaches to variable selection. All these methods have common advantages over subset selection procedures; they are computationally simpler, the derived sparse estimators are stable, and they facilitate higher prediction accuracies.

The sparse penalized methods are useful techniques for selection of relevant variables in many practical problems. For example, Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO) and found that it can perform parameter estimation and variable selection simultaneously. Another popular method, the smoothly clipped absolute deviation (SCAD) penalized estimation, was proposed by Fan and Li (2001) and Fan and Peng (2004). They proved that the SCAD estimator has the oracle property-the asymptotic equivalence of the SCAD estimator with the oracle estimator. Here, the oracle estimator is an estimator obtained by deleting all irrelevant predictive variables (i.e., variables whose true regression coefficients are zero) in advance.

Several theoretical results about sparse penalized approaches have been studied.

For the LASSO, Knight and Fu (2000) studied asymptotic properties of LASSO-type estimators with a fixed number of parameters. Zou (2006) developed the adaptive LASSO that has the oracle property when the weights over the shrinkage parameters are controlled properly.

For high-dimensional cases, where the number of parameters exceeds the sample size, the sign consistency of the LASSO estimator was proved by Zhao and Yu (2006) and Meinshausen and Buhlmann (2006), respectively. For the SCAD, Fan and Li (2001) and Fan and Peng (2004) proved that the SCAD estimator achieves the oracle property for the case of a diverging number of parameters, and this result is extended to high-dimensional cases by Kim et al. (2008a).

Computational complexities should be considered in using sparse penalized methods. For the LASSO, Efron et al. (2004) developed the least angle regression (LARS) algorithm which can find the entire solution path of the LASSO estimator exactly.

A similar path-finding algorithm was proposed by Rosset and Zhu (2007) for the families of regularized problems that have the piecewise quadratic property. For generalized linear models, Kim et al. (2008b) suggested a gradient decent algorithm and Park and Hastie (2007) introduced an approximated path-finding algorithm using the idea of the LARS algorithm. For the SCAD, computational techniques are more involved since the SCAD penalty is non convex. Fan and Li (2001) suggested an iterative local quadratic approximation (LQA) algorithm to apply a modified Newton-Raphson algorithm. Kim et al. (2008a) and Wu and Liu (2009) proposed concave-convex procedure (CCCP) techniques to find an exact local minimizer of the SCAD penalized loss function, and Zou and Li (2008) introduced a local linear approximation algorithm and proved that their approximation is the tightest convex upper bound of the SCAD penalty function.

Recently, Wang and Leng (2007) proposed a method of least squares approximation (LSA) which provides a simple unified framework applicable to most LASSO estimations. The LSA estimator still possesses most of properties of the original LASSO estimator and can be calculated easily by adapting the LARS algorithm.

Theoretical properties of the sparse penalized approaches have been studied by many authors. For a finite number of parameters, Knight and Fu (2000) studied the properties of LASSO-type estimators. Fan and Li (2001) proved that there exists a local maximizer of the SCAD-penalized log-likelihood that achieves the oracle property. Here, the oracle property means that a penalized maximum likelihood estimator (MLE) is asymptotically equivalent to the oracle MLE that is an ideal non-penalized MLE obtained by deleting all irrelevant parameters in advance. Zou (2006) proposed the adaptive LASSO that achieves the oracle property by varying the weights on the tuning parameter.

Patrick Waldmann et.al, (2013) compared the statistical performance of two methods (the least absolute shrinkage and selection operator-LASSO and the elastic net) on two simulated datasets and one real dataset from a $50K$ genome-wide single nucleotide polymorphism(SNP) panel of 5,570 Fleckvieh bulls. They used cross validation to find the optimal value of regularization parameter λ with both minimum MSE and minimum $MSE + 1SE$ of minimum MSE. The optimal λ values were used for variable selection. Based on the first simulated data, they found that the $minMSE$ in general picked up too many SNPs. At min $MSE + 1SE$, the LASSO didn't acquire any false positives, but selected too few correct SNPs. The elastic net provided the best compromise between few false positives and many correct selections when the penalty weight α was around 0.1.

However, in their simulation setting, this α value didn't result in the lowest

$min_{MSE} + 1_{SE}$. The number of selected SNPs from the QTLMAS 2010 data was after correction for population structure 82 and 161 for the LASSO and the elastic net, respectively. In the Fleckvieh dataset after population structure correction lasso and the elastic net identified from 1291 to 1966 important SNPs for milk fat content, with major peaks on chromosomes 5, 14, 15, and 20. Hence, conclude that it is important to analyse GWAS data with both the lasso and the elastic net and an alternative tuning criterion to minimum MSE is needed for variable selection.

For a diverging number of parameters, Fan and Peng (2004) proved that the results of Fan and Li (2001) hold when the number of parameters is less than the sample size. Kim, Choi, and Oh (2008) studied the asymptotic properties of the SCAD-penalized least square estimator (LSE) in linear regression when the number of parameters exceeds the sample size. They proved that the oracle LSE asymptotically becomes a local minimizer of the SCAD-penalized residual sum of squares. They also proved that the oracle LSE asymptotically becomes the global minimizer of the SCAD-penalized residual sum of squares when the design matrix is non-singular. Zhao and Yu (2006) and Meinshausen and Buhlmann (2006) proved the sign consistency of the LASSO when the number of parameters exceeds the sample size. The sure independence screening method, a type of correlation learning, was proposed by Fan and Lv (2010) for ultrahigh-dimensional model selection problems. For a detailed overview of current research on variable selection in high-dimensional models, see Fan and Lv (2010).

In their article titled " Asymptotic oracle properties of SCAD-penalized least squares estimators", Jian Huang and Huiliang Xie (2007) studied the asymptotic properties of the SCAD-penalized least squares estimator in sparse, high-dimensional, linear regression models when the number of covariates may increase with the sample size. They were particularly interested in the use of this estima-

tor for simultaneous variable selection and estimation. They showed that under appropriate conditions, the SCAD-penalized least squares estimator is consistent for variable selection and that the estimators of non zero coefficients have the same asymptotic distribution as they would have if the zero coefficients were known in advance. Simulation studies indicate that this estimator performs well in terms of variable selection and estimation. What this study seeks to do is to compare the number of significant variables selected by each of the specified methods and propose the best amongst them based on the one that gives the least penalized cross-validated errors (PCVE).

2.6 Joint-GLM and Hierarchical Generalized Linear Models

In addition to variable selection, this thesis is interested in modelling crop yield in the three northern regions of Ghana using the recently developed random-effect models known as hierarchical Generalized Linear Models (HGLMs; Lee and Nelder, 1996; Lee, Nelder and Pawitan, 2006). Fan and Li (2002) proposed the penalized marginal likelihood method using the SCAD penalty function for gamma frailty model, and very recently Androulakis et al. (2012) extended it to other frailty distributions such as inverse Gaussian distribution. The models they considered are the shared models with only one frailty component, using frailty distributions which give an explicit marginal likelihood (Andersen et al., 1997). However, the marginal likelihood function of such models involves analytically intractable integrals when eliminating the frailties. The hierarchical likelihood (h-likelihood; Lee and Nelder, 1996) obviates the need for the marginalization over the frailty distribution and provides a statistically efficient procedure in various random-effect models such as hierarchical GLMs (HGLMs; Lee and Nelder, 1996; Lee, Nelder and Pawitan, 2006).

HGLMs consist of the three objects, namely the data, fixed unknown constants (parameters) and unobserved random variables (unobservable s). Traditional Bayesian models consist of the two objects, the data and unobservable s, while frequentist's (or Fisher's) models consist of the data and parameters. By allowing all three objects in the statistical modeling it is possible to describe various features in the data, for example, within-subject correlation in longitudinal studies, smooth spatial and temporal trends, function fittings, and factor analysis, heteroskedasticity, heavy-tailed distributions, robust modelling and sparse variable selections.

In the statistical literature, unobservable s appear with various names such as random effects, latent processes, factor, missing data, unobserved future observations, potential outcomes etc. Handling of such unobservable s is the key to new extended likelihood inferences. Lee and Nelder (1996, 2006) and Lee et al., (2006) have shown how to model and make inferences using the h-likelihood. Inferences about unobservables can be made without resorting to an empirical Bayes framework (Lee and Nelder, 2010). A single algorithm, iterative weighted least squares, can be used throughout all new models and requires neither prior distributions of parameters nor multi-dimensional quadrature. The h-likelihood plays a key role in the synthesis of the computational algorithms needed for broad class of new models.

The hierarchical generalized linear models is a synthesis of three widely used existing model classes; generalized linear models (McCullagh and Nelder, 1989), linear mixed models having both fixed and random effects (Longford, 1993) and models with structured dispersion as used in the analysis data from quality improvement experiments (Nelder and Lee, 1991, 1998). It uses the h-likelihood (Lee and Nelder, 1996) for inference about fixed and random effects given dispersion components and an adjusted profile h-likelihood for inference about dispersion

components given fixed and random effects. This method therefore leads to reliable and useful estimators. It shares properties with those derived from marginal likelihoods, while having the considerable advantage of not requiring the integrating out of random effects.

The algorithm for fitting these models can be reduced to the fitting of two-dimensional set of generalized linear models, one dimension been mean and dispersion, and the other been fixed and random effects, so that no special code is needed for the estimation of dispersion components. This formulation implies that, the models-checking techniques derived for generalized linear models (McCullagh and Nelder, 1989, chapter 12), can be carried over to the wider class. The hierarchical generalized linear models method does not require the use of prior probabilities. The model, it's fitting methods as well as theoretical justification are detailed in chapter three of this thesis.

Jiao H et.al (2005) in their article titled "Modelling local item dependence with the hierarchical generalized linear model", proposes a three-level hierarchical generalized linear model (HGLM) to model LID when LID is due to such contextual effects. The proposed three-level HGLM was examined by analysing simulated data sets and was compared with the Rasch-equivalent two-level HGLM that ignores such a nested structure of test items. The results demonstrated that the proposed model could capture LID and estimate its magnitude. Also, the two-level HGLM resulted in larger mean absolute differences between the true and the estimated item difficulties than those from the proposed three-level HGLM. Furthermore, it was demonstrated that the proposed three-level HGLM estimated the ability distribution variance unaffected by the LID magnitude, while the two-level HGLM with no LID consideration increasingly underestimated the ability variance as the LID magnitude increased. Noh et al., (2005) modelled heavy tailed distributions for random effects to take ascertainment into account in hu-

man QTL studies. Noh et al., (2006) used HGLM to minimize bias in heritability estimation for binary traits in human family data. HGLM has also been successfully applied in survival analysis with random effects (Noh et al., 2006).

2.7 Modelling Crop Yield

Statistical models, in which historical data on crop yields and weather are used to calibrate relatively simple regression equations, provide a common alternative to process-based models. Three main types of statistical approaches are found in the literature: those based purely on time series data from a single point or area (time series methods), those based solely on variations in space (cross-section methods) and those based on variations both in time and space (panel methods). Time-series models are generally believed to have the advantage of capturing the behaviour particular to the given area, whereas panel and cross-section methods must assume common parameter values for all locations, and cross-section methods in particular are prone to errors from omitted variables such as soil quality or fertilizer inputs that vary spatially. On the other hand, time-series models are often limited by data whereas panel and cross-section methods can aggregate data from multiple sites. A further discussion of the strengths and limits of particular methods in the context of predicting yield responses to climate change can be found in Lobell and Burke (2009).

The main advantages of statistical models are their limited reliance on field calibration data, and their transparent assessment of model uncertainties. For example, if a model does a poor job of representing crop yield responses to climate, this will be reflected in a low coefficient of determination (R^2) between modelled and observed quantities, as well as a large confidence interval around model coefficients and predictions. Although process-based models could in theory be accompanied with similar statistics, in practice they rarely are.

Statistical models are not without serious shortcomings, however, and in particular they are subject to problems of co-linearity between predictor variables (e.g., temperature and precipitation), assumptions of stationarity (e.g., that past relationships will hold in the future, even if management systems evolve), and low signal-to-noise ratios in yield or weather records in many locations.

An example of the co-linearity problem was highlighted by Sheehy et al. (2006) in response to the statistical models of Peng et al. (2004), which showed a 10 per cent decline of Philippine rice yields with a 1^0C increase in average minimum temperature (T_{min}). Sheehy et al. (2006) argued that solar radiation was a strong negative correlate of T_{min}, and thus an apparent negative effect of warming could easily arise from a positive effect of higher solar radiation. Similarly, Lobell and Ortiz-Monasterio (2007) showed that historical correlations between T_{min} and wheat yields in Mexico arose in part because of a negative correlation between solar radiation and T_{min}.

Despite the frequent caveats to results from statistical approaches (e.g., White, 2009), little work has been done to systematically evaluate their performance for predicting yield responses to climate. As their widespread use continues, it would be useful to know the specific conditions under which these models are most likely to mislead, and to quantify the errors incurred by adopting this convenient if imperfect approach. Moreover, because the aforementioned factors that challenge statistical approaches (e.g., co-linearity, signal-to-noise) will vary with scale, it is useful to evaluate statistical models at a range of different spatial scales.

As a step toward these goals, this dissertation evaluates the ability of statistical models to predict yield responses to some nine (9) variables and their interaction terms for nearly 790 sites in the three Northern regions of Ghana. Since the 'true' yield responses are unknown, we invoke the 'perfect model' approach whereby a

statistical model is tested for its ability to recreate the underlying relationships between the factors used and yields. Such a study is very relevant in our country Ghana since Agriculture is our main backbone.

The World Bank in 2009 presents Ghana's GDP shares of agriculture, industry, manufacturing (as part of industry) and services between 1965 and 2008. The report indicates that before the late 1980s when the economy growth rate was negative, agricultural growth rate, which was also negative, was less negative than the other sectors in the economy. Thus, our GDP share of agriculture rose in this period and peaked at 60 per cent in a few years in the late 1970s and early 1980s. When growth started to recover and turned into positive after 1983, the non-agricultural sector needed more recovery time as it declined more in the previous period. While growth in the agricultural sector also turned to become positive, its share in GDP fell back to its level in the 1960s immediately after the independence.

Improved understanding of the potential effects of climate change on crop yields is central to planning appropriate and timely responses. Analysts wishing to anticipate these effects must inevitably rely on some conceptual or numerical model of how crop yields respond to climate. A widely used approach to this prediction problem is to rely on numerical models that emulate the main processes of crop growth and development. These process-based models are typically developed and tested using experimental trials and thus offer the distinct advantage of leveraging decades of research on crop physiology and reproduction, agronomy, and soil science, among other disciplines. Yet these models also require extensive input data on cultivar, management, and soil conditions that are unavailable in many parts of the world.

More significantly, even in the presence of such data these models can be very

difficult to calibrate because of a large numbers of uncertain parameters. Often this parameter uncertainty is ignored and a subjective decision is made to proceed with a single set of parameter values that produces acceptable agreement with observations. When uncertainties in parameter values are explicitly considered, however, the uncertainty estimates for model projections can widen substantially.

For example, Iizumi et al. (2009) and Tao et al. (2009) describe efforts to estimate distributions of parameter values for a simplified process-based model from data on yields of rice and maize, respectively. Both studies employed a Markov Chain Monte Carlo technique to retrieve parameter distributions, with the width of these distributions reflecting the inability of historical datasets to completely constrain parameter values. Parameter uncertainties then translated to large uncertainties in projecting responses to climate change, particularly for future scenarios that exceeded those in the calibration period (Iizumi et al., 2009).

This is a common technique, for instance, in climate modelling studies where one model is used as 'observations' and the others are tested for their ability to reproduce observations (Murphy et al.,2004; Tebaldi and Knutti, 2007). This approach does not rely on the 'perfect model' actually being perfect (which no model is), but rather tests the ability of a given model and calibration technique to recreate the behaviour of a reference model. In this case, they used the well-established and widely used process-based model CERES-Maize as our 'perfect model' to simulate historical yields, and then fit statistical regressions to the simulated data. They then evaluated the performance of the statistical models for different sites, level of spatial aggregation, and number of years used to calibrate the model.

Food security is a crucial issue in sub-Saharan Africa as a consequence of unreliable rainfall, marginal soil fertility and a low level of inputs leading to declining crop yields. As a case study, Braimoh and Vlek, 2005 investigated the most

important variables affecting maize yield in northern Ghana. They combined a soil quality index on a continuous scale with a social data set to model maize yield using linear multiple regression. Five significant variables were identified ($P < 0.05$): soil quality index, fertilizer use, household size, distance from main market, and the interaction between fallow length and soil quality index. The effect of the interaction between soil quality and fallow on maize yield is negative, suggesting the influence of litter quality and N immobilization in the soils. Their conclusion was that, Research and policy should focus on the development of site-specific, legume-based cropping, and the integration of crop and livestock farming in Northern Ghana and similar areas in sub-Saharan Africa.

Evenson and Mwabu in 2001 studied land productivity effects of the training and visit (T and V) systems of agricultural extension in Kanya, taking into account other determinants of crop yields such as the schooling of farmers and characteristics of agro-ecology. The T and V system was incorporated into Kanya's system of agricultural extension in 1982 as a strategy for raising farm yields. The data they used evaluate the performance were collected by the government of Kanya in 1982 and 1990, but the estimation results reported in their paper were based primarily on the 1982 data set. The sample used for estimation contains information about crop production, agricultural extension workers, educational attainment of farmers, usage of farm inputs among others. A quantile Regression technique was used to investigate productivity effects of agricultural extension and other farm inputs over the entire conditional distribution of farm yield residuals.

Their results showed that, productivity effect of agricultural extension is highest for farmers at the extreme ends of distribution of yield residuals. Complementary of unobserved farmer ability with extension service at higher yield residuals and the diminishing returns to the extension inputs, which are uncompensated for by the ability at the lower tail of the distribution, are hypothesized to account

for this U-shaped pattern of the extension effect. Their findings suggests that for a given level of extension input, unobserved factors such as farm management abilities, affect crop yield differently. Effects of schooling on farm yields are positive but statistically insignificant. Other determinants of farm yields they analysed included labour input, farmers experience, agro-ecological characteristics of farms, fallow acreage, and type of crop grown (Evenson and Mwabu, 2001).

Lobell and Burke in 2010 studied the use of statistical models to predict crop yield responses to climate change. In their study, predicting the potential effects of climate change on crop yields requires a model of how crops respond to weather. As predictions from different models often disagree, understanding the sources of this divergence is central to building a more robust picture of climate changes likely impacts. They used a perfect model approach to examine the ability of statistical models to predict yield responses to changes in mean temperature and precipitation, as simulated by a process-based crop model. The CERES-Maize model was first used to simulate historical maize yield variability at nearly 200 sites in Sub-Saharan Africa, as well as the impacts of hypothetical future scenarios of 2^0C warming and 20 per cent precipitation reduction. Statistical models of three types (time series, panel, and cross-sectional models) were then trained on the simulated historical variability and used to predict the responses to the future climate changes.

The agreement between the process-based and statistical models' predictions was then assessed as a measure of how well statistical models can capture crop responses to warming or precipitation changes. The performance of statistical models differed by climate variable and spatial scale, with time-series statistical models ably reproducing site-specific yield response to precipitation change, but performing less well for temperature responses. In contrast, statistical models that relied on information from multiple sites, namely panel and cross-sectional

models, were better at predicting responses to temperature change than precipitation change.

The models based on multiple sites were also much less sensitive to the length of historical period used for training. For all three statistical approaches, the performance improved when individual sites were first aggregated to country-level averages. Results suggest that statistical models, as compared to CERES-Maize, represent a useful, even if imperfect, a tool for projecting future yield responses, with their usefulness higher at broader spatial scales. It is also at these broader scales that climate projections are most available and reliable, and therefore statistical models are likely to continue to play an important role in anticipating future impacts of climate change (Lobell and Burke, 2010).

The Alliance for a Green Revolution in Africa (AGRA) with the primary goal of easing the flow of produce from the farm-gate to the market by linking smallholder farmers to commercial buyers and processors (FtM Grant Narrative Report, 2011) is one of the key agencies presently undertaking broad range of activities, including the provision of education and advisory services to farmers, expansion of farmer institutions, and development of agribusiness and improving market linkages, aimed at raising the productivity of Ghanaian farmers. Their primary goal it to ease the flow of produce from the farm-gate to the market by linking smallholder farmers to commercial buyers and processors (FtM Grant Narrative Report, 2011). However, recent analysis show that for a majority of staple crops, agricultural productivity is decreasing and any output gains if sub-Saharan Africa are attributed primarily to the expansion of cultivated land (Kraybill, Bashsaasha, and Betz, 2009). These practices have contributed to Ghana to having one of the highest rates of soil depletion in all of Sub-Saharan Africa.

Improved farming technologies such as high yield crop varieties, chemical fertilizers, and irrigation techniques have been central in raising yields in other parts of the world; however, African farmers have been much slower in adopting these new methods. One reason that farmers cite for not adopting the new technologies is a lack of information regarding how to apply the improved inputs (Morris, Kelly, Kopicki, and Byerlee, 2007). In many cases if the improved inputs are not applied correctly yields will be lower than traditional crop varieties, leading farmers to abandon the new technologies. Consequently, access to reliable information is an integral part in any farmer's ability to raise productivity. Information about improved methods or new technologies come through a variety of mechanisms such as formal government extension, mass media such as radio, and as often is the case, through other farmers.

Agricultural extension is the primary mechanism that developing country governments use to assist farmers in expanding their ability to adopt and implement new methods and to relay information concerning new technologies. Throughout Africa extension programs have the reputation of being largely ineffective (Dejene, 1989; Gautam, 2000), adding very little to the productivity of farmers. This reputation is no exception in Ghana.

Previous studies have investigated the relationship between agricultural extension and productivity with varying results. Birkhaeuser, Evenson, and Feder (1991) review 26 studies that use linear regression to determine the relationship between extension contact and farm productivity, with only 11 statistically significant at the 90 per cent level. Evenson (1997) points out that because of large variation in program design and field worker skill it is not feasible to make broad generalizations about the economic contribution of agricultural extension.

Birkhaeuser, Evenson, and Feder (1991) also point out two major difficulties of

including extension variables in the estimation of agricultural production functions. First, most studies use a farm-level extension contact variable that does not account for knowledge spill overs occurring when farmers talk to each other and exchange information. In this case a farmer that has not been visited by an extension agent, but has obtained the same potentially output increasing information from a neighbour, has received the treatment without any statistical accounting of it, biasing the results upward.

The second difficulty with using a farm-level extension variable is that there is possible endogeneity within the farmer-extension worker interactions. That is, more productive farmers may have some unobservable quality, such as a desire for the best farming methods, which would also lead them to seek out extension agents. Owens, Hoddinott, and Kinsey (2003) control for the endogeneity of the extension variable by including farm plot characteristics, location dummies, and a variable representing farmer ability into the regression equation.

This thesis attempts to correct for both the endogeneity and spill over effects by including control variables for farmer ability and information exchange between farmers. Another relevant question with respect to agricultural extension in Ghana is whether the farmer-extension worker interaction has differential effects on farms of different size. As is the case in most developing countries, the Ghanaian government can only devote limited resources to agricultural extension programs and so most programs are only administered to a limited proportion of the population. Because there is significant variation of farm size throughout the three northern regions and Ghana as a whole, and likely significant variation in the determinants of output for different sized farms, it is critical for all stakeholders, the academia and the general public to understand which support services and policies will benefit farms and improve crop yield for the different farmer based organizations of different farm sizes.

Past research has found relationships between farm size and factors of production and also farm size and output. Larger farms are more likely to use advanced farming inputs such as fertilizer, irrigation, and improved seed varieties (Feder, Just, and Zilberman, 1985) when compared to smaller farms. This has led many agricultural programs to solely target larger, more sophisticated farms that are viewed as better equipped to make use of additional resources.

Conversely, a vast literature exists showing an inverse relationship between land productivity and farm size (Sen 1962; Berry and Cline 1979; Rosenzweig and Binswanger, 1993) suggesting that smaller farms are more productive and would be better targets of available resources. It may prove advantageous for the Ministry of Agriculture to provide assistance to farms of all sizes simultaneously, in which case it is important to understand how extension enters the production technology of various farm sizes differently.

This study further examines the relationship between farm size and crop type, with particular attention given to six support services.

Chapter 3

Methodology

3.1 Introduction

In this chapter, a review the concept of regularization in statistics and penalized methods, such as Lasso, SCAD and the H-likelihood are presented. Also, a review of existing methods for fixed effects selection such as GLMs and the proposed Joint GLM method are presented. For both fixed and random effects modelling, the HGLM method is proposed and well discussed in this chapter.

3.2 The Concept of Regularization in Statistics

The concept of penalization was first introduced in the context of solving integral equation numerically by Tikhonov (1943). As is well known, if $f \in L_2(\mathbb{R})$ and $K(x, y)$ is a smooth kernel, the range of the operator A, $R(A)$, $A : L_2(\mathbb{R}) \to L_2(\mathbb{R})$ with $(Af)(y) \equiv \int K(x, y)f(x)dx$ is dense in $L_2(\mathbb{R})$ but not onto. Thus, the inverse A^{-1} is ill-posed. The solution to the equation

$$Af = g \tag{3.1}$$

is hard to determine since approximations to g easily lie outside $R(A)$. Tikhonov's solution was to replace 3.1 by the minimization of

$$\|Af - g\|^2 + \gamma W(f)$$

, where the Tikhonov's factor $\gamma > 0$ is a regularization parameter and $W(f)$ is a smoothness penalty such as $\int [f'(x)]^2 dx$. Numerical (finite dimensional) approx-

imations to this problem are much stable. Note that unless $\gamma = 0$, the solution will not satisfy (3.1).

There has been an enormous amount of work in statistics dealing with regularization in a wide spectrum of problems. An exhaustive survey is beyond the scope of this dissertation. We therefore present a unifying view encompassing more recent developments. The main features of most current data are both size and complexity. The size may permit us to non-parametrically estimate quantities which are 'unstable' and 'discontinuous' functions of the underlying distribution of the data, with the density being a typical example.

Complexity of the data, which usually corresponds to high dimensionality of observations, makes us attempt more and more complex models to fit the data. The fitting of models with a large number of parameters is also inherently unstable (Breiman, 1996). Both of these features, force us to regularize in order to get sensible procedures. For recent discussions of these issues from different points of view, see Donoho (2000) and Fan and Li (2006). We will consider only the asymptotic of regularization and only in the simplest context, i.i.d samples of size n of p dimensional vectors. The main issues are already quite clear in this context.

Loosely, regularization is the class of methods needed to modify maximum likelihood to give reasonable answers in unstable situations. There are also a number of generic issues that will arise such as the reasons for choosing particular forms of regularization, how to determine the analogue of the Tikhonov factor γ which, as we shall see, is somewhat driven by our particular statistical goals, and last but not least, computational issues which are also critical nowadays.

3.2.1 Variable selection

In statistics, the first instance of this type of problem arose in the context of multiple linear regression with continuous predictor variables, when the number of predictor variables is larger than the sample size. Suppose we observe an i.i.d sample $(Z_i, Y_i),\ i = 1, \ldots, n$, where $Z_i = (Z_i^1, \ldots, Z_i^p)$. The resulting model is

$$Y_i = Z_i^T \beta + \varepsilon_i \tag{3.2}$$

Where $\varepsilon_i,\ i = 1, \ldots, n$ are i.i.d $N(0, \sigma^2)$. In the case of $p > n$, the usual least squares equations 'over-fit'. All observations are predicted perfectly, but there are many solutions to the coefficients of the fit and new observations become not uniquely predictable.

The classical solution to this problem was to try to reduce the number of variables by processes such as forward and backward regression with reduction in variables determined by hypothesis tests, see Draper and Smith (1998), for example. An alternative strategy that emerged (Hoerl and Kennard, 1970) was ridge regression, adding to the residual sum of squares $\sum_{i=1}^{n}(Y_i - Z_i^T)^2$ a plenty, $\lambda \sum_{j=1}^{p} \beta_j^2$, which now yields a unique solution. These methods, often actually have two aims; to construct a good predictor (the values of coefficients in the regression are then irrelevant - goal 1) and to give causal interpretations of the factors and determine which variables are 'important' (goal 2).

Regularization is important for both aims. But, as we shall see, the appropriate magnitude of the regularization parameter (tuning parameter) may be governed by which aim is more important. Goal 1 is the one which is primary in machine learning theory. The model postulated is non-parametric,

$$Y = m(Z) + \varepsilon \tag{3.3}$$

Where $E(\varepsilon/Z) = 0$ and m is essentially unknown. A fundamental approach is to consider a family of basis functions $g_j(Z),\ j = 1, 2, \ldots,$ such that m is arbitrarily well approximated in, for instance, the L_2 sense, $\inf_\beta E(m(Z) - \sum_{j=1}^p \beta_j g_j(Z))^2 \to \infty$ as $p \to \infty$, where $\beta = (\beta_1, \ldots, \beta_p)^T$. A parametric model postulation with $g_j(Z) = Z^(j),\ j = 1, \ldots, p,$ corresponds to the linear models specification. Then, since, as we have seen, minimizing $\sum_{i=1}^n (Y_i - \sum_{j=i}^p \beta_j g_j(Z_i))^2$ is unreasonable for $p \ll n$, it is consistent with the penalty point of view to minimize

$$\sum_{i-1}^{n} (Y_i - \sum_{j=i}^{p} \beta_j g_j(Z_i))^2 + \gamma Pen(\beta) \tag{3.4}$$

The regression choice of $Pen(\beta) = \sum_{j=1}^p \beta_j^2$ is not nowadays the one attracting the greatest attention theoretically, but the LASSO, $Pen(\beta) = \sum_{j=1}^p |\beta_j^2|$ (Tibshirani, 1996) is being studied extensively. This stems from the idea that, at least to a high degree of approximation, most $|\beta_j^2|$ in the best representation of $m(Z)$ as a linear combination of p basis elements $g_j(Z)$ in the L_2 sense are 0. That is, the representation is "sparse" in the sense of Donoho and Johnstone (1998). Then the "natural" penalty is

$$Pen(\beta) = \sum_{j=1}^{p} 1(|\beta_j| > 0) \tag{3.5}$$

An unpleasant function of β. Evidently, $\sum_{j=1}^p |\beta_j|$ is the closest convex member of the family of penalties $\sum_{j=1}^p |\beta_j|^\alpha,\ \alpha > 0$ to 3.5.

Minimizing subject to penalty 3.5 may also be seen as selecting a model including the variables with $\beta_j \neq 0$, following Goal (2). This approach and its generalization to generalized linear and other models as well as related penalties has been developed by Fan and coworkers and others, see Fan and Li (2001), Fan and Peng (2004), Fan and Li (2006) and Zou and Hastie (2005). Note that, at least implicitly, this point of view implies that we believe a meaningful (sparse)

representation in basis functions g_j.

$$m(Z) = \sum_{j=1}^{p^*} \beta_j g_j(Z) \tag{3.6}$$

is true for some $p^* \ll p$.

Penalization is far from the only form of regularization that has arisen in statistics. In the context of density estimation, binning in histograms is the oldest method, and kernel methods were proposed by Rosenblatt (1956) and Parzen (1962). In turn these methods led to Nadaraya-Watson estimation (Nadaraya, 1964; Watson, 1964) in non-parametric regression. There are also methods which have appeared outside non-parametric regression contexts, where formulations such as semi parametric or generalized linear models do not capture the necessary structure.

Throughout this dissertation, we limit ourselves to the case where our observations $X_1, \ldots, X_n$ are i.i.d, taking values in a space χ, typically $\mathbb{R}^p$. We assume that their common distribution $P \in \mathcal{P}$, our model, which through most of our discussion, we assume is non parametric, effectively all P, although we can and shall impose smoothness or other general properties on the members of $\mathcal{P}$. We let P_n denote the empirical distribution, placing mass n^{-1} at each observation.

For our treatment of covariance estimation it may be convenient to think of $X = (X_1, X_2, \ldots, X_p, \ldots)^T$, as a stochastic process for which we have data of size n on the first p coordinates, and of the unknown P as living on $\mathbb{R}^\infty$. However, we will only be interested in estimating the covariance matrix of these first p coordinates.

Most statistical activities centre around estimation or testing hypotheses or putting

confidence regions on parameters, which we define as functions $\theta(P)$, mapping $\mathcal{P}$ into Θ. Θ is not necessarily just $\mathbb{R}$ or a Euclidean space. We shall limit ourselves almost exclusively to function valued parameters.

For instance, suppose $P \in \mathcal{P}$ are characterized as having densities $f(.)$, which are continuous. Then $\theta(P) = f(.)$ is a parameter. If P is the joint distribution of (Z, Y), then $\theta(P) = E(Y|Z = .)$, the regression function is a parameter. It will also be convenient to think of parameters which themselves vary with n and p, $\theta^{(n,p)}(P)$. Thus, the covariance matrix $\sum$ of $(X_1, \ldots, X_p)^T$, which we are interested in studying is, $\theta^{(p)}(P)$ if we think of our observation as being $(X_1, X_2, \ldots)^T$.

Similarly, the extreme percentile of the distribution of $X \in \mathbb{R}$, $F^{-1}(1)$ where F is the empirical distribution function of X, typically equals ∞ and cannot be estimated, but $F^{-1}(1 - \frac{1}{n})$, the quantile corresponding to the maximum of $X_1, \ldots, X_n$ can. We will usually suppress such dependence on p and n. Any estimate $\hat{\theta}(X_1, \ldots, X_n)$ of $\theta(P)$ may, by sufficiency of the P_n, be thought of as a function $\theta_n(P_n)$, where the domain of θ_n is at least the possible empirical distributions and typically includes at least all finite discrete distributions on χ. The least we can require of an estimate (really a sequence of estimates) is consistency:

$$\rho(\hat{\theta}, \theta(P)) \xrightarrow{P} 0 \tag{3.7}$$

where ρ is Euclidean distance if Θ is Euclidean and ρ is a suitably defined metric, e.g., the L_2 distance, if Θ is a function space.

If $\mathcal{P}$ contains all discrete distribution, then the natural thing to use as an estimate of $\theta(P)$ is the "plug-in" estimate $\theta(P_n)$. For instance, if $\chi \in \mathbb{R}$, and $\theta(P)$ is the mean, which we represent as $\theta(P) = \int x dP(x)$, then $\theta(P_n) = \int x dP_n(x) = \bar{X}$, the sample mean. If $\theta(P) = F(.)$, where $F(x) = P(X \leq x)$, the cdf of X, then $\theta(P_n)$ is the empirical cdf,$\theta(P_n) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$. Consistency for plug-in estimates

follows if

1. θ is continuous in ϱ, for a given metric ϱ on $\mathcal{P}$

2. P_n is consistent with respect to ϱ. That is, $\varrho(P_n, P) \xrightarrow{P} 0$ if P is true.

In the usual situations, where Θ is Euclidean, $\theta \mapsto (., \theta)$ is smoothly invertible, and $\theta(P_n)$ makes sense, consistency holds. But, consider the situation we have discussed, $\theta(P) = f(.)$. Now the density. $\theta(P_n)$ doesn't make sense, since the discrete distributions do not belong to P. What is done, in this case, and implicitly in all such situations we know about is regularization. A generic regularization process is summarized as,

1. A sequence of approximations.

   (i) We construct a sequence θ_k defined on $\mathcal{P}$ and the discrete distributions, say on $\mathcal{M}$ such that $\rho(\theta_k(P), \theta(P)) \to 0$, that is, $\theta_k(P) \to \theta(P)$, or more generally $\varrho(\theta_k(P), \theta^{(n,p)}(P)) \to 0$ as $k, n, p \to \infty$, for each $P \in \mathcal{P}$.

   (ii) $\theta_k(P_n) \xrightarrow{P} \theta_k(P)$ for all k.

2. Selection of approximations. We select a data determined value $\hat{k}_n(X_1, \ldots, X_n)$ and uses as estimate, $\theta_{\hat{k}_n}(P_n)$.

That is, we approximate $\theta(P)$ by a 'nice', call it regular, parameter θ_k which can be estimated by plug-in and then determine how fine an approximation we will use. Of course, k need not be an integer, but could be a continuous parameter such as the bandwidth. It is often useful to decompose the

$$\theta_k(P_n) - \theta(P) = [\theta_k(P_n) - \theta_k(P)] + [\theta_k(P) - \theta(P)] \tag{3.8}$$

The first term is naturally identified with variance, the second with bias, and the choice of k is the choice of best balance between the two. In this review, we necessarily mention only a small subset of the many ways the approximations have been chosen, but do stress the importance of the choice of k in many instances.

3.2.2 Sequence Approximation

We return to model 3.1, which could equally well be written that we observe (Z, Y) with a completely unknown joint distribution (subject possibly to moment and smoothness conditions). Our goal is estimation in the $L_2(P)$ sense of the function valued parameter $\theta(P) = m(.) = E(Y|Z = .)$. This goal makes sense if we wish, knowing P, to predict a new Y given a new Z. If we use the predictor $\delta(Z)$, our loss is

$$\ell(P, \delta(Z)) = \int (y - \delta(Z))^2 dP(z, y) \tag{3.9}$$

The best choice of $\delta(Z)$ if, of course, $m(Z)$. Since we don't know P, we must use our "training sample" $(X_1, \ldots, X_n)$ to construct $\hat{\delta}(Z; X_1, \ldots, X_n)$. Since $m(Z)$ cannot be estimated by plug-in if Z is continuous, we need to apply regularization. The first step is to select a sequence of approximation $\theta_k(P)$ which are meaningful if $P = P_n$.

As we mentioned, there are many ways of selecting the sequence $\theta_k(P) = m_k(.)$, penalization as in (3.4), see, for instance, Zhang et al.(2004), or in a more structured way, sometimes referred to as the method of sieves, which we now explain. We consider the models $\mathcal{P}_k = P : m(Z) = \sum_{j=1}^{k} \beta_j g_j(Z)$ for some β, and define an estimate appropriate to the parametric model $\mathcal{P}_k$. Least squares is the natural choice here. Compute $\hat{\beta}_k$, the least squares estimate and $\hat{m}_k(z) = \hat{\beta}_k^T g(Z)$, where $g(Z) = (g_1(Z), \ldots g_k(Z))^T$. The corresponding population $m_k(.)$ is just $\sum_{j=1}^{k} \beta_j g_j(Z)$, where $\beta = (\beta_1, \ldots, \beta_k)^T = \arg\min_\beta \int (\int (y - \delta(Z))^2 dP(z, y))$.

3.3 Choice of regularization parameter

We want to select $\hat{k} = k(P_n)$, which is optimal in terms of our loss function,

$$R(P, \delta) = E_p(Y - \delta(Z; X_1, \ldots, X_n))^2 \tag{3.10}$$

the expected squared error integrated out with respect to Z and $(X_1, \ldots, X_n)$. And so our first goal is consistency, $R(P, \hat{m}_k(.)) \to R(P, m(.))$. It is easy to see that, by orthogonality, this is equivalent to $\int (\hat{m}_k(z) - m(z))^2 dP(z) \xrightarrow{P} 0$. This is equivalent to choose ρ to be $L_2(P)$ distance in the range of $\theta(P)(.)$, which we identify as all square integrable functions of Z. Consistency corresponds to what we have called Goal (I).

As a concrete example, suppose that we believe that $\mathcal{P}_k$ is correct for some k, and our goal is to find the correct model or smallest correct model if the P_k are nested, as in our case, and then estimate β. The type (I) goal formulation leads, after construction of an unbiased estimator of the $MSE_k = E(\hat{m}_k(Z) - m(Z)^2)$, where $m(z)$ is the true population parameter, to a solution due to Akaike (1970), Mallows (1973) and others, "choose $\hat{k}$ to minimize $\sum_{i=1}^{n}(Y_i - \hat{m}_k(Z_i))^2 + 2k$. This choice comes from the representation

$$E(Y_i^0 - \hat{m}_k(Z_i))^2 = E(Y_i - \hat{m}_k(Z_i))^2 + 2Cov(\hat{m}_k(Z_i), Y_i) \tag{3.11}$$

where $Y_i^0 = m(Z_i) + \varepsilon_i^0$ is a new independent observation, and

$$2\sum_{i=1}^{n} Cov(\hat{m}_k(Z_i), Y_i) = k$$

under the normality assumption on ε, see Efron (2004) for more details. On the other hand, pursuit of the type (II) goal puts great importance on identifying $k_0(P) = \min(k : P \in \mathcal{P}_k)$, the smallest model containing P first and then estimating β for purposes of interpretation. A Bayesian argument (Schwarz, 1978) to choose k by maximizing the posterior probability of $\mathcal{P}_k$ leads to the penalty $k \log n$ which evidently leads to much lower values of $\hat{k}$.

The Akaike/Mallows criterion does choose a model which is "correct" but not of smallest size. Readers are referred to Shao (1997) for more discussion on this

issue. When p is allowed to increase with n, Bunea et al. (2006) show that consistent variable selection can also be achieved via multiple testing. Much more general choices of k involving types of cross validation are given later in this section.

3.3.1 Selection of regularization parameter via cross validation

Cross-validation is a method that uses part of data to fit the model and the rest part to test the performance of the fitted model. Cross-validation and bootstrapping are the two classes of resampling methods currently recommended for prediction error measures for variable selection. Cross-validation procedures partition the data into two disjoint sets. The model is fit with one set (the training set), which is subsequently used to predict the responses for the observations in the second set (assessment set). Bootstrap procedures form many samples of the original data by sampling with replacement. Details of Cross-validation procedures and their application to the variable selection problem are outlined below.

We have touched several ways to select γ(or k) in our previous discussion. We now address cross validation, as a most general model selection rule. An extensive review of model selection has been given by Wang (2004). Shao (1997) provided an interesting taxonomy of various model selection schemes in linear regression context. A general approach is leave one out cross validation.

1. Leave-one-out cross validation.

   Let $X_{(-i)} = X_j : j \neq i$ and consider the predictor of Y_i, $\hat{m}_\gamma^{(-1)}(Z_i)$, trained from $X_{(-i)}$ by penalizing with $\gamma Pen(\beta)$. Then the cross validation estimate of error is just

   $$CV(\gamma) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_\gamma^{(-1)}(Z_i))^2 \tag{3.12}$$

   The "optimal" $\hat{\gamma}$ is defined as giving the smallest cross validation error.

The motivation here is reasonably clear and goes back to the work of Stone (1974). $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_\gamma^{(-1)}(Z_i))^2$ is an unbiased estimate of the actual risk of $\hat{m}_\gamma^{(-1)}(Z_i)$ which we expect is very close to that of $\hat{m}_\gamma(X_1, \ldots, X_n; Z_{n+1}) = \hat{m}_\gamma(Z_{n+1})$ for which we want to compute $E(Y_{n+1} - \hat{m}_\gamma(Z_{n+1}))^2$. For a linear estimator $(\hat{m}_\gamma(X_1), \ldots, \hat{m}_\gamma(X_n))^T = H(\gamma)(Y_1, \ldots, Y_n)^T$.

Shao (1993) proves with asymptotic results and simulations that the model with the minimum value for the leave-one-out cross-validation estimate of prediction error is often over specified. That is, too many insignificant variables are contained in set β_1. He recommends using a method that leaves out a subset of observations, called n-fold cross-validation.

2. Generalized cross validation: Generalized cross-validation minimizing

$$GCV(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \hat{m}_\gamma(Z_i))^2}{(1 - tr(H(\gamma))/n)^2} \tag{3.13}$$

was proposed by Craven and Wahba (1979) for computational reasons, as an approximation to leave-one-out cross validation, since the computation of $\hat{m}_\gamma^{(-1)}(i = 1, \ldots, n)$ multiplies computation time by a factor of n.

Efron (2004) showed that all the methods we have discussed in this section so far correspond to the estimation of the expected optimism,

$$E(Y_i^0 - \hat{m}_\gamma(Z_i))^2 - E(Y_i - \hat{m}_\gamma(Z_i))^2 \tag{3.14}$$

in an approximately unbiased fashion. Using a Rao-Blackwell type argument, he further showed that the model-based penalty methods (C_p, AIC, SURE) outperformed the non-parametric methods such as leave 1 out CV, assuming the model is believable. They also gave similar connections between parametric and non-parametric bootstrapping methods.

Again, For linear models, we have

$$GCV = \frac{RSS}{N}\frac{1}{(1-p/N)^2} \tag{3.15}$$

By Taylor expansion,

$$GCV \approx \frac{RSS}{N} + 2\hat{\sigma}\frac{p}{N} \tag{3.16}$$

Since $\frac{RSS}{N} \to \sigma^2$ as $N \to \infty$, GCV yields the same result as AIC and Mallow's C_p asymptotically.

The extent to which the use of CV and GCV yield procedures satisfying our optimality criteria has been studied (Li, 1985, 1986, 1987). Birge and Massart (1997) showed that leave one out cross validation is equivalent to Mallows C_p in regression, making it optimal for nested models but selecting too large a model if all 2^p sub-models are considered.

3. V-fold cross validation.

   In fact, few of these methods for selecting have been used in machine learning practice. The standard approach is to choose V dividing n, divide the sample into V disjoint parts of size $m = n/V$ say $\Psi^{(1)}, \ldots, \Psi^{(V)}$, and then use the $n-m$ observations in $V-1$ of the parts to calculate $\hat{m}_\gamma(\Psi^{(-l)})$ and evaluate

$$Q_t(\gamma) = \frac{1}{m}\sum_{j \in \Psi^{(t)}(\hat{m}_\gamma, t)} (Z_j - Y_j)^2 \tag{3.17}$$

   an unbiased estimate of the risk of the prediction based on n - m observations. Then, although looking at more than a single partition is not necessary for theory, form $Q(\gamma) = \frac{1}{V}\sum_{t=1}^{V} Q_t(\gamma)$, and choose $\hat{\gamma}$ by minimizing $Q(\gamma)$. Leave 1 out CV is also of this form with $V = n$. However, taking, say, $V = \frac{n}{\Omega(n)}$, where $\Omega(n)$ is slowly varying, can be shown to work very generally to establish both oracle and minimax results, see Gyorfi et al. (2002), Bickel et al. (2006). Some further discussion is in Dudoit and

Van der Veen (2005).

A great advantage of both leave 1 out CV and V-fold CV is that they immediately generalize to any prediction question, such as generalized linear model prediction as in Fan and Li (2006), or more general model selection. V-fold cross validation is closely related to the m out of n bootstrap and sub-sampling.But that is not within the scope of this thesis. This discussion of the choice of γ in classification has been entirely in the context of Goal (I). When we turn to Goal (II), in which we assume there is a true model $\mathcal{P}_k$, the situation is different. If we choose via BIC, or in more complex situations, the closely related Bayesian, MDL criterion of Rissanen (1984), we can obtain the true k with probability tending to 1 and thus safely act as if $\hat{k}$ gave us the true model. On the other hand, as we have noted previously, AIC and the Goal (I) oriented criteria end up picking models that are larger than necessary.

Precisely in this thesis, the researcher suggests using a repeated 10-fold CV. A repeated 10-fold CV consisting of 100 runs of the 10-fold CV procedure with different random splits into 10 disjoint groups. It is a balanced version of LMO-CV, since every object is used exactly 100 times for assessing the candidate model. The study introduced the PCVE (Penalized Cross Validated Errors) which is based on a repeated 10-fold CV. This PCVE is used for the comparison of different penalized methods of variable selection and was found to perform better than LOO-CV. The parameter n of the repeated n-fold CV was set to n=10 in this study. This corresponds to leaving out 10 per cent of the data during cross-validation. The number of repetitions was set to 100.

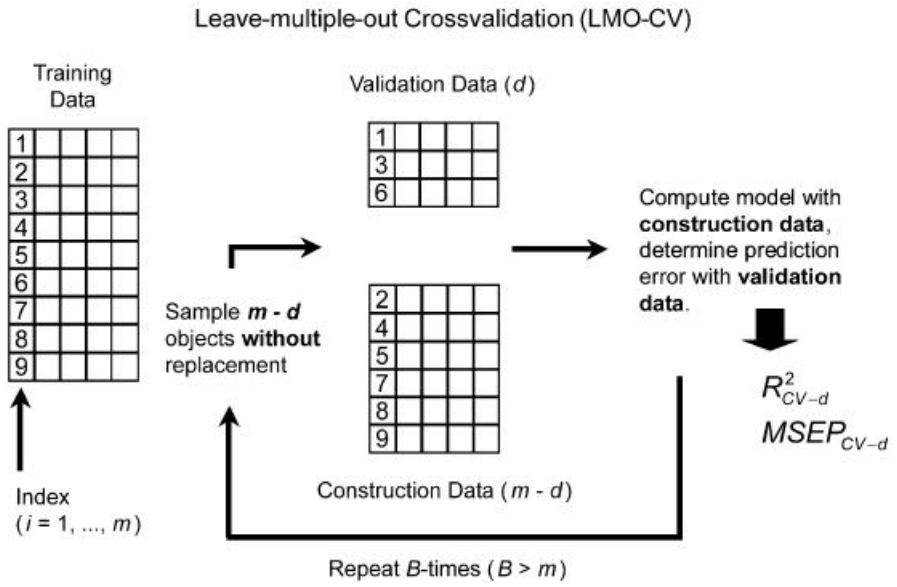

Figure 3.1: Schematic illustration of leave-multiple-out cross-validation. m: number of objects; d: number of objects left out; B: number of splits into construction and validation data; R^2_{CV-d}: leave-d-out cross-validated squared correlation coefficient; and, $MSEP_{CV-d}$: leave-d-out cross-validated mean squared error of prediction.(Flow chart of the proposed Repeated n-fold Cross validation)

3.4 Computational Methods for Penalized Variable selection

The main idea of penalized approaches is to impose a certain type of penalty to the regression coefficients, such that they are shrunk towards zeros, and some small coefficients will become exactly zero, achieving the purpose of variable selection. As a continuous variable selection procedure, the penalized method gives better prediction and small variance than traditional searching methods when the model is properly tuned. In general, the penalized likelihood function is of the form $-2\log(likelihood) + P_\lambda(\beta)$ where $P_\lambda(\beta)$ is the penalty and $\lambda \geq 0$. As λ increases, the penalty term also increases and and impose more shrinkage on the coefficients.

3.4.1 Least Absolute Shrinkage and Selection Operator (LASSO)

We consider the setting where we have observed data $(y_1, x_1), \ldots, (y_n, x_n)$ with each y_i a realisation of a scalar random variable Y_i, and each $x_i = (x_{i1}, \ldots, x_{ip})^T$ a p-vector of explanatory variables. Let X be a matrix whose ith row is given by x_i^T. Without loss of generality, the study shall require that the columns of X are centred. It assumes that

$$Y_i = \mu + (X\beta)_i + \varepsilon_i, \tag{3.18}$$

where each ε_i is i.i.d $N(0, \sigma^2)$. In the classical linear model, the study assumes X has full column rank, and so $p < n$.

The tuning parameter λ controls the sparsity of the estimate, with large values of λ resulting in estimates with many components set to 0. Unfortunately, this optimisation problem is hard, and to the best of our knowledge, it is computationally intractable for $p > 50$.

The Lasso (Tibshirani, 1996) solves the related problem:

$$(\hat{\mu}, \hat{\beta}(\lambda)) = \arg\min_{m,b} \{\frac{1}{2n}\|Y - m - Xb\|^2 + \lambda\|b\|_1\} \tag{3.19}$$

The non-differentiability of the ℓ_1 norm at 0 ensures that the resulting estimator is sparse, and its convexity makes the overall optimisation problem convex. There exist very efficient algorithms for solving this problem, even when $p > 105$ (see for example the R package glmnet of Friedman et al.).

Theoretical Properties for Variable Selection

In this section the researcher presents some necessary and sufficient conditions for the Lasso estimator to correctly estimate the sign of β. This conditions are

so for the noiseless case, where

$$y = \mu + X\beta \tag{3.20}$$

The case with noise is similar. For convenience we define $N = 1, \ldots, p_S$, and for a set of variables J, we let X_J denote the matrix formed from the columns of X indexed by J. The study assumes that X_S has full column rank.

Theorem 1:

Let $\lambda > 0$, and

$$\theta = X_N^T X_S (X_S^T X_S)^{-1} sgn(\beta_S) \tag{3.21}$$

If $\|\theta\|_\infty \leq 1$, and for $j \in S$

$$|\beta_j| > \lambda |sgn(\beta_S)^T \{(\frac{1}{n} X_S^T X_S)^{-1}\}^{(j)}|, \tag{3.22}$$

then there exist a Lasso solution with $sgn(\hat{\beta}(\lambda)) = sgn(\beta)$. As a partial converse, if there exist a Lasso solution with $sgn\hat{\beta}(\lambda) = sgn(\beta)$, then $\|\theta\|_\infty \leq 1$.

Remark 1:

$\|\theta\|_\infty$ can be interpreted as the maximum in absolute value over $j \in N$ of the dot product of $sgn(\beta_S)$ and the coefficient vector obtain by regressing $X(j)$ on X_S. That is

$$\|\theta\|_\infty = \max_{j \in N} |sgn(\beta_S)^T (X_S^T X_S)^{-1} X_S^T X^{(-j)}|. \tag{3.23}$$

The condition $\|\theta\|_\infty \leq 1$ is known as (a form of) the irrepresentable condition in literature.

Proof. By considering sub-gradients or simply directional derivatives, the LASSO estimator satisfies

$$\frac{1}{n} X^T \{X(\beta - \hat{\beta}) + (\mu - \hat{\mu})1\} = \frac{1}{n} X^T X(\beta - \hat{\beta}) = \lambda\tau, \tag{3.24}$$

where $\|\tau\|_\infty \leq 1$ and $\tau_j = sgn(\hat{\beta}_j)$ for j such that $\hat{\beta}_j \neq 0$ (and the dependence of $\hat{\beta}$ on λ is suppressed). These are known as the KKT conditions for LASSO (in the noiseless case) in the literature. This equation is expanded into

$$\frac{1}{n}\begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X_N^T X_N \end{pmatrix}\begin{pmatrix} \beta_S - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} = \lambda \begin{pmatrix} \tau_S \\ \tau_N \end{pmatrix} \tag{3.25}$$

The converse is to be proved first.

Suppose $sgn(\beta) = sgn(\hat{\beta})$ (so $\hat{\beta}_N = 0$), then since X_S has full column rank, the top block of 3.25 can be re-written as

$$\beta_S - \hat{\beta}_S = \lambda(\frac{1}{n}X_S^T X_S)^{-1}\tau_S \tag{3.26}$$

We can substitute this into the second block of equations of 3.25 to get

$$\frac{1}{n}X_N^T X_S(\beta_S - \hat{\beta}_S) = \lambda X_N^T X_S(\frac{1}{n}X_S^T X_S)^{-1}\tau_S = \lambda\tau_N \tag{3.27}$$

But if $sgn(\beta_S) = sgn(\hat{\beta}_S)$ then $\tau_S = sgn(\beta_S)$. Thus observing that $\|\tau_N\|_\infty \leq 1$ completes the proof of the converse. Now to the positive statement. We claim that taking

$$(\hat{\beta}_S, \hat{\beta}_N) = (\beta_S - \lambda(\frac{1}{n}X_S^T X_S)^{-1}sgn(\beta_S), 0)$$

$$(\tau_S, \tau_N) = (sgn(\beta_S), X_N^T X_S(X_S^T X_S)^{-1}sgn(\beta_S))$$

satisfies the KKT condition 3.25 or equivalently, since we are taking $\hat{\beta}_N = 0$, equation 3.26 and 3.27. Indeed, the assumption that

$$|\beta_j| > \lambda|sgn(\beta_S)^T((\frac{1}{n}X_S^T X_S)^{-1}))^{(j)}|$$

for $j \in S$ ensures that $sgn(\hat{\beta}_S) = sgn(\beta_S)$, so the condition for τ_S is satisfied.

Then checking 3.26 and 3.27 is easy.

Prediction and Estimation

In order to understand the sort of results we should expect for the prediction and estimation properties of the LASSO, we first imagine that S is known. If we truly knew S, we could simply apply the least squares estimator where we take the design matrix as X_S. If we let $\beta^* = (X_S^T X_S)^{-1} X_S^T Y$ and write $\Omega_{jj} = (\frac{1}{n} X_S^T X_S)^{-1}_{jj}$, we have

$$\mathbb{E}(\frac{1}{n}\|X(\beta^* - \beta)\|^2) = \frac{\sigma^2 s}{n} \tag{3.28}$$

$$\mathbb{E}\|\beta^* - \beta\|_1 = \frac{s\sigma}{\sqrt{n}} \times \frac{1}{s} \sum_{j \in S} \sqrt{\frac{2\Omega_{jj}}{n\phi^2}} \tag{3.29}$$

It is shown that the LASSO achieves these rates for prediction and estimation up to a $log(p)$ factor, and subject to some conditions on the design matrix. It requires that there exists a $\phi > 0$ such that for all b satisfying $\|b_N\|_1 \leq 4\|b_S\|_1$, it holds that

$$\|b_S\|_1^2 \leq \frac{s\|Xb\|^2}{n\phi^2} \tag{3.30}$$

this type of condition is known as the compatibility condition. It can be noted that the constant 4 appearing in the definition is quite arbitrary and could be replaced by any constant greater than 1. Furthermore, it will require that the columns of X are scaled such that $\|X^{(j)}\|^2 = n$ for $j = 1, \ldots, p$.

Theorem 2. Let

$$\tau = A\sigma\sqrt{\frac{log(p)}{n}}$$

Then with probability at least $1 - (p^{1-A^2/8} + p^{-5sA^2/(2\phi^2)})$,

$$\frac{1}{n}\|X(\hat{\beta} - \beta)\|^2 + \lambda\|\hat{\beta} - \beta\|_1 \leq 25\lambda^2 s/\phi^2 = \frac{25A^2}{\phi^2}\frac{\phi^2 s log(p)}{n}$$

Proof. By the definition of $(\hat{\mu}, \hat{\beta})$, it is clear that

$$\frac{1}{2n}\|\mu 1 + X\beta + \varepsilon - \hat{\mu}1 - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|\varepsilon\|^2 + \lambda\|\beta\|_1$$

$$\frac{1}{2n}\|X(\hat{\beta}-\beta)\|^2+\lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta}-\beta)+\frac{1}{n}\varepsilon^{-2}+\lambda\|\beta\|_1 \tag{3.31}$$

Define the following events.

$$\Omega_1 = \frac{1}{n}\|X^T\varepsilon\|_\infty \leq \lambda/2$$

$$\Omega_2 = \varepsilon^{-2} \leq 5\lambda^2 s/\phi^2$$

It is straightforward to show that $\mathbb{P}(\Omega_1 \cap \Omega_2) \leq 1-(p^{1-A^2/8}+p^{-5sA^2/(2\phi^2)})$. In all of the following, we work on $\Omega_1 \cap \Omega_2$. Since

$$\frac{1}{n}|\varepsilon^T X(\hat{\beta}-\beta)| \leq \frac{1}{n}\|X^T\varepsilon\|_\infty\|\hat{\beta}-\beta\|_1 \leq \frac{\lambda}{2}|\hat{\beta}-\beta\|_1$$

it can be seen from 3.31 that

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|^2+2\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta}-\beta\|_1+2\lambda\|\beta\|_1+5\lambda^2 s/\phi^2$$

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|^2+2\lambda\|\hat{\beta}_N\|_1+2\lambda\|\hat{\beta}_S\|_1 \leq \lambda\|\hat{\beta}_S-\beta_S\|_1+\lambda\|\hat{\beta}_N\|_1+2\lambda\|\beta_S\|_1+5\lambda^2 s/\phi^2$$

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|^2+\lambda\|\hat{\beta}_N\|_1 \leq 3\lambda\|\hat{\beta}_S-\beta_S\|_1+5\lambda^2 s/\phi^2 \tag{3.32}$$

First suppose that $\|\hat{\beta}_S-\beta_S\|_1 \leq 5\lambda s/\phi^2$. Then from 3.32 we have,

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|^2+\lambda\|\hat{\beta}_N\|_1 \leq 4\lambda\|\hat{\beta}_S-\beta_S\|_1, \tag{3.33}$$

so in particular

$$\|\hat{\beta}_N-\beta_N\|_1 \leq 4\|\hat{\beta}_S-\beta_S\|_1$$

Thus

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|^2+\lambda\|\hat{\beta}-\beta\|_1 = \frac{1}{n}\|X(\hat{\beta}-\beta)\|^2+\lambda\|\hat{\beta}_N\|_1+\lambda\|\hat{\beta}_S-\beta_S\|_1$$

$$\leq 5\lambda\|\hat{\beta}_S-\beta_S\|_1$$

$$\leq \frac{5\lambda\sqrt{s}\|X(\hat{\beta}-\beta)\|}{\phi\sqrt{n}} \leq 25\lambda^2 s/\phi^2$$

where in the last line we made use of the compatibility condition 3.30.

3.4.2 Smoothly Clipped Absolute Deviation (SCAD)

Again, consider the setting where $(X_i, Y_i), i = 1, \ldots, n$,as n observations satisfying

$$Y_i = \beta_o + X_i'\beta + \varepsilon_i,\ i = 1, \ldots, n \tag{3.34}$$

where $Y_i \in R$ is a response variable, X_i is a $p_n \times 1$ covariates vector and ε_i has mean 0 and variance σ^2. Here the superscripts are used to make it explicit that both the covariates and parameters may change with n. For simplicity, we assume $\beta_o = 0$

In sparse models. the p_n covariates can be classified into two categories: the important ones whose corresponding coefficients are non-zero and the trivial ones whose coefficients are zero. For notational convenience, we write

$$\beta = (\beta_1', \beta_2')', \tag{3.35}$$

where $\beta_1' = (\beta_1, \ldots, \beta_{kn})$ and $\beta_2' = (0, \ldots, 0)$. Here $k_n (\leq p_n)$ is the number of non trivial covariates. Let $m_n = p_n - k_n$ be the number of zero coefficients. Let $Y = (Y_1, \ldots, Y_n)'$ and let $\mathbb{X} = (X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p_n)$ be the $n \times p_n$ design matrix. According to the partition of β, write $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)$, where $\mathbb{X}_1$ and $\mathbb{X}_2$ and $n \times k_n$ and $n \times m_n$ matrices, respectively.

Given $a > 2$ and $\lambda > 0$, the SCAD penalty at θ is

$$p\lambda(\theta; a) = \begin{cases} \lambda|\theta|, & |\theta| \leq \lambda, \\ -(\theta^2 - 2a\lambda|\theta| + \lambda^2)/[2(a-1)], & \lambda < |\theta| \leq a\lambda, \\ (a+1)\lambda^2/2, & |\theta| > a\lambda. \end{cases} \tag{3.36}$$

More insight into it can be gained through its first derivative:

$$p'\lambda(\theta;a) = \begin{cases} sgn(\theta)\lambda, & |\theta| \leq \lambda, \\ sgn(\theta)(a\lambda - |\theta|)/(a-1), & \lambda < |\theta| \leq a\lambda, \\ 0, & |\theta| > a\lambda. \end{cases} \tag{3.37}$$

The SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but not differentiable at 0. Its derivative vanishes outside $[-a\lambda, a\lambda]$. As a consequence, SCAD penalized regression can produce sparse solutions and unbiased estimates for large coefficients. More detailed discussions of this penalty can be found in Fan and Li (2001).

The penalized least squares objective function for estimating β with the SCAD penalty is

$$Q_n(b;\lambda_n,a) = \|Y - \mathbb{X}b\|^2 + n\sum_{j=1}^{p_n} p\lambda_n(b_j;a) \tag{3.38}$$

where $\|.\|$ is the L_2 norm. Given penalty parameters λ_n and a, the LS-SCAD estimator of β is

$$\hat{\beta}_n \equiv \hat{\beta}(\lambda_n;a) = \arg\min Q_n(b;\lambda_n,a) \tag{3.39}$$

We write $\hat{\beta}_n = (\hat{\beta}'_{1n}, \hat{\beta}'_{2n})'$ the way we partition β into β_1 and β_2.

Asymptotic properties of the LS-SCAD estimator

Below, the results on the asymptotic properties of the LS-SCAD estimator are stated. Results for the case of fixed design are slightly different from those for the case of the random design. For convenience, the main assumptions required for conclusions in this section are listed $A1$ to $A5$ for fixed covariates. Let $\rho_{n,1}$ be the smallest eigenvalue of $n^{-1}\mathbb{X}'\mathbb{X}$. τ_n, k_n and ω_n, m_n are the largest eigenvalues of $n^{-1}\mathbb{X}'_1\mathbb{X}_1$ and $n^{-1}\mathbb{X}'_2\mathbb{X}_2$, respectively. Let $X'_{i1} = (X_{i1}, \ldots, X_{ik_n})$ and $X'_{i2} = (X_{i,k_n+1}, \ldots, X_{ip_n})$.

1. (a) ε_i's are i.i.d with mean 0 and variance σ^2

(b) For any $j \in \{1, \ldots, p_n\}$, $\|\mathbb{X}_j\|^2 = n$

2. (a) $\lim_{n\to\infty} \sqrt{k_n \lambda_n}/\sqrt{\rho_n, 1} = 0$; (b) $\lim_{n\to\infty} \sqrt{p_n}/\sqrt{n\rho_n, 1} = 0$.

3. (a) $\lim_{n\to\infty} \sqrt{k_n \lambda_n}/(\sqrt{\rho_n, 1} \min_{1\le j\le k_n} |\beta_j| = 0$; (b) $\lim_{n\to\infty} \sqrt{k_n \lambda_n}/(\sqrt{n\rho_n, 1} \min_{1\le j\le k_n} |\beta_j|$ 0;

   (c) $\lim_{n\to\infty} \sqrt{p_n}/n/\rho_n, 1 = 0$

4. $\lim_{n\to\infty} \sqrt{\max(\pi_n, k_n, \omega_n, m_n)p_n}/(\sqrt{n\rho_{n,1}\lambda_n}) = 0$

5. $\lim_{n\to\infty} \max_{1\le i\le n} X'_{i1}(\sum_{i=1}^n X_{i1}X'_{i1})^{-1}X_{i1} = 0$

For random covariates, we require conditions $(B1)$ through $(B4)$. Suppose (X'_i, ε_i)'s are i.i.d as $(X', \varepsilon) = (X_1, \ldots, X_{p_n}, \varepsilon)$. Analogous to the fixed design case, ρ_1 denotes the smallest eigenvalue of $E[XX']$. Also π_{k_n} and ω_{m_n} are the largest eigenvalues of $E[X_{i1}X'_{i1}]$ and $E[X_{i2}X'_{i2}]$, respectively.

1. $(X'_i, \varepsilon_i) = (X_{i1}, \ldots, X_{ip_n}, \varepsilon_i)$, $i = 1, \ldots, n$ are i.i.d. with
   (a) $E[X_{ij}] = 0$, $Var(X_{ij}) = 1$
   (b) $E[\varepsilon|X] = 0$, $Var(\varepsilon|X) = \sigma^2$

2. (a) $\lim_{n\to\infty} p_n^2/(n\rho_1^2) = 0$
   (b) $\lim_{n\to\infty} k_n\lambda_n^2/\rho_1 = 0$

3. (a) $\lim_{n\to\infty} \sqrt{p_n}/(\sqrt{n\rho_1} \min_{1\le j\le k_n} |\beta_j|) = 0$
   (b) $\lim_{n\to\infty} \lambda_n\sqrt{k_n}/(\sqrt{\rho_1} \min_{1\le j\le k_n} |\beta_j|) = 0$

4. $\lim\limits_{n\to\infty} \frac{\sqrt{\max(\pi_{k_n}, \omega_{m_n})p_n}}{\sqrt{n\rho_1}\lambda_n} = 0$

Theorem 1: (Consistency in fixed design setting).
Under $(A1) - (A2)$,

$$\|\hat{\beta}_n - \beta_n\| \xrightarrow{P} 0 \; as \; n \to \infty$$

A similar results hold for random design case.

Theorem 2: (Consistency in the random design setting).
Suppose that there exists an absolute constant M_4, such that for all n, $\max_{1\leq j\leq p_n} E[X_j^4] \leq M_4 < \infty$, then under $(B1)-(B2)$,

$$\|\hat{\beta}_n - \beta_n\| \xrightarrow{P} 0 \; as \; n \to \infty$$

For consistency, λ_n has to be kept small so that the SCAD penalty would not introduce any bias asymptotically. Note that in both design settings, the restriction on the penalty parameter λ_n does not involve m_n, the number of trivial covariates.

This is shared by the $L_q(0 < q < 1)$-penalized estimators in Huang, Horowitz and Ma (2006). However, unlike the bridge estimators, no upper bound requirement is imposed on the components of β_1, since the derivative of the SCAD penalty vanishes beyond a certain interval while that of the L_q penalty does not. The next two theorems state that the LS-SCAD estimator is consistent for variable selection.

Theorem 3: (Variable selection in the fixed design setting). Under $(A1)-(A4)$, $\hat{\beta}_{2n} = 0_{m_n}$ with probability tending to 1.

Theorem 4: (Variable selection in the random design setting).
Suppose there exists an absolute constant M such that $\max_{1\leq j\leq p_n} |X_j| \leq M < \infty$. Then under $(B1)-(B4)$, $\hat{\beta}_{2n} = 0_{m_n}$ with probability tending to 1. $(A3.a)$ and $(A3.b)$ are identical to $(A2.a)$ and $(A2.b)$, respectively, provided that

$$\liminf_{n\to\infty} \min_{1\leq j\leq k_n} |\beta_j| > 0$$

$(B2)$ has a requirement for $\max_{1\leq j\leq p_n} |\beta_j|$ similar to $(A3)$. $(A4)$ concerns the largest eigenvalues of $n^{-1}\mathbb{X}_1'\mathbb{X}_1$ and $n^{-1}\mathbb{X}_2'\mathbb{X}_2$. Due to the standardization of

covariates,

$$\pi_n, k_n \leq k_n$$

and

$$\omega_n, m_n \leq m_n$$

So $(A4)$ is implied by

$$\lim_{n\to\infty} \frac{p_n}{\sqrt{n\rho_{n,1}}\lambda_n} = 0 \tag{3.40}$$

Likewise, $(B4)$ can be replaced with

$$\lim_{n\to\infty} \frac{p_n}{\sqrt{n\rho_1}\lambda_n} = 0 \tag{3.41}$$

Both $(A4)$ and $(B4)$ require λ_n not to converge too fast to 0 in order for the estimator to be able to "discover" the trivial covariates. It may be of concern if there are λ_n's that simultaneously satisfy $(A2) - (A4)$ (in the random design setting $(B2) - (B4)$) under certain conditions. When $\liminf \rho_{n,1} > 0$ and $\liminf_{n\to\infty}$, it can be checked that there exists λ_n that meets both $(A3)$ and $(A4)$ as long as $p_n = 0(n^{1/3})$. If we further know either that k_n is fixed, or that the largest eigenvalue of $n^{-1}\mathbb{X}'\mathbb{X}$ is bounded from above, as is assumed in Fan and Peng (2004),$p_n = 0(n^{1/2})$ is sufficient. When both of these are true, $p_n = 0(n)$ is adequate for the existence of such λ_n's. Similar conclusions hold for the random design case except that $p_n = 0(n^{1/2})$ is indispensable there.

The advantage of the SCAD penalty is that once the trivial covariates have been correctly picked out, regression with or without the SCAD penalty will make no difference to the non-trivial covariates. So it is expected that $\hat{\beta}_{1n}$ is asymptotically normally distributed.

Let A_n, $n = 1, 2, \ldots$ be a sequence of matrices of dimension $d \times k_n$ with full row rank.

Theorem 5: (Asymptotic normality in the fixed design setting). Under $(A1) - (A5)$,

$$\sqrt{n}\sum_{n}^{-1/2} A_n(\hat{\beta}_{1n} - \beta_1) \xrightarrow{D} N(0_d, I_d), \tag{3.42}$$

where $\sum_n = \sigma^2 A_n(\sum_{i=1}^n X_{i1}X_{i1}'/n)^{-1}A_n'$

Theorem 6: (Asymptotic normality in the random design setting).

Suppose that there exists an absolute constant M such that $\max_{1\leq j\leq p_n}\|X_j\| \leq M < \infty$ and $a\sigma_4$ such that $E[\varepsilon^4|X_{11}] \leq \sigma_4 < \infty$ for all n. Then under $(B1)-(B4)$,

$$n^{-1/2}\sum_{n}^{-1/2} A_n E^{-1/2}[X_{i1}X_{i1}']\sum_{i=1}^{n} X_{i1}X_{i1}'(\hat{\beta}_{1n} - \beta_1) \xrightarrow{D} N(0_d, I_d), \tag{3.43}$$

where $\sum_n = \sigma^2 A_n A_n'$.

For the random design the assumptions for asymptotic normality are no more than those for variable selection. While for the fixed design, a Lindeberg-Feller condition $(A5)$ is needed in addition to $(A1) - (A4)$.

Computing the SCAD

The algorithm of Hunter and Li (2005) is used to compute the LS-SCAD estimator for a given λ_n and a. This algorithm approximates a non-convex target function with a convex function locally at each iteration step. Steps to compute the approximate standard error of the estimator are also described.

1. Computation of the LS-SCAD estimator

   Given λ_n and a the target function to be minimized is

   $$Q_n(b; \lambda_n, a) = \sum_{i=1}(Y_i - X_i'b)^2 + n\sum_{j=1}^{p_n} p\lambda_n(b_j; a). \tag{3.44}$$

   Hunter and Li (2005) proposes to minimize its approximation

$$Q_{n,\xi}(b;\lambda_n,a) = \sum_{i=1}^{n}(Y_i - X_i'b)^2 + n\sum_{j=1}^{p_n} p\lambda_n,\xi(b_j;a)$$

$$= \sum_{i=1}^{n}(Y_i - X_i'b)^2 + n\sum_{j=1}^{p_n}(p\lambda_n(b_j;a) - \xi\int_0^{|b_j|}\frac{p'\lambda_n(t;a)}{\xi+t}dt). \tag{3.45}$$

Around $b_{(k)} = (b_{(k),1},\ldots,b_{(k),p_n})'$, it can be approximated by

$$S_{n,\xi}(b;\lambda_n,a) = \sum_{i=1}^{p_n}[p\lambda_n,\xi(b_{(k),j};a) + \frac{p'\lambda_n(|b_{(k),j}|;a)}{2(\xi+|b_{(k),j}|)}(b_j^2 - b_{(k),j}^2)] \tag{3.46}$$

where ξ is a very small perturbation to prevent any component of the estimate from getting stuck at 0. Therefore the one-step estimator starting from $b_(k)$ is

$$b_{(k+1)} = (\mathbb{X}'\mathbb{X} + nD_\xi(b_{(k)};\lambda_n,a))^{-1}\mathbb{X}'Y, \tag{3.47}$$

where $D_\eta(b_{(k)};\lambda_n,a)$ is the diagonal matrix whose diagonal elements are $\frac{1}{2}p'_{\lambda_n} \times (|b_{(k),j}|;a)/(\xi+|b_{(k),j}|),\ j=1,\ldots,p_n$. Given the tolerance τ, convergence is claimed when

$$|\frac{\partial Q_{n,\xi}(b)}{\partial b_j}| < \frac{\tau}{2},\ \forall j = 1,\ldots,p_n \tag{3.48}$$

And finally the b_j's that satisfy

$$|\frac{\partial Q_{n,\xi}(b)}{\partial b_j} - \frac{\partial Q_n(b)}{\partial b_j}| = \frac{n\xi p'\lambda_n(|b_j|;a)}{\xi+|b_j|} > \frac{\tau}{2} \tag{3.49}$$

are set to 0. A good starting point would be $b_{(0)} = \hat{\beta_{LS}}$, the least squares estimator. The perturbation ξ should be kep(t small so that difference between $Q_{n,\xi}(.)$ and $Q_n(.)$ is negligible. Hunter and Li 2006 suggests using

$$\xi = \frac{\tau}{2n\lambda_n}\min(|b_{(0),j}| :\ b_{(0),j} \neq 0). \tag{3.50}$$

2. Standard Error

The standard errors for the non-zero coefficient estimates can be obtained via the approximation

$$\frac{\partial S_\xi(\hat{\beta}_{1n};\lambda,a)}{\partial\hat{\beta}_{1n}} \approx \frac{\partial S_\xi(\hat{\beta}_1;\lambda_n,a)}{\partial\hat{\beta}_1} + \frac{\partial^2 S_\xi(\hat{\beta}_1;\lambda_n,a)}{\partial\hat{\beta}_1\partial\hat{\beta}_1'}(\hat{\beta}_{1n}-\beta_1) \tag{3.51}$$

So

$$\begin{aligned}\hat{\beta}_{1n}-\beta_1 &\approx -(\frac{\partial^2 S_\xi(\hat{\beta}_1;\lambda_n,a)}{\partial\hat{\beta}_1\partial\hat{\beta}_1'})^{-1}\frac{\partial S_\xi(\hat{\beta}_1;\lambda_n,a)}{\partial\hat{\beta}_1}\\ &\approx -(\frac{\partial^2 S_\xi(\hat{\beta}_{1n};\lambda_n,a)}{\partial\hat{\beta}_{1n}\partial\hat{\beta}_{1n}'})^{-1}\frac{\partial S_\xi(\hat{\beta}_{1n};\lambda_n,a)}{\partial\hat{\beta}_{1n}} \qquad (3.52)\end{aligned}$$

Since

$$\frac{\partial S_\xi(\hat{\beta}_{1n};\lambda_n,a)}{\partial\hat{\beta}_j} = -2\mathbb{X}_j'Y + 2\mathbb{X}_j'\mathbb{X}_1\hat{\beta}_{1n} + n\frac{\hat{\beta}_j p_{\lambda_n}'(|\hat{\beta}_j|;a)}{\xi+|\hat{\beta}_j|}$$

$$= \sum_{i=1}^{n}[-2X_{ij}Y_i + 2X_{ij}X_{i1}'\hat{\beta}_{1n} + \frac{\hat{\beta}_j p_{\lambda_n}'(|\hat{\beta}_j|;a)}{\xi+|\hat{\beta}_j|}]$$

,

$$\triangleq 2\sum_{i=1}^{n}U_{ij}(\xi;\lambda_n,a)$$

letting $U_{ij}=U_{ij}(\xi;\lambda_n,a)$, we have, for $j,l=1,\ldots,k_n$,

$$Cov(n^{-1/2}\frac{\partial S_\xi(\hat{\beta}_{1n};\lambda_n,a)}{\partial\hat{\beta}_j}, n^{-1/2}\frac{\partial S_\xi(\hat{\beta}_{1n};\lambda_n,a)}{\partial\hat{\beta}_l}) \tag{3.53}$$

$$\approx \frac{4}{n}\sum_{i-1}^{n}U_{ij}U_{il} - \frac{4}{n^2}\sum_{i-1}^{n}U_{ij}\sum_{i-1}^{n}U_{il}. \tag{3.54}$$

Let $\mathbb{C}=C_{jl}, j,l=1,\ldots,k_n$, where

$$C_{jl} = \frac{1}{n}\sum_{i-1}^{n}U_{ij}U_{il} - \frac{1}{n^2}\sum_{i-1}^{n}U_{ij}\sum_{i-1}^{n}U_{il}. \tag{3.55}$$

The variance-covariance matrix of the estimates can be approximated by

$$Cov(\hat{\beta}_{1n}) \equiv n(\mathbb{X}_1'\mathbb{X}_1 + nD_\xi(\hat{\beta}_{1n};\lambda_n,a))^{-1}\mathbb{C}(\mathbb{X}_1'\mathbb{X}_1 + nD_\xi(\hat{\beta}_{1n};\lambda_n,a))^{-1} \tag{3.56}$$

3. Selection of λ_n

   The above computational algorithm is for the case when λ_n and a are specified. In data analysis, they can be selected by minimizing the generalized cross validation score, which is defined to be

$$GCV(\lambda_n, a) = \frac{\|Y - \mathbb{X}_1\hat{\beta}_{1n}\|^2/n}{(1 - p(\lambda_n, a)/n)^2} \tag{3.57}$$

   where

$$p(\lambda_n, a) = tr[\mathbb{X}_1(\mathbb{X}_1'\mathbb{X}_1 + nD_0(\hat{\beta}_{1n}; \lambda_n, a))^{-1}\mathbb{X}_1'] \tag{3.58}$$

   is the number of effective parameters and $D_0(\hat{\beta}_{1n}; \lambda_n, a)$ is a sub-matrix of the diagonal matrix $D_\xi(\hat{\beta}_n; \lambda_n, a)$ with $\xi = 0$. By sub-matrix, we mean the diagonal of $D_0(\hat{\beta}_{1n}; \lambda_n, a)$ only contains the elements corresponding to the non-trivial components in $\hat{\beta}$. Note that here $\mathbb{X}_1$ also only includes the columns of which the corresponding elements of $\hat{\beta}_n$ are non-vanishing. The requirement that $a > 2$ is implied by the SCAD penalty function. Simulation suggests that the generalized cross validation score does not change much with a given λ. So to improve computing efficiency, we fix $a = 3.7$, as suggested by Fan and Li (2001).

3.5 Hierarchical Likelihood (HL)

Classical likelihood and its extensions that we have discussed so far are defined for fixed parameters. We may say confidently that we understand their properties quite well, and there is a reasonable consensus about their usefulness. Statisticians have disagreed on a general definition of likelihood that also covers unobserved random variables, e.g. Bayarri et al. (1987). Many ask if there exist a theoretical basis for choosing a particular form of general likelihood? We can actually ask a similar question about the classical likelihood, and the answer seems to be provided by the likelihood principle (Birnbaum, 1962) that the likelihood

contains all the evidence about a (fixed) parameter.

Bjornstad (1996) established the extended likelihood principle, showing that a particular definition of general likelihood contains all the evidence about both fixed and random parameters and this formed the basis for the definition of extended likelihood and h-likelihood of Lee and Nelder (1996). Lee and Nelder (1996) introduced the h-likelihood for inferences in hierarchical GLMs' but being fundamentally different from classical likelihood, it generated some controversies. One key property of likelihood inference that people expect is an invariance with respect to transformations.

The extended likelihood for estimation lacks invariance, so that different scales of the random parameters can lead to different estimates. The dependence on scale makes the extended likelihood immediately open to criticism. In fact, this has been the key source of the controversies. This problem is resolved for the h-likelihood, as it is defined as an extended likelihood for a particular scale of the random parameters with special properties, i.e., it is not defined on an arbitrary scale, so that transformation is not an issue.

For uniformity of notations and as a reminder, we use $f_\theta()$ to denote probability density functions of random variables with fixed parameters θ; the arguments within the brackets determine what the random variable is, and it can be conditional or unconditional. Thus, $f_\theta(y, v)$, $f_\theta(v)$, $f_\theta(y|v)$ or $f_\theta(v|y)$ correspond to different densities even though we use the same basic notation $f_\theta()$. Similarly, the notation $L(a; b)$ denotes the likelihood of parameter a based on data or model b, where a and b can be of arbitrary complexity. For example, $L(\theta; y)$ and $L(\theta; v|y)$ are likelihoods of θ based on different models. The corresponding loglihood is denoted by $\ell(.)$.

3.5.1 Fisher's Likelihood

The classical likelihood framework has two types of object, a random outcome y and an unknown parameter θ, and two related processes on them:

1. Data generation: Generate an instance of the data y from a probability function with fixed parameters θ

$$f\theta(y)$$

2. Parameter inference: Given the data y, make statements about the unknown fixed θ in the stochastic model by using the likelihood

$$L(\theta; y)$$

.

The connection between these two processes is

$$L(\theta; y) \equiv f\theta(y)$$

where L and f are algebraically identical, but on the left-hand side y is fixed while θ varies, while on the right-hand side θ is fixed while y varies. The function $f\theta(y)$ summarizes, for fixed θ, where y will occur if we generate it from $f\theta(y)$, while $L(\theta; y)$ shows the distribution of 'information' as to where θ might be, given a fixed dataset y. Since θ is a fixed number, the information is interpreted in a qualitative way.

Fisher's likelihood framework has been fruitful for inferences about fixed parameters. However, a new situation arises when a mathematical model involves random quantities at more than one level. Consider the simplest example of a

2-level hierarchy with the model

$$y_{i,j} = \beta + v_i + \varepsilon_{i,j},$$

where $v_i \approx N(0, \lambda)$ and $\varepsilon i, j \approx N(0, \phi)$ with v_i and $\varepsilon_{i,j}$ being uncorrelated. This model leads to a specific multivariate distribution. Classical analysis of this model concentrates on estimation of the parameters β, λ and ϕ.

From this point of view, it is straightforward to write down the likelihood from the multivariate normal distribution and to obtain estimates by maximizing it. However, in many recent applications the main interest is often the estimation of $\beta+v_i$. These applications are often characterized by a large number of parameters. Although the v_i are thought of as having been obtained by sampling from a population, once a particular sample has been obtained they are fixed quantities and the likelihood based upon the marginal distribution provides no information on them.

3.5.2 Extended Likelihood

There have been many efforts to generalize the likelihood, e.g., Lauritzen (1974), Butler (1986), Bayarri et al. (1987), Berger and Wolpert (1988) or Bjornstad (1996), where the desired likelihood must deal with three types of object: unknown parameters θ, unobservable random quantities v and observed data y. The previous two processes now take the forms:

1. Data generation:

   (i) Generate an instance of the random quantities v from a probability function $f\theta(v)$ and then with v fixed,

   (ii) generate an instance of the data y from a probability function $f\theta(y|v)$. The combined stochastic model is given by the product of the two proba-

bility functions

$$f\theta(v)f\theta(y|v) = f\theta(y, v) \tag{3.59}$$

2. Parameter inference: Given the data y, we can

   (i) make inferences about θ by using the marginal likelihood $L(\theta; y)$, and

   (ii) given θ, make inferences about v by using a conditional likelihood of the form

$$L(\theta, v; v|y) \equiv f\theta(v, y)$$

The extended likelihood of the unknown (θ, v) is defined by

$$L(\theta, v; y, v) \equiv L(\theta; y)L(\theta, v; v|y). \tag{3.60}$$

The connection between these two processes is given by

$$L(\theta, v; y, v) \equiv f_\theta(y, v) \tag{3.61}$$

so the extended likelihood matches the definition used by Butler (1986), Berger and Wolpert (1988) and Bornstad (1996). On the left-hand side y is fixed while (θ, v) vary, while on the right-hand side θ is fixed while (y, v) vary. In the extended likelihood framework the v appear in data generation as random instances and in parameter estimation as unknowns.

In the original framework there is only one kind of random object y, while in the extended framework there are two kinds of random objects, so that there may be several likelihoods, depending on how these objects are used. The h-likelihood is a special kind of extended likelihood, where the scale of the random parameter v is specified to satisfy the following conditions;

3.5.3 Canonical scale, h-likelihood and joint inference

Let θ_1 and θ_2 be an arbitrary pair of values of fixed parameter θ. The evidence about these two parameter values is contained in the likelihood ratio

$$\frac{L(\theta_1; y)}{L(\theta_2; y)}$$

Suppose there exists a scale v, such that the likelihood ratio is preserved in the following sense

$$\frac{L(\theta_1, \hat{v}_{\theta_1}; y, v)}{L(\theta_2, \hat{v}_{\theta_2}; y, v)} = \frac{L(\theta_1; y)}{L(\theta_2; y)} \tag{3.62}$$

where $\hat{v}_{\theta_1}$ and $\hat{v}_{\theta_2}$ are the MLE's of v for θ at θ_1 and θ_2, so that $\hat{v}_\theta$ is information-neutral concerning θ. Alternatively, 3.62 is equivalent to

$$\frac{L(\theta_1, \hat{v}_{\theta_1}; v|y)}{L(\theta_2, \hat{v}_{\theta_2}; v|y)} = 1,$$

which means that neither the likelihood component $L(\theta, \hat{v}_\theta; v|y)$ nor $\hat{v}_\theta$ carry any information about θ, as is required by the classical likelihood principle. Such a v-scale shall be called and referred to as the canonical scale of the random parameter, and we make an explicit definition to highlight this special situation:
If the parameter v in $L(\theta, v; y, v)$ is canonical we call L an h-likelihood.

To call $L(\theta, v; y, v)$ an h-likelihood assumes that v is canonical, and we shall use the notation $H(\theta, v)$ to denote h-likelihood and $h(\theta, v)$ the h-loglihood. The h-loglihood can be treated like an ordinary loglihood, where, for example, we can take derivatives and compute Fisher information for both parameters (θ, v), etc.

In an arbitrary statistical problem it may not be obvious what the canonical scale is. However, it is quite easy to check whether a particular scale is canonical. When it exists, the canonical scale has many interesting properties that make it the most convenient scale to work with. Let $I_m(\hat{\theta})$ be the observed Fisher information of

the MLE $\hat{\theta}$ from the marginal likelihood $L(\theta; y)$ and let the partitioned matrix

$$I_h^{-1}(\hat{\theta}, \hat{v}) = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}$$

be the inverse of the observed Fisher information matrix of $(\hat{\theta}, \hat{v})$ from the h-likelihood $H(\theta, v; y, v)$, where I^{11} corresponds to the θ part. Then

1. The MLE $\hat{\theta}$ from the marginal likelihood $L(\theta; y)$ coincides with the $\hat{\theta}$ from the joint maximizer of the h-likelihood $L(\theta, v; y, v)$.

2. The information matrices for $\hat{\theta}$ from the two likelihoods also match, in the sense that

   $$I_m^{-1} = I^{11}$$

   This means that (Wald-based) inference on the fixed parameter θ can be obtained directly from the h-likelihood framework.

3. Furthermore, I^{11} yields an estimate of $var(\hat{v} - v)$. If $\hat{v} = E(v|y)|_{\theta=\hat{\theta}}$ this estimates

   $$var(\hat{v} - v) \geq E(var(v|y)),$$

   accounting for the inflation of variance caused by estimating θ.

There are many models which do not have a canonical scale. Maintaining invariance of inferences from the joint maximization of the extended loglihood for trivial re-expressions of the underlying model leads to a definition of the scale of random parameters for the h-likelihood, which covers the broad class of GLM models. It is regarded that this scale as a weak canonical scale and study models allowing such scale. However, models exist which cannot be covered by such a condition. For such models, the adjusted profile likelihoods is proposed for inferences for fixed parameters, which often gives satisfactory estimations.

3.5.4 Variable selection using the Penalized H-Likelihood

This section discusses useful penalty functions for variable selection. Consider variable selection of fixed effects β by maximizing a penalized profile h-likelihood h_p using $h_w^*(\beta, v, \theta)$ and a penalty; it is defined by

$$h_p(\beta, \nu, \theta) = h_\omega^* - n \sum_{j=1}^{p} J_\gamma \left(|\beta_j|\right) \tag{3.63}$$

where $J\gamma\left(|\cdot|\right)$ is a penalty function that controls model complexity using the tuning parameter γ. Note here that no penalty was imposed on the frailty parameter θ. Typically, setting $\gamma = 0$ result in the sub-hazard frailty model, whereas the regression coefficient estimates $\hat{\beta}$ tend to 0 as $\gamma \to \infty$ is inclined to choose a complex model (Fan and Lv, 2010).

Various penalty functions have been used in the literature on the variable selection in the statistical models including Cox-type PH models (Fan and Li, 2001, 2002; Fan and Lv,2010). This dissertation mainly consider the following three penalty functions, but our results can be applied to other penalty functions which are not discuss here.

(i) LASSO (Tibshirani, 1996)

$$J_\gamma \left(|\beta|\right) = \gamma \left|\beta\right|, \tag{3.64}$$

(ii) SCAD (Fan and Li, 2001)

$$J'_\gamma \left(|\beta|\right) = \gamma \left(|\beta| \leq \gamma\right) + \frac{(a\gamma - |\beta|)}{a-1} I \left(|\beta| > \gamma\right), \tag{3.65}$$

where $a = 3.7$ and x_+ denotes the positive part of x, i.e. x_+ is x if $x > 0$, zero otherwise.

(iii) HL (Lee and Oh, 2009)

$$J_\gamma(|\beta|) \equiv J_{(a,b)}(|\beta|) = \log\Gamma(1/b) + \frac{\log b}{b} + \frac{\beta^2}{2au(|\beta|)} + \frac{(b-2)\log u(|\beta|)}{2b} + \frac{u(|\beta|)}{b}, \tag{3.66}$$

where $u(|\beta|) = [\{8b\beta^2/a + (2+b)^2\}^{1/2} + 2 - b]/4$

A good penalty function should estimates that satisfy unbiasedness, scarsity, and continuity (Fan and Li, 2001, 2002). The LASSO in (3.64) is the most common penalty as L_1 penalty, but it does not simultaneously satisfy these three properties. Fan and Li (2001) showed that SCAD in (3.65) satisfy all the these properties and that it can perform well as the orcale procedure in terms of selecting the correct subset model and estimating the true non-zero coefficient, simultaneously. Lee and Oh (2009) proposed a new penalty unbounded at the origin within the framework of a random effect model. The new unbound HL penalties in (3.66), $J_{a,b}(\beta)$, at various values of $b = 0, 2$ and 30 and $a = 1$ are shown in Figure 1.

The form of the penalty changes from a quadratic shape ($b = 0$) for ridge regressions to a cusped form ($b = 2$) for LASSO and then to an unbounded form ($b > 2$) at the origin. In case of $b = 2$, it allow for an infinite gain at zero. The SCAD provides oracle ML estimates (least squares estimators), whereas the HL gives oracle shrinkage estimates.;When there is multi-collinearity, shrinkage estimation is better than the ML estimation. Lee et al. (2010, 2011a,b) have shown its advanages of the HL approach over LASSO and SCAD methods, especially when the number of covariates is larger than the sample size (i.e $p > n$); it actually has a property for a variable selection without losing prediction power. Since a in (3.66) has a greater sensitivity to change of penalty than b, we consider only a few values for b, e.g. $b = 2.1, 3, 10, 30, 50$ representing small, medium and large.

3.5.5 Penalized h-likelihood procedure

By maximizing the penalized h-likelihood h_p in (3.63), we need to screen variable and estimate their associated regression coefficients simultaneously. In other words, those variable whose regression coefficients are estimated as zero are automatically deleted. To achieve the goal, using h_p, the estimation procedures of the fixed parameters(β, θ) and random effects ν are required. First, the maximum penalized h-likelihood (MPHL) estimates of (β, ν), given frailty parameter θ, are obtained by solving the joint estimating of β and ν:

$$\partial h_p/\partial \beta_j = \partial h_w^*/\partial \beta_j - n\sum_{j=1}^{p}[J_\gamma\left(|\beta_j|\right)]' = 0 \tag{3.67}$$

and

$$\partial h_p/\partial v = \partial h_p^*/\partial v = 0 \tag{3.68}$$

Note that (3.67) is an adjusted estimating equation induced by adding the penalty term, whereas (3.68) is the same as the standard estimating equation without penalty. However, for the three penalty functions considered in (3.64)-(3.66), J_γ in (3.67) becomes non-differentiable at the origin and it does not have continuous second-order derivatives. To overcome this difficulty in solving (3.67) we use local quadratic approximation (referred to as LQA, Fan and Li, 2001) to such penalty functions. That is, given an initial value of β^0 close to the true value of β, the penalty function J_γ can be locally approximated by a quadratic function as

$$[J_\gamma\left(|\beta_j|\right)]' = J_\gamma'\left(|\beta_j|\right) sgn\left(|\beta_j|\right) \approx \{J_\gamma'\left(\left|\beta_j^0\right|\right)/[\left|\beta_j^0\right|\}\beta_j \quad for \quad \beta_j \approx \beta_j^0 \tag{3.69}$$

Then the negative Hessian matrix of β and ν based on h_p can be explicitly written as a simple matrix from (Ha and Lee, 2003):

$$H(h_p; \beta, v) = -\partial^2 h_p/\partial(\beta, v)^2 = \begin{pmatrix} X^T W^* X + n\sum_\gamma & X^T W^* Z \\ Z^T W^* X & Z^T W^* Z + U \end{pmatrix} \tag{3.70}$$

Where $\sum_\gamma = diag\{J'_\gamma(|\beta_j|)/\beta_j\}$. Here X and Z are $n \times q$ and $n \times q_*$ model matrices for β and v whose ijth row vectors are x_{ij}^T and z_{ij}^T respectively, $W^* = W^*(\beta, v) = -\partial^2 h_w/\partial\eta^2$ is a form of the symmetric matrix given in Appendix 2 of Ha and Lee(2003) and Ha et al. (2013), $\eta = X\beta + Z\nu$ and $U = -\partial^2\ell_2/\partial v^2$ is a $q^* \times q^*$ matrix that takes a form of $U = BD(\sum^{-1}, ..., \sum^{-1})$ if $\nu \sim N(0, \sum)$, where $q^* = q \times r$ and BD$(\cdot)$ denotes a block diagonal matrix.

Following Ha and Lee (2003) and (3.68), it can be shown that given θ, the MPHIL estimates of (β, ν) are obtained from the
following scores equations:

$$\begin{pmatrix} X^TW^*X + n\sum_\gamma & X^TW^*Z \\ Z^TW^*X & Z^TW^*Z + U \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X^Tw^* \\ Z^Tw^* + R^* \end{pmatrix} \tag{3.71}$$

where $w^* = W^*\eta + (\delta - \mu)$ with $\mu = \exp(\log w + \log\Lambda_{01}^s + \eta)$ and $R^* = Uv + (\partial\ell/\partial v)$. Here w is the weight w_{ij} and Λ_{01}^s is the baseline cumulative sub-hazard function. In particular, $R^* = 0$ if the log-frailty v follows $N(0, \sum)$. The scores equations (3.69) are extensions of the existing estimation procedures. For example under no penalty (i.e.,$\sum_\gamma$) they become the score equations of Ha et al. (2003) for the standard sub-hazard frailty models. for variable selection under the Fine-Gray model (1999) without frailty, they also reduce to

$$(X^TW^*X + n\sum_\gamma)\hat{\beta} = X^Tw^*, \tag{3.72}$$

implying that the new equations (3.69) gives a special case of the penalized equation (3.70) for the Fine-Gray model. Notice that to avoid some numerical difficulty in solving (3.69), we employ $\sum_{\gamma,\epsilon} = dia\{J'_\gamma(|\beta_j|)/(|\beta_j| + \epsilon)\}$, for a small positive value of ϵ (e.g.$\epsilon = 10^{-8}$), instead of $\sum_\gamma$,to assure the existence of $\sum_{\gamma,\epsilon}$ (Lee and Oh, 2009). As long as ϵ is small, the diagonal element of $\sum_{\gamma,\epsilon}$ are close to those of $\sum_\gamma$. In fact, this algorithm is identical to that of Hunter and Li (2005)

for improving the LQA; see also Johnson et al. (2008).

This dissertation reports $\hat{\beta} = 0$ if all five printed decimals are zero. In case of the SCAD and HL penalties, there exist several local maximums. Thus, a good initial value is essential to obtain a proper estimate $\hat{\beta}$. Also in this thesis, a LASSO solution is used as the initial value for the SCAD and HL penalties.
Next, for estimation of θ, an adjusted profile h-likelihood $p_\tau(h_p)$ is used (Ha and Lee, 2003; Lee et al., 2006) which eliminates (β, v) from h_p in (3.64), defined by

$$p_\tau(h_p) = \left[h_p - \frac{1}{2} \log det\{ \frac{H(h_p;\tau)}{(2\pi)} \} \right] \tag{3.73}$$

where $\tau = (\beta^T, v^T)^T$ and $\hat{\tau} = \hat{\tau}(\theta) = (\hat{\beta}^T(\theta), \hat{v}^T(\theta))^T$. The estimates of θ are obtained by solving the score equations $\partial p_\tau(h_p)/\partial\theta = 0$ as in Ha et al. (2013). Accordingly,it is seen that the proposed procedure is easily implemented via a slight modification to the existing h-likelihood procedures (Ha and Lee,2003;Ha et al., 2011, 2013).

3.5.6 Standard error and selection of tuning parameter

This subsection first shows that the standard error (SE) of $\hat{\beta}$ can be obtained by computing an approximated covariance estimate of
$\hat{\beta}$. For this, consider a further penalized profile h-likelihood after estimating v in h_p of (3.64), defined by

$$\hat{h}_p(\beta, \theta) \equiv h_p|_{v=\hat{v}} = \hat{h} - n \sum_{j=1}^{p} J_\gamma \left(|\beta_j| \right), \tag{3.74}$$

where $\hat{h} = \hat{h}(\beta, \theta) = h_w^*(\beta, \theta, v)|_{v=\hat{v}}$ In frailty models, regression parameters β frailty parameter θ are asymptotically orthogonal (Lee and Nelder, 1996; Ha and Lee, 2003; Ha et al, 2011), so that, in estimating θ is minimal. Thus, the SEs of

$\hat{\beta}$ from a sandwich formula (Fam amd Li, 2002; Cai et al, 2005) based on $\hat{p}$.

$$cov(\hat{\beta}) = H(\hat{h}_p;\beta)^{-1} cov(\partial \hat{h}_p/\partial\beta) H(\hat{h}_p;\beta)^{-1} \tag{3.75}$$

where $H(\hat{h}_p;\beta) \equiv -\partial^2 \hat{h}_p/\partial\beta^2 = H_{\beta\beta} + n\sum_{\gamma}$. Here, $H_{\beta\beta} \equiv H(\hat{h};\beta) \equiv -\partial^2 \hat{h}_p/\partial\beta^2$ explicitly computed as follows:

$$H_{\beta\beta} = \{(\partial^2 h_w^*/\partial\beta^2) - (-\partial^2 h_w^*/\partial\beta\partial v)(-\partial^2 \hat{h}_p/\partial v^2)(-\partial^2 h_w^*/\partial v\partial\beta)\}|_{v=\hat{v}}$$

$$= \{(X^T W^* X) - (X^T W^* Z)(Z^T W^* Z + U)^{-1}(Z^T W^* X)\}|_{v=\hat{v}} \tag{3.76}$$

since $\partial\hat{h}/\partial\beta = (\partial h_w^*/\partial\beta) + (\partial h_w^*/\partial v)(\partial\hat{v}/\partial\beta)|_{v=\hat{v}}$ (Ha and Lee, 2003; Ha et al., 2011). Here, the researcher use $H_{\beta\beta}$ to estimate cov$(\partial\hat{h}_p/\partial\beta)$. This study investigates the performance of the proposed SE using (3.72) by simulation studies in the next section.

Selecting important variables, using the penalized likelihood approaches also depends on an appropriate choice of the tuning parameters (Wang et al., 2007; Zhang et al., 2010). For the choice of the tuning parameters γ, a generalized cross-validation(GCV) statistic has been extensively used (Fan and Li, 2001, 2002; Androulakis et al., 2012). However, Wang et al. (2007) showed that the GCV approach can not select the tuning parameters satisfactorily, with a non-ignorable over-fitting effect in the resulting model (Fan and Lv, 2010; Zhang et al., 2010) Thus, they proposed to use a BIC-based selection criterion. In spit of Wang et al. (2007), this study propose to use a BIC-type criterion based on the h-likelihood for selecting tuning parameters γ, defined by

$$BIC(\gamma) = -2p_v(h_w^*)(\hat{\beta},\hat{\theta}) + e(\gamma)\log(n), \tag{3.77}$$

where $p_v(h_w^*) = [h_w^* - (1/2)\log\det H(h_w^*;v)/(2\pi)]$ with $H(h_w^*;v) = -\partial^2/\partial v^2$ is the first-order Laplace approximation to the marginal partial likelihood $m_w^*(\beta,\theta) = \log\{\int \exp(h_w^*)dv\}$ (Therneau et al., 2003; Ha et al.,2011). and it is evaluated at

(β, θ), and

$$e(\gamma) = tr\{H(\hat{h}_p : \beta)^{-1} H(\hat{h}; \beta)\} = tr[\{H_{\beta\beta} + n \sum_{\gamma}\}^{-1} H_{\beta\beta}]$$

Note that $\hat{\gamma} = argmin_{\gamma}\{BIC(\gamma)\}$ is calculated using a simple grid search method as in Fan and Li (2002)

In summary, in the inner loop we maximize h_p for $\tau = (\beta^T, v^T)$ (i.e, the researcher solves (3.69)) and the adjusted profile h-likelihood $p_{\tau}(h_p)$ in (3.71) for θ, respectively. In the outer loop, the study finds γ that minimizes BIC(γ) in (3.73). After convergence has occurred, the study computes the estimates of SEs for $\hat{\beta}$ using (3.72).

3.6 Generalized Linear Models (GLM's)

A classical statistical Linear model is define as

$$y = X\beta + \varepsilon$$

where y is a response variable, X is a model matrix with elements usually depending on some predictor variables, the ε are random variables. β is a vector of unknown parameters, and the purpose of statistical inference with a linear model is to learn about β from the data. This definition of a linear model is based on several important assumptions:

For the systematic part of model a first assumption is additivity of effects; the individual effects of the explanatory variables are assumed to combine additively to form the joint effect. The second assumption is linearity: the effect of each explanatory variable is assumed to be linear, in the sense that doubling the value of x will double the contribution of that x to the mean μ.

For the random part of the model a first assumption is that the errors associated with the response variable are independent. Secondly that the variance of the response is constant, and, in particular, does not depend upon the mean. The assumption of normality, although important as the basis for an exact finite-sample theory, becomes less relevant in large samples. The theory of least squares can be developed using assumptions about the first two moments only, without requiring a normality assumption. The first-moment assumption is the key to the unbiasedness of estimates of β, and the second moment to their optimality.

Generalized linear models (GLM) was first introduced by Nelder and Wedderburn (1972, JRSSA) as extensions of classical linear models. It is derived by two extensions, one to the random part and one to the systematic part. Random elements may now come from a one-parameter exponential family, of which the normal distribution is a special case. Distributions in this class include Poisson, binomial, gamma and inverse Gaussian as well as normal. A generalized linear model consists of three components:

1. A random component, specifying the conditional distribution of the response variable, y_i given the explanatory variables.

2. A linear function of the regressors, called the linear predictor,

$$\eta_i = \alpha + \beta_i X_{i1} + \ldots + \beta_k X_{ik} = X_i' \beta$$

   on which the expected value μ_i of y_i depends.

3. An invertible link function

$$g(\mu_i) = \eta_i,$$

   which transforms the expectation of the response to the linear predictor. The inverse of the link function is sometimes called the mean function:

$g^{-1}(\eta_i) = \mu_i$.

3.6.1 Exponential Family of Distributions

The response variable in a GLM can have any distribution from the exponential family. A distribution belongs to the exponential family of distributions if its probability density function, or probability mass function, can be written as

$$f_\theta(y) = exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)]; \tag{3.78}$$

where b, a and c are arbitrary functions, ϕ an arbitrary scale parameter, and θ is known as the canonical parameter of the distribution (in the GLM context, θ will completely depend on the model parameters β, but it is not necessary to make this explicit yet). It is possible to obtain general expressions for the mean and variance of exponential family distributions, in terms of a, b and ϕ. The log likelihood of θ, given a particular y, is simply $\log[f_\theta(y)]$ considered as a function of ϕ. That is

$$l(\theta) = [\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)]$$

And also

$$\frac{\partial \ell}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$$

Treating ℓ as a random variable, by replacing the particular observation y by the random variable Y, enables the expected value of $\frac{\partial \ell}{\partial \theta}$ to be evaluated:

$$E(\frac{\partial \ell}{\partial \theta}) = \frac{E(y) - b'(\theta)}{a(\phi)}$$

Using the general result that $E(\frac{\partial \ell}{\partial \theta}) = 0$ (at the true value of θ) and re-arranging implies that

$$E(y) = b'(\theta). \tag{3.79}$$

i.e. the mean, of any exponential family random variable, is given by the first derivative of b w.r.t. θ, where the form of b depends on the particular distribution. This equation is the key to linking the model parameters, β, of a GLM to the canonical parameters of the exponential family. In a GLM, the parameters β determine the mean of the response variable, and, via (3.79), they thereby determine the canonical parameter for each response observation. Differentiating the likelihood once more yields

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{-b''(\theta)}{a(\phi)}$$

And plugging this into the general result, $E(\frac{\partial^2 \ell}{\partial \theta^2}) = -E[(\frac{\partial \ell}{\partial \theta})^2]$ (the derivatives are evaluated at the true θ value), gives

$$\frac{b''(\theta)}{a(\phi)} = \frac{E[(Y - b'(\theta))^2]}{a(\phi)^2}$$

Which re-arranges to the second useful general result:

$$var(Y) = b''(\theta)a(\phi)$$

a could in principle be any function of ϕ, and when working with GLMs there is no difficulty in handling any form of a, if ϕ is known. However, when ϕ is unknown matters become awkward, unless we can write $a(\phi) = \phi/\omega$, where ω is a known constant. This restricted form in fact covers all the cases of practical interest here. $a(\phi) = \phi/\omega$ allows the possibility of, for example, unequal variances in models based on the normal distribution, but in most cases ω is simply 1. Hence we now have

$$var(Y) = b''(\theta)\phi/\omega \tag{3.80}$$

In subsequent sections it will often be convenient to consider var(Y) as a function of $\mu \equiv E(Y)$, and, since μ and θ are linked via (3.79), we can always define a variance function $V(\mu) = b''(\theta)/\omega$, such that var($Y$) $= V(\mu)\phi$.

3.6.2 Fitting Generalized Linear Models

Recall that in GLM's, an n-vector of independent response variables, Y, where $\mu \equiv E(Y)$, via

$$g(\mu_i) = X_i\beta$$

and

$$Y_i \approx f_{\theta_i}(y_i)$$

where $f_{\theta_i}(y_i)$ indicates an exponential family distribution, with canonical parameter θ_i, which is determined by μ_i (via equation 3.79) and hence ultimately by β. Given vector y, an observation of Y, maximum likelihood estimation of β is possible. Since the Y_i are mutually independent, the likelihood of β is

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i)$$

And hence the log-likelihood of θ is

$$\ell(\beta) = \sum_{i=1}^{n} log[f_{\theta_i}(y_i)]$$

$$= \sum_{i=1}^{n} \frac{y_i\theta_i - b_i(\theta_i)}{(\phi)} + c_i(y_i, \phi)$$

where the dependence of the right hand side on β is through the dependence of the θ_i on β. Notice that the functions a, b and c may vary with i - this allows different binomial denominators, n_i, for each observation of a binomial response, or different (but known to within a constant) variances for normal responses, for example. ϕ, on the other hand, is assumed to be the same for all i. As discussed in the previous section, for practical work it suffices to consider only cases where we can write $a_i(\phi) = \phi/\omega_i$, where ω_i is a known constant (usually 1), in which case

$$\ell(\beta) = \sum_{i=1}^{n} \frac{\omega_i[y_i\theta_i - b_i(\theta_i)]}{(\phi)} + c_i(y_i, \phi) \tag{3.81}$$

Maximization proceeds by partially differentiating ℓ w.r.t. each element of β, setting the resulting expressions to zero and solving for β. However, these equations are exactly the equations that would have to be solved in order to find β by non-linear weighted least squares, if the weights $V(\mu_i)$ were known in advance and were independent of β.

3.6.3 Iterative Weighted Least Squares

The underlying procedure for fitting GLMs by maximum likelihood takes the form of iterative weighted least squares (IWLS) involving an adjusted dependent variable z, and an iterative weight W. Given a starting value of the mean $\hat{\mu}_0$ and linear predictor $\hat{\eta}_0 = g(\hat{\mu}_0)$, z and W are computed as

$$z = \hat{\eta}_0 + (y - \hat{\mu}_0)(\frac{\partial \eta}{\partial \mu})_0$$

where the derivative is evaluated at $\hat{\mu}_0$, and

$$W^{-1} = (\frac{\partial \eta}{\partial \mu})_0^2 V_0$$

where V_0 is the variance function evaluated at $\hat{\mu}_0$. z is now regressed on the covariates $x_1, x_2, x_3, \ldots, x_p$ with weight W to produce revised estimates $\hat{\beta}_1$ of the parameters, from which we get a new estimate $\hat{\eta}_0$ of the linear predictor. Iteration then starts and continuous until the changes are sufficiently small. Although non-linear, the algorithm has a simple starting procedure by which the data themselves are used as a first estimate of $\hat{\mu}_0$. Simple adjustments to the starting values are needed for extreme values such as zeros in count data. Given the dispersion parameter ϕ, the ML estimators for β are obtained by solving the IWLS equation

$$X^t \Sigma^{-1} X \hat{\beta} = X^t \Sigma^{-1} z \tag{3.82}$$

where $\Sigma = \phi W^{-1}$, and the variance-covariance estimators are obtained from

$$var(\hat{\beta}) = (X^t \Sigma^{-1} X)^2 = \phi (X^t W X)^{-1} \tag{3.83}$$

In IWLS equations $1/\phi$ plays the part of a prior weight. We may view the IWLS equations (3.82) as WLS equations from the linear model

$$z = X\beta + e,$$

where $e = (y - \mu)(\frac{\partial \eta}{\partial \mu}) \approx N(0, \Sigma)$. Note here that $I = X^t \Sigma^{-1} X$ is the expected Fisher information and the IWLS equations (3.82) are obtained by the Fisher scoring method, which uses the expected Fisher information matrix in the Newton-Raphson method. The Fisher scoring and Newton-Raphson methods reduce to the same algorithm for the canonical link, because here the expected and observed informations coincide. Computationally the IWLS procedure provides a numerically stable algorithm. For a detailed derivation of this algorithm see McCullagh and Nelder (1989, section 2.5).

3.6.4 Deviance for Goodness of fit

For a measure of goodness of fit, analogous to the residual sum of squares for normal models, two such measures are in common use: the first is the generalized Pearson χ^2 statistic, and the second the log likelihood-ratio statistic, called the deviance in GLMs. These take the form

$$\chi^2 = \sum (y - \hat{\mu})^2 / V(\hat{\mu})$$

and

$$D = 2\phi[\ell(y; y) - \ell(\hat{\mu}; y)]$$

where ℓ is the loglihood of the distribution. For normal models the scaled deviances χ^2/ϕ and D/ϕ are identical and become the scaled residual sum of squares,

having an exact χ^2 distribution with $n - p$ degrees of freedom. In general they are different and we rely on asymptotic results for other distributions. When the asymptotic approximation is doubtful, for example for binary data with $\phi = 1$, the deviance cannot be used to give an absolute goodness-of-fit test.

For grouped data, e.g. binomial with large enough n, we can often justify assuming that χ^2 and D are approximately χ^2. The deviance has a general advantage as a measure of discrepancy in that it is additive for nested sets of models, leading to likelihood-ratio tests. Furthermore, the χ^2 approximation is usually quite accurate for the differences of deviances even though it could be inaccurate for the deviances themselves. Another advantage of the deviance over the χ^2 is that it leads to the best normalizing residuals (Pierce and Schafer, 1986).

3.6.5 Estimation of the Dispersion Parameter

It remains to estimate the dispersion parameter ϕ for those distributions where it is not fixed (ϕ is fixed at 1 for the Poisson and binomial distributions). If the term $c(y, \phi)$ in the loglihood is available explicitly, the full likelihood can be used to estimate β and ϕ jointly. But often $c(y, \phi)$ is not available, so estimation of ϕ needs a special consideration. One can simply state that ϕ may be estimated using either χ^2 or D, divided by the appropriate degrees of freedom. While χ^2 is asymptotically unbiased (given the correct model) D is not. However, D often has smaller sampling variance, so that, in terms of MSE, neither is uniformly better (Lee and Nelder, 1992). If ϕ is estimated by the REML method (Chapter 3) based upon χ^2 and D, the scaled deviances $\chi^2/\hat{\phi}$ and $D/\hat{\phi}$ become the degrees of freedom $n-p$, so that the scaled deviance test for lack of fit is not useful when ϕ is estimated, but it can indicate that a proper convergence has been reached in estimating ϕ.

3.6.6 Residuals

In GLMs the deviance is represented by sum of deviance components

$$D = \sum d_i,$$

where the deviance component

$$d_i = 2 \int_{\hat{\mu}}^{y_i} (y_i - s)/V(s)ds.$$

The forms of the deviance components for our preferred GLM distributions are as follows;

1. Normal - $(y_i - \hat{\mu}_i)^2$
2. Gamma - $2(-\, log(\frac{y_i}{\hat{\mu}_i}) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i})$

Two forms of residual are based on the signed square-root of the components of χ^2 or D. One is the Pearson residual

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

and the other is the deviance residual

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

Deviance residuals as a set are usually more nearly normal with non-normal GLM distributions than Pearson residuals (Pierce and Schafer, 1986) and are therefore to be preferred for normal plots etc. Other definitions of residuals have been given.

3.6.7 Model Checking

Model checking is perhaps the most important part of applied statistical modelling. In the case of ordinary linear models, this is based on examination of the model residuals, which contain all the information in the data, not explained by the systematic part of the model. Examination of residuals is also the chief means for model checking in the case of GLMs, but in this case the standardization of residuals is both necessary and a little more difficult.

For GLMs the main reason for not simply examining the raw residuals, $\hat{e}_i = y_i - \hat{\mu}_i$, is the difficulty of checking the validity of the assumed mean variance relationship from the raw residuals. For example, if a Poisson model is employed, then the variance of the residuals should increase in direct proportion to the size of the fitted values ($\hat{\mu}_i$). However if raw residuals are plotted against fitted values it takes an extraordinary ability to judge whether the residual variability is increasing in proportion to the mean, as opposed to, say, the square root or square of the mean. For this reason it is usual to standardize GLM residuals, in such a way that, if the model assumptions are correct, the standardized residuals should have approximately equal variance, and behave, as far as possible, like residuals from an ordinary linear model.

The analysis process consists of two main activities: the first is model selection, which aims to find parsimonious well-fitting models for the basic responses being measured, and the second is model prediction, where the output from the primary analysis is used to derive summarizing quantities of interest together with their uncertainties (Lane and Nelder, 1982). In this formulation it is clear that summarizing statistics are quantities of interest belonging to the prediction stage, and thus that they cannot be treated as a response in model selection. Discrepancies between the data and the fitted values produced by the model fall into two main classes, isolated or systematic.

1. **Isolated Discrepancy** - Isolated discrepancies appear when a few observations only have large residuals. Such residuals can occur if the observations are simply wrong, for instance where 129 has been recorded as 192. Such errors are understandable if data are hand recorded, but even automatically recorded data are not immune. Robust methods were introduced partly to cope with the possibility of such errors. Observations with large residuals are systematically down-weighted so that the more extreme the value the smaller the weight it gets. Total rejection of extreme observations (outliers) can be regarded as a special case of robust methods. Robust methods are data driven, and to that extent they may not indicate any causes of the discrepancies.

   A useful alternative is to seek to model isolated discrepancies as being caused by variation in the dispersion, and to seek covariates that may account for them. The techniques of joint modelling of mean and dispersion developed in this thesis makes such exploration straightforward.

   Furthermore if a covariate can be found which accounts for the discrepancies this gives a model-based solution which can be checked in the future. Outliers are observations which have large discrepancies on the y-axis. For the x-axis there is a commonly used measure, the so-called leverage. Outliers or data points with large leverage tend to be potentially influential.

2. **Systematic Discrepancy** - Systematic discrepancies in the fit of a model imply that the model is deficient rather than the data. There is a variety of types of systematic discrepancy, some of which may mimic the effects of others. For this reason it is hard, perhaps impossible, to give a fool proof set of rules for identifying the different types. Consider, for example, a simple regression model with a response y and a single covariate x. Fitting

a linear relation with constant-variance normal errors: discrepancies in the fit might require any of the following:

(a) x should be replaced by $f(x)$ to produce linearity,

(b) the link for y should not be the identity,

(c) both (1) and (2): both should be transformed to give linearity,

(d)the errors are non-normal and require a different distribution,

(e) the errors are not independent and require specification of some kind of correlation between them

(f) an extra term in the model should be added, and so on.

GLMs allow for a series of checks on different aspects of the model. Thus we can check the assumed form of the variance function, of the link, or of the scale of the covariates in the linear predictor. A general technique is to embed the assumed value of, say, the variance function in a family indexed by a parameter, fit the extended model and compare the best fit with respect to the original fit for a fixed value of the parameter.

3.6.8 Model Checking Plots

Residuals based on $r = y - \hat{\mu}$ play a major role in model checking for normal models. Different types of residual have been extended to cover GLMs. These include standardized (Studentized) and deletion residuals. We propose to use standardized residuals from component GLMs for checking assumptions about components. Note that $\text{var}(r) = \phi(1-q)$, so that a residual with a high leverage tends to have large variance.

The standardized residuals are

$$r = \frac{y - \hat{\mu}}{\sqrt{\phi(1-q)}}$$

The standardized Pearson residual is given by

$$r_p^s = \frac{r_p}{\sqrt{\phi(1-q)}} = \frac{y-\hat{\mu}}{\sqrt{\phi V(\hat{\mu})(1-q)}}$$

Similarly, the standardized deviance residual is given by

$$r_d^s = \frac{r_D}{\sqrt{\phi(1-q)}}$$

In this thesis we use deviance residuals since they give a good approximation to Normality for all GLM distributions (Pierce and Schafer, 1986), excluding extreme cases such as binary data. With the use of deviance residuals the normal-probability plot can be used for model checking.

The model-checking plots of Lee and Nelder (1998) are applied to GLMs. In a normal probability plot ordered values of standardized residuals are plotted against the expected order statistics of the standard normal sample. In the absence of outliers this plot is approximately linear. Besides the normal probability plot for detecting outliers, two other plots are used:

1. the plot of residuals against fitted values on the constant-information scale (Nelder, 1990), and

2. the plot of absolute residuals similarly.

For a satisfactory model these two plots should show running means that are approximately straight and flat. If there is marked curvature in the first plot, this indicates either an unsatisfactory link function or missing terms in the linear predictor, or both. If the first plot is satisfactory, the second plot may be used to check the choice of variance function for the distributional assumption. If, for example, the second plot shows a marked downward trend, this implies that the residuals are falling in absolute value as the mean increases, i.e. that the assumed variance function is increasing too rapidly with the mean.

The study also used the histogram of residuals. If the distributional assumption is right it shows symmetry provided the deviance residual is the best normalizing transformation. In GLMs responses are independent, so that these model-checking plots assume that residuals are almost independent. Care will be necessary when we extend these residuals to correlated errors in later techniques employed in this thesis.

3.7 Proposed Joint Generalized Linear Model (JGLM)

Given a statistical model we prefer to use likelihood inferences. However, there are many practical problems for which a complete probability mechanism (statistical model) is too complicated to specify fully or is not available, except perhaps for assumptions about the first two moments, hence precluding a classical likelihood approach. Typical examples are structured dispersions of non-Gaussian data for modelling jointly, the mean and dispersion. Wedderburn's (1974) quasi-likelihood approach deals with this problem, and the analyst needs to specify only the mean-variance relationship rather than a full distribution for the data.

Suppose we have independent responses $y_1, \ldots, y_n$ with means $E(y_i) = \mu_i$ and variance $\text{var}(y_i) = \phi V(\mu_i)$, where μ_i is a function of unknown regression parameters $\beta = (\beta_1, \ldots, \beta_p)$ and $V()$ is a known function. Wedderburn defined the quasi-likelihood (QL, strictly a quasi-loglihood) as a function $q(\mu_i; y_i)$ satisfying

$$\frac{\partial q(\mu_i; y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)}, \tag{3.84}$$

and, for independent data, the total quasi-likelihood is $\sum_i q(\mu_i; y_i)$.

The regression estimate $\hat{\beta}$ satisfies the GLM-type score equations

$$\sum_i \frac{\partial q(\mu_i; y_i)}{\partial \beta} = \sum_i \frac{\partial \mu_i}{\partial \beta} \frac{y_i - \mu_i}{\phi V(\mu_i)} = 0, \tag{3.85}$$

Within the context of Wedderburn's (1974) quasi-likelihood approach;

1. There exists an implied probability structure, a quasi-distribution from a GLM family of distributions, that may not match the underlying distribution. For example, the true distribution may be the negative-binomial, while the quasi-distribution is Poisson. Also, a quasi-distribution might exist on a continuous scale, when the true distribution is supported on a discrete scale, or vice versa.

2. There does not exist an implied probability structure, but a quasilikelihood is available, i.e. there exists a real valued function $q(\mu_i; yi)$, whose derivatives are as in equation 3.85.

3. The estimating equations 3.85 can be further extended to correlated responses. Then, a real valued function $q(\mu_i; yi)$ may not even exist.

The original quasi-likelihood approach was developed to cover the first two contexts and has two notable features:

1. In contrast to the full likelihood approach, we are not specifying any probability structure, but only assumptions about the first two moments. This relaxed requirement increases the flexibility of the QL approach substantially.

2. The estimation is for the regression parameters for the mean only. For a likelihood-based approach to the estimation of the dispersion parameter ϕ some extra principles are needed.

With the general quasi-likelihood approach, for a response y_i and predictor x_i,

the study specify, using known functions $f(.)$ and $V(.)$

$$E(y_i) = \mu_i = f(x_i^t \beta)$$

or

$$g(\mu_i) = x_i^t \beta$$

where $g(\mu_i)$ is the link function, and $\text{var}(y_i) = \phi V(\mu_i) \equiv V_i(\beta, \phi)$. It is possible to generate a GLM using either the quasi- or full likelihood approach. The QL extends the standard GLM by

1. allowing a dispersion parameter ϕ to common models. and

2. allowing a more flexible and direct modelling of the variance function.

3.7.1 Iterative weighted least squares

The main computational algorithm for QL estimates of the regression parameters can be expressed as iterative weighted least squares (IWLS). It can be derived as a Gauss-Newton algorithm to solve the estimating equation. The study solves

$$\sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} (y_i - \mu_i) = 0$$

by first linearizing μ_i around an initial estimate β^0 and evaluating V_i at the initial estimate. Let $\eta_i = g(\mu_i) = x_i^t \beta$ be the linear predictor scale. Then

$$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} x_i$$

so

$$\mu_i \approx \mu_i^0 + \frac{\partial \mu_i}{\partial \beta} (\beta - \beta^0)$$

$$= \mu_i^0 + \frac{\partial \mu_i}{\partial \eta_i} x_i^t (\beta - \beta^0)$$

and

$$y_i - \mu_i = y_i - \mu_i^0 - \frac{\partial \mu_i}{\partial \eta_i} x_i^t (\beta - \beta^0)$$

Putting these into the estimating equation results into

$$\sum_i \frac{\partial \mu_i}{\partial \eta_i} V_i^{-1} x_i \{ y_i - \mu_i^0 - \frac{\partial \mu_i}{\partial \eta_i} x_i^t (\beta - \beta^0) \} = 0 \tag{3.86}$$

which is solve for β as the next iterate, giving an updating formula

$$\beta^1 = (X^t \overset{-1}{\sum} X)^{-1} X^t \overset{-1}{\sum} z, \tag{3.87}$$

where X is the model matrix of the predictor variables, $\sum$ is a diagonal matrix with elements

$$\sum_{ii} = (\frac{\partial \eta_i}{\partial \mu_i})^2 V_i$$

where $V_i = \phi V(\mu_i^0)$, and z is the adjusted dependent variable

$$z_i = x_i^t \beta^0 + \frac{\partial \eta_i}{\partial \mu_i} (y_i - \mu_i^0) \tag{3.88}$$

The constant dispersion parameter ϕ is not used in the IWLS algorithm.

3.7.2 Extended Quasi-likelihood

Wedderburn's original theory of quasi-likelihood (QL) assumes the dispersion parameter ϕ to be known, so his quasi-distribution belongs to the one parameter exponential family. For unknown ϕ, the statement that 'QL is a true loglihood if and only if the distribution is in the exponential family' is not generally correct. In practice, the dispersion parameter is rarely known, except for standard models such as the binomial or Poisson, and even in these cases the assumption that $\phi = 1$ is often questionable. However, the classical QL approach does not tell us how to estimate ϕ from the QL. This is because, in general, the quasi-distribution

implied by the QL, having log-density

$$logf(y_i;\mu_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi), \tag{3.89}$$

contains a function $c(y_i,\phi)$ which may not be available explicitly. Jorgensen (1987) called this GLM family the exponential dispersion family, originally investigated by Tweedie (1947).

Although the standard QL formulation provides consistent estimators for the mean parameters provided the assumed first two-moment conditions hold, it does not include any likelihood-based method for estimating ϕ. Following Wedderburn's original paper, one can use the method of moments, giving

$$var(\frac{y_i - \mu_i}{V(\mu_i)^{1/2}}) = \phi$$

so we expect a consistent estimate

$$\hat{\phi} = \frac{1}{n-p}\sum_i \frac{(y_i-\mu_i)^2}{V(\mu_i)},$$

where μ_i is evaluated using the estimated parameters, and p is the number of predictors in the model. Alternatively, one might consider the so-called pseudo-likelihood (PL)

$$PL(\phi) = -\frac{n}{2}log\{\phi V(\hat{\mu_i})\} - \frac{1}{2\phi}\sum_i \frac{(y_i-\mu_i)^2}{V(\mu_i)}, \tag{3.90}$$

where $\hat{\mu_i}$ is computed using the QL estimate. The point estimate of ϕ from the PL is the same as the method-of-moments estimate. In effect, it assumes that the Pearson residuals

$$r_{pi} = \frac{y_i - \hat{\mu_i}}{V(\hat{\mu_i})^{1/2}} \tag{3.91}$$

are normally distributed.

The PL cannot be used to estimate the regression parameters, so that if we use it in conjunction with the quasi-likelihood, we are employing two distinct likelihoods. However, if we want to use the GLM family (3.89) directly, estimation of ϕ needs an explicit $c(y_i, \phi)$. Nelder and Pregibon (1987) defined an extended quasi-likelihood (EQL) that overcomes this problem. The contribution of y_i to the EQL is

$$Q_i(\mu_i, \phi; y_i) = -\frac{1}{2} log(\phi V(y_i)) - \frac{1}{2\phi} d(y_i, \mu_i) \tag{3.92}$$

and the total is denoted by $q^+ = \sum_i Q_i$, where $d(y_i, \mu_i)$ is the deviance function defined by

$$d_i \equiv d(y_i, \mu_i) = 2 \int_{\mu_i}^{y_i} \frac{y_i - u}{V(u)} du. \tag{3.93}$$

In effect, EQL treats the deviance statistic as $\phi\chi^2_1$-variate, a gamma variate with mean ϕ and variance $2\phi^2$. This is equivalent to assuming that the deviance residual

$$r_{di} \equiv sign(y_i - \mu_i)\sqrt{d_i}$$

is normally distributed. For one-parameter exponential families, the deviance residual has been shown to be the best normalizing transformation (Pierce and Schafer, 1986). Thus, it can be expected that the EQL works well under GLM family. The EQL approach allows a GLM for the dispersion parameter using the deviance as 'data'. In particular, in simple problems with a single dispersion parameter, the estimated dispersion parameter is the average deviance

$$\hat{\phi} = \frac{1}{n} \sum d(y_i, \mu_i)$$

which is analogous to the sample mean $\bar{d}$ for the parameter ϕ. In contrast with PL, the EQL is a function of both the mean and variance parameters. More generally, the EQL forms the basis for joint modelling of structured mean and dispersion parameters, both within the GLM framework.

3.7.3 Joint GLM of Mean and Dispersion

Suppose that we have two interlinked models for the mean and dispersion based on the observed data y and the deviance d:

$$E(y_i) = \mu_i,\ \eta_i = g(\mu_i) = x_i^t\beta,\ var(y_i) = \phi_i V(\mu_i)$$
$$E(d_i) = \phi_i,\ \xi_i = h(\phi_i) = g_i^t\gamma,\ \mathrm{var}(d_i) = 2\phi_i^2$$

where g_i is the model matrix used in the dispersion model, which is a GLM with a gamma variance function. Now the dispersion parameters are no longer constant, but can vary with the mean parameters. One key implication is that the dispersion values are needed in the IWLS algorithm for estimating the regression parameters, and that these values have a direct effect on the estimates of the regression parameters. The EQL q^+ yields a fitting algorithm, which can be computed iteratively using two interconnected IWLS:

1. Given $\hat{\gamma}$ and the dispersion estimates ϕ_is, use IWLS to update $\hat{\beta}$ for the mean model.

2. Given $\hat{\beta}$ and the estimated means $\hat{\mu}_i$s, use IWLS to update $\hat{\gamma}$ with the deviances as data.

3. Iterate Steps 1-2 until convergence.

For the mean model in the first step, the updating equation is

$$X^t \sum^{(-1)} X\beta = X^t \sum^{(-1)} z, \tag{3.94}$$

where

$$z_i = X_i^t\beta + \frac{\partial \eta_i}{\partial \mu_i}(y_i - \mu_i), \tag{3.95}$$

is the adjusted dependent variable and $\sum$ is diagonal with elements

$$\sum_{ii} = \phi_i(\frac{\partial \eta_i}{\partial \mu_i})^2 V(\mu_i)$$

As a starting value, we can use $\phi_i \equiv \phi$, so no actual value of ϕ is needed. Thus, this GLM is specified by a response variable y, a variance function $V(.)$, a link function $g(.)$, a linear predictor $X\beta$ and a prior weight $1/\phi$.

For the dispersion model, first compute the observed deviances $d_i = d(y_i, \hat{\mu}_i)$ using the estimated means. For a moment, we let $d_i^* = d_i/(1 - q_i)$ with $q_i = 0$. For the REML adjustment we use the GLM leverage for q_i.
The updating formula for $\hat{\gamma}$ is

$$G^t \sum_d^{-1} G\gamma = G^t \sum_d^{-1} z_d,$$

Where the dependent variables are defined as

$$z_{di} = g_i^t \gamma + \frac{\partial \xi_i}{\partial \phi_i}(d_i^* - \phi_i) \tag{3.96}$$

and $\sum_d$ is diagonal with elements

$$\sum_{dii} = 2(\frac{\partial \xi_i}{\partial \phi_i})^2 \phi_i^2$$

This GLM is characterized by a response d, a gamma error, a link function $h(.)$, a linear predictor $G\gamma$ and a prior weight $(1 - q)/2$.

At convergence one can compute the standard errors of $\hat{\beta}$ and $\hat{\gamma}$. If the GLM deviance is used, this algorithm yields estimators using the EQL, while with the Pearson deviance it gives those from the PL.

The deviance components d^* become the responses for the dispersion GLM. Then the reciprocals of the fitted values from the dispersion GLM provide prior weights of the next iteration for the mean GLM; these connections are marked in figure 3.2. The resulting see-saw algorithm is very fast to converge. This means that all

the inferential tools used for GLMs can be used for the GLMs for the dispersion parameters. For example, the model-checking techniques for GLMs can be applied to check the dispersion model.

Components	β (fixed)	γ (fixed)
Response	y	d^*
Mean	μ	ϕ
Variance	$\phi V(\mu)$	$2\phi^2$
Link	$\eta = g(\mu)$	$\xi = h(\phi)$
Linear Pred.	$X\beta$	$G\gamma$
Dev. Comp.	d	$gamma(d^*, \phi)$
Prior Weight	$1/\phi$	$(1-q)/2$

Figure 3.2: GLM attributes for joint GLMs.

$d_i = 2\int_{\hat{\mu}_i}^{y_i}(y_i - s)/V(s)ds$

$d^* = d/(1-q)$, $gamma(d^*, \phi) = 2\{-log(d^*/\phi) + (d^* - \phi)/\phi\}$

This gives the EQL procedure if $q = 0$, and the REML procedure if q is the GLM leverage (Lee and Nelder, 1998).

3.7.4 REML Procedure for QL Models and JGLM's allowing true likelihood

In estimating the dispersion parameters, if the size of β is large relative to the sample size, the REML procedure is useful in reducing bias. Because

$$E(\partial^2 q^+/\partial\beta\partial\phi_i) = 0,$$

Lee and Nelder (1998) proposed to use the adjusted profile loglihood

$$p_\beta(q^+) = [q^+ - \{log\ det(I(\hat{\beta}_\gamma)/2\pi)\}/2]|_{\beta=\beta_\gamma} \tag{3.97}$$

where $I(\hat{\beta}_\gamma) = X^t \sum^{-1} X$ is the expected Fisher information, $\sum = \Phi W^{-1}$, $W = (\partial\mu/\partial\eta)^2 V(\mu)^{-1}$, and $\Phi = diag(\phi_i)$. In GLMs with the canonical link - satisfying $d\mu/d\theta = V(\mu)$ - the observed and expected information matrices are the same. In general they are different. For confidence intervals, the use of observed information is better because it has better conditional properties, see Pawitan (2001, Section 9.6), but the expected information is computationally easier to implement.

The interconnecting IWLS algorithm is as before, except for some modification to the adjusted deviance

$$d_i^* = d_i/(1 - q_i)$$

where q_i is the ith diagonal element of

$$X(X^t\Sigma^{-1}X)^{-1}X^t\Sigma^{-1}.$$

(The adjusted deviance also leads to a standardized deviance residual

$$r_{d_i} = sign(y_i - \mu_i)\sqrt{d_i^*/\phi_i}.$$

which can be compared with the theoretical standard normal). Suppose that the responses y have a normal distribution, i.e. $V(\mu) = 1$. If the β were known each $d_i^* = (y_i - x_i\beta)^2 = d_i$ would have a prior weight 1/2, which is reciprocal of the dispersion parameter. This is because

$$E(d_i^*) = \phi_i$$

and

$$var(d_i^*) = 2\phi_i^2$$

and 2 is the dispersion for the $\phi\chi_1^2$ distribution, a special case of the gamma.

With β unknown, the responses $d_i^* = (y_i - x_i\hat{\beta})^2/(1-q_i)$ would have a prior weight $(1-q_i)/2$ because $E(d_i^*) = \phi_i$, and $\text{var}(d_i^*) = 2\phi_i^2/(1-q_i)$. Another intuitive interpretation would be that d_i^*/ϕ_i has approximately χ^2 distribution with $1-q_i$ degrees of freedom instead of 1, because they have to be estimated. For normal models our method provides the ML estimators for β and the REML estimators for ϕ. For the dispersion link function $h()$ we usually take the logarithm.

The REML algorithm using EQL gives a unified framework for joint GLMs (JGLMs) with an arbitrary variance function $V()$. However, since the EQL is an approximation to the GLM likelihood, we use the true likelihood for that variance function, if it exists. For example, suppose that the y component follows the gamma GLM such that $E(y) = \mu$ and $\text{var}(y) = \phi\mu^2$; we have

$$-2\ log\ L = \sum\{\frac{d_i}{\phi_i} + \frac{2}{\phi_i} + \frac{2log\ (\phi_i)}{\phi_i} + 2\ log\ \Gamma(\frac{1}{\phi_i})\}, \tag{3.98}$$

Where

$$d_i = 2\int_{\hat{\mu}_i}^{y_i} \frac{(y_i - s)}{s^2} ds = 2\{\frac{(y-\mu)}{\mu} - log\frac{y}{\mu}\}$$

The corresponding EQL is

$$-2\ log\ q^+ = \sum\{\frac{d_i}{\phi_i} + log(2\pi\phi_i y_i^2)\} \tag{3.99}$$

Here, we notice that $\log f(y)$ and $\log q(y)$ are equivalent up to the Stirling approximation

$$log\Gamma(\frac{1}{\phi_i}) \approx -\frac{log\ \phi_i}{\phi_i} + \frac{log\ \phi_i}{2} + \frac{log(2\pi)}{2} - \frac{1}{\phi_i}. \tag{3.100}$$

Thus, the EQL can give a bad approximation to the gamma likelihood when the value of ϕ is large. It can be shown that $\partial p_\beta(L)/\partial\gamma_k = 0$ leads to the REML method with

$$q_i^* = q_i + 1 + \frac{2\ log\ \phi_i}{\phi_i} + \frac{2\ dg(1/\phi_i)}{\phi_i},$$

where $dg()$ is the di-gamma function.

3.8 Generalized Linear Mixed Models

Let y be an N-vector of responses, and X and Z be an $N \times p$ and $N \times q$ model for the fixed-effect parameters β and random-effect parameters ν. The standard linear mixed model specifies

$$y = X\beta + Z\nu + e \tag{3.101}$$

Where $e \sim MVN(0, \Sigma), \nu \sim MVN(0, D)$, and ν and e are independent. The variance matrices Σ and D are parametrized by an unknown variance-component parameter τ, so random-effect models are also known as variance-component models. The random-effect term ν is sometimes assumed to be $MVN(0, \sigma_\nu^2 I_q)$, and the error term $MVN(0, \sigma_e^2 I_N)$, where I_k is a $k \times k$ matrix, so the variance-component parameter is $\tau = (\sigma_e^2, \sigma_\nu^2)$.

If inferences are required about the fixed parameters only, they can be made from the implied multivariate normal model.

$$y \sim MVN(X\beta, V)$$

Where

$$V = ZDZ' + \Sigma$$

for known variance components, the **MLE**

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y \tag{3.102}$$

is the BLUE and BUE. When the variance components are unknown, we plug in the variance component estimators, resulting in a non-linear estimator for the mean parameters. The simplest random-effect model is the classical one-way

layout

$$y_{ij} = \mu + \nu_i + e_{ij}, \quad i = 1, ..., q, \quad j = 1, ..., n_i \tag{3.103}$$

where μ is the overall mean parameter. The index i typically refers to a cluster and the vector $y_i = (y_{i1}, ...y_{in_i})$ to a set of measurements taken from the cluster. Thus, a cluster may define a person, a family or an arbitrary experimental unit on which we obtain multiple measurements.

3.8.1 Likelihood estimation of fixed parameters

If the interest is only about fixed parameters, marginal likelihood inferences can be made from multivariate normal model

$$y \sim MVN(X\beta, V)$$

It is instructive to look closely at the theory of the simplest random-effect model. Consider the one-way random-effect model

$$y_{ij} = \mu + \nu_i + e_{ij}, \quad i = 1, ..., q, \quad j = 1, ..., n_i \tag{3.104}$$

Where for simplicity we shall assume that the data are balanced in the sense that $n_i \equiv n$. Measurements within a cluster are correlated according to

$$Cov(y_{ij}, y_{ik}) = \sigma^2, \quad j \neq k$$

and $var(y_{ij}) = \sigma_\nu^2 + \sigma_e^2$.

So, $y_i = (y_{i1}, ..., y_{in})^t$ is multivariate normal with mean μ_1, and the variance matrix has the so-called compound- symmetric structure

$$V_i = \sigma_e^2 I_n + \sigma_\nu^2 J_n \tag{3.105}$$

where J_n is an $n \times n$ matrix of ones. Setting $\tau = (\sigma_e^2, \sigma_\nu^2)$, the loglihood of the fixed parameters is given by

$$\ell(\mu, \tau) = -\frac{q}{2} \log |2\pi V_i| - \frac{1}{2} \sum_i (y_i - \mu_1)^t V_i^{-1} (y_i - \mu_1)$$

Where μ is subtracted element-by-element from y_i. To simplify the likelihood, we use the formulae (e.g.,Rao 1973)

$$|V_i| = \sigma_e 2(n-1)(\sigma_e^2 + n\sigma_\nu^2)$$

$$V_i^{-1} = \frac{I_n}{\sigma_e^2} - \frac{\sigma_\nu^2}{\sigma_e^2(\sigma_e^2 + n\sigma_\nu^2)} J_n \tag{3.106}$$

where I_n is an $n \times n$ matrix of ones.

thus,

$$\ell(\mu, \tau) = -\frac{q}{2}[(n-1)\log(2\pi\sigma_e^2) + \log(2\pi n\sigma_\nu^2)]$$
$$-\frac{1}{2}\{\frac{SSE}{\sigma_e^2} + \frac{SSV + q_n(\bar{y}.. - \mu)^2}{\sigma_e^2 + n\sigma_\nu^2}\} \tag{3.107}$$

Where we have defined the error and cluster sums of squares respectively as

$$SSE = \sum_i \sum_i (y_{ij} - \bar{y}_i)^2$$
$$SSE = n \sum_i (\bar{y}_i. - \bar{y}..)^2$$

It is clear that for any fixed $(\sigma_e^2, \sigma_\nu^2)$, the **MLE** of μ is the overall mean $\bar{y}..$, so the profile Likelihood of the variance components is given by

$$\ell_p(\tau) = -\frac{q}{2}[(n-1)\log(2\pi\sigma_e^2) + \log\{2\pi(\sigma_e^2 + \sigma_\nu^2)\}]$$
$$- \frac{1}{2}\left(\frac{SSE}{\sigma_e^2} + \frac{SSV}{\sigma_e^2 + n\sigma_\nu^2}\right) \tag{3.108}$$

3.8.2 Inferences about the fixed effects

From the multivariate normal model, the marginal loglihood of the fixed parameters is given by (β, τ) in the form

$$\ell(\beta, \tau) = -\frac{1}{2}\log|2\pi V| - \frac{1}{2}(y - X\beta)^t V^{-1}(y - X\beta), \tag{3.109}$$

Where the dispersion parameter τ enters through the marginal variance First we show that, conceptually, multiple-component models are no more complex than single-component models. Extensions of the profile likelihood to include more random components take the form

$$y = X\beta + Z_1\nu_1 + ... + Z_m\nu_m + e,$$

Where Z_i are $N \times q_i$ model matrices, and the ν_i are independent $MVN_{qi}(0, D_i)$.

In some applications the random effects are iid, so the variance matrix is given by

$$D = \sigma_\nu^2 I_q$$

It is also quite common to see a slightly more general variance matrix

$$D = \sigma_\nu^2 R$$

with known matrix R. This can be reduced to the simple iid form by re-expressing the model in the form.

$$y = X\beta + ZR^{1/2}R^{-1/2}\nu + e$$

$$= X\beta + Z^\star\nu^\star + e$$

by defining $Z^\star \equiv ZR^{\frac{1}{2}}$ and $\nu^\star = R^{\frac{-1}{2}}\nu$, where $R^{\frac{1}{2}}$ is the square root matrix of R. Now $\nu^\star$ is $MVN(0, \sigma_\nu^2 I_q)$. This means that methods developed for the iid case

can be applied more generally. for fixed τ, taking the derivative of the loglihood with respect to β gives

$$\frac{\partial \ell}{\partial \beta} = X^t V^{-1}(y - X\beta)$$

So that the **MLE** of β is the solution of

$$X^t V^{-1} X\hat{\beta} = X^t V^{-1} y,$$

The well known generalized least-squares formula. Hence the profile likelihood of the variance parameter τ is given by

$$\ell_p(\tau) = -\frac{1}{2} log|2\pi V| - \frac{1}{2}(y - X\hat{\beta}_\tau)^t V{-}1(y - X\hat{\beta}_\tau), \tag{3.110}$$

and the fisher information of β is the solution of

$$I(\hat{\beta}_\tau) = X^t V^{-1} X.$$

In practice, the estimated value of τ is plugged into the information formula, from which we can find the standard error for the **MLE** $\hat{\beta}$ in the form.

$$\hat{\beta} = \hat{\beta}_\tau$$

$$I(\hat{\beta}) = X^t V_\tau^{-1} X$$

Where the dependence of **V** on the parameter estimate is made explicit. The standard errors computed from this plug-in formula do not take into account the uncertainty in the estimation of τ, but this is nevertheless commonly used. Because $E(\partial^2/\partial\beta\partial r) = 0$, i.e. the mean and dispersion parameters are orthogonal(Pawitan 2001), this variance inflation caused by the estimation of τ is fortunately asymptotically negligible. However, it could be non-negligible if the design is very unbalanced in small samples. In such cases numerical methods such as Jackknife method is useful to estimate the variance inflation in finite samples

(Lee,1991). For finite sample adjustment of t- and F-test see Kenward and Roger (1997). In the linear Models it is not necessary to have distributional assumptions about y, but only that

$$E(Y) = X\beta \qquad and \qquad var(Y) = V$$

So that the **MLE** above is the **BLUE** for given dispersion parameters. Then the dispersion Parameters are estimated by the method of moments using **ANOVA**. However, this simple technique is difficult to extend to more complex models.

3.8.3 Estimation of variance components

If we include the **REML** adjustment to account for the estimation of the fixed effect β, because $E(\partial^2 l/\partial\beta\partial r) = 0$, from profile likelihood we get an adjusted profile likelihood

$$p_\beta(\ell\backslash\tau) = \ell(\hat{\beta}_\tau, \tau) - \frac{1}{2}\log|X^t V^{-1} X/(2\pi)|$$

In normal linear mixed models, this likelihood can be derived as an exact likelihood either by conditioning or marginalizing.

3.8.4 Conditional likelihood

Let $\hat{\beta} = Gy$ where $G = (X^t V^{-1} X)^{-1} X^t V{-}1$ From

$$f(y) = |2\pi V|^{-1/2} \exp\{-\frac{1}{2}(\beta \hat{-} \beta)^t V^{-1(y-X\beta)}\}$$

and for fixed τ, $\hat{\beta} \sim MVN(\beta, (X^t V^{-1} X)^{-1})$, so

$$f(\hat{\beta}) = |2\pi(X^t V^{-1} X)^{-1}|^{-1/2} \exp\{-\frac{1}{2}(\hat{\beta} - \beta)^t X^t V^{-1} X(\hat{\beta} - \beta)\}$$

giving the conditional likelihood

$$f(y\backslash\hat{\beta}) = |2\pi V|^{-1/2}|X^tV{-}1X/(2\pi)|^{-1/2}\exp\{-\frac{1}{2}(y - X\hat{\beta})^tV^{-1}(y - X\beta)\}$$

The loglihood gives $p_\beta(l/t)$

3.8.5 Marginal likelihood

The marginal likelihood is constructed from the residual vector. Let

$$P_X \equiv X(X^tX)^{-1}X^t$$

be the hat matrix with rank p. Let A be an $n \times (n - p)$ matrix satisfying $A^tA = I_{n-p}$ and $AA^t = I_n - Px$. Now R $= A^ty$ spans the space of residuals, and satisfies

$$E(r) = 0$$

Then, R and $\hat{\beta}$ are independent because.

$$cov(R, \hat{\beta})$$

Let $T = (A, G)$. Then, matrix manipulation shows that

$$f(y) = f(R, \hat{\beta})|T|$$

$$= f(R, \hat{\beta})|T^tT|^{1/2}$$

$$= f(R)f(\hat{\beta})|X^tX|^{-1/2}$$

This residual density $f(R)$ is proportional to the conditional density $f(y/\hat{\beta})$, and the corresponding loglihood is, up to a constant term, equal to the adjusted profile loglihood $p_\beta(l/\tau)$.

3.8.6 Classical estimation of random effects

Since the study deals with random parameters, the classical approach is based on optimising the mean-square error

$$E||\hat{\nu} - \nu||^2,$$

which gives the **BUE** $\hat{\nu} = E(\nu/y)$. In the general normal linear mixed model (3.101) we have

$$E(\nu \backslash y) = (Z^t \sum^{-1} Z + D{-}1^{-1} Z^t \sum^{-1} (y - X\beta)) \tag{3.111}$$

If the data are not normal, the formula is **BLUE**. If β is unknown, one can use its **BLUE** (3.102) and the resulting estimator of ν is still **BLUE**

For the record, the researcher emphasis that Henderson (1959) recognized that the estimates (3.102) and (3.111) derived for optimal estimation can be obtained by maximizing the joint density function [our emphasis of] y and ν:

$$\log f(y, \nu) \propto -\frac{1}{2}(y - X\beta - Z\nu)^t \sum^{-1} (y - X\beta - Z\nu) - \tfrac{1}{2}\nu^t D^{-1}\nu, \tag{3.112}$$

with respect to β and ν. In the 1950 he called these the *joint maximum likelihood estimates*. It is known that such a joint optimization works only if the random effects ν are the canonical scale for β and this is so here. However, the result is not invariant with respect to non-linear transformations of ν. Later in 1973 Henderson wrote that these estimates should not be called maximum likelihood estimates, since the function being maximized is not a likelihood. It is thus clear that he used the joint maximization only as an algebraic device, and did not recognize the theoretical implications in terms of extended likelihood inference

the derivative of $f(y, \nu)$ with respect to β is

$$\frac{\partial \log f}{\partial \beta} = X^t \sum^{-1} (y - X\beta - Z\nu)$$

Combining this with the derivative with respect to ν and setting them to zero gives

$$\begin{pmatrix} X^{t-1}X & X^t \sum^{-1} Z \\ Z^t \sum^{-1} X & Z^t \sum^{-1} Z + D^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\nu} \end{pmatrix} = \begin{pmatrix} X^t \sum^{-1} \\ Z^t \sum^{-1} y \end{pmatrix} \tag{3.113}$$

The estimates resulting from these simultaneous equations are exactly those we get from (3.102) and (3.111). The joint equation, which forms the basis for the most algorithms in the mixed models, is often called Henderson's mixed model equation. When D^{-1} goes to zero the resulting estimating equation is the same as that treating ν as fixed. Thus, the so-called intra-block estimator can be obtained by taking $D^{-1} = 0$.

3.8.7 Inference for mean parameters

From the optimization of the log-density, given D and Σ, the h-likelihood estimates β And υ satisfy the mixed model equation 3.113. Let **H** be the square matrix of the left hand side of the equation,$V = ZDZ^t + \Sigma$ and $\Lambda = Z^t \Sigma^{-1} Z + D^{-1}$. The solution for β gives the **MLE**, satisfying

$$X^t V^{-1} X\hat{\beta} = X^t V^{-1} y \tag{3.114}$$

and the solution for υ gives the empirical **BUE**

$$\hat{\upsilon} = E(\hat{\upsilon} \backslash y) = E(\upsilon \backslash y)|_{\beta=\hat{\beta}}$$

$$= DZ^t V^{-1} (y - X\hat{\beta})$$

$$= \Lambda^{-1} Z^t \Sigma^{-1} (y - X\hat{\beta})$$

This yields $(X^t V^{-1} X)^{-1}$, as a variance estimate for $\hat{\beta}$, which coincides with that for the **ML** estimate.we now show that H^{-1} also gives the correct estimate for $E\{(\hat{\upsilon} - \upsilon)(\hat{\upsilon} - \upsilon)^t\}$, one that accounts for the uncertainty in $\hat{\beta}$. When β is known the random-effect estimate is given by

$$\tilde{\upsilon} = E(\upsilon \backslash y)$$

So we have

$$var(\tilde{\upsilon} - \upsilon) = E(\tilde{\upsilon} - \upsilon)(\tilde{\upsilon} - \upsilon)^t = E\{var(\upsilon|y)\}$$

where

$$var(\upsilon|y) = D - DZ^t V^{-1} ZD = \Lambda^{-1}$$

So when β is known Λ^{-1}. Gives a proper estimate of the variance of $\tilde{\upsilon} - \upsilon$. However, when β is unknown, the plugged-in empirical Bayes estimate $\Lambda^{-1}|_{\beta=\hat{\beta}}$ for $var(\hat{\upsilon} - \upsilon)$ does not properly account for the extra uncertainty due to estimating β. By contrast, the h-likelihood computation gives a straight forward correction. Now we have

$$var(\hat{\upsilon} - \upsilon) = E\{var(\upsilon \backslash y)\} + E\{(\hat{\upsilon} - \tilde{\upsilon})(\hat{\upsilon} - \tilde{\upsilon})^t\} \tag{3.115}$$

where the second term shows the variance inflation caused by estimating the unknown β as an estimate for $var(\hat{\upsilon} - \upsilon)$. The appropriate component of H^{-1} gives

$$\{\Lambda^{-1} + \Lambda^{-1} Z^t \Sigma^{-1} X (X^t V^{-1} X)^{-1} X^t \Sigma^{-1} Z \Lambda^{-1}\}|_{\beta=\hat{\beta}}$$

Because $\hat{\upsilon} - \tilde{\upsilon} = -DZ^tV^{-1}X(\hat{\beta} - \beta)$ it can be shown that

$$\{(\hat{\upsilon} - \tilde{\upsilon})(\hat{\upsilon} - \tilde{\upsilon})^t\} = \Lambda^{-1}Z^t\Sigma^{-1}X(X^tV^{-1}X)^{-1}X^t\Sigma^{-1}Z\Lambda^{-1}.$$

Thus, the h-likelihood approach correctly handles the variance inflation caused by estimating the fixed effects. From this we can construct confidence bounds for unknown υ.

3.8.8 Estimation of variance components

We have previously derived the profile likelihood for the variance component parameter τ, but the resulting formula 3.110 is complicated by terms involving $|V|$ or V^{-1}. In practice these matrices are usually too unstructured to deal with directly. Instead we can use formulae derived from the h-likelihood. First, the marginal likelihood of (β, τ) is

$$L(\beta,\tau) = |2\pi\Sigma|^{-\frac{1}{2}} \int exp{-\frac{1}{2}}(y - X\beta - Z\upsilon)^t\Sigma^{-1}(y - X\beta - Z\upsilon)$$

$$\times|2\pi D|^{-\frac{1}{2}}exp\{-\frac{1}{2}\upsilon^tD^{-1}\upsilon\}d\upsilon$$

$$= |2\pi\Sigma|^{-\frac{1}{2}}exp\{-\frac{1}{2}(y-X\beta-Z\hat{\upsilon}_{\beta,\tau})^t\Sigma^{-1}(y-X\beta-Z\hat{\upsilon}_{\beta,\tau})\}\times|2\pi D|^{-\frac{1}{2}}exp\{-\frac{1}{2}\hat{\upsilon}^t_{\beta,\tau}D^{-1}\hat{\upsilon}_{\beta,\tau}\}d\upsilon$$
$$\times \int exp\{-\frac{1}{2}(\upsilon - \hat{\upsilon}_{\beta,\tau})^tI(\hat{\upsilon}_{\beta,\tau})(\upsilon - \hat{\upsilon}_{\beta,\tau})\}d\upsilon$$

$$= |2\pi\Sigma|^{-\frac{1}{2}}exp\{-\frac{1}{2}(y-X\beta-Z\hat{\upsilon}_{\beta,\tau})^t\Sigma^{-1}(y-X\beta-Z\hat{\upsilon}_{\beta,\tau})\}\times|2\pi D|^{-\frac{1}{2}}exp\{-\frac{1}{2}\hat{\upsilon}^t_{\beta,\tau}D^{-1}\hat{\upsilon}_{\beta,\tau}\}$$
$$\times\, |I(\hat{\upsilon}_{\beta,\tau}/2\pi)|^{-\frac{1}{2}}.$$

(Going from the first to the second formula involves tedious matrix algebra.) One can obtain the marginal loglihood in terms of the adjusted profile likelihood:

$$\iota(\beta,\tau) = h(\beta,\tau,\hat{\tau}_{\beta,\tau}) - \frac{1}{2}log|I(\hat{\upsilon}_{\beta,\tau}/2\pi)|,$$

$$= p_{\upsilon}(h|\beta,\tau) \tag{3.116}$$

where, from before,

$$I(\hat{\upsilon}_{\beta,\tau}) = \frac{\partial^2}{\partial\upsilon\partial\upsilon^t}|_{\upsilon=\hat{\upsilon}_{\beta,\tau}} = Z^t\Sigma^{-1}Z + D^{-1}.$$

The constant (2π) is kept in the adjustment term to make the loglihood an exact log-density; this facilitates comparisons between models as in the example below. Thus the marginal likelihood in the mixed effects models is equivalent to an adjusted profile likelihood obtained by profiling out the random effects.

In the one-way random-effect model

$$y_{ij} = \mu + \nu_i + e_{ij}, i = 1, ..., q, j = 1, ..., n \tag{3.117}$$

From our previous derivations, given the fixed parameters (μ,τ),

$$\hat{\upsilon} = (\frac{n}{\sigma_e^2} + \frac{1}{\sigma_{\upsilon}}^2)^{-1}\frac{n}{\sigma_e^2}(\overline{y}_i - \mu)$$

$$= \frac{n\sigma_{\upsilon}^2}{\sigma_e^2 + n\sigma_{\upsilon}^2}(\overline{y}_i - \mu)$$

$$I(\hat{\upsilon}) = \frac{n}{\sigma_e^2} + \frac{1}{\sigma_\upsilon^2}$$

So the adjusted profile loglihood becomes

$$p_\upsilon(h\ \mu, \tau) = -\frac{qn}{2} log(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2} \Sigma_{i=1}^{q} \Sigma_{j=1}^{n} (y_{ij} - \mu - \hat{\upsilon}_i)^2$$

$$-\frac{q}{2} log(2\pi\sigma_\upsilon^2) - \frac{1}{2\sigma_\upsilon^2} \Sigma_{i=1}^{q} \hat{\upsilon}_i^2 - \frac{q}{2} log \frac{\sigma_e^2 + n\sigma_\upsilon^2}{2\pi\sigma_\upsilon^2\sigma_e^2}$$

$$= \frac{-q}{2}[(n-1)log(2\pi\sigma_e^2) + log2\pi(\sigma_e^2 + n\sigma_\upsilon^2)] - \frac{1}{2\sigma_e^2} \Sigma_{i=1}^{q} \Sigma_{j=1}^{n} (y_{ij} - \mu - \hat{\upsilon}_i)^2 - \frac{1}{2\sigma_e^2} \Sigma_{i=1}^{q} \hat{\upsilon}_i^2$$

$$= \frac{-q}{2}[(n-1)log(2\pi\sigma_e^2) + log2\pi(\sigma_e^2 + n\sigma_\upsilon^2)] - \frac{1}{2}\{\frac{SSE^2}{\sigma_{\ e}} + \frac{SSV + qn(\overline{y}_{\cdots} - \mu)^2}{\sigma_e^2 + n\sigma_\upsilon^2}\},$$

Note that the h-loglihood $h(\mu, \tau, \upsilon)$ and information matrix $I(\hat{\upsilon}_i)$ are unbounded as σ_υ^2 goes to zero, even though the marginal loglihood $\iota(\mu, \sigma_e^2, \sigma_\upsilon^2 = 0)$ exists. The theoretical derivation here shows that the offending terms cancels out. Numerically, this means that we cannot use $p_{\upsilon(h)}$ at $(\mu, \sigma_e^2, \sigma_\upsilon^2 = 0)$. This problem occurs more generally when we have several variance components. In these cases we should instead compute $p_\upsilon(h)$ based on the h-likelihood of the reduced model when one or more of the random components is absent. For this reason the constant 2π should be kept in the adjusted profile loglihood. (Lee and Nelder, 1996)

3.8.9 REML estimation of variance components

In terms of the h-likelihood, the profile likelihood oft the variance components 3.110 can be rewritten as

$$\iota_p(\tau) = \iota(\hat{\beta}_\tau, \tau)$$

$$= h(\hat{\beta}_\tau, \tau, \hat{\upsilon}_\tau) - \frac{1}{2} log|I(\hat{\upsilon}_\tau)/(2\pi)| \tag{3.118}$$

where τ enters the function through $\Sigma, D, \hat{\beta}_\tau$ and $\hat{\upsilon}_\tau$, and as before $I(\hat{\upsilon}_\tau) = Z^t\Sigma^{-1}Z + D^{-1} = \Lambda$ since $I(\hat{\upsilon}_\beta, \tau)$ is not a function β. The joint estimation of $\hat{\beta}$ and $\hat{\upsilon}$ as a function of τ was given previously by (3.116) If we include **REML** adjustment for the estimation of the fixed effect β, results in

$$p_\beta(\iota\ \tau) = \iota(\hat{\beta}_\tau, \tau) - \frac{1}{2} log|X^t V^{-1X/(2\pi)}|)$$

$$= h(\hat{\beta}_\tau, \tau, \hat{\upsilon}_\tau) - \frac{1}{2} log|I(\hat{\upsilon}_\tau)/(2\pi)| - \frac{1}{2} log|X^t V^{-1X/(2\pi)}|)$$

$$= p_{\beta,\upsilon}(h|\tau) \tag{3.119}$$

where here the $p(.)$ notation allows the representation of the adjusted profiling of both fixed and random effects simultaneously. Hence in the normal case, the different forms of likelihood of the fixed and random effects simultaneously. Hence, in the normal case, the different forms of likelihood of the fixed parameters match exactly the adjusted profile likelihood derived from the h-likelihood. Since $\iota(\beta, \tau) = p_\upsilon(h)$, also the equation

$$p_{\beta,\upsilon}(h) = p_\beta p_\upsilon(h) \tag{3.120}$$

A useful result that would be only approximately true in non-normal cases.

3.8.10 fitting algorithm

The h-likelihood approach provides an insightful fitting algorithm, particularly with regard to the estimation of the dispersion parameters. The normal case is a useful prototype for the general case dealt with in the next chapter. Consider an augmented classical linear model

$$y_a = T\delta + e_a \tag{3.121}$$

where $e_a \sim MVN(0, \Sigma_a)$, and

$$y_a = \begin{pmatrix} y \\ \psi_M \end{pmatrix}, T \equiv \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}, \delta = \begin{pmatrix} \beta \\ v \end{pmatrix}$$

$$e_a = \begin{pmatrix} e \\ e_M \end{pmatrix}, \sum_a \equiv \begin{pmatrix} \sum & 0 \\ 0 & D \end{pmatrix}$$

It can be seen that, $\Sigma = \sigma^2 I$ and $D = \sigma_v^2 I$ Because the augmented linear model is a **GLM** with a constant variance function it is possible to apply the **REML** methods for the joint **GLM** in section 3.7 to fit the linear mixed models. Here the deviance components corresponding to e are the squared residuals

$$d_i = (y_i - X_i\hat{\beta} - Z_i\hat{v})^2$$

and those corresponding to $e_M are$

$$d_{Mi} = (\psi_M - \hat{v}_i)^2 = \hat{v}_i,$$

and the corresponding leverages are diagonal elements of

$$T(T^t\Sigma_a^t T)^{-1}T^t\Sigma_a^{-1}$$

The estimation of (β, τ, v). in the linear mixed model can be done by **IWLS** for the augmented linear model as follows, where for the clarity we show all the required equations:

0. Start with an estimate of the variance parameter τ.

1. Given the current estimate of τ update $\hat{\delta}$ by solving the augmented generalized least squares equation:
   $T^t \Sigma_a^{-1} \Sigma_{-1}^a T\hat{\delta} = T^t \Sigma_a^{-1} y_a$

2. Given the current value of δ get an update ofτ;the **REML** estimators can be obtained by fitting a gamma **GLM** as follows: the estimator for σ^2 is obtained from the **GLM**, characterized by a response $d^* = d/(1-q)$ a gamma error, a link $h()$, a linear predictor γ (intercept only model), and a prior weight $(1-q)/2$ and estimator, for $\sigma_v^2$is obtained by the **GLM**, characterize by a response $d_M^* = d_M/(1-qM)$ a gamma error, a link $h_M()$, a linear predictor γ_M (intercept only model) and a prior weight $(1-q_M)/2$. Note here that
   $$E(d^*) = \sigma^2$$
   and
   $$var(d^*) = 2\sigma^2/(1-q_i),$$
   and
   $$E(d_{Mi}^*) = \sigma_v^2$$
   and
   $$var(d_{Mi}^*) = 2\sigma_v^2/(1-q_{Mi}),$$
   This algorithm is often much faster than the ordinary **REML** procedure of the previous section. The **MLE** can be obtained by taking the levelages to be zero.

3. Iterate between 1 and 2 until convergence. At convergence, the standard

error of $\hat{\beta}$ and $\hat{v} - v$ can be computed from the inverse of the information matrix H^{-1} from the h-likehood and the standard errors of $\hat{\tau}$ are computed from the Hessian of $p_{\beta,v}(h|\tau)$ at τ. Typically there is no explicit formula for this quantity.

This is an extension of the **REML** procedure for joint **GLMs** to linear mixed models. Fitting involves inter-connected component **GLMs**. Each connected **GLM** can be viewed as a joint **GLM**. Then, these joint **GLMs** are connected by an augmented linear model for β and v components.

3.9 Hierarchical Generalized Linear Models (HGLM)

The Hierarchical Generalized Linear Models is a synthesis of three widely used existing model classes; the Generalized Linear Models (McCullagh and Nelder, 1989), Mixed Linear Models having both fixed and random effects (Longford, 1993), and models with structured dispersions (Nelder and Lee, 1991, 1998). The h-likelihood (Lee and Nelder, 1996) is used for inference about fixed and random effects given dispersion components, and an adjusted profile h-likelihood for inference about dispersion components given fixed and random effects. This leads to a reliable and useful estimators; these share properties with those derived from marginal likelihoods, while having the considerable advantage of not requiring the integrating out of random effects.

The algorithm for fitting these models can be reduced to the fitting of a two-dimensional set of generalized linear models; one dimension being mean and dispersion, and the other fixed and random effects, so that no special code is needed for the estimation of the dispersion components. This formulation implies that the model-checking techniques derived for generalised linear models (McCullagh and Nelder, 1989, Chap 12) can be carried over to the wider class. This method

does not require the use of prior probabilities.

3.9.1 The Model

The Hierarchical Generalized Linear Models of Lee and Nelder (1996) are defined as follows. Conditional on random effects u, the response y follows a GLM family, satisfying $E(y/u) = \mu$ and $\text{var}(y/u) = \phi V(\mu)$, for which the kernel of the likelihood is given by

$$\sum \frac{y(\theta) - b(\theta)}{\phi}$$

Where $\theta = \theta(\mu)$ is the canonical parameter. The linear predictor takes the form

$$\eta = g(\mu) = X\beta + Zv, \tag{3.122}$$

Where $v = v(u)$, for some monotone functions $v(.)$, are the random effects and β are the fixed effects. The random component u follows a distribution conjugate to a GLM family of distributions with parameters λ. For clarity of purpose and for the sake of this thesis; suppose that we have responses $y = (y_1, y_2, \ldots, y_n)^T$ and unobserved random variables $u = (u_1, u_2, \ldots, u_q)^T$, having $\text{E}(y_i/(u)) = \mu_{0i}$ and $\text{var}(y_i/(u)) = \phi_i V_0(\mu_{0i})$.

1. Given random effects u, the elements y_i of y follow a generalized linear model, which has likelihood

$$\ell_0(\theta(\mu_0), \phi; y/u) = \sum (\frac{[y_i\theta(\mu_{0i}) - b\theta(\mu_{0i})]}{\phi_i} + k(y_i, \phi_i)) \tag{3.123}$$

   where $\theta(\mu_{0i})$ denotes the canonical parameter and ϕ_i is the dispersion parameter. The linear predictor takes the form

$$\eta_0 = g(\mu_0) = X\beta + Zv \tag{3.124}$$

   Where $u_0 = (u_{01}, u_{02}, \ldots, u_{0n})^T$, $g()$ is the link function, X is the $n \times p$ model

matrix for fixed effects β, and Z is the $n \times q$ model matrix for random effects $v = g_1() = (v_1, v_2, \ldots, v_q)^T$, where $v_i = g_1(u_i)$, for some strictly monotonic function of u_i.

2. The random effects u_i, are independent with dispersion parameters λ_i.

For simplicity of presentation, we use the subscript 0 for y/v components and $1, 2, \ldots$ for the components. We suppress scripts when unnecessary. For simplicity of argument, we first consider a model with one extra random component, though there is no difficulty in generalising this to two or more such components. We let ϕ_i and λ_i vary over units to allow for structured dispersions.

For example, the normal linear mixed models is an HGLM because

1. y/u follows a GLM distribution with $\text{Var}(y/u) = \phi$, with $\phi = \sigma^2$ and $V(\mu) = 1$, $\eta = \mu = X\beta + Zv$, where $v = u$

2. $u \approx N(0, \lambda)$ with $\lambda = \sigma_v^2$.

We call this model the normal-normal HGLM, where the first adjective refers to the distribution of the $y|u$ component and the second to the u component.

3.9.2 H-Likelihood Approach

The h-likelihood, denoted by h, is defined by

$$h = \ell\left(\theta', \phi; y \backslash \nu\right) + \ell(\alpha; \nu), \tag{3.125}$$

where $\ell(\alpha; \nu)$ is the logarithm of the density function for ν with parameter α, and $\ell(\theta', \phi; y|\nu)$ is that for $y|\nu$. The random component ν is the scale on which the random effect u occurs linearly in the linear predictor. It is possible to derive the h-likelihood from density functions of u and $y|u$ as well; $\ell(\alpha; \nu)$ can be derived from the density function of u with differential element $dv(u)$ and $\ell(\theta', \phi; y|v(u))$, the logarithm of the density function for $y|u$, since ν is the strictly monotonic

function of u. The h-likelihood is the logarithm of the joint density function for ν and y. When both distributions are normal the h-likelihood is Henderson's joint likelihood. When one or both of the distributions are non-normal, the h-likelihood is an obvious generalization of the joint likelihood.

Clearly the h-likelihood is not an orthodox likelihood because the ν are not observed. Estimates derived from maximizing the h-likelihood is known as maximum h-likelihood estimates (MHLEs); these are obtained by solving.

$$\begin{aligned} \partial h/\partial \beta &= 0 \\ \partial h/\partial \nu &= 0 \end{aligned} \tag{3.126}$$

From the definition of the h-likelihood (3.125) it is easy to see that the MHLEs for, β given u are obtained by the GLM equations with $\nu(u)$ as an offset. As with maximum likelihood (ML) estimates, the MHLEs for random effects are invariant with respect to the transformation of random components u; for example, estimating equations $\partial h/\partial \nu = 0$ and $\partial h/\partial u = 0$ result in the same random effect estimate.

For normal linear mixed models the classical approach provides sensible inferences about β and the random parameters v; for further discussion, see e.g. Robinson (1991). However, its extension to non-normal models is not straightforward. To prepare for the necessary extensions later, we study here h-likelihood inference for linear mixed models. The general model can be stated equivalently as follows: conditional on v the outcome y is normal with mean

$$E(y|v) = X\beta + Zv \tag{3.127}$$

and variance Σ, and v is normal with mean zero and variance D. From above, the extended loglihood of all the unknown parameters is given by

$$\ell_e(\beta, \tau, v) = log\ f(y|v) + log\ f(v) \tag{3.128}$$

$$= \frac{-1}{2} log\ |2\pi\Sigma| - \frac{1}{2}(y - X\beta - Zv)^t\Sigma^{-1}(y - X\beta - Zv)$$
$$-\frac{1}{2} log\ |2\pi D| - \frac{1}{2}v^t D^{-1}v, \tag{3.129}$$

where the dispersion parameter τ enters via Σ and D. To use the h-likelihood framework, first we need to establish the canonical scale for the random effects. Given the fixed parameters, by maximizing the extended likelihood, we obtain

$$\hat{v} = (Z^t\Sigma^{-1}Z + D^{-1})^{-1}Z^t\Sigma^{-1(y-X\beta)} \tag{3.130}$$

and from the second derivative ℓ_e with respect to v, we get the Fisher information

$$I(\hat{v}) = (Z^t\Sigma^{-1}Z + D^{-1}) \tag{3.131}$$

Since the Fisher information depends on the dispersion parameter τ, but not on β, the scale v is not canonical for τ, but it can be for β. In fact it is the canonical scale. This means that the extended likelihood is an h-likelihood, allowing us to make joint inferences about β and v, but estimation of τ requires a marginal likelihood. Note that $\hat{v}$ is a function of fixed parameters, so that we use notations $\hat{v}$, $\hat{v}(\beta, \tau)$ and $\hat{v}\beta, \tau$ for convenience. This is important when we need to maximize adjusted profile likelihoods. The canonical scale v is unique up to linear transformations.

For non-linear transformations of the random effects, the h-likelihood must be derived following the invariance principle; i.e.,

$$H(\beta, \tau, u(v)) \equiv H(\beta, \tau, v) \tag{3.132}$$

With this, joint inferences of β and v from the h-likelihood are invariant with respect to any monotone transformation (or re-expression) of v. The study compares the h-likelihood inference with the classical approach: All inferences - including those for the random effects - are made within the (extended) likelihood framework, Joint estimation of β and v is possible because v is canonical for β, Estimation of the dispersion parameter requires an adjusted profile likelihood.

For inferences from HGLMs we should define the h-loglihood of the form

$$h \equiv log\ f_{\beta,\phi}(y|v) + log\ f_{\lambda}(v) \tag{3.133}$$

where (ϕ, λ) are dispersion parameters. It can be seen that in normal mixed linear models v is the canonical scale for β. However, this definition is too restrictive because, for example, there may not exist a canonical scale for non-normal GLMMs.

3.9.3 Conjugate Hierarchical Generalized Linear Models

Much current work on this area assumes that the distribution of the random component ν is normal. The normality assumption is convenient when the random components ν are correlated. However, the distribution of ν, or equivalently u, is better decided by the properties of the data or the purposes of inference. So a broader class of hierarchical models is of interest. In generating a new class of HGLMs, we first try for a simple form of random effect estimates. Secondly, if possible, we want to avoid a difficulty with the inference about the population mean $E(y)$ with fixed effects, B. In the GLMM, $t' = E(y|u) = g^{-1}(\eta + \nu)$. So that $\mu = g^{-1}(\eta) \neq Eg^{-}1(\eta+\nu)) = E(y)$, unless the link function η is the identity function with $E(\nu) = 0$.

In multiplicative models, where, $\mu' = \mu u$, this bias can be avoided by having a distribution of u satisfying $E(u) = 1$

For simplicity of argument, let the response be y_{ij} for $i = 1, ..., t$ and $j = 1, ..., n_i$, with $n = \sum n_i$, and u_i the unobserved random components. We define the conjugate HGLM as follows. Consider the canonical link model such that $\theta'_{ij} = \theta_{ij} + \nu_i$, where $\theta'_{ij} = \theta_{ij}(\mu_{ij})$, $\theta_{ij} = \theta(\mu_{ij})$ and $\nu_i = \theta(u_i)$. Then, we have

$$\partial h/\partial \beta_k = \sum_y (y_{ij} - \mu'_{ij}) x_{kij}/\phi. \tag{3.134}$$

Assume that the kernel of $\ell(\alpha; \nu)$

$$\sum_i \{a_1(\alpha)\nu_i - a_2(\alpha)b(\nu_i)\} \tag{3.135}$$

where $a_1()$ and $a_2()$ are some functions of dispersion parameters a. Even though expression (3.135) takes the form of the Bayesian conjugate prior (Cox and Hinkley, 1974), that prior is for θ' itself, whereas for our conjugate distribution it is for ν only; we do not specify priors for ,β, ϕ or α. Then the kernel of the h-likelihood becomes

$$\sum_{ij} \{\theta' y - b(\theta')\} /\phi + \sum_i \{a_1(\alpha)\nu - a_2(\alpha)b(\nu)\}.$$

Since $\partial b(\theta(\mu))/\partial\theta = \mu$ so that $\partial b(\nu)/\partial\nu = u$ we have

$$\partial h/\partial \nu_i = \left\{ \sum_j (y_{ij} - \mu'_{ij}) + \phi a_1(\alpha) \right\} /\phi - a_2(\alpha) u_i.$$

thus equating $\partial h/\partial \nu_i$ to 0 gives an estimate of the random effect

$$\hat{u}_i = \frac{y_i - \mu'_{i+} + \phi a_1(\alpha)}{\phi a_2(\alpha)} \tag{3.136}$$

where $y_{i+} = \sum_j y_{ij}$ and $\mu'_{i+} = \sum_j \mu'_{ij}$. This shows that, in the conjugate HGLMs, the MHLE for the random effects has a simple form on the $u - scale$. If $E(u) = a_1(\alpha)/a_2(\alpha)$, and the fixed effects have an intercept term, then from equations (3.134) and (3.136) we have $\sum \hat{u}_i = a_l(\alpha)/a_2(\alpha)$, analogously to the result for residuals in normal linear models. This section considers various HGLMs in more

detail, in particular the HGLMs with conjugate distributions.

3.9.4 GLM family for the Random Components

A key aspect of HGLMs is the flexible specification of the distribution of the random effects u, which can come from an exponential family with log-density proportional to

$$\Sigma[k_1 c_1(u) + k_2 c_2(u)]$$

for some functions $c_1(.)$ and $c_2(.)$, and parameters k_1 and k_2. The weak canonical scale gives a nice representation of loglihood for random effects, which can be written as

$$\sum \frac{\psi_M \theta_M(u) - b_M(\theta_M(u))}{\lambda} \tag{3.137}$$

for some known functions $\theta_M(u)$ and $b_M(.)$, so that it looks conveniently like the kernel of the GLM family, and choosing a random-effect distribution becomes similar to choosing a GLM model. Examples of these functions based on common distributions are given in Table 3.1. (We use the label M to refer to the mean structure). Allowing for the constraint on $E(u)$ as discussed above, the constant ψ_M takes a certain value, so the family (3.137) is actually indexed by a single parameter λ. Table 3.1 also provides the corresponding values for ψ_M in the different families. As to be demonstrated later, in conjugate distributions we have $E(u) = \psi_M$ and $\text{var}(u) = \rho V_M(\psi_M)$ Recall that the loglihood based on $y|v$ is

$$\sum \frac{y\theta(u) - b(\theta(u))}{\phi}$$

Now, by choosing the specific functions $\theta_M(u) = \theta(u)$ and $b_M(\theta_M) = b(\theta)$, we obtain the conjugate loglihood

$$\sum \frac{\psi_M \theta(u) - b(\theta(u))}{\lambda} \tag{3.138}$$

Table 3.1: GLM family of random components for HGLM

y/v distribution	y/v link	u distribution	u link	model
Normal	identity	Normal	identity	conjugate model
Poisson	log	Gamma	log	conjugate model
Binomial	compl.-log-log	Beta	logit	conjugate model
Gamma	reciprocal	Inverse-gamma	reciprocal	conjugate model
Gamma	log	Inverse-gamma	log	conj with canonical link
Poisson	log	Normal	identity	Ext conjugate model*
Binomial	logit	Normal	identity	Ext conjugate model*
Binomial	comp.-log-log	Gamma	log	Ext conjugate model
Gamma	log	Gamma	log	Ext conjugate model

for the random effects.

Cox and Hinkley (1974) defined the so-called conjugate distribution. We call (3.138) the conjugate loglihood to highlight that it is not a log-density for ψ_M. The corresponding HGLM is called a conjugate HGLM, but there is of course no need to restrict ourselves to such models. It is worth noting that the weak canonical scale of v leads to this nice representation of conjugacy. In conjugate distributions the scale of random effects is not important when they are to be integrated out, while in conjugate likelihood the scale is important, leading to nice inferential procedures.

In principle, various combinations of GLM distribution and link for $y|v$ and a conjugate to any GLM distribution and link for v can be used to construct HGLMs. Examples of useful HGLMs are shown in Table 3.1. Note that the idea allows a quasi-likelihood extension to the specification of the random effects distribution, via specification of the mean and variance function.

$*$ = A generalised linear model with a normal distribution and identity link for the random effects u.

Attribute	Mean model	Dispersion model
	y/v Component	
Response	y	d_0^*
Mean	μ_0	ϕ
Variance	$\phi V_0(\mu_0)$	$2\phi^2$
Link	$g_0(\mu_0)$	$f_0(\phi)$
Linear predictor	$X\beta + Zv$	$G_0\gamma_0$
Deviance components	d_0	$2\{-\log(d_0^*/\phi) + (d_0^* - \phi)/\phi\}$
Prior weight	$1/\phi$	$1 - q_0$
	v Component	
Response	ψ	d_1^*
Mean	u	λ
Variance	$\lambda V_1(u)$	$2\lambda^2$
Link	$g_1(u)$	$f_1(\lambda)$
Linear predictor	v	$G_1\gamma_1$
Deviance components	d_1	$2\{-\log(d_1^*/\lambda) + (d_1^* - \lambda)/\lambda\}$
Prior weight	$1/\lambda$	$1 - q_1$

Figure 3.3: Generalised linear model attributes for hierarchical generalised linear models

3.9.5 Gamma-Inverse Gamma Model

We assume that the conditional distribution of y given u is the gamma distribution with dispersion parameter $\phi = 1/\nu'$ whose log-likelihood has kernel $-\nu \sum_{ij}(y_{ij}\theta'_{ij} - \log \theta'_{ij})$, where $\theta' = 1/\mu'$.

Canonical link models

The conjugate HGLM leads to the model $\theta'_{ij} = \theta_{ij} + \nu_i$, where $\theta_{ij} = 1/\mu_{ij}$, and $\nu_i = l/u_i$. We choose the conjugate distribution (inverse gamma) such that $E(u_i) = 1$. This gives a log-likelihood for ν of the form

$$\ell(\alpha; \nu) = \sum \{-\alpha\nu_i + \alpha \log \nu_i + (\alpha + 1) \log \alpha - \log \Gamma(\alpha + 1)\}. \tag{3.139}$$

The estimating equations for $\hat{u}_i$ are given by

$$\hat{u}_i = \{\nu(y_{i+} - \mu_{i+} + \alpha)\} / \alpha \tag{3.140}$$

and those for β by the standard GLM equations for a canonical link,

$$\sum_{ij}(y_{ij} - \mu'_{ij})x_{kij} = 0 \tag{3.141}$$

As with the corresponding GLMs, we require the linear predictor to be positive to ensure that the mean remains positive.

Log-link Models

With the canonical link, because of the requirement that $\mu'_{ij} > 0$, care should be taken in computing $\hat{\beta}$ and ν'. So, with gamma errors, the log-link is often used. Consider the multiplicative model $\mu' = \mu u$ with $E(u) = 1$. Let $\eta'_{ij} = \log \mu'_{ij} = \log \mu_{ij} + \nu_i$, where $\log \mu = X\beta$ and $\nu = \log u$. Suppose that u_i has the inverse

gamma density function,

$$\frac{1}{\Gamma(\alpha+1)}\left(\frac{\alpha}{u_i}\right)^{\alpha+1}\exp\left(-\frac{\alpha}{u_i}\right)d(\log u_i) \tag{3.142}$$

so that the kernel of h becomes

$$\nu\sum_{ij}\left\{-\frac{y_{ij}}{\mu'_{ij}}+\log\left(\frac{\nu y_{ij}}{\mu'_{ij}}\right)\right\}+\sum_{i}\left\{-(\alpha+1)\nu_i-\frac{\alpha}{u_i}\right\}.$$

The MHL equations for β are

$$\frac{\partial h}{\partial\beta_k}=\nu\sum_{ij}\left(\frac{y_{ij}-\mu'_{ij}}{\mu'_{ij}}\right)x_{kij}=\nu\sum_{ij}\left(\frac{y_{ij}/u_i-\mu_{ij}}{\mu_{ij}}\right)x_{kij}=0 \tag{3.143}$$

from $\partial h/\partial\nu_i=\sum_j(\nu y_{ij}/\mu_{ij}u_i-\nu)-(\alpha+1)+\alpha/u_i$, we obtain

$$\hat{u_i}=\frac{\nu\sum\limits_j y_{ij}/\mu_{ij}+\alpha}{\sum\limits_j\nu+\alpha+1} \tag{3.144}$$

The next Section shows that the MHLE for β is the same as the marginal ML estimator. Here it is the $\frac{1}{\mu}$ scale on which the random effects μ occur linearly in the estimating equations(3.143) and

$$E(\frac{1}{\mu_i}|y)=\frac{\sum_j\upsilon+\alpha+1}{\upsilon\sum_j\frac{y_{ij}}{\mu_{ij}}+\alpha} \tag{3.145}$$

In the normal-normal mixed model, $\mu=X\beta$ are location parameters and the random effects estimators are location invariant. The above multiplicative model can be written as $logy=log\mu+logu+loge$, where $e\sim\Gamma(1,\upsilon)$; here μ are scale parameters and equation(3.144) shows that the random effects estimators are also scale invariant.

3.9.6 Inverse Gaussian-Gamma Model

The researcher assume that the conditional distribution of y given μ is the inverse Gaussian distribution with dispersion parameter $\phi = 1/\nu$ whose log-likelihood has kernel $-\nu\Sigma_{ij}\{\frac{y_{ij}\theta_{ij}^{1}}{2} - \theta_{ij}^{,1/2}\}$, where $\theta^{'} = \frac{1}{\mu^{'2}}$

3.9.7 Canonical link models

The conjugate HGLM leads to the model $\theta_{ij}^{'} = \theta_{ij} + \nu_i$, where $\theta_{ij} = 1/\mu_{ij}^2$ and $\nu_i = 1/\mu_i^2$. Here the kernel of $\ell(\alpha;\nu)$ for the conjugate likelihood is $\sum\{-\alpha_1\nu_i + \alpha_2(2\nu_i)^{-1/2}\}$, giving $\hat{\mu_i} = \{\alpha_1 - \frac{1}{2}\nu(y_{i+} - \mu_{i+}^{'})\}/\alpha_2$. For the inverse Gaussian distribution, the conjugate distribution is not unique but depends on the parametrization see Consonni and Veronese(1992). Suppose that ν_i has a gamma distribution, giving

$$\ell(\alpha;\nu) = \sum\{(\alpha-1)log\nu + \alpha log\alpha - \alpha\nu - log\Gamma(\alpha)\} \tag{3.146}$$

Here $\mu^{'2} = \mu^2/(1+\mu^2\nu)$ so that the MHL equation for ν becomes

$$\frac{\partial h}{\partial \nu_i} = -\nu(y_{i+} - \mu_{i+}^{'})/\nu_i - \alpha = 0 \tag{3.147}$$

whence $1/\hat{\nu_i} = \{\frac{1}{2}\nu(y_{i+} - \mu_{i+}^{'}) + \alpha\}/(\alpha - 1)$. Thus a simple form of random effect estimators is possible with distributions other than the conjugate distribution.

3.9.8 Log-link models

With inverse Gaussian errors, the log-link may often be used, as with gamma errors. Consider the multiplicative model $\mu^{'} = \mu$ with log-link so that we have $\eta_{ij}^{'} = log\mu_{ij}^{'} = \eta^{ij} + \nu_i$, where $\eta = log\mu = X\beta$, $\nu = log\mu$. Suppose that μ_i has the inverse gamma distribution with $E(\mu_i) = 1$ so that $\ell(\alpha;\nu)$ has kernel $\sum\{-(\alpha+1)\nu_i - \alpha/\mu_i\}$. Now let $r_i = l/\mu_i$; then $\mu_{ij}^{'} = \mu_{ij}/r_i$. Since $\ell(\beta,\phi;y/\mu)$ has kernel $-\nu\sum_{ij}(y_{ij}r_i^2/2\mu_{ij}^2 - r_i/\mu_{ij})$, the MHLE for r is easy to derive and is

given by

$$\frac{\partial h}{\partial r_i} = -\nu \sum_j (\frac{r_i y_{ij}}{\mu_{ij}^2} - \frac{1}{\mu_{ij}}) + \frac{\alpha+1}{r_i} - \alpha$$

$$= -\nu \sum_j \frac{(y_{ij} - \mu_{ij}^{'})/\mu_{ij}^{'2}}{r_i} + \frac{\alpha+1}{r_i} - \alpha$$

$$= 0$$

so we have

$$\hat{r_i} = \frac{1}{\hat{\mu}_i} = \frac{\alpha + 1 - \nu \sum_j (y_{ij} - \mu_{ij}^{'})/\mu_{ij}^{'2}}{\alpha} \tag{3.148}$$

Now suppose that $r_i \sim N(d, 1/\alpha)$. Desmond and Chapman (1993) considered a model similar to this, since it allows an explicit marginal likelihood. Here $\nu = -logr$ is not defined for $r \leq 0$ and the Jacobian appears awkward. Further, both $\ell(\beta, \phi, y/\mu)$ and $\ell(\alpha; r)$, the likelihood derived from the density of r directly, are quadratic functions of r so here it seems natural to use $\ell(\alpha; r)$ instead of $\iota(\alpha; \nu)$ in forming the h-likelihood, because then we have a $\partial h/\partial r_i = -\sum_j v(r_i y_{ij}/\mu_{ij}^2 - 1/\mu_{ij}) - \alpha r_i + d\alpha$, so that

$$\hat{r}_i = \frac{d\alpha + \sum_j \nu/\mu_{ij}}{\nu \sum_j y_{ij}/\mu_{ij}^2 + \alpha} \tag{3.149}$$

In this model we can show that the conditional likelihood of $r|y$ is a normal distribution such that $E(r_i|y)$ is the right-hand side of the above equation and

$$var(r_i/y) = \frac{1}{\nu \sum_j y_{ij}/\mu_{ij}^2 + \alpha} \tag{3.150}$$

3.10 Properties of Maximum h-likelihood estimates

Consider the hierarchical model

$$y|\nu \sim f_1(y|\nu, \beta, \phi) \tag{3.151}$$

$$\nu \sim f_2(\nu|\alpha),$$

where f_1 and f_2 are arbitrary density functions of $y|\nu$ and ν respectively. Assume that ϕ and α are given and, β are parameters of interest. The h-likelihood $h(\beta, \phi, \alpha; y, \nu)$ has components $\ell(\beta, \phi; y|\nu) = log f_1(y|\nu, \beta, \phi)$ and $\ell(\alpha; \nu) = log f_2(\nu|\alpha)$. It can also be written in the form

$$h = \ell(\beta, \phi; y|\nu) + \ell(\alpha; \nu) = L + \ell(\beta, \phi, \alpha; \nu|y) \tag{3.152}$$

where L is the marginal likelihood and

$$\ell(\beta, \phi, \alpha; \nu|y) = log\{f_1(y|\nu, \beta, \phi) f_2(\nu|\alpha) / \int f_1(y|\nu, \beta, \phi) f_2(\nu|\alpha) d\nu\} \tag{3.153}$$

is the logarithm of the density function of v/y.

For GLMMs, Breslow and Clayton(1993) investigated the Laplace approximation for the marginal likelihood L:

$$\hat{L}\alpha A(B) + h(\beta, \phi, \alpha; y, \tilde{\nu}). \tag{3.154}$$

Here $A(\beta) = -\frac{1}{2} log\{det(D^*)\}$ where D^* is the matrix whose ij^{th} element is $\partial^2 h / \partial \nu_i \partial \nu_j|_{\nu=\tilde{\nu}}$, and ν is a solution of the equations $\partial h / \partial \nu = 0$ given β. Breslow and Clayton(1993) showed that the Laplace approximation(3.154) is very accurate for likelihood-based inferences in GLMMs with common group means and one ν component. Breslow and Clayton(1993) again extended these results to models with arbitrary fixed effects and more than one ν-component. Consider the approximate marginal ML estimating equations

$$\partial \hat{L} / \partial \beta = \partial A(\beta) / \partial \beta + \partial h(\beta; y, \tilde{\nu}) / \partial \beta = 0 \tag{3.155}$$

In GLMMs, Breslow and Clayton(1993) showed that $A(\beta)$ depends on β through

the GLM weight function. Assuming that this GLM weight varied slowly as a function of β, the y proposed to ignore the term $\partial A(\beta)/\partial\beta$ in obtaining the marginal ML estimate. Then the second term in the above equations becomes

$$\partial h(\beta; y, \tilde{\nu})/\partial\beta \approx \partial h/\partial\beta|_{\nu=\tilde{\nu}} + \partial h/\partial\nu|_{\nu=\tilde{\nu}}(\partial\tilde{\nu}/\partial\beta) = 0 \tag{3.156}$$

Here they ignored the second term and justified the method as jointly maximizing Green's(1987) penalized quasi-likelihood and also as maximizing the Bayes posterior distribution. Thus for GLMMs, Breslow and Clayton(1993) recommended the MHLEs for β as an approximate marginal ML estimator.

3.10.1 Fixed Vrs Random Effects

Fixed effects can describe systematic mean patterns such as trend, while random effects may describe either correlation patterns between repeated measures within subjects or heterogeneities between subjects or both. The correlation can be represented by random effects for subjects, and heterogeneities by saturated random effects. In practice, it is often necessary to have both types of random components. However, sometimes it may not be obvious whether effects are to be treated as fixed or random. For example, there has been much debate among econometricians about two alternative specifications of fixed and random effects in mixed linear models: see Baltagi (1995) and Hsiao (1995).

When y_i are random, the ordinary least-square estimator for β, treating v_i as fixed, is in general not fully efficient, but is consistent under wide conditions. By contrast, estimators for β, treating v_i as random, can be biased if random effects and covariates are correlated (Hausman, 1978). Thus, even if random effects are an appropriate description for v_i one may still prefer to treat the v_i as fixed unless the assumptions about the random effects can be confirmed. Without sufficient

random effects to check their assumed distribution it may be better to treat them as fixed.

This produces what is known as the intra-block analysis, and such an analysis is robust against assumptions about the random effects in normal linear mixed models. Econometrics models are mainly based upon the normality assumption. However, with binary data the robustness property of intra-block estimators no longer holds. In general there is no guarantee that the intra-block analysis will be robust.

3.10.2 Random Effect Estimation

For any function $r()$ let $r = r(u)$. The quantity $\delta = E(r|y)$ is the best unbiased predictor for a random effect r in the sense that

$$E(\delta)E_y\{E(r|y)\} = E(r) \tag{3.157}$$

and it has the minimum mean-square error of prediction

$$E(\delta - r)^{'} P(\delta - r)$$

for any positive definite matrix P; see Searle et al. (1992). Given β, let $\tilde{\mu}$ be a solution of the equations $\partial h/\partial \nu = 0$. Under the normal-normal mixed model $\nu = E(v|y)$ and ν is linear in y; hence ν is called the best linear unbiased predictor(BLUP).

We have seen that $\frac{1}{\tilde{\mu}}$ is the best unbiased predictor for the gamma-inverse gamma and inverse Gaussian-normal model with log-link; see equation(3.141) and the sentence following equation (3.149). Now, under appropriate conditions, we show that $\tilde{r}$ and $\hat{r} = \tilde{r}|_{\beta=\hat{\beta}}$ converge to $E(r|y)$, so that the MHLEs for r are asymptotical best unbiased predictors. Asymptotic arguments can be derived for any strictly

monotonic transformation of u. Given ϕ and α, the Laplace approximation for expression(3.154) is based on the expansion of h:

$$\hat{h}\alpha h(\beta,\phi,\alpha;y,\tilde{\nu})-(\tilde{\nu}-\nu)^{'}D^{*}(\tilde{\nu}-\nu)/2=\hat{L}+\hat{\ell}(\beta,\phi,\alpha;\nu|y) \tag{3.158}$$

Ignoring the constant term, which depends only on y and not on v, expressions (3.154) and (3.158) imply that

$$\nu|y\sim N(\tilde{\nu},D^{*-1}) \tag{3.159}$$

would be a good approximation for the distribution of $v|y$. If so, given, β, the solution $\tilde{\nu}$ of $\partial h/\partial v=0$ is approximately $E(v|y)$. The convergence rate of D^{*-1} in expression (3.159) plays a crucial role in the asymptotic properties of the MHLEs obtained from equations(3.126). We assume earlier that $D_i^{*-1}=O_p(n^{-1})$ for all i, where n is the sample size, the standard condition for the Laplace approximation. For simplicity of argument, we consider again HGLMs with one v-component where $D^{(}*)$ is a diagonal matrix with the ith element $D_i^{*}=-\partial^2 h/\partial\nu_i^2|_{\mu=\tilde{\mu}}$ and $n=\sum n_i$. It is true that $D_i^{*-1}=O_p(n^{-1})$ if the number of groups t remains the same but within-group sample sizes $n_i\longrightarrow\infty$ at the same rate. Lee and Nelder (1996) show that

$$\tilde{\nu}_i=E(\nu_i|y)+0_p(n^{-1}) \tag{3.160}$$

Similarly, they showed that

$$var(\nu_i|y)=D_i^{*-1}\{1+0_p(n^{-1})\} \tag{3.161}$$

and that equations (3.160) and (3.161) often hold exactly on some scale of μ. ; for example, on the μ-scale, $iu=E(\mu|y)$ and $var(\mu|y)=D^{*-1}$ with $D_i^{*}=-\partial^2 h/\partial\mu_i^2|_{\mu=\tilde{\mu}}$, for the Poisson-gamma model, and, on the $r=1/\mu$ scale, $\tilde{r}=E(r|y)$ and $var(r|y)=D^{*-1}$ with $D_i^{*}=-\partial^2 h/\partial r_i^2|_{\mu=\tilde{\mu}}$, for the multiplicative inverse Gaussian-normal models.

3.10.3 Fixed Effect Estimation

According to the EM algorithm of Dempster et al. (1977), in which the us are the missing data, the (marginal) ML estimate for, β can be obtained by solving

$$E\{\partial h/\partial\beta|y,\hat{\beta}^{(p)}\}=0, \tag{3.162}$$

where $\beta^{(p)}$ is the estimate of β from the previous iteration. However, in general the EM algorithm is difficult to apply because the conditional expectation is difficult to evaluate. Having shown that the estimating equations(3.162) become

$$\partial h/\partial\beta|_{\nu=\hat{\nu}}=0$$

for the Poisson-gamma model, we consider now the gamma-inverse gamma model with log-link. From equation (3.143), the term l/μ_i appears in the equation $\partial h/\partial\beta=0$. From equation (3.145) it is obvious that equation(3.143) becomes the EM equation (3.162): see equation (3.144). This equivalence is well known for normal-normal models. Thus in these three models the marginal ML estimators for, β are the same as the MHLEs. The alternation between the pair of estimating equations in-equation (3.126) is of the EM type for the marginal ML estimator for, β, in which

1. given μ, solving $\partial h/\partial\beta=0$ is the M-step and

2. given, β, solving $\partial h/\partial\beta=0$ is the E-step.

Actually (1) is the first-order approximation to equation(3.162) at $\hat{\nu}$. This argument shows that the MHLE, $\hat{\beta}$ differs from the marginal ML estimator only slightly. Lee and Nelder (1996) showed that the difference is $O_p(n^{-1})$ and hence both have asymptotically a common variance. Since the(marginal) ML estimator for, β is asymptotically most efficient, so is the MHLE for, β.

3.10.4 Covariance Estimators of Maximum h-likelihood Estimates

Let

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = n \begin{pmatrix} var(\hat{\beta}) & cov(\hat{\beta}, \hat{\nu} - \nu) \\ cov(\hat{\nu} - \nu, \hat{\beta}) & var(\hat{\nu} - \nu) \end{pmatrix}, \tag{3.163}$$

$$M = \frac{1}{n} \begin{pmatrix} B & C \\ C^T & D \end{pmatrix}, \tag{3.164}$$

where B, C and D are matrices such that the ij^{th} element of B is

$$-\partial^2 h / \partial \beta_i \partial \beta_j |_{\beta=\hat{\beta}, \nu=\hat{\nu}}$$

the jk^{th} element of C is

$$-\partial^2 h / \partial \beta_j \partial \nu_k |_{\beta=\hat{\beta}, \nu=\hat{\nu}}$$

and the ij^{th} element of D is

$$-\partial^2 h / \partial \nu_i \partial \nu_j |_{\beta=\hat{\beta}, \nu=\hat{\nu}}$$

If E(M) is non-singular, under appropriate regularity conditions M^{-1} converges to V as $n \longrightarrow \infty$. This holds when entries of M are replaced by corresponding expectations since B, C and D are sums of matrices. If the model matrix X has full column rank E(M) is non-singular; see the Hessian matrix (in next Section). It is always possible to select the columns of X to have full column rank.

Given data y, the realized values of random effects v are fixed constants, so the h-likelihood can be treated as if it were an orthodox likelihood in which ν are regarded as fixed parameters for realized but unobserved values of random ef-

fects. This result leads to the first-order asymptotics for the estimators of β and ν. According to Harville (1976), the estimates of $cov(\hat{\beta}, \hat{\nu} - \nu)$ and $var(\hat{\nu} - \nu)$ are useful for making inferences about realized or sample values of ν.

Now consider the likelihood-ratio-type test statistics for fixed effects, testing $/beta = \beta_o$. Given dispersion parameters ϕ and α, we may use the test statistic

$$T_f = 2[h(\hat{\beta}, \phi, \alpha; y, \hat{\nu}) - h\{\beta_o, \phi, \alpha; y, \hat{\nu}(\beta_o)\}], \tag{3.165}$$

where $\hat{\nu}(\beta_o)$ is the solution of

$$\partial h/\partial \nu|_{\beta=\beta_o} = 0$$

An alternative, using the Laplace approximation of L, expression (3.154), is the test statistic

$$T_a = 2\{\hat{L}(\hat{\beta}) - \hat{\beta}_o\} = T_f + 2\{A(\hat{\beta}) - A(\beta_o)\}. \tag{3.166}$$

As Liu and Pierce (1994) have pointed out for GLMMs, the term $A(\hat{\beta}) - A(/\beta_o)$ is the nuisance parameter adjustment of Cox and Reid (1987) when v are fixed nuisance parameters orthogonal to β. But the results of this section show that they are not orthogonal so we are uncertain which of T_f and T_a is better. Under the hypothesis $\beta = \beta_o$ the adjustment term in-equation (3.167) is asymptotically negligible: see the definition of $A(\hat{\beta})$ in expression (3.154) and the discussion in Breslow and Clayton (1993). Using arguments from Cox and Hinkley (1974), the expansion of T_f leads to asymptotically equivalent test statistics

$$(\beta - \beta_o)^T I^{\beta,\beta}(\hat{\beta}, \hat{\nu})^{-1}(\hat{\beta} - \beta_o)$$

and

$$(\hat{\beta} - \beta_o)^T I^{\beta\beta}(\beta_o, \hat{\nu}(\beta_o))^{-1}(\hat{\beta} - \beta_o),$$

where $I^{\beta,\beta}$ is a consistent estimator of $var(\hat{\beta})$ in the result. So if the limiting distribution of β is normal the X^2-test would be based on T_f . As shown in earlier sections, $\hat{\beta}$ differs from the marginal ML estimator by $O_p(n^{-1})$ so the asymptotic normality for β holds when the marginal ML estimator is asymptotically normal.

3.10.5 Inference Procedure

From the h-loglihood we have two useful adjusted profile loglihoods: the marginal loglihood and the restricted loglihood.

$$log\ f_{\phi,\lambda}(y|\hat{\beta}),$$

where $\hat{\beta}$ is the marginal ML estimator given $\tau = (\phi, \lambda)$. Following Cox and Reid (1987), the restricted loglihood can be approximated by

$$p_\beta(\ell|\phi, \lambda).$$

In principle we should use the h-loglihood h for inferences about v, the marginal-loglihood ℓ for β and the restricted loglihood $log\ f_{\phi,\lambda}(y|\hat{\beta})$, for the dispersion parameters. If the restricted loglihood is hard to obtain we may use the adjusted profile likelihood $p_\beta(\ell)$. When ℓ is numerically hard to obtain, Lee and Nelder (1996, 2001) proposed to use $p_v(h)$ as an approximation to ℓ and $p_{\beta,v}(h)$ as an approximation to $p_\beta(\ell)$, and therefore to $log\ f_{\phi,\lambda}(y|\hat{\beta})$, $p_{\beta,v}(h)$ gives approximate restricted ML (REML) estimators for the dispersion parameters and $p_v(h)$ approximate ML estimators for the location parameters. Because $log\ f_{\phi,\lambda}(y|\hat{\beta})$ has no explicit form except in normal mixed models, in this thesis we call dispersion estimators that maximize $p_{\beta,v}(h)$ the REML estimators.

3.11 Score Equations for Fixed and Random Effect Estimators

In this section the study seeks an efficient score algorithm for, β and ν given ϕ and α. One reason for developing an algorithm for the ν-scale rather than for the μ-scale is that ν can often assume any real value where as μ usually has range restrictions which may cause problems in convergence.

Let h have kernel

$$\sum_{ij}\{\phi^{'}y - b(\phi^{'})\}/\phi + \sum_{i}\iota(\alpha;\nu_i) \tag{3.167}$$

and

$$\eta_{\prime} = g(u^{'}) = X\beta + Z\nu \tag{3.168}$$

where Z is the $n \times t$ group indicator matrix whose $(ij,k)^{th}$ element is $\partial\eta_{ij}^{'}/\partial\nu k$ and ν is the $t \times 1$ vector whose i^{th} element is ν_i. The score equations become

$$\phi(\partial h/\partial\beta_k) = \sum W(y-\mu^{'})(\partial\eta^{'}/\partial\mu^{'})x_k = 0, \tag{3.169}$$

$$\phi(\partial h/\partial\nu_i) = \sum W(y-\mu^{'})(\partial\eta^{'}/\partial\mu^{'})z_i + \phi\{\partial\iota(\alpha;\nu/\partial\nu_i)\} = 0, \tag{3.170}$$

where W is the GLM weight function, $W = (\partial\mu^{'}/\partial\eta^{'})^2 V(\mu^{'})^{-1}$. By taking conditional expectations of the second derivatives analogously to Fisher scoring, we have

$$-\phi E(\partial h/\partial\beta_k\partial\beta_s|\nu) = \sum W x_s x_k,$$

$$-\phi E(\partial h/\partial\beta_k\partial\beta_s|\nu) = \sum W x_k x_s,$$

$$-\phi E(\partial^2 h/\partial\nu_s\partial\nu_k|\nu) = \sum W z_s z_k - \phi\{\partial^2\iota(\alpha;\nu)/\partial\nu_s\partial\nu_k\}.$$

The corresponding expected Hessian matrix H/ϕ can be written as

$$H = \begin{pmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z + U \end{pmatrix} \tag{3.171}$$

where U is a $t \times t$ diagonal matrix whose i^{th} element is $-\phi\{\partial^2 \iota(\alpha;\nu)/\partial \nu_i^2\}$, with the off-diagonal elements being 0 because the ν_i are independent. So analogously to the derivation on page 42 of McCullagh and Nelder (1989) we have score equations

$$\begin{pmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z \end{pmatrix} \begin{pmatrix} \beta + \delta\beta \\ v + \delta v \end{pmatrix} = \begin{pmatrix} X^T W w \\ Z^T W w \end{pmatrix} \tag{3.172}$$

where w is the GLM adjusted dependent variable

$$w = \eta\prime + (y - \mu\prime)(\partial\eta\prime/\partial\mu\mu\prime) \tag{3.173}$$

and

$$R = Uv + \phi\{\partial l(\alpha; v)/\partial(v)\} \tag{3.174}$$

When the v have a normal distribution with mean 0, $R = 0$, and the score equations become Henderson's (1975) mixed model equations for normal-normal models. On the basis of various approximations, the MHLE score equations (3.173) for GLMMs have been derived by many researchers, e.g.Breslow and Clayton (1993), Wolfinger (1993), Engel and Keen (1994) and McGilchrist (1994). Equations (3.173) are equivalent to

$$\begin{pmatrix} X^{*T} W^* X^* & X^{*T} W^* Z^* \\ Z^{*T} W^* X^* & Z^{*T} W^* Z^* \end{pmatrix} \begin{pmatrix} \beta + \delta\beta \\ v + \delta v \end{pmatrix} = \begin{pmatrix} X^{*T} W^* w^* \\ Z^{*T} W^* w^* \end{pmatrix} \tag{3.175}$$

where

$$X^* = \begin{pmatrix} X \\ 0 \end{pmatrix}, \quad Z^* = \begin{pmatrix} Z \\ I \end{pmatrix}, \quad w^* = \begin{pmatrix} w \\ U^{-1}R \end{pmatrix}, \quad W^* = \begin{pmatrix} W & 0 \\ 0 & U \end{pmatrix} \tag{3.176}$$

This is equivalent to augmenting the model matrices X and Z with extra rows, one for each random component, with corresponding extensions to the adjusted dependent variable and weight matrix. In conjugate HGLMs, from the equation preceding equation (3.136), we have

$$\partial l(\alpha : v_i)/\partial v_i = a_1(\alpha) - a_2(\alpha)u_i = -a_2(\alpha)u_i - E(u_i) \tag{3.177}$$

and

$$\partial^2 l(\alpha : v_i)/\partial v_i^2 = -a_2(\alpha)V(u_i)$$

since

$$E(u) = a_1(\alpha)/a_2(\alpha)$$

and

$$\partial(u_i)/\partial v_i = V(u_i)$$

.

Here H/ϕ is the actual not the expected Hessian matrix, and U is the $t \times t$ diagonal matrix with the ith element $\phi a_2(\alpha)V(u_i)$ and

$$R = Uv - \phi a_2\{u - E(u)\} = U[v - (\partial v/partialu)\{u - E(u)\}] \tag{3.178}$$

Note that

$$v - (\partial v/partialu)\{u - E(u)\}$$

is $E(v)$ to a first-order approximation so that R may be negligible when $E(v =$ $)=0$. When $a_2(\alpha) = 0$, both U and R become null so that equations (3.173) become the ordinary GLM score equations. When $a_2(\alpha) = \infty, \hat{v} = \theta\{E(u)\}$; see the discussions below equations (3.136). In conjugate HGLMs the equations (3.176) have the following nice interpretation.

Consider artificial data $y^{*T} = (y^T, E(u)^T)$ from GLMs such that

$$\mu^{*T} = E(y^{*T}) = (\mu'^T), \theta^* = X^*\beta + Z^*v \tag{3.179}$$

$$var(y^*) = \phi diag\{V(\mu\prime), V(u)/\phi a_2(\alpha)\} \tag{3.180}$$

Then we can show that

$$w^* = \theta^* + (y^* - \mu^*)(\partial\theta^*/\partial\mu^*)$$

is an augmented adjusted dependent variable where

$$\partial\mu^*/\partial\theta^* = diag\{V(\mu\prime), V(u)\}$$

.

From the arguments above we have $\hat{\phi}H^{-1}$ as the estimator of the covariance matrix of $\hat{\beta}$ and $\hat{v} - v$; see the result in subsection 3.10.4. When the realized value of v are known, $\hat{\phi}(X^TWX)^{-1}$ is the estimate of var(β) takes account of the information loss caused by estimating the random effects, but not that caused by estimating the variance components α. However, this information loss will be negligible because of the asymptotic orthogonality to be shown in the next two subsections. Numerical studies for GLMMs by Breslow and Clayton (1993) and McGilchrist (1994) support this near-orthogonality.

3.11.1 Scaled Deviance Test

For the goodness-of-fit criterion,the scaled deviance is defined by

$$D(y, \hat{\mu}) = -2\{l(\hat{\mu\prime}, \phi; y|v) - l(y, \phi; y|v)\} \tag{3.181}$$

with the estimated degrees of freedom $n - trace(H^{-1}H^*)$ where

$$H^* = \begin{pmatrix} X^T W X & X^T W Z \\ Z^T W Z & Z^T W Z \end{pmatrix} \tag{3.182}$$

In the absence of random components v, this becomes the scaled deviance of GLMs, $D(y, \hat{\mu}) = -2\{l(\hat{\mu}, \phi; y) - l(y, \phi; y)\}$ with its degrees of freedom $n - rank(X)$.

Suppose that v_i have a distribution such that $\text{var}(v_i) = \sigma_v^2$ and $E(v) \longrightarrow 0$ as $\sigma_v^2 \longrightarrow 0$. Earlier we show by examples that the MHLEs for v becomes the fixed effect estimates when $\sigma_v^2 = \infty$ and the zero estimates when $\sigma_v^2 = 0$, i.e there are no random effects. Here, as $\sigma_v^2 \longrightarrow \infty$, the estimated degrees of freedom go to $n - rank(X)$ and as $\sigma_v^2 \longrightarrow \infty$ to $n - rank(X, Z)$. Lee and Nelder (1996) we show that $E\{D(y, \hat{\mu\prime})\}$ can be estimated by the estimated degrees of freedom. Thus, if the computed scaled deviance is much larger than the estimated degrees of freedom we may suspect the absence of some necessary fixed or random effects in the linear predictor $\eta\prime$, or over dispersion in the $y|v$ distribution. The scaled deviance uses the distribution of $y|v$ only, so it cannot be used for testing dispersion components.

3.12 Estimation of Dispersion Components

For estimation of dispersion components, the (marginal) ML estimator may be substantially biased owing to the estimation of β. For normal-normal mod-

els,Patterson and Thompson (1971) introduced restricted (marginal) likelihood to yield the REML estimator. Breslow and Clayton (1993) extended this approach to GLMMs by using the normal likelihood. We consider an adjusted h-likelihood

$$h_A = h + \frac{1}{2}\log\{det(2\pi\phi H^{-1})\} \tag{3.183}$$

It can be shown that in normal-normal models Patterson and Thompson's (1971) restricted likelihood is equivalent to the adjusted profile h-likelihood (APHL)

$$h_P = h_A|_{\beta=\hat{\beta},v=\hat{v}}; \tag{3.184}$$

This is also Cox and Reid's (1987) adjusted profile likelihood of dispersion components (ϕ, α) with nuisance parameters (β, v), when they are asymptotically orthogonal. From equations (3.170) and (3.171), we have

$$n^{-1}E(\partial^2 h/\partial\beta\partial\phi) = n^{-1}E(\partial^2 h/\partial\beta\partial\alpha) = n^{-1}E(\partial^2 h/\partial v\partial\phi) = 0 \tag{3.185}$$

Note that

$$n^{-1}E(\partial^2 h/\partial v\partial\alpha) = n^{-1}E(\partial^2 l(\alpha, v)/\partial v\partial\alpha) = 0(1/n)$$

and is 0 in conjugate HGLMs: see equation (3.182). So in the h-likelihood (ϕ, α) and (β, v) are at least asymptotically orthogonal.

We can derive the maximum adjusted profile h-likelihood estimators (MAPHLEs) for dispersion parameters by solving iterative

$$\partial h_A/\partial\alpha|_{\beta=\hat{\beta},v=\hat{v}} = 0$$

and

$$\partial h_A/\partial\phi|_{\beta=\hat{\beta},v=\hat{v}} = 0$$

where $\hat{\beta}$ and $\hat{u}$ are re-evaluate in each iteration. Since

$$\partial[\log\{det(H)\}]/\partial\theta = trace\{H^{-1}(\partial H/\partial\theta)\}, \tag{3.186}$$

where $\theta = \alpha$ or $\theta = \phi$ and K is the matrix given by the bottom right-hand corner of H^{-1}, the score equations are

$$\partial h_A/\partial\alpha = \partial l(\alpha; v)/\partial\alpha - \frac{1}{2}trace\{K(\partial U/\partial\alpha)\} \tag{3.187}$$

$$\partial h_A/\partial\phi = \partial l(\beta, \phi; y|v)/\partial\phi + \frac{1}{2}[\{t + rank(X)\}]/\phi - trace\{K(\partial U/\partial\phi)\} \tag{3.188}$$

The score equations can be solved by the Newton method because the Hessian matrix can be easily obtained by using the fact that

$$\frac{\partial[trace\{H^{-1}(\partial H/\partial\theta)\}]}{\partial\theta} = trace\{H{-}1(\partial^2 H/\partial\theta^2)\}{-}trace\{H^{-1}(\partial H/\partial\theta)H^{-1}(\partial H/\partial\theta)\}$$

In GLMs, expression (3.188) gives the ML equations with a degrees-of-freedom adjustment. For example, if $h(\beta, \phi; y)$ is normal it provides the unbiased estimator for ϕ. In GLMMs where $v \sim N(O, \sigma^2)$, if we let $\alpha = \sigma_v^2$, it can be shown that

$$\partial h_A/\partial\alpha|_{\beta=\hat{\beta},v=\hat{\beta}} = (t - v)/\sigma_v - \sum \hat{v}_i^2/\sigma_v^3 \tag{3.189}$$

with $v = (\phi/\sigma_v^3)trace(K)$; this yields McGilchrist's (1994) REML estimator for σ_v^2. Further, we have

$$\partial h_A/\partial\phi = \partial l(\beta, \phi : y|v)/\partial\phi|_{v=\hat{v}} + \frac{1}{2}\{t + rank(X) - v\}/\phi \tag{3.190}$$

If we now approximate $l(\beta, \phi; y|v)$ by a pseudo-likelihood

$$P = -\frac{1}{2}S/\phi - \sum \frac{1}{2}\log\{2\pi\phi V(\mu\prime_{ij})\} \tag{3.191}$$

where S is the Pearson χ^2. We obtain the MeGitchrist (1994) REML estimator for ϕ. This is equivalent to assuming that the adjusted dependent variables w have normal distributions; see also Wolfinger (1993). Nelder and Lee (1992) showed that the dispersion estimator based on the Pearson χ_2 is often inefficient unless the underlying distribution is normal. So we may expect equation (3.153) to provide a better estimator.

Harville (1977) and Speed (1991) pointed out that in normal-normal models the REML estimators can be derived by equating observed and expected sums of squares. Note that the first terms in equations (3.187) and (3.188) evaluated at $\hat{\beta}$ and $\hat{v}$ are proportional to the observed sum of squares in normal-normal models. Lee and Nelder (1996) showed that the remaining terms in equations (3.187) and (3.188) are corresponding expectations of the first terms so that their result generalizes.

3.13 Generalizations

Consider now HGLMs with more than one extra random component, so that

$$\eta\prime = X\beta + Z_1 v^{(1)} + Z_2 v^{(2)} + ... + Z_k v^{(k)} \tag{3.192}$$

where Z_i is the $n \times q_i$ model matrix, $v^{(i)}$ are the $q_i \times 1$ random effects and $v^{(i)}$ and $v^{(j)}$ are independent if $i \neq j$. Let $Z = (Z_1, Z_2, ..., Z_k)$, $v = (v^{(1)T}, v^{(2)T}, ..., v^{(k)T})^T$ and $q = \sum q_i$; then model (3.154) be written as $\eta\prime = X\beta + Zv$ as in previous sections. Here $l(\alpha; v)$ in the h-likelihood becomes $\sum l(\alpha_i; v^i)$, If the distributions of $v^{(i)}$ are specified the generalizations of the procedures developed above are straightforward. For example, as in the score equations (3.173), the score

equations for β and v become

$$\begin{pmatrix} X^TWX & X^TWZ_1 & \dots & X^TWZ_k \\ Z^TWX & Z_1^TWZ_1+U_1 & \dots & Z_1^TWZ_k \\ \vdots & \vdots & \ddots & \vdots \\ Z_k^TWX & Z_k^TZ_1 & \cdots & Z_k^TWZ_k+U_k \end{pmatrix} \begin{pmatrix} \beta+\delta\beta \\ v^{(1)}+\delta v^{(1)} \\ \vdots \\ v^{(k)}+\delta v^{(k)} \end{pmatrix} = \begin{pmatrix} X^TWw \\ Z_1^TWw+R_1 \\ \vdots \\ Z_k^TWw+R_k \end{pmatrix}$$

where

$$U_i = -\phi\{\partial^2 l(\alpha_i; v^i)/\partial v^{(i)2}\}$$

and

$$R_i = U_i v^i + \phi\{\partial l(\alpha_i; v^{(i)})\}$$

Extensions of MAPHLEs and the scaled deviance with estimated degrees of freedom are also straightforward.

3.13.1 Test Criterion for Random Components

We propose a test statistic for random components, of the form

$$T_d = -2h_P \tag{3.193}$$

Note the following useful properties of the APHL h_P. First, h_P is invariant with respect to linear transformations of random effects v. Suppose that $\eta\prime = \eta + Zv = \eta + Z_a r$ where $Z_a = ZA^{-1}$ and $r = Av$for some non-singular matrix A.

The h-likelihood can be based on either v or r since both v and r can appear linearly in η.Since here the Jacobian term is constant this does not cause any problem for inferences based on the h-likelihood. The value of h_p remains the same since the Jacobian term cancels; see equation (3.184). Thus the value of h_p remains fixed for equivalent formulations of random effects. Suppose that the random components $v^{(i)}$ have a distribution such that $\text{var}(v^{(i)}) = \sigma_i^2 D_i$ for some

matrix D_i and that $E(v^{(i)}) \longrightarrow 0$ as $\sigma_i^2 \longrightarrow 0$. The test for $\sigma_i^2 = 0$ is then equivalent to the test for the absence of random components $v^{(i)}$. For various HGLMs, including GLMMs, Lee and Nelder (1996) showed that

$$\lim_{\sigma_i^2 \longrightarrow 0}(h_P) = h_{P(-i)}, \tag{3.194}$$

where $h_{P(-i)} = h_P|_{\sigma_i^2} = 0$ is the APHL for model (3.192) without the $v^{(i)}$-component. Thus values of h_P change smoothly with respect to the absence of random effects. The term 2π in equation (3.183) plays an important role in this. Owing to the invariance of h_P with respect to linear transformations we have

$$\lim_{\rho \longrightarrow 1}(h_P(\hat{v})) = \lim_{\rho \longrightarrow 1}(h_P(\hat{r})), \tag{3.195}$$

where $h_P(\hat{v})$ and $h_P(\hat{r})$ are the APHLs with respect to random effects v and r respectively. The left-hand side is not computable, owing to the singularity of the covariance matrix at $\rho = 1$, but the right-hand side can be easily computed since it is equivalent to the absence of random effects r_{j2}. Thus with h_p we may test not only the HGLMs can be defined in the framework of Bayesian hierarchical models. For GLMMs, Bayesian procedures for posterior distributions have been developed by Zeger and Karim (1991) by using the Gibbs sampler; for some comparisons with GLMM analysis see Breslow and Clayton (1993). Bayesian models with conjugate priors have been studied by George *et al.* (1993).

3.13.2 Deviances in HGLMs

Lee and Nelder (1996) proposed to use three deviances based upon $f_\theta(y, v)$, $f_\theta(y)$ and $f_\theta(y|\hat{\beta})$ for testing various components of HGLMs. For testing random effects they proposed to use the deviance $-2h$, for fixed effects -2ℓ and for dispersion parameters $-2\ log\ f_\theta(y|\hat{\beta})$. When ℓ is numerically hard to obtain, they used $p_v(h)$ and $p_{\beta,v}(h)$ as approximations to ℓ and $log\ f_\theta(y|\hat{\beta})$. When testing hypotheses on the boundary of the parameter space, for example for $\lambda = 0$, the critical value is

$\chi^2_{2\alpha}$ for a size-α test. This results from the fact that the asymptotic distribution of likelihood-ratio test is a $50:50$ mixture of χ^2_0 and χ^2_1 distributions (Chernoff, 1954; Self and Liang, 1987): for application to random-effect models see Stram and Lee (1994), Vu et al. (2001), Vu and Knuiman (2002), Verbeke and Molenberghs (2003) and Ha and Lee (2004).

Based upon $log\ f_\theta(y|v)$, Lee and Nelder (1996) proposed the use of the scaled deviance for the goodness-of-fit test, defined by

$$D = D(y, \hat{\mu}) = -2\{\ell(\hat{\mu}; y|v) - \ell(y; y|v)\}, \tag{3.196}$$

where $\ell(\hat{\mu}; y|v) = log\ f(y|v; \hat{\beta})$ and $\mu = E(y|v)$, having the estimated degrees of freedom, $d.f. = n - p_D$, where

$$p_D = trace\{(T_m^t \Sigma_m^{-1} T_m)^{-1} T_m^t \Sigma_0^{-1} T_m\} \tag{3.197}$$

and $\Sigma_0^{-1} = W_{ma} diag(\Phi^{-1}, 0)$. Lee and Nelder (1996) showed that $E(D)$ can be estimated by the estimated degrees of freedom; $E(D) \approx n - p_D$ under the assumed model. Spiegelhalter et al. (2002) viewed p_D as a measure of model complexity. This is an extension of the scaled deviance test for GLMs to HGLMs.

If ϕ is estimated by the REML method based upon $p_{\beta,v}(h)$, the scaled deviances $D/\hat{\phi}$ become the degrees of freedom $n - p_D$ so that the scaled deviance test for lack of fit is not useful when ϕ is estimated, but it can indicate that a proper convergence has been reached in estimating ϕ.

For model selection for fixed effects ϕ the information criterion based upon the deviance ℓ, and therefore $p_v(h)$, can be used, while for model selection for dispersion parameters, the information criterion based upon the deviance $p_\beta(\ell)$, and therefore $p_{v,\beta}(h)$, can be used. However, these information criteria cannot be used

for models involving random parameters. For those Spiegelhalter et al. (2002) proposed to use in their Bayesian framework an information criterion based upon D.

We claim that one should use the information criterion based upon the conditional loglihood $logf_\theta(y|v)$ instead of D. Suppose that $y \approx N(X\beta, \phi I)$, where the model matrix X is $n \times p$ matrix with rank p. Then, there are two ways of constructing the information criterion; one is based upon the deviance and the other is based upon the conditional loglihood. First suppose that ϕ is known. Then, the AIC based upon the conditional loglihood is

$$AIC = n\ log\ \phi + \Sigma(y_i - x_i^t\hat{\beta})^2/\phi + 2p_D, \tag{3.198}$$

while the information criterion based upon the deviance D is

$$DIC = \Sigma(y_i - x_i^t\hat{\beta})^2/\phi + 2p_D. \tag{3.199}$$

Here the two criteria differ by a constant and both try to balance the sum of the residual sum of squares, $\Sigma(y_i - x_i^t\hat{\beta})^2$ and the model complexity p_D. Now suppose that ϕ is unknown. Then,

$$DIC = \Sigma(y_i - x_i^t\hat{\beta})^2/\hat{\phi} + 2p_D, \tag{3.200}$$

which becomes $n + 2p_D$ if the ML estimator is used for ϕ and $n + p_D$ if the REML estimator is used. So it always chooses the simplest model of which the extreme is the null model, having $p_D = 0$. Here

$$AIC = n\ log\ \hat{\phi} + \Sigma(y_i - x_i^t\hat{\beta})^2/\hat{\phi} + 2p_D, \tag{3.201}$$

which becomes $n\ log\ \hat{\phi} + n + 2p_D$ if the ML estimator is used for ϕ and $n\ log\ \hat{\phi} + n + p_D$ if the REML estimator is used. Thus, the AIC still tries to balance the

residual sum of squares $\Sigma(y_i - x_i^t\hat{\beta})^2$ and the model complexity p_D. This means that we should always use the conditional likelihood rather than the deviance. Thus, we use $-2\ log\ f_\theta(y|v) + 2p_D$ for model selection involving random parameters. In this thesis, four deviances, based upon h, $p_v(h)$, $p_{\beta,v}(h)$ and $log\ f_\theta(y|v)$, are used for model selection and for testing different aspects of models.

In GLMs, the Wedderburn (1974) quasi-likelihood equations provide estimators for, β given an arbitrary variance function $V(\mu)$. However, quasi-likelihood does not provide estimates of dispersion parameters. Nelder and Pregibon (1987) developed an extended quasi-likelihood Q defined

$$Q = -\sum d_{ij}/2\phi - \sum \frac{1}{2}\log\{2\pi\phi V(y_{ij})\} \tag{3.202}$$

where

$$d_{ij} = 2\int_y^{\mu\prime} \frac{y-x}{V(x)}dx$$

is the deviance component. Extended quasi-likelihood provides the Wedderburn (1974) quasi-likelihood equations $\partial Q/\partial\beta = 0$ for β and also estimating equations $\partial Q/\partial\phi = 0$ for ϕ. Nelder and Lee (1992) showed by numerical studies that the extended quasi-likelihood dispersion estimator performs well in finite samples and may even on occasion be better than the ML estimator. Let the extended quasi-h-likelihood

$$h_E = Q + l(\alpha; v) \tag{3.203}$$

For any variance function $V(\mu\prime)$, the estimating equations $\partial h_E/\partial\beta = 0$ and $\partial h_E/\partial v = 0$ can be solved by the score equations and we call solutions $\hat{\beta}$ and $\hat{v}$ the (maximum) quasi-h-likelihood estimators; this is an extension of Wedderburn's (1974) quasi-likelihood equations to HGLMs. With the adjustment (3.183), the quasi- MAHPL equations $\partial h_A/\partial\theta = 0$ with $\theta = \alpha$ or $\theta = \phi$ give the corresponding equations for dispersion parameters. It would be interesting to develop a quasi-

likelihood for $l(\alpha; v)$, in particular a quasi-conjugate distribution for the power variance function family $V(\mu\prime) = \mu'^{\gamma}$. When $l(\alpha; v)$ is the normal likelihood, the score equations (3.178) for the quasi-h-likelihood estimators depend only on the first two moments of $y|v$ and v, as Schall (1991) and Breslow and Clayton (1993) noted. This allows the class of HGLMs to be extended to models defined only by the first two moments of $y|v$ and v.

3.13.3 Asymptotic Properties of Maximum h-likelihood Estimate

A crucial condition for the asymptotic properties of MHLEs discussed in Section 3 to hold is that $D^{*-1} = O_P(n^{-1})$. When this is so, our results in Section 3 can be easily modified to give the asymptotic properties of the MHLEs for an HGLM with more than one extra random component. For $D^{*-1} = O_P(n^{-1})$ to be true, we require that the total number of random effects t remains fixed, which is not always realistic. As the sample size n increases t may often increase as well. We now drop the assumption of a fixed t.

In normal-normal models, the properties of MHLEs for β and v hold for fixed sample size n. We showed in earlier sections that for some conjugate HGLMs with one random component the fixed effect estimators are the same as the marginal likelihood estimators and the random effect estimates are the best unbiased predictors on some scale of random effects. On a model-by-model basis, some other properties of MHLEs can be shown. In GLMMs, the MHLEs for β and v are the same as theirs. So the procedures developed here may be reliable and useful as an approximate inference even in the worst situations.

Chapter 4

Results and Discussion

4.1 Preliminary (Exploratory) Analysis

In all, data from 800 Maize and Soybean farmer based organizations (FBOs) were gathered by means of a structured questionnaire. This was later cleaned to 790 distinct observations. The FBOs were randomly selected through a multi-stage random procedure.

Fixed effect variables measured include; crop type (Maize or Soybean), Financial Credit (Acquired or Not), Training (Acquired or Not), Study tour (Acquired or Not), Demonstrative Practicals (Acquired or Not), Networking Events (Acquired or Not), Post-harvest Equipment (Acquired or Not), Number of farmers in the FBO and Plot size cultivated. These fixed factors constitute the physical factors that contributes to crop yield. Other factors such as rainfall, climate change, fertilizer use, soil nature, etc. are not considered to be physical factors to crop yield. Beside these 9 fixed effects, 36 two-way interaction terms are also generated as fixed interaction terms. This brings the total number of fixed covariates to 45. Dependent variable measured is Total Crop Yield. The regions and the particular communities are treated as Random variables.

The target population consists of mainly Maize and Soybeans Farmer based organizations in selected communities in the three Northern regions of Ghana. Northern Region = 7 communities Upper East Region = 3 communities Upper West Region = 3 communities FBO's interviewed = 800 with 10 missing data Hence total FBO's used = 790

Firstly, the raw data is plotted and the patterns of Crop yield against some selected covariates are observed. Figure 4.1 presents the observed scatter-plot of the crop yield against Plot size, Figure 4.2 presents the observed scatter-plot of the crop yield against number of Farmers, Figure 4.3 presents the observed scatter-plot of the crop yield against Regions while Figure 4.4 presents the observed scatter-plot of the crop yield against the 13 communities. From Figure

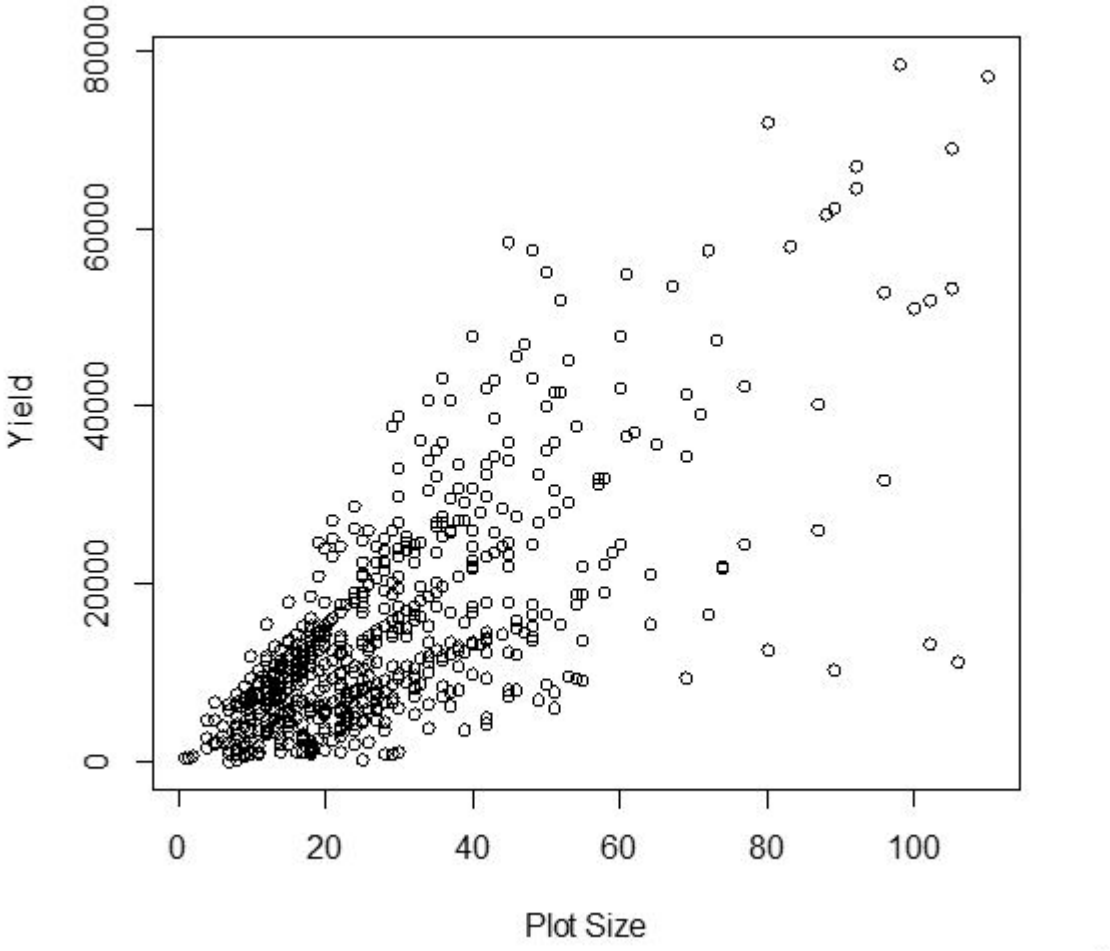

Figure 4.1: Scatter-plot of Crop yield against Plot size

4.1, crop yield is strongly related to plot size positively. Also from Figure 4.2, crop yield appears to be strongly related to No. of farmers positively.

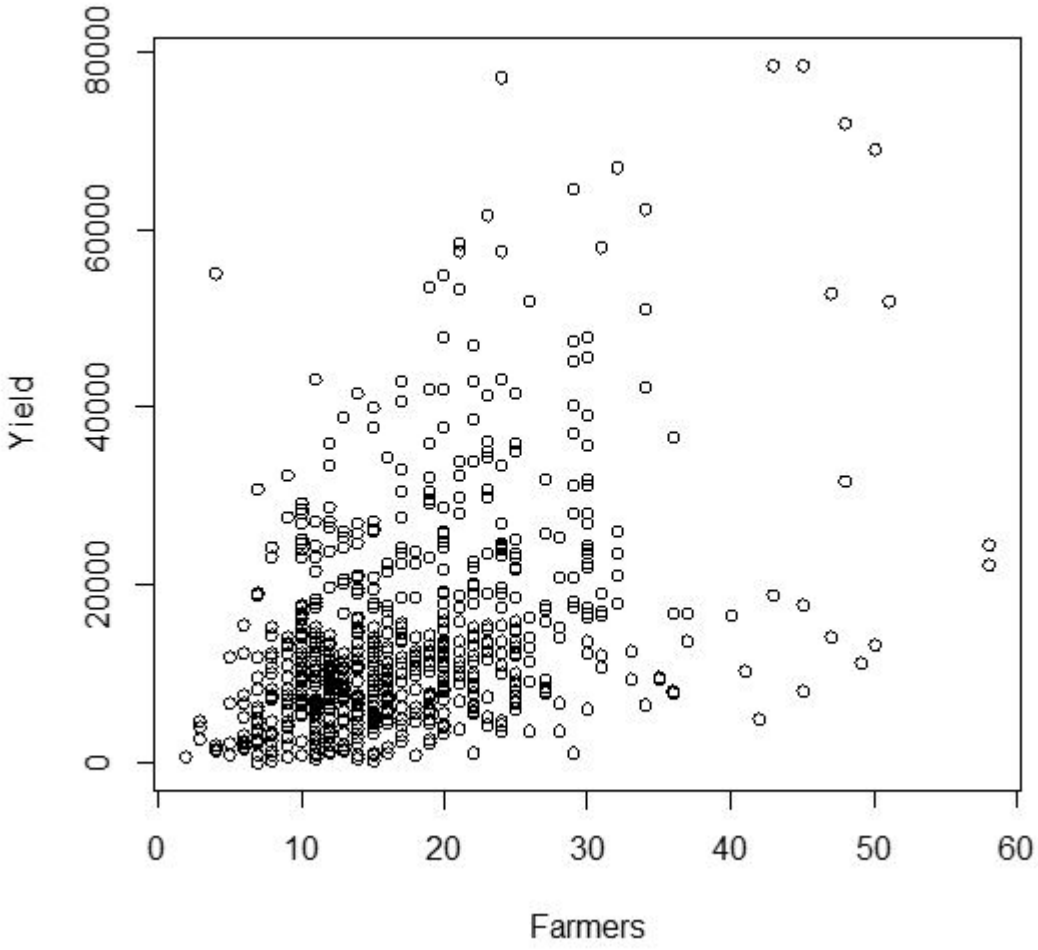

Figure 4.2: Scatter-plot of Crop yield against No. of farmers

The community as well as regional outlook of the crop yield is found in appendix A.

4.2 Penalized variable selection

4.2.1 Simulation studies

In this section, the researcher first investigate the performance of the HL method through simulated data, and compare HL methods to existing methods including the LASSO and SCAD. For each method the researcher selects optimal tuning parameters that maximize the log-likelihood obtained from an independent validation dataset of size $n/2$, where n is the size of the training set. The number

of covariates (p) and fixed coefficients (q) in two simulation were varied. In one simulation we use $n = 200$ while in the other $n = 100$.

For the simulation, the study considered the following GLM:

$$y/x \sim N(\mu(X'\beta), 2) \tag{4.1}$$

with linear link function $\mu(X'\beta) = X'\beta$ where the linear predictor $X'\beta = \sum_{j=1}^{pk} x_{jk}\beta_{jk}$ consist of p covariates. To generate covariate x'_{jk}s, we first generate $p = \sum_{j=1}^{k} p_k$ random variables x'_{jk}s independently from the standard normal distribution. Then z'_ks are simulated with a multivariate normal distribution.
The covariate x'_{jk}s are generated from

$$x_{kj} = (z_k + \varepsilon_{kj})/\sqrt{2}\; k = 1, \ldots, pk \tag{4.2}$$

where $z = (z_1, \ldots, z_k)' \sim N(0, \sum)$ with covariance structure $\sum_{kl} = cov(z_k, z_l) = 0.5^{|k-l|}$ and $\varepsilon_{kj} \sim N(0, I_p)$ that of independent of z. The true non-zero coefficients are

$$\beta_{kj} = c/j,\; j = 1, \ldots, qk,\; k \neq A$$

where qk is the number of non-zero coefficients in the kth group, and A is the set of the non-null groups. A group is said to be non-null if at least one coefficient in the group is estimated to be non-zero. The constant c is chosen so that the signal-to-noise ratio is equal to 5 in the linear model.

For each model setting the study considered one dimensionality level only, the one with $p < n$. So, overall we have 4 simulation scenarios, where each is replicated 100 times with sample size $n = 200$ and $n = 100$. The cross validation errors which are defined as Equation 3.13 based on independent test sample of size $N = 5000$ forms the basis for performance comparison. The results are shown in table below.

Table 4.1: Comparative simulation results for penalized variable selection methods

N=100 P= 10 Q =3										
Method	sim 1	sim 2	sim 3	**sim 4**	**sim 5**	sim 6	sim 7	sim 8	sim 9	sim 10
LASSO	12.078814	9.154382	8.09983	**8.075933**	**11.001972**	11.456325	13.042181	12.49598	8.606529	10.729943
SCAD	12.161675	9.304411	8.193715	**8.134833**	**11.079497**	11.558803	13.031146	12.53982	8.662975	11.357576
H-L	11.424703	8.989876	7.700814	**9.586048**	**11.239171**	10.810527	12.891827	10.58886	8.532304	10.799194
N=200 P= 10 Q = 3										
LASSO	16.08259	23.72576	18.89797	**25.91064**	**18.28419**	22.8221	21.5849	18.24477	24.7486	19.22165
SCAD	16.12418	23.67929	18.88515	**25.92822**	**18.22037**	22.82895	21.63649	18.31884	24.75796	19.19523
H-L	15.86426	22.57885	18.33407	**23.89222**	**18.02888**	21.61746	20.70557	18.04133	25.58047	18.44424
N=100 P= 8 Q =5										
LASSO	12.477717	9.976822	9.567962	7.672933	11.656413	13.577737	10.798311	16.09166	14.212732	12.474876
SCAD	12.664895	10.003768	9.849049	7.660239	11.845172	13.811124	10.989915	17.00159	14.318873	12.343221
H-L	11.912251	9.440881	9.317837	7.763213	11.043935	16.451458	10.599491	13.9028	13.701678	11.351132
N=200 P= 8 Q =5										
LASSO	23.62724	21.68409	22.38465	20.89009	17.62312	18.37307	22.97981	21.5812	26.90833	25.05369
SCAD	23.64128	21.64646	22.44532	20.86011	17.91219	18.43819	22.95806	21.53924	27.03923	25.16297
H-L	22.78054	21.70741	22.26482	20.29142	17.29375	18.20228	21.64576	21.33577	25.43343	24.61937

For variable selection quality, cross validation errors for the three methods are compared and the method with the smallest cv errors is preferred. The HL estimator performs better generally than the other methods for prediction accuracy as evident by its smallest cross validated errors comparative to the other methods.

4.2.2 Real Data Analysis (Crop yield data)

We analyse the crop yield datasets: The Crop yield data consists of a numeric response variable, the 2013 main season yield measured in kilograms, and 9 covariates obtained from 790 farmer based organizations in the three northern regions of Ghana. The researcher dropped a covariate which has too many missing values and exclude 10 observations (FBO's) due to missing values. The dataset has 7 categorical covariates crop type (Maize or Soybean), Financial Credit (Acquired or Not), Training (Acquired or Not), Study tour (Acquired or Not), Demonstrative Practical (Acquired or Not), Networking Events (Acquired or Not), Post harvest Equipment (Acquired or Not)) and 2 continuous variables, including the so called plot size in acres and number of farmers. Beside these 9 fixed effects, 36 two-way interaction terms are also generated as fixed interaction terms. This brings the total number of fixed covariates to 45. To allow possible non-linear effects, a third-degree polynomial is used for each continuous covariate, and dummy variables are used for categorical variables.

The results are obtained by 100 random partitions of the data set split into training (70 percent) and test sets (30 percent). For each random partition, the tuning parameters are selected by the 10-fold cross validation within the training set, and the prediction errors are computed on the test set. Table 4.3 presents averages of cross validated errors, the number of significant variables and number of insignificant variables.

The HL estimator performs better than the other methods for prediction accuracy as evident by its smallest cross validated errors comparative to the other methods.

Table 4.2: Standardized Penalized Coefficients of Crop Yield Data

Selected Variables	LASSO	SCAD	H-L
Crop	-1.39	-2.38	-0.74
Credit			1.23
Training	0.82	1.91	2.29
Study Tour			
Demo. Practical	3.65	5.14	5.04
Networking Events			0.69
Post harvest Equipment	-5.56	-7.21	-7.21
No. of farmers	0.29		1.88
plot size	11.53	12.42	12.83
Crop*Credit	-1.72	-2.41	-2.74
Crop*Training			-0.67
Crop*Study Tour	0.86	1.47	1.53
Crop*Demo. Practical	-1.68	-2.76	-2.63
Crop*Networking Events			
Crop*Post-harvest Equipment	2.74	4.60	4.50
Crop*No. of farmers	-0.96		-2.11
Crop*plot size			
Credit*Training			
Credit*Study Tour	-0.963	-1.14	-1.23
Credit*Demo. Practical	0.20		
Credit*Networking Events	0.89	1.37	1.34
Credit*Post-harvest Equipment			
Credit*No. of farmers	-0.74		-1.5
Credit*plot size	2.80	2.30	2.86
Training*Study Tour	0.81	0.62	0.11
Training*Demo. Practical	-1.10	-1.41	-1.33
Training*Networking Events	-0.06		-0.69
Training*Post-harvest Equipment	0.32		0.48
Training*No. of farmers	-1.91	-3.11	-1.97
Training*plot size	-0.53		-0.82
Study Tour*Demo. Practical	0.14		0.13
Study Tour*Networking Events			
Study Tour*Post-harvest Equipment	-1.38	-1.27	-1.22
Study Tour*No. of farmers	0.46	0.02	1.08
Study Tour*plot size	-0.14		-0.74
Demo. Practical*Networking Events	-0.29		-0.46
Demo. Practical*Post-harvest Equipment	1.65	1.68	1.69
Demo. Practical*No. of farmers			
Demo. Practical*plot size	-3.53	-3.89	-3.58
Networking Events*Post-harvest Equipment			
Networking Events*No. of farmers			
Networking Events*plot size	-1.04	-2.04	-2.09
Post-harvest Equipment*No. of farmers			-0.38
Post-harvest Equipment*plot size	4.43	4.59	5.06
No. of farmers*plot size			-0.67

Table 4.3: Performance of Penalized methods on Crop Yield Data

Method	LASSO	SCAD	H-L
No. of Significant Variables Selected	31	21	35
No. of Variables Ignored	14	24	10
Cross validated Errors	24.097	24.043	23.543

4.3 Crop yield models for fixed covariates

4.3.1 Generalized Linear Models

The same crop yield data set, consisting of 790 units with nine (7) categorical explanatory variables Credit, Crop, Training, Tour, Practical, Networking, Equipment, representing our support services (factors) and two (2) continuous variables No. of farmers, Plot size and a response Crop Yield was used. From the original analysis below, The researcher fitted a Gaussian GLM with linear predictors; Credit 1 + Crop 2 + Training 1+ Tour 1+ Practical 1+ Networking 1+ Equipment 1+ No. of farmers + Plot size, dropping Credit 1, Tour 1, Practical 1, Networking 1 and Equipment 1 after t-tests (Approximated to Z due to large sample size) on individual parameters. Before using the distributional results for inference, it is always necessary to check that the model meets its assumptions well enough that the results are likely to be valid. Figure 4.3 shows the model-checking plots for this model.

From Figure 4.3, the diagnostic plots have only few satisfactory features. The running mean in the plot of residuals against fitted values shows some form of marked increasing trend, and the plots of absolute residuals has a relatively unstable slop. The normal plots shows no discrepancy but with few trade-off. In addition, the histogram of residuals is almost symmetric to the left. These are not so good indications of an appropriate model.

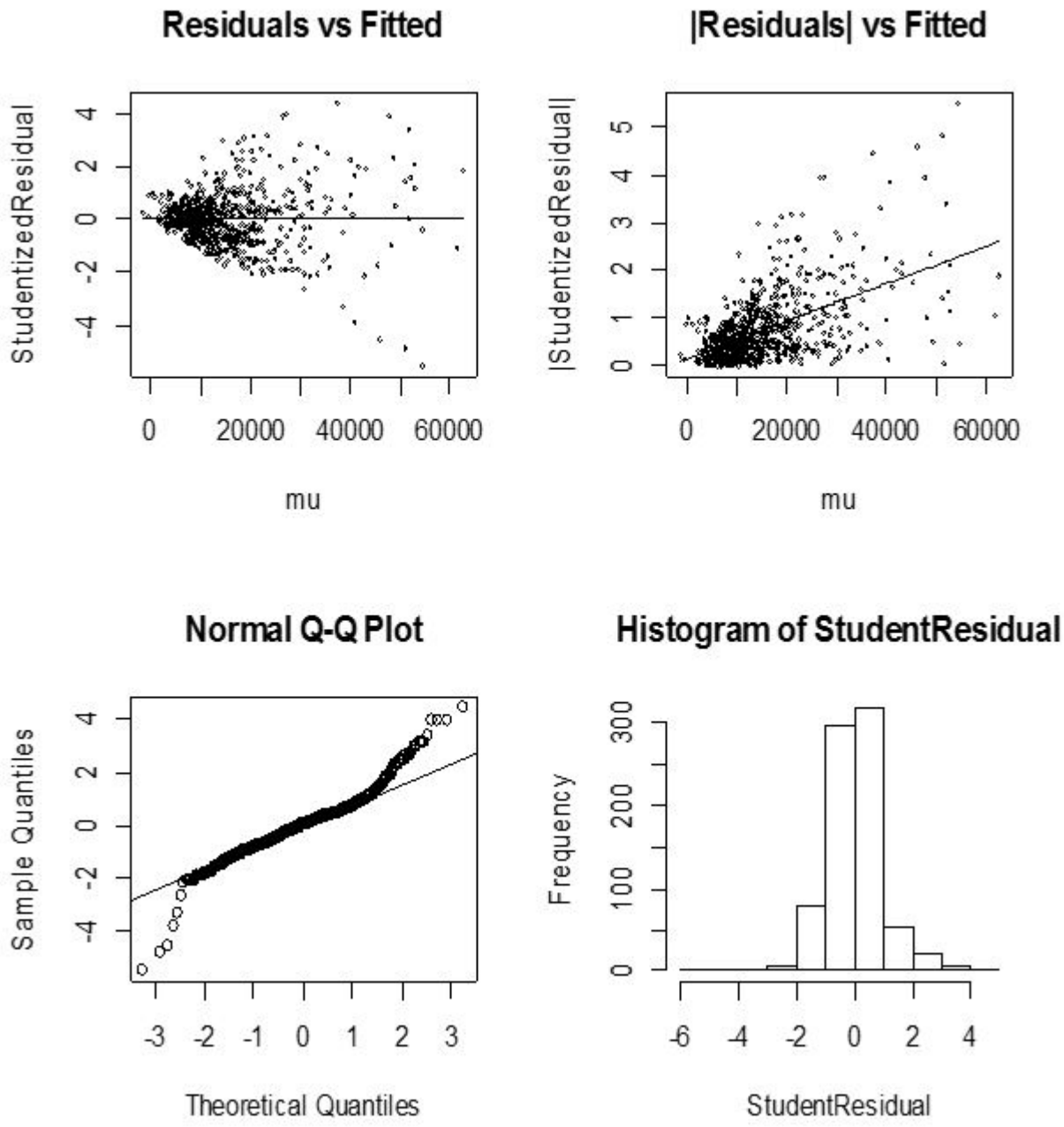

Figure 4.3: Diagnostic plots of Gaussian GLM for crop yield

However this thesis seeks to present the very best of models. The defects present in the histogram suggests that something can be done to improve the model. We sort to remove any likely defects by moving to a GLM with gamma errors and a log link. The additive model is still satisfactory, with linear predictors Credit 1 + Crop 2 + Training 1 + Tour 1 + Practical 1 + Networking 1 + Equipment 1 + No. of farmers + Plot size, dropping Credit 1, Tour 1, Practical 1, and Networking 1 after t-tests on individual parameters.

The model-checking plots are appreciably better than for the normal Gaussian model and more improved. The resulting plots are shown in Figure 4.4.

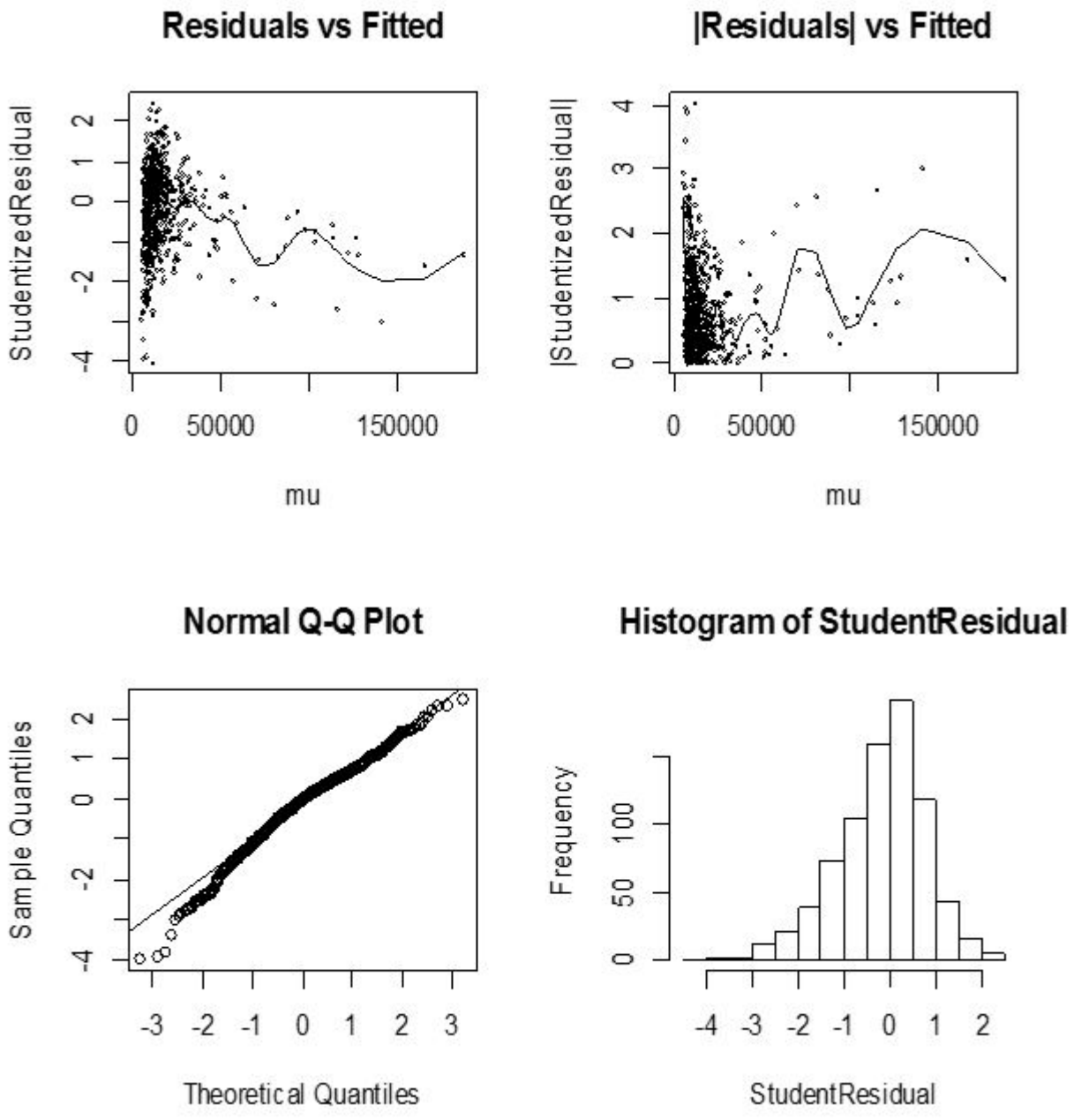

Figure 4.4: Diagnostic plots of Gamma GLM for crop yield

Here in the Gamma GLM, the running means in the plot of residuals against fitted values shows no form of marked trend, and the plots of absolute residuals has a relatively stable slop compared to the Gaussian GLM. The normal plots shows no discrepancy but with few trade-off far better than the Gaussian GLM. However,

the histogram of residuals is almost skewed to the left. Even though these are not so good indications of an appropriate model, it indicates an improvement in the earlier Gaussian GLM. The approach in this analysis has been to include as much variation in the model as possible, as distinct from down weighting individual yields to make an unsuitable model fit.

4.3.2 Model Interpretation

Once a model has been selected and checked, we then have to examine and interpret its estimated coefficients. For complex models such as we have here, it cannot be less easy to interpret, especially with factor variables when identified ability constraints are needed. This can look intimidating, although in fact the parameter names are pretty helpful here. They basically tell us the circumstance under which the coefficient of a factor were added to the model. Table 4.4 presents the estimates of the individual covariates and model selection criteria.

Table 4.4: Model Estimates for Gaussian and Gamma distributed GLM's

		GAUSSIAN GLM				GAMMA GLM			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
μ /log(μ)	(Intercept)	5869.6	1104.53	5.3141	0.00017	8.917414	0.07397	120.5543	<0.00001
	(Credit) 1								
	(Crop) 2	-3489.1	627.32	-5.562	0.00012	-0.193212	0.042012	-4.599	0.00049
	(Training) 1	-2598	710.71	-3.6555	0.002213	-0.134143	0.047596	-2.8184	0.00911
	(Tour) 1								
	(Practical) 1								
	(Networking) 1								
	(Equipment) 1					-0.108639	0.043526	-2.496	0.01583
	Farmers	-236.6	50.78	-4.6599	0.000448	-0.006341	0.003401	-1.8648	0.04596
	Plot size	577.2	23.27	24.8103	<0.00001	0.032139	0.001558	20.627	<0.00001
				Selection Criterion	Gaussian GLM	Gamma GLM			
				-2ML(-2 h)	16421.62	16104.67			
				-2RL(-2$p_{beta}(h)$)	16468.19	16151.25			
				cAIC	16441.62	16124.67			

In the Gaussian model for example if Crop type is 2 (soy bean) for some response measurement, then we include the (Crop) 2 term in the model (which just amounts to adding -3489.1 to the linear predictor in this case, since (Crop) 2 is a factor). If (Crop) 2 and (Training) 1 are in the model for some response measurement, then terms (Crop) 2, (Training) 1 and their interaction (Crop) 2: (Training) 1 are included, and so on. To make things completely clear, however, one can also look at the model matrix (and original data frame). The final Gaussian GLM for crop yield is given as

$$\eta = \mu = \beta_0 - \beta_{(Crop2)} - \beta_{(Training1)} - \beta_{(farmers)} + \beta_{(Plotsize)}$$

$$mean(Yield) = 5869.6 - 3489.1(soybean) - 2598.0(Training1)$$

$$-236.6(farmers) + 577.2(plotsize) \tag{4.3}$$

For a unit change in no. of farmers and plot sized, the predicted Yield of a soy bean growing FBO who received training support would therefore be given as

$$mean(Yield) = 5869.6 - 3489.1 - 2598.0 - 236.6(1) + 577.2(1)$$

$$mean(Yield) = 123(kg)$$

of soy bean in a typical main season per FBO in the three regions.

Similarly, for a unit change in no. of farmers and plot sized, the predicted Yield of a maize growing FBO who received training support would be given as

$$mean(Yield) = 5869.6 - 3489.1(0) - 2598.0(1) - 236.6(1) + 577.2(1)$$

$$mean(Yield) = 3,612.2(kg)$$

of maize in a typical main season per FBO in the three regions. Notice that there are now important differences between the two models. It is observed that

(Equipment) 1 is statistically significant in the gamma GLM, but not in the Gaussian normal model. Some of the coefficients are quite different. The final Gamma GLM for crop yield is given as

$$mean(Yield) = exp^{(8.917414 - 0.193212(soybean) - 0.134143(Train1) - 0.108639(Equipt1)} \\ -0.006341(farmers) + 0.032139(plot)) \quad (4.4)$$

4.3.3 Joint-Generalized Linear Models

The analysis in this section (as detailed in the methodology), supposes that we have two interlinked models for the mean and dispersion based on the observed data y and the deviance d. In other words, the algorithm for fitting these models can be reduced to the fitting of two-dimensional set of generalized linear models; one dimension being mean and the other being dispersion, so that no special code is needed for the estimation of dispersion components. The dispersion model is a GLM with a gamma variance function.

Here, the dispersion parameters are no longer constant, as we know from the usual Generalized Linear Models, but can vary with the mean parameters. The deviance components d^* become the responses for the dispersion GLM. Then the reciprocals of the fitted values from the dispersion GLM provide prior weights of the next iteration for the mean GLM.

This formulation implies that, the models-checking techniques derived for generalized linear models (McCullagh and Nelder, 1989, chapter 12), can be carried over to this wider class. Figures 4.5 and 4.6 represents the diagnostic plots for the Gaussian and Gamma joint-GLM's respectively.

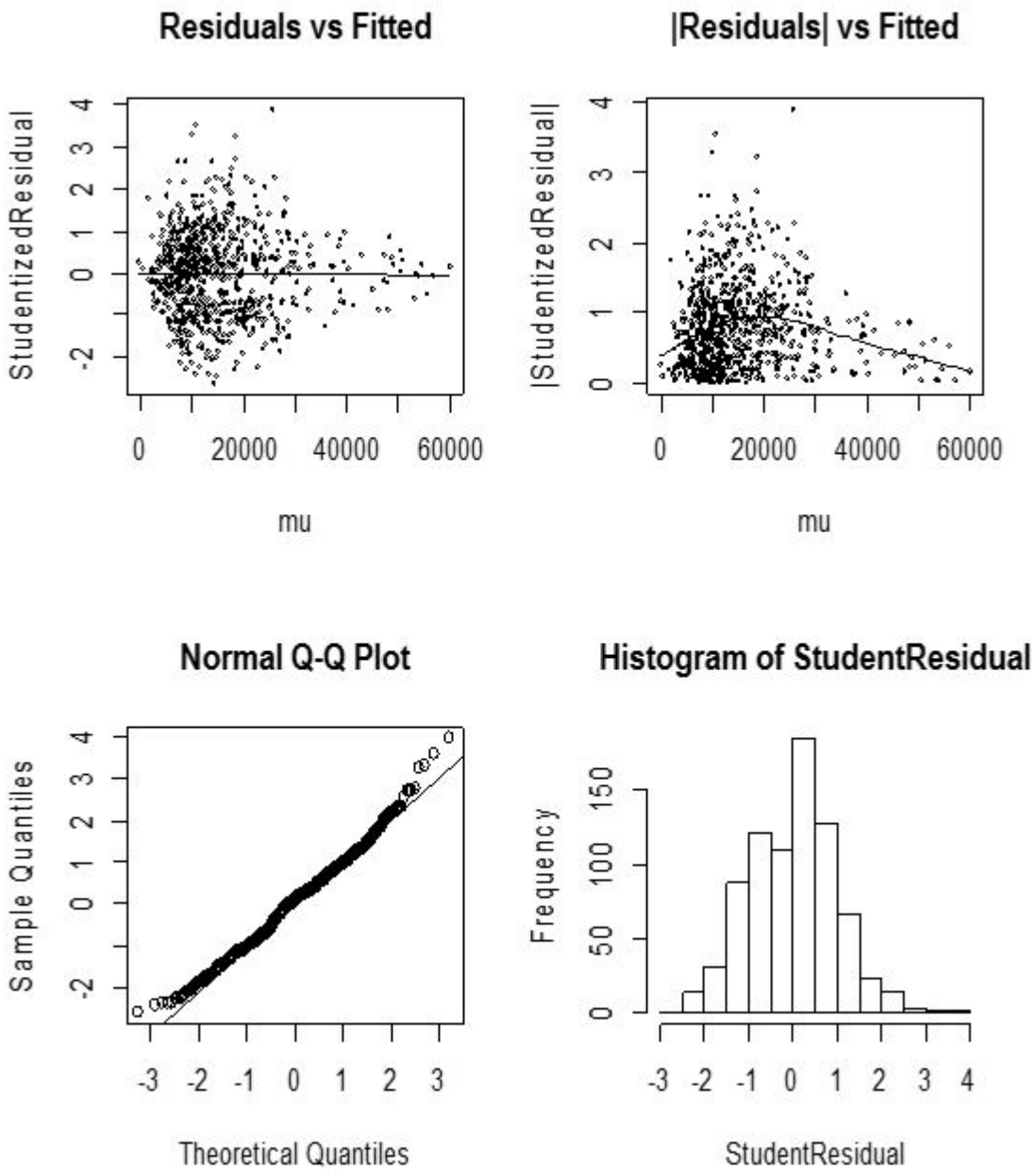

Figure 4.5: Diagnostic plots of Gaussian Joint-GLM for crop yield

From Figure 4.5 the diagnostic plots have several excellent features compared to the Gaussian ordinary GLM diagnostic plots in Figure 4.3. The running mean in the plot of residuals against fitted values shows no form of marked trend at all, and the plot of absolute residuals has a very stable slope, indicating that the variance is constant and satisfies the independence assumption, that the right link function was specified and also indicates no missing dependency. The normal plot also shows no discrepancy. In addition, the histogram of residuals is

almost symmetric. These are very good indications of an appropriate model and an excellent improvement over the counterpart Gaussian GLM in Figure 4.3.

The gamma joint GLM diagnostic plots of Figure 4.6 as below, also shows an incredible performs over the first gamma GLM of Figure 4.4.

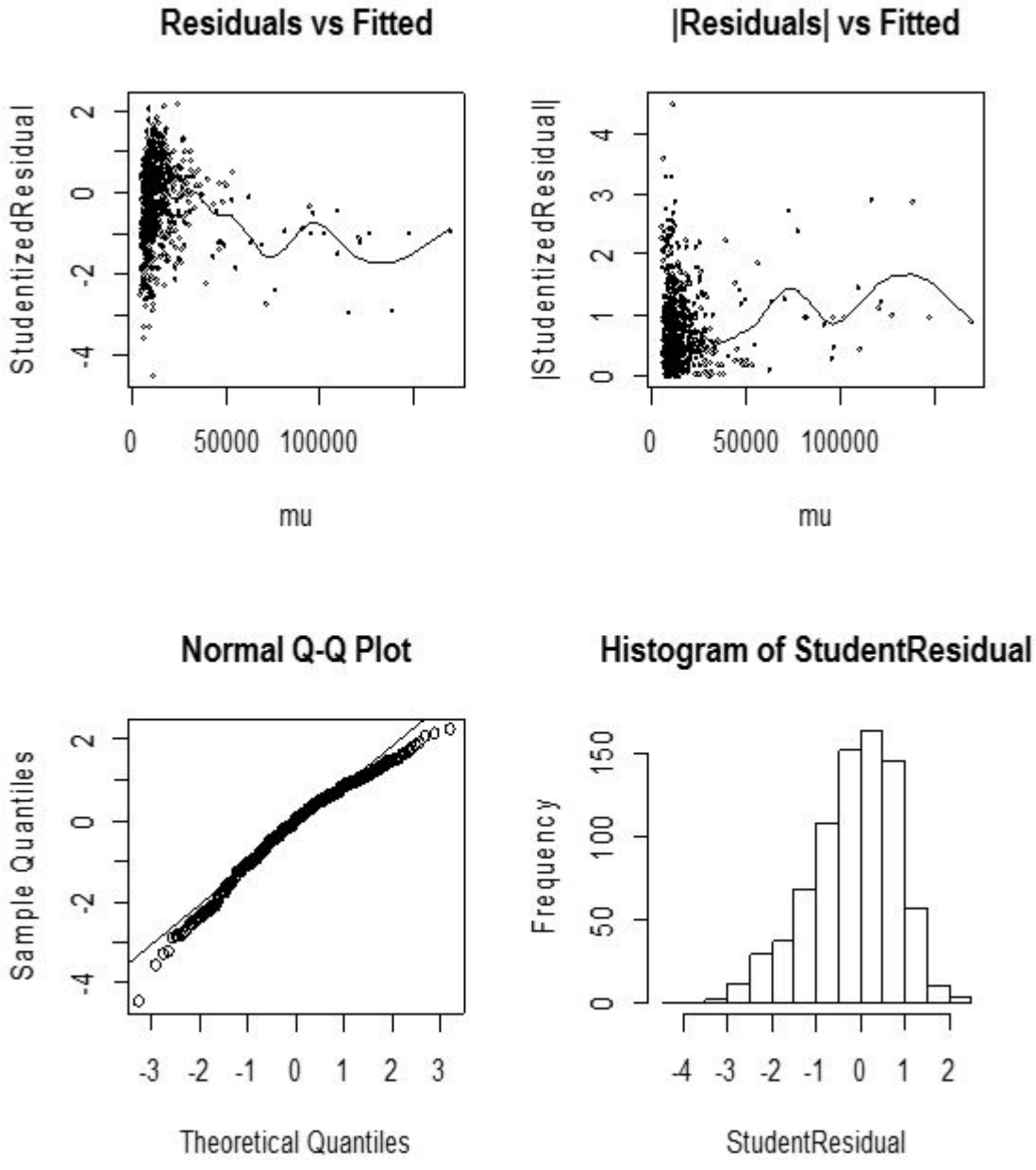

Figure 4.6: Diagnostic plots of Gamma Joint-GLM for crop yield

The model-checking plots are appreciably similar to the normal Gaussian joint GLM and both demonstrate an excellent improvement of their GLM's. The

resulting plots are shown in Figure 4.4.

4.3.4 Model Interpretation

Table 4.5 represents the model parameter estimates for both the Gaussian and the Gamma joint-GLM's. Log(μ) or μ on the table represents the mean model whereas log(ϕ) represents the dispersion model. The final mean model for the Gaussian joint-GLM does not include access to Training, Study tour and demonstrative practicals where as the dispersion model excludes Post harvest equipments and number of farmers. In the final mean model for the Gamma joint-GLM, access to Credit, Networking events and number of farmers excluded where as the dispersion model excludes Crop type, Study tour and number of farmers.

Apart from plot size and networking events which increase the mean crop yield in the Gaussian model, we observe that all other support services rather tends to reduce crop yield. In a similar scenario, only study tour and plot size increases crop yield significantly from the Gamma mean model.

Table 4.5: Model Estimates for Gaussian and Gamma distributed Joint-GLM's

		GAUSSIAN JOINT-GLM				GAMMA JOINT-GLM			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
log(μ)	(Intercept)	2789.82	630.85	4.4223	0.000645	8.91886	0.056282	158.467	<0.00001
	(Credit) 1	-694.45	368.33	-1.8854	0.044395				
	(Crop) 2	-13444.65	361.52	-3.7194	0.001991	-0.149376	0.036715	-4.0685	0.001129
	(Training) 1					-0.158761	0.035628	-4.4561	0.000612
	(Tour) 1					0.05197	0.037108	1.4005	0.09581
	(Practical) 1					-0.084918	0.033583	-2.5286	0.014987
	(Networking) 1	831.65	358.33	2.3209	0.021354				
	(Equipment) 1	-983.56	339.32	-2.8986	0.007944	-0.129216	0.032611	-3.9624	0.001339
	Farmers	-87.66	34.80	-2.5187	0.015246				
	Plot size	541.36	22.68	23.8712	<0.00001	0.031049	0.001343	23.1155	<0.00001
log(ϕ)	(Intercept)	14.916823	0.208605	71.5075	<0.00001	-2.37071	0.235844	-10.052	<0.00001
	(Credit) 1	0.444352	0.125076	3.55266	0.002623	0.32474	0.141297	2.29828	0.022195
	(Crop) 2	-0.585502	0.11825	-4.9514	0.000289				
	(Training) 1	0.887435	0.134344	6.60569	0.00003	0.98346	0.15179	6.47908	0.000035
	(Tour) 1	0.323429	0.121485	2.66229	0.011905				
	(Practical) 1	0.414736	0.115033	3.60536	0.002403	0.509524	0.129886	3.92286	0.001427
	(Networking) 1	-0.470067	0.127219	-3.6949	0.002075	-0.401953	0.143677	-2.7976	0.009446
	(Equipment) 1					0.43765	0.13861	3.15742	0.005102
	Farmers								
	Plot size	0.069479	0.004379	15.8664	<0.00001	0.009238	0.004967	1.85988	0.046273
				Selection Criterion	Gaussian GLM	Gamma GLM			
				-2ML(-2 h)	15894.87	15984.99			
				-2RL(-2$p_{beta}(h)$)	15772.61	16045.60			
				cAIC	15914.87	16004.99			

4.3.5 Joint-Generalized Linear Models for Quality Improvement

Table 4.6 below reveals that the initial Gaussian GLM even though was satisfactory mean model, modelling both mean and dispersion (Joint-GLM) improves the quality of the same Gaussian distributed model significantly.

Table 4.6: Model criteria for Gaussian GLM and Gaussian Joint-GLM

Selection Criterion	Gaussian Joint-GLM	Gaussian GLM
-2ML(-2 h)	15894.87	16422.00
-2RL(-2$p_{beta}(h)$)	15772.61	16468.00
cAIC	15914.87	16442.00

Even though similar can be said of the Gamma GLM and Joint-GLM as evident in Table 4.7, we observe but for the conditional AIC, all the other two selection criteria confirms that modelling both mean and dispersion (Joint-GLM) improves model quality.

Table 4.7: Model criteria for Gamma GLM and Gamma Joint-GLM

Selection Criterion	Gaussian Joint-GLM	Gaussian GLM
-2ML(-2 h)	15984.99	16104.67
-2RL(-2$p_{beta}(h)$)	16045.60	16151.25
cAIC	16124.67	15914.87

Detailed model estimates for GLM and Joint-GLM are in appendix A

4.4 Crop yield models for fixed and random covariates

4.4.1 Hierarchical Generalized Linear Models (HGLM 1)

In an unpublished technical report, Pierce and Sands (Oregon State University, 1975) introduced generalized linear mixed models (GLMMs), where the linear predictor of a GLM is allowed to have, in addition to the usual fixed effects, one or more random components with assumed normal distributions. Although the normal distribution is convenient for specifying correlations among the random effects, the use of other distributions for the random effects greatly enriches the class of models. Lee and Nelder (1996) extended GLMMs to hierarchical GLMs (HGLMs), referred to in this thesis as HGLM 1, in which the distribution of random components are extended to conjugates of arbitrary distributions from the GLM family. Figure 4.7 and 4.8 represent diagnostic plots for the Gaussian and Gamma HGLM's respectively.

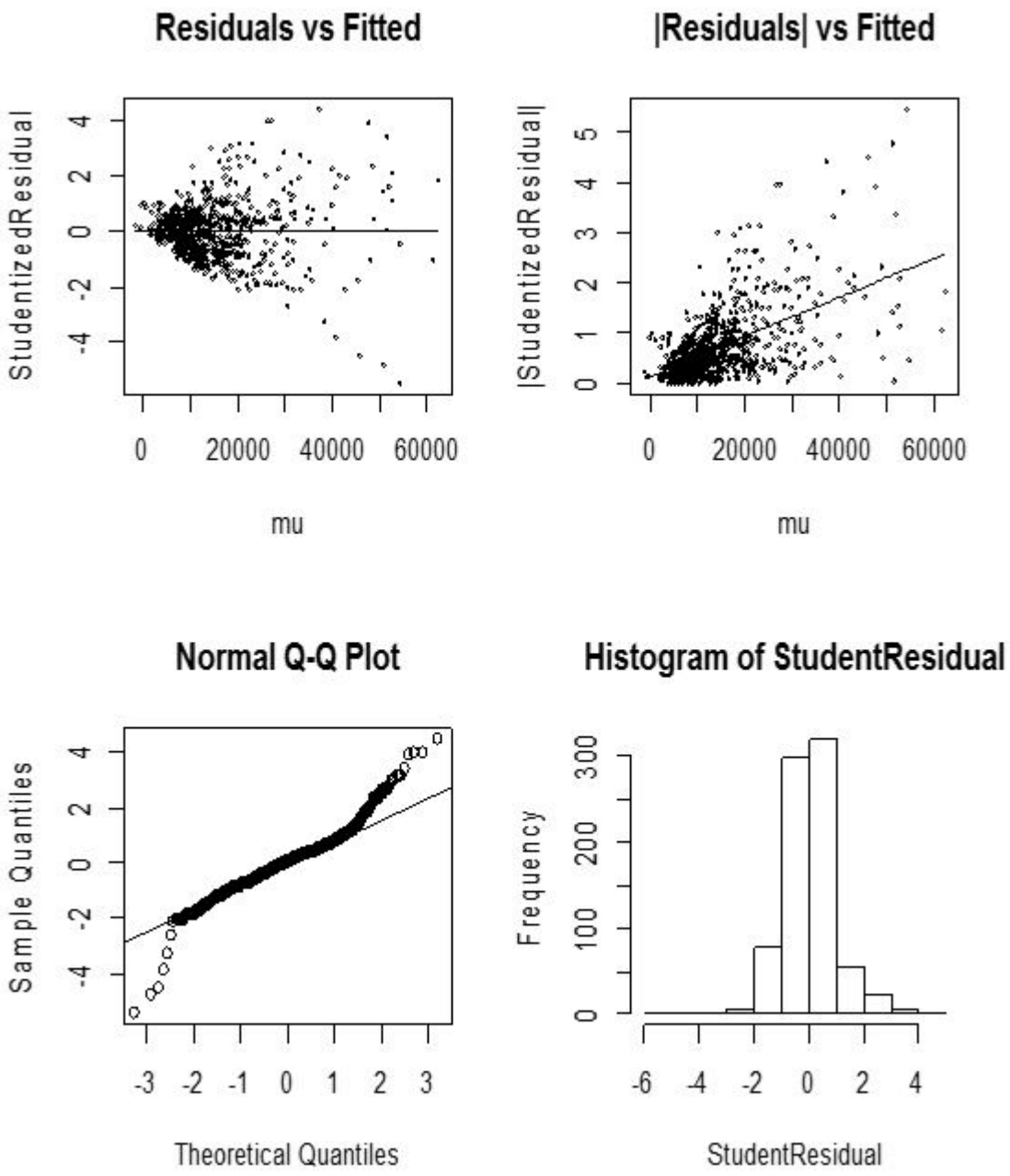

Figure 4.7: Diagnostic plots of Gaussian HGLM 1 for crop yield

The Gaussian diagnostic plots have some unsatisfactory features although not a worse case scenario. The normal plot shows some discrepancy. The running means in the plot of residuals against fitted values shows a form of outward trend. In addition, the histogram of residuals is almost symmetric. These may indicate an unsatisfactory and inappropriate model. The researcher therefore tried to remove any likely defects by moving to a HGLM with gamma errors and a log link. The model-checking plots does not appear appreciably better than for the

Gaussian model. The resulting plots are shown in Figure 4.8

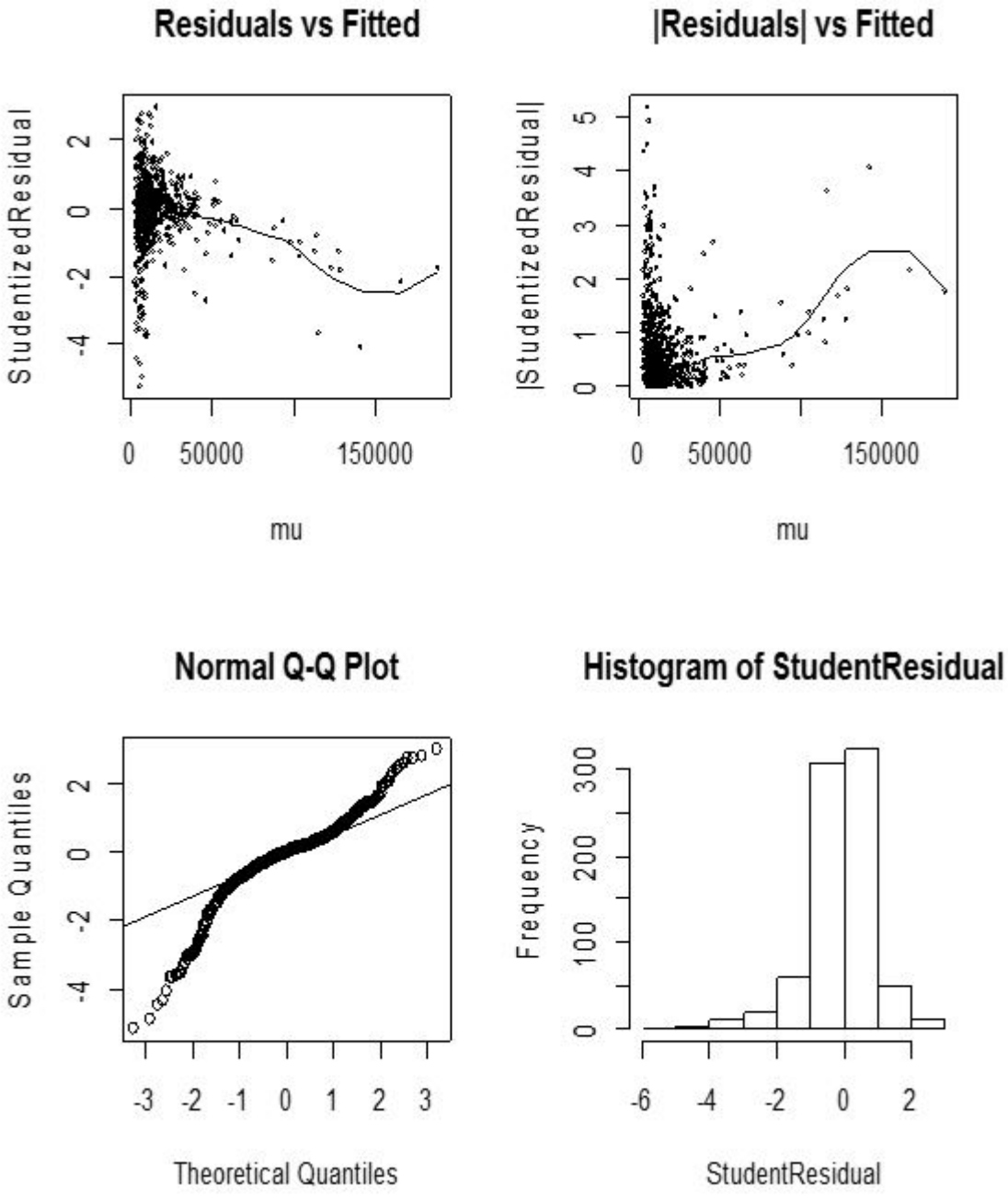

Figure 4.8: Diagnostic plots of Gamma HGLM 1 for crop yield

4.4.2 Model Interpretation

Table 4.8 represents the model parameter estimates for both the Gaussian and the Gamma HGLM's. Log(μ) or μ on the table represents the mean model. Considering the random effects of Regions and the specific farming communities, the final mean model for the Gaussian HGLM does not include access to credit, Study

tour, demonstrative practicals, Networking events and post-harvest equipments. In the counterpart model for the Gamma HGLM, only number of farmers and the cultivated plot size were significant contributors to crop yield when the random effects of Regions and the specific farming communities are considered in the model.

Table 4.8: Comparative Model Estimates for Gaussian HGLM and Gamma HGLM

		GAUSSIAN HGLM 1				GAMMA HGLM 1			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
log(μ)	(Intercept)	5869.6	1098.21	5.3447	0.000163	8.425926	0.46619	18.07397	<0.00001
	(Credit) 1								
	(Crop) 2	-3489.1	623.73	-5.594	0.000115				
	(Training) 1	-2598	706.64	-3.6765	0.002137				
	(Tour) 1								
	(Practical) 1								
	(Networking) 1								
	(Equipment) 1								
	Farmers	-236.6	50.49	-4.6867	0.00043	-118	38.52	-3.064	0.005981
	Plot size	577.2	23.13	24.9531	<0.00001	521.1	23.55	22.1256	<0.00001
log(λ)	(Intercept)	17.95	0.0855	209.9415	<0.00001	-1.624	0.09102	-17.8422	<0.00001
	Province	-13.96	0.8563	-16.3027	<0.00001	-3.958	0.8563	-46222	0.000064
	Community	-11.94	0.3922	-30.4436	<0.00001	-1.596	0.3922	-4.0694	0.000246
				Selection Criterion	Gaussian HGLM-1	Gamma HGLM-1			
				-2ML(-2 h)	16421.56	15678.71			
				-2RL(-2$p_{beta}(h)$)	16288.60	15729.01			
				cAIC	16441.60	15649.61			

4.4.3 Hierarchical Generalized Linear Models (HGLM 2)

HGLM 2 is an extension of the above discussed Hierarchical Generalised model (HGLM 1). In section 4.3.3 of this chapter, we introduce and demonstrate a useful alternative to modelling isolated discrepancies as being caused by variation in the dispersion, and to seek covariates that may account for them with the help of the techniques of joint modelling of mean and dispersion (Lee and Nelder, 2010). With the success stories of the HGLM (Lee and Nelder, 2010), there was the need to extend the HGLM to enable models with structured dispersion as used in the analysis data from quality improvement experiments (Nelder and Lee, 1991, 1998).

HGLM 2 therefore comprises of a fixed effects model from a known distribution, a random effects model allowed to follow conjugates of arbitrary distributions from the GLM family and a dispersion model as described in section 4.3.3. Figures 4.9 and 4.10 represents the diagnostic plots for the Gaussian and Gamma H-GLM's (2) respectively.

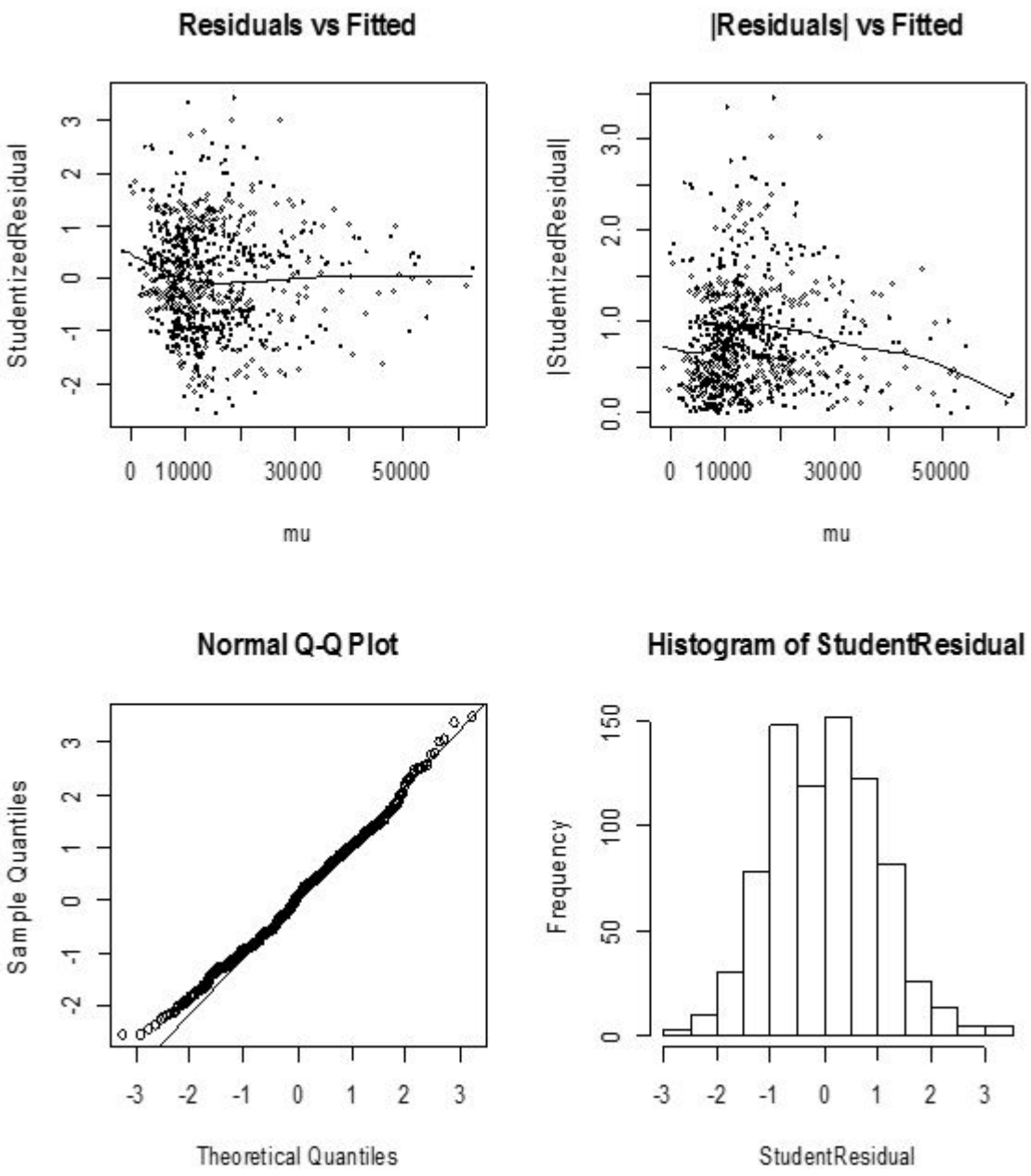

Figure 4.9: Diagnostic plots of Gaussian HGLM 2 for crop yield

From Figure 4.9 the diagnostic plots have several excellent features compared to the Gaussian HGLM (1) diagnostic plots in Figure 4.7. The gamma HGLM 2 diagnostic plots of Figure 4.10 also shows an incredible performs over the first gamma HGLM 1 of Figure 4.8. In addition, the histogram of residuals is highly symmetric. These are very good indications of an appropriate model. However this thesis seeks to present the very best of models hence the very minor defects present in the histogram may suggest something can be done to improve the

model.

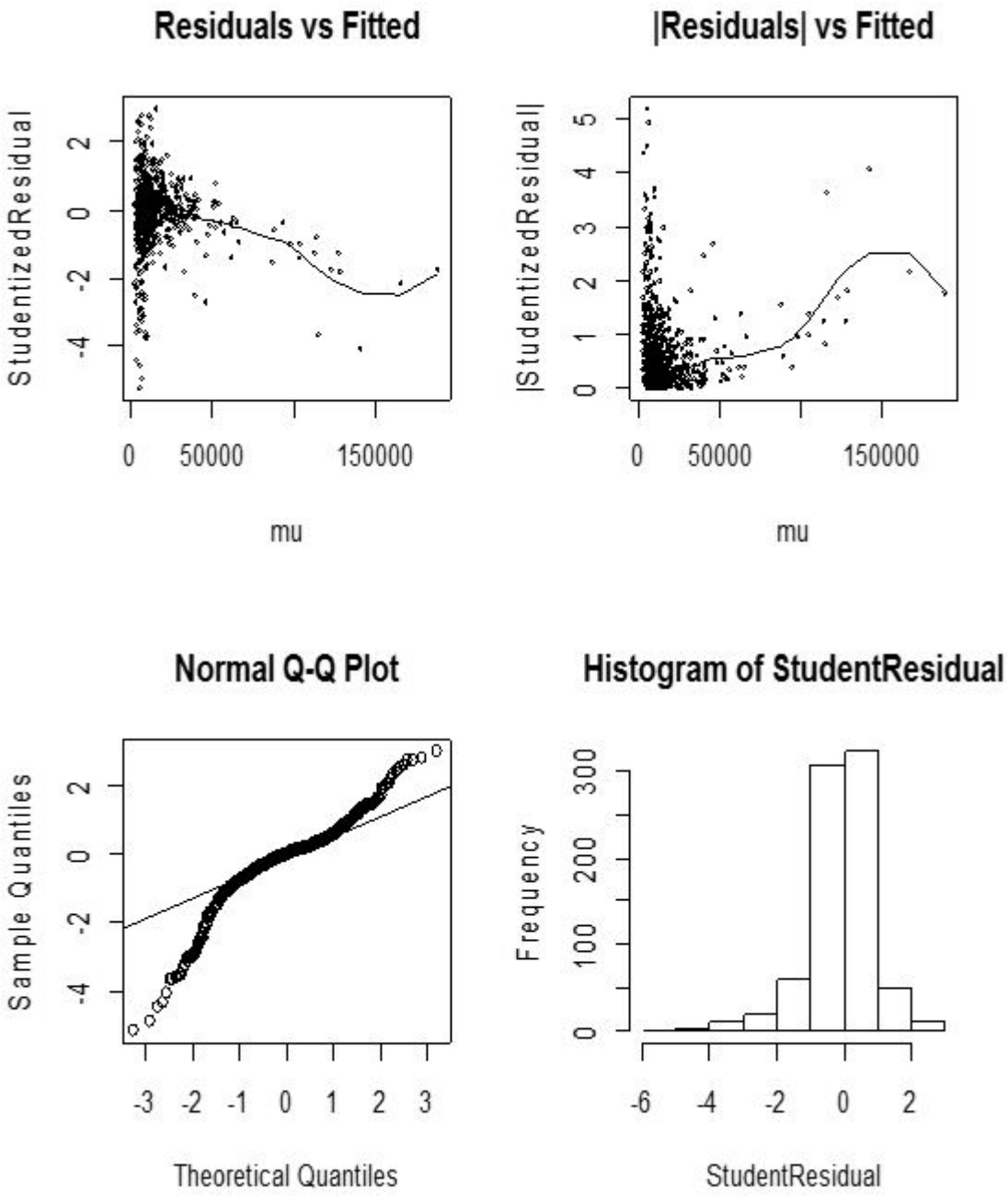

Figure 4.10: Diagnostic plots of Gamma HGLM 2 for crop yield

4.4.4 Model Interpretation

Table 4.9 represents the model parameter estimates for both the Gaussian and the Gamma HGLM 2. Log(μ) or μ on the table represents the mean model whereas log(ϕ) represents the dispersion model. The final mean model for the Gaussian HGLM 2 does not include access to Credit, networking events as well as

post-harvest equipments where as the dispersion model excludes only number of farmers, suggesting that this variable does not introduce any form of discrepancy. In the final mean model for the Gamma HGLM 2, demonstrative practicals, Networking events and post harvest equipments are excluded where as the dispersion model includes access to credit, training and post harvest equipments, excluding the rest of the variables.

Table 4.9: Model Estimates for Gaussian and Gamma distributed HGLM 2

		GAMMA HGLM 2				GAUSSIAN HGLM 2			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
$\log(\mu)$	(Intercept)	8.3421	0.461262	18.0854	<0.00001	4753.8	697.2	6.8184	0.000023
	(Credit) 1								
	(Crop) 2					-2398.3	405.08	-5.9204	0.000074
	(Training) 1					-1776.5	398.34	-4.4597	0.000609
	(Tour) 1					1164.8	416.4	2.7974	0.009439
	(Practical) 1	-0.05225	0.029128	-1.7939	0.043189	-726	367.79	-1.974	0.038317
	(Networking) 1								
	(Equipment) 1								
	Farmers	0.006923	0.002516	2.7514	0.005741	-118	38.52	-3.064	0.005981
	Plot size	0.030736	0.001151	26.7071	<0.00001	521.1	23.55	22.1256	<0.00001
$\log(\phi)$	(Intercept)	-3.50947	0.45707	-7.6782	<0.00001	15.32357	0.195893	78.224	<0.00001
	(Credit) 1	0.46636	0.20552	2.26917	0.016571	0.350278	0.117581	2.979	0.006916
	(Crop) 2					-0.46499	0.111278	-4.179	0.000945
	(Training) 1	1.16659	0.22775	5.12224	0.000018	0.458066	0.126042	3.6342	0.00229
	(Tour) 1					0.643937	0.11427	5.6352	0.000108
	(Practical) 1					0.303105	0.108174	2.802	0.009365
	(Networking) 1					-0.51501	0.119565	-4.307	0.000772
	(Equipment) 1	0.44948	0.20936	2.14692	0.021391	0.294979	0.115258	2.5593	0.014204
	Farmers								
	Plot size					0.056217	0.004127	13.622	<0.00001
$\log(\lambda)$	(Intercept)								
	Province	-3.627	0.8563	-4.2356	0.000162	-11.1	0.8563	-12.9627	<0.00001
	Community	-1.641	0.3922	-4.1841	0.000184	-10.95	0.3922	-27.9194	<0.00001

		Selection Criterion	Gaussian HGLM-1	Gaussian HGLM-2
		-2ML(-2 h)	15509.20	15982.81
		-2RL(-2$p_{beta}(h)$)	15564.50	15858.20
		cAIC	15477.20	16002.80

From the dispersion model in Table 4.9, it is observed that, in relying on the Gaussian mean model for crop yield, we record a dispersion of 15.323. However we also observe that the contribution of some of the covariates in the dispersion model to this dispersion value increases it while others tend to decrease it. Once a covariate which accounts for the discrepancies can be found, we get a model-based solution which can be checked in the future.

In the Gaussian model for example, covariates such as access to credit, Training, study tour, demonstrative practicals, post harvest equipments and plot size increases the dispersion significantly and should be carefully dealt with or checked once we aim at reducing the discrepancies between the data and the fitted values produced by the crop yield model.

Also in the Gamma HGLM 2 dispersion model, covariates such as access to credit, Training, and Post harvest equipments tends to increases the dispersion significantly and should be carefully dealt with or checked once we aim at reducing the discrepancies between the data and the fitted values produced by the crop yield model.

By model fitness criteria, the Gamma HGLM 2 performed far better than the Gaussian distributed HGLM 2 by both the AIC ($-2ML(-2h)$), BIC ($-2RL(-2p_{beta}(h))$ as well as the cAIC as evident in the last row of Table 4.9

4.4.5 Hierarchical Generalized Linear Models for Quality Improvement

The study again seeks to strongly recommend that, if we really aim at controlling significantly, the effects of structured dispersions, even in the presence of correlated random errors, the techniques of HGLM 2 as a means of improving the quality should be the number one option. This the researcher has demonstrated

using the crop yield data with two random effects resulting from the regional and community variations in this thesis. Table 4.10 below reveals that the initial Gaussian HGLM even though was satisfactory mixed model (HGLM 1), modelling both mean and dispersion (HGLM 2) improves the quality of the same Gaussian distributed model significantly.

Table 4.10: Model criteria for Gaussian HGLM 1 and Gaussian HGLM 2

Selection Criterion	Gaussian HGLM 1	Gaussian HGLM 2
-2ML(-2 h)	16421.56	15982.81
-2RL(-2$p_{beta}(h)$)	16288.60	15858.20
cAIC	16441.60	16002.80

Similar can be said of the Gamma HGLM 1 and HGLM 2 as evident in Table 4.11 below confirming the fact that HGLM 2 improves model quality of mixed models with structured dispersions and significantly reduces the large standard errors resulting from the correlated random effects.

Table 4.11: Model criteria for Gamma HGLM 1 and Gamma HGLM 2

Selection Criterion	Gamma HGLM 1	Gamma HGLM 2
-2ML(-2 h)	15678.71	15509.20
-2RL(-2$p_{beta}(h)$)	15729.01	15564.50
cAIC	15649.61	15477.20

4.5 Discussion

4.5.1 Variable selection

In section 4.2, the study sort to select significant variables among many potential ones to be included in a model via penalized methods. The researcher have compare the sparsity and number of significant crop yield variables selected by the three penalized methods; LASSO, SCAD, and H-likelihood all through simulation studies and by the real data (See Table 4.1 and 4.2). All these methods

have common advantages over the classical selection procedures; they are computationally simpler, the derived sparse estimators are stable, and they facilitate higher prediction accuracies. The study have shown how to select important variables in general semi-parametric models through those penalized methods. The study have demonstrated via numerical studies and data analysis that the proposed procedure with H-Likelihood performs best followed by SCAD, with LASSO coming last (See Table 4.1 and Table 4.3).

Basically, all these penalized methods of variable selection were developed in the wake of the two fundamental limitations traditional variable selection procedures; First, when the number of predictors p is large, it is computationally infeasible to perform subset selection. Second, subset selection is extremely variable because of its inherent discreteness (Breiman, 1996; Fan and Li,2001). To overcome these difficulties, several other penalties have been proposed. The L_2-penalty yields a ridge regression estimation, but it does not perform variable selection. With the L_1- penalty, specifically, the PLS estimator becomes the least absolute shrinkage and selection operator (LASSO), which thresholds predictors with small estimated coefficients (Tibshirani, 1996).

LASSO is a popular technique for simultaneous estimation and variable selection, ensuring high prediction accuracy, and enabling the discovery of relevant predictive variables. Donoho and Johnstone (1994) selected significant wavelet bases by thresholding based on an L_1 penalty.

LASSO has been criticized on the grounds that a single tuning parameter λ is used for both variable selection and shrinkage. It typically ends up selecting a model with too many variables to prevent over shrinkage of the regression coefficients (Radchenko and James, 2008); otherwise, regression coefficients of selected variables are often over-shrunken. This assertion is highly confirmed by the re-

sults of this in Table 4.2.

To overcome this problem, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty for oracle variable selection. More recently, Zou (2006) showed that LASSO does not satisfy Fan and Li's (2001) oracle property, and proposed the adaptive LASSO. Based on the findings of this study, we also propose the H-likelihood approach by Lee and Nelder (2009), as the best in crop yield variable selection and we do so on the basis that, compared to other forms of penalized methods ie. LASSO and SCAD, the H-likelihood approach (Lee and Nelder 2009) facilitates higher prediction accuracy since it has least estimated penalized cross validated errors (see table 4.3)

An additional advantage of this method is that it can be easily implemented by a slight modification to the existing h-likelihood estimation procedure. Thus our method can be straightforwardly applied to variable selection in practical random-effect models such as generalized linear mixed models or HGLMs (Lee et al., 2006), etc.

4.5.2 Crop yield models for fixed covariates

Generalized linear models (GLMs) of Nelder and Wedderburn (1972) are a standard tool for analyzing data in various types of responses, continuous quantities, counts, proportions and positive quantities. However, GLMs allow the regression (or fixed effect) model only for the mean of independent responses. Model checking in the case of ordinary linear models is based on examination of the model residuals, which contain all the information in the data, not explained by the systematic part of the model. Examination of residuals is also the chief means for model checking in the case of GLMs, but in this case the standardization of residuals is both necessary and a little more difficult.

For GLMs the main reason for not simply examining the raw residuals, is the difficulty of checking the validity of the assumed mean variance relationship from the raw residuals. However if raw residuals are plotted against fitted values it takes an extraordinary ability to judge whether the residual variability is increasing in proportion to the mean, as opposed to, say, the square root or square of the mean. For this reason it is usual to standardize GLM residuals, in such a way that, if the model assumptions are correct, the standardized residuals should have approximately equal variance, and behave, as far as possible, like residuals from an ordinary linear model. Once we have standardized residuals we plot them to try and find evidence that the model assumptions are not met.

The main useful plots are: Standardized residuals against fitted values. A trend in the mean of the residuals violates the independence assumption and often implies that something is wrong with the model from the mean of the response - perhaps a missing dependence, or the wrong link function. A trend in the variability of the residuals is diagnostic of a problem with the assumed mean variance relationship - i.e. with the assumed response distribution. Standardized residuals against all potential predictor variables (selected or omitted from the model). Trends in the mean of the residuals can be very useful for pinpointing missing dependencies of the mean response on the predictors.

Normal QQ plots can be useful for highlighting problems with the distributional assumptions, in cases where the response distribution can be well approximated by a normal distribution (with appropriate non-constant variance). For example Poisson residuals for a response with a fairly high mean fall into this category. Plots of standardized residuals against leverage are useful for highlighting single points that have a very high influence on the model fitting. Leverage is a measure of how influential a data point could be, based on the distance of its predictor variables from the predictors of other data. We also used the histogram of resid-

uals. If the distributional assumption is right it shows symmetry provided the deviance residual is the best normalizing transformation. In GLMs responses are independent, so that these model-checking plots assume that residuals are almost independent. Care will be necessary when we extend these residuals to correlated errors in later techniques employed in this thesis.

From Figure 4.3 the diagnostic plots have several satisfactory features although not the best. The running mean in the plot of residuals against fitted values shows no form of marked trend, and the plot of absolute residuals has a relatively stable slope, indicating that the variance is constant and satisfies the independence assumption, that the right link function was specified and also indicates no missing dependency. The normal plot shows no discrepancy. These are very good indications of an appropriate model. However this thesis seeks to present the very best of models hence the very minor defects present in the histogram may suggest something can be done to improve the model.
We sort to remove any likely defects by moving to a GLM with gamma errors and a log link. The model-checking plots (See Figure 4.4)are appreciably better than for the normal Gaussian model and more improved.

These two models (Gaussian and gamma GLM) are not nested and have different distributions for the response, which makes direct comparison problematic. The AIC criterion, which is minus twice the maximized likelihood plus twice the number of parameters, has often been used as a way to choose between models. Smaller values are preferred (See Table 4.4). However, when computing a likelihood, it is common practice to discard parts that are not functions of the parameters.

This has no consequence when models with same distribution for the response are compared since the parts discarded will be equal. For responses with differ-

ent distributions, it is essential that all parts of the likelihood be retained. The large difference in AIC for these two models indicate that this precaution was not taken. Nevertheless, we note that the null deviance for both models is almost the same while the residual deviance is smaller for the gamma GLM. This improvement relative to the null indicates that the gamma GLM should be preferred here. Note that purely numerical comparisons such as this are risky and that some attention to residual diagnostics, scientific context and interpretation is necessary.

Statistics from the Gaussian and gamma GLM are given in Table 4.4. Because the gamma model is not a standard linear model the researcher used the AIC for model comparison, and Table 4.4 strongly indicates the improvement in fit from the gamma GLM over the normal Gaussian models. Referring back to the summary, it seems that when there is a unit change in no. of farmers and the FBO grew soy beans and received Training, the yield is lower than you would expect from those growing Maize, for each unit change in plot size (although all the factor coefficients are significantly different from 0). The coefficients gives the expected increase in Crop yield when one unit in the plot size (referring back to the summary, plot size seem to lead to a significant increase in Crop yield, on their own).

4.5.3 Joint-Generalized Linear Models for Quality Improvement

Nelder and Lee (1991) defined joint GLMs (JGLMs), which allow regression models for both the mean and dispersion. See Aitkin (1987) and Smyth (1989) for earlier treatment of models of this type. The dispersion model for both distributions are of very useful importance in determining the actual variable that are accounting for the discrepancies that may exist between observed crop yield and the estimated crop yield. Discrepancies between the data and the fitted values produced by the model fall into two main classes, isolated or systematic.

Systematic discrepancies in the fit of a model imply that the model is deficient rather than the data. There is a variety of types of systematic discrepancy, some of which may mimic the effects of others. For this reason it is hard, perhaps impossible, to give a foolproof set of rules for identifying the different types. This type even though important in statistical modelling, the technique of Joint GLM demonstrated in this dissertation seeks to correct the possible presence of an isolated discrepancy.

Isolated discrepancies appear when a few observations only have large residuals. Such residuals can occur if the observations are simply wrong, for instance where 129 has been recorded as 192. Such errors are understandable if data are hand recorded, but even automatically recorded data are not immune. Robust methods were introduced partly to cope with the possibility of such errors; for a description of robust regression in a likelihood context see, e.g. Pawitan (2001, Chapters 6 and 14). Observations with large residuals are systematically down weighted so that the more extreme the value the smaller the weight it gets. Total rejection of extreme observations (outliers) can be regarded as a special case of robust methods. Robust methods are data driven, and to that extent they may not indicate any causes of the discrepancies.

A useful alternative is to seek to model isolated discrepancies as being caused by variation in the dispersion, and to seek covariates that may account for them. This techniques of joint modelling of mean and dispersion (Lee and Nelder, 2010) is what we have proposed and demonstrated by this study.

Figures 4.5 and 4.6 represents the diagnostic plots for the Gaussian and Gamma joint-GLM's respectively. From Figure 4.5 the diagnostic plots have several excellent features compared to the Gaussian ordinary GLM diagnostic plots in Figure

4.3. The plot of absolute residuals has a very stable slope, indicating that the variance is constant and satisfies the independence assumption, that the right link function was specified and also indicates no missing dependency. The normal plot also shows no discrepancy. In addition, the histogram of residuals is almost symmetric. These are very good indications of an appropriate model and an excellent improvement over the counterpart Gaussian GLM in Figure 4.3. The gamma joint GLM diagnostic plots of Figure 4.6 as, also shows an incredible performs over the first gamma GLM of Figure 4.4. The model-checking plots are appreciably similar to the normal Gaussian joint GLM and both demonstrate an excellent improvement of their GLM's.

From the dispersion model in Table 4.5, it is observed that, in relying on the Gaussian mean model for crop yield, the study recorded a dispersion of 14.916. However we also observe that the contribution of some of the covariates in the dispersion model to this dispersion value increases it while others tend to decrease it. Once a covariate which accounts for the discrepancies can be found, we get a model-based solution which can be checked in the future.

In the Gaussian model for example, covariates such as access to credit, Training, study tour, demonstrative practicals and plot size increases the dispersion significantly and should be carefully dealt with or checked once we aim at reducing the discrepancies between the data and the fitted values produced by the crop yield model. Also in the Gamma Joint-GLM dispersion model, covariates such as access to credit, Training, demonstrative practicals, Post harvest equipments and plot size increases the dispersion significantly and should be carefully dealt with or checked once we aim at reducing the discrepancies between the data and the fitted values produced by the crop yield model.

In the model fitness criteria, the Gaussian Joint-GLM performed far better than

the Gamma distributed Joint-GLM by both the AIC, BIC as well as the cAIC as evident in the last row of Table 4.5

JGLMs have been broadly used for the analysis of quality-improvement experiments (Lee and Nelder, 1998). The study also seeks to strongly recommend this techniques of joint modelling of mean and dispersion as a means of improving the quality of all forms of models that fall under the general class of generalized linear model and its extensions. This the researcher have demonstrated using the crop yield data in this thesis. Table 4.6 reveals that the initial Gaussian GLM even though was satisfactory mean model, modelling both mean and dispersion (Joint-GLM) improves the quality of the same Gaussian distributed model significantly. Similar can be said of the Gamma GLM and Joint GLM as evident in Table 4.7.

4.5.4 Crop yield models for fixed and random covariates

GLMs are extended to generalized linear mixed models (GLMMs), in which the linear predictor of a GLM is allowed to have, in addition to the usual fixed effects, random effects following a normal distribution (Breslow and Clayton, 1993; Molenberghs and Verbeke, 2005).

The Gaussian HGLM 1 diagnostic plots (See Figure 4.7) have some satisfactory features although not the best. The running mean in the plot of residuals against fitted values shows no form of marked trend, even though the plot of absolute residuals has a relatively unstable slope. This does not indicate that the variance is not constant and may not satisfies the independence assumption strictly. It rather suggest the presence of some correlated random effect in the fitted model as expected. The histogram of residuals is almost symmetric. These are satisfactory indications of an appropriate model. Similar is said of the Gamma distributed model in Figure 4.8.

In both models however (See Table 4.8), plot size cultivated remains the only positive significant contributor to crop yield. High standard errors are observed in the HGLM 1 compared to the GLM and the JGLM and this is due to the presence of correlated random errors resulting from the inclusion of the two random effects; Regions and Communities.

4.5.5 Hierarchical Generalized Linear Models for Quality Improvement

Although the normal distribution is convenient for specifying correlations among the random effects, the use of other distributions for the random effects greatly enriches the class of models. Lee and Nelder (1996) introduced hierarchical generalized linear models (HGLMs), in which the distribution of random effects can be any conjugate distribution for the GLM family of distributions. Dispersion parameters of the random components and the residual variance (overdispersion) can be further modelled as regression models with random effects.

In the statistical literature unobservables appear with various names such as random effects, latent processes, factor, missing data, unobserved future observations, potential outcomes etc. Handling of such unobservables is the key to new extended likelihood inferences. Lee and Nelder (1996, 2006) and Lee et al. (2006) have shown how to model and make inferences using the h-likelihood. Inferences about unobservables can be made without resorting to an empirical Bayes framework (Lee and Nelder, 2010). A single algorithm, iterative weighted least squares, can be used throughout all new models and requires neither prior distributions of parameters nor multi-dimensional quadrature.

From Figure 4.9 the diagnostic plots have several excellent features compared to the Gaussian HGLM (1) diagnostic plots in Figure 4.7. The running mean

in the plot of residuals against fitted values shows no form of marked trend at all, and the plot of absolute residuals has an almost stable slope, indicating that the variance is constant and satisfies the independence assumption, that the right link function was specified and also indicates no missing dependency. The normal plot also shows no discrepancy. In addition, the histogram of residuals is almost symmetric. These are very good indications of an appropriate model and an excellent improvement over the counterpart Gaussian HGLM (1) in Figure 4.7. The gamma HGLM 2 diagnostic plots of Figure 4.10 also shows an incredible performs over the first gamma HGLM 1 of Figure 4.8.

HGLMs consist of the three objects, namely the data, fixed unknown constants (parameters) and unobserved random variables (unobservables). Traditional Bayesian models consist of the two objects, the data and unobservables, while frequentist's (or Fisher's) models consist of the data and parameters. By allowing all three objects in the statistical modelling it is possible to describe various features in the data, for example, within-subject correlation in longitudinal studies, smooth spatial and temporal trends, function fittings, and factor analysis, heteroskedasticity, heavy-tailed distributions, robust modellings and sparse variable selections.

Table 4.10 reveals that the initial Gaussian HGLM 1 even though was satisfactory mixed model (HGLM 1), modelling both mean and dispersion (HGLM 2) improves the quality of the same Gaussian distributed model significantly. Similar can be said of the Gamma HGLM 1 and HGLM 2 as evident in Table 4.11 confirming the fact that HGLM 2 improves model quality of mixed models with structured dispersions and significantly reduces the large standard errors resulting from the correlated random effects.

In summary, HGLMs provide a rich class of hierarchical models, giving many inferential tools for testing and checking models, and are especially helpful for the

analysis of data from multi-centre field trials. Inferences about both population-average and subject-specific responses can be effectively drawn from a common HGLM.

Chapter 5

Conclusions

5.1 Introduction

This study proceeded on two paths; to select significant variables among many potential ones to be included in a model via penalized methods and to also propose and demonstrate the excellent performance of higher levels and very recent extensions of the Generalized Linear Models (GLM); Joint Generalized Linear Models (JGLM) and Hierarchical Generalized Linear Models (HGLM) in the global quest to developing Statistical Models with highest model accuracy.

The researcher sought to propose the H-Likelihood method of penalized variable selection as well as the unified JGLM and HGLM with gamma fixed and mixed effects as best methods useful for variable selection and modelling crop yield in the three Northern regions of Ghana respectively. After the highly rigorous processes and data analysis, the study concludes on the following

5.2 Conclusion

1. H-Likelihood method of penalized variable selection is the best method so far existing. It does both selection of significant variables and estimation of their coefficients simultaneously with the least penalize cross-validated errors compared to the SCAD and the LASSO

2. In modelling the effects of fixed physical support services given to farmer based organizations on the crop yield, the GLM with assumed fixed dispersion is highly not recommended. The study concludes that the initial

GLM even though was a satisfactory mean model, modelling both mean and dispersion (Joint-GLM) improves the quality of the models significantly.

3. Also in the case of modelling both fixed and random effects, the HGLM 2 which has the ability of specifying different suitable fixed effects model from a known distribution, a random effects model allowed to follow conjugates of arbitrary distributions from the GLM family and a dispersion model is highly recommended. This study concludes that the GLMM and HGLM 1 are still highly satisfactory statistical models but the HGLM 2 performs far better, gives a more fitting models and improves the quality of the models significantly.

5.3 Recommendation

The researcher strongly recommend this technique of joint modelling of mean and dispersion as a means of improving the quality of all forms of models that fall under the general class of generalized linear model and its extensions.

The study strongly recommend the h-likelihood method of variable selection

The researcher strongly recommend the unified JGLM and HGLM with gamma fixed and mixed effects as best methods useful for modelling crop yield in the three Northern regions of Ghana.

The study finally recommend that a deliberate effort be put into strengthening the Agricultural support systems as a form of strategy for increasing crop production in Northern Ghana. Access to credit, training, access to post harvest equipments, access to demonstrative practicals and access to large plot size are the physical support services highly recommended by this study.

5.4 Areas of Further Research

The crop yield model considered in this study suggest that some physical support services; Access to credit facility, Training, Study tour, Demonstrative practical, Networking events and Post harvest Equipments, plays an important role in determining crop yields even though their individual and interaction effects on yield is not uniform across farmer base organizations. The researcher admit that but for the unavailability of data, as frequently the case in many parts of our world, extensive input data on farm management practices, soil condition, climate and other non-physical contributors to yield would have enriched our models. The study therefore suggest further research that would consider these.

Even though some work have started already in the HGLM laboratory at the Seoul National University - South Korea in this regard, the researcher suggest that other researchers consider extending the HGLM 2 to include the introduction of random effects in the structured dispersion model as a means of further improving the performance of the powerful HGLM 2. As at the completion of this dissertation, work on the fitting algorithm of such extension and the R-programming codes were at various stages of completion.

The method of JGLM and HGLM have not been applied in many areas of applied statistics despite its extraordinary ability of quality improvement. Researchers in the areas of financial modelling and medical and epidemiology should consider this methodology since it has the ability to model the volatility characteristic of data used in those areas.

REFERENCES

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium Information Theory, Ed. B. N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademia Kiado.

2. Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60, 255-265.

3. AlHassan, R. M., and X. Diao. 2007. Regional disparities in Ghana: policy options and public investment implications. International Food Policy Research Institute Discussion Paper No. 00693. http://www.ifpri.cgiar.org/sites/default/files/publications/gsspwp02.pdf.

4. Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society B, 32, 283-301.

5. Andersen, P.K., Klein, J.P., Knudsen, K. and Palacios, R.T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. Biometrics, 53, 1475-1484.

6. Anderson, T.W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. Journal of the American Statistical Association, 52, 200-203.

7. Andreini, M., van de Giesen, N., van Edig, A., Fosu, M. and Andah, W. 2000. Volta Basin water balance. ZEF Discussion Papers on Development Policy No 21. Center for Development Research, Bonn, Germany.

8. Androulakis, E., Koukouvinos, C. and Vonta, F. (2012), "Estimation and variable selection via frailty models with penalized likelihood," Statistics in Medicine, 31, 2223-2239.

9. Baltagi, B.H. (1995). Economic analysis of panel data. New York: Wiley.

10. Bayarri, M.J., DeGroot, M.H. and Kadane, J.B. (1988). What is the likelihood function? (with discussion). Statistical Decision Theory and Related Topics IV. Vol. 1, eds S.S. Gupta and J.O. Berger, New York: Springer.

11. Berger, J.O. and Wolpert, R. (1984). The Likelihood Principle. Hayward: Institute of Mathematical Statistics Monograph Series.

12. Berry, R.A., and W.R. Cline. 1979. Agrarian Structure and Productivity in Developing Countries. Baltimore: Johns Hopkins University Press.

13. Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002). A Stability Based Method for Discovering Structure in Clustered Data. Pacific Symposium on Biocomputing, 6-17.

14. Bickel, P. J. and Levina, E. (2006). Regularized estimation of large covariance matrices. Technical Report 716, Department of Statistics, University of California, Berkeley, CA.

15. Bickel, P. J., Ritov, Y., and Zakai, A. (2006). Some theory for generalized boosting algorithms. Journal of Machine Learning Research.

16. Bindlish, V. and R.E. Evenson (1997), 'The Impact of T and V Extension in Africa: The Experience of Kenya and Burkina Faso', The World Bank Research Observer, Vol. 12, No. 2, 183-201.

17. Birge, L. and Massart, P. (1997). From model selection to adaptive estimation. In D. Pollard, E. Torgessen, and G. Yang, eds., A Festschrift for Lucien Le Cam: Research papers in Probability and Statistics, pp. 55-87. Springer-Verlag, New York.

18. Birge, L. and Massart, P. (2001). Gaussian model selection. Journal of the European Mathematical Society, 3(3):203-268.

19. Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). Journal of the American Statistical Association, 57, 269-306.

20. Birkhaeuser, D., R.E. Evenson and G. Feder. 1991. "The Economic Impact of Agricultural Extension: A Review." Economic Development and Cultural Change 39:607-650.

21. Bjornstad, J.F. (1990). Predictive likelihood principle: a review (with discussion). Statistical Science. 5, 242-265.

22. Bjornstad, J.F. (1996). On the generalization of the likelihood function and likelihood principle. Journal of the American Statistical Association, 91, 791-806.

23. Bogetic, Y., M. Bussolo, X. Ye, D. Medvedev, Q. Wodon, and D. Boakye. 2007 (April). Ghana's growth story: How to accelerate growth and achieve MDGs? Background paper for Ghana Country Economic Memorandum. Washington, DC: World Bank.

24. Bovelstad,H.M. (2007). Predicting survival from microarray data-a comparative study. Bioinformatics, 23, 2080-2087.

25. Braimoh, A.K. and Vlek, P.L.G. 2005. Land-cover change trajectories in Northern Ghana. Environmental Management, 36, 356-373.

26. Breiman, L. (1996) Heuristics of instability and stabilization in model selection, Ann. Statist., 24, 2350-2383.

27. Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. International Statistical Review 60, 291-319.

28. Breslow, N.E. (1974). Covariance analysis of censored survival data. Biometrics, 30, 89-99.

29. Breslow, N.E. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88, 9-25.

30. Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373-384.

31. Breiman, L., 1996. Bagging predictors. Mach. Learning 24, 123-140.

32. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classi1cation and Regression Trees. Wadsworth and Brooks/Cole, Paci1c Grove, CA.

33. Breisinger, C., X. Diao, J. Thurlow, B. Yu, and S. Kolavalli. 2008. Accelerating Growth and Structural Transformation: Ghana's Options for Reaching Middle-Income Country Status. IFPRI Discussion Paper 00750. Washington, DC: IFPRI.

34. Breisinger, C. and Diao X. 2008. Economic Transformation in Theory and Practice: What are the Messages for Africa? IFPRI Discussion Paper 797. Washington, DC: IFPRI.

35. Buhlmann, P. (2006). Boosting for high-dimensional linear models. The Annals of Statistics, 34(2):559-583.

36. Buhlmann, P. and Yu, B. (2006). Sparse boosting. Journal of Machine Learning Research, 7:1001-1024.

37. Bunea, F.,Wegkamp, M. H., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. Journal of Statistical Planning and Inference, 136(12):4349-4364.

38. Bunea, F.,Wegkamp, M. H., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. Journal of Statistical Planning and Inference, 136(12):4349-4364.

39. Butler, R.W. (1990). Comment on "Predictive likelihood inference with applications" by J.F. Bjornstad. Statistical Science., 5, 255-259.

40. Butler, R.W. (1986). Predictive likelihood inference with applications (with discussion). Journal of the Royal Statistical Society B, 48, 1-38.

41. CAADP (2009). Comprehensive African Agriculture Development Programme: Framework for African Food Security, Pillar III.

42. Casella, G. and Berger, R. L. (1990) Statistical Inference. Belmont, California: Wadsworth & Brooks/Cole.

43. Casella, G. (1985). "An introduction to empirical Bayes data analysis," The American Statistician, 39, 83-87.

44. CGIAR Technical Advisory Committee, 1996. Report of the Fourth External Programme and Management Review of the International Institute of Tropical Agriculture (IITA). TAC Scretariat, Food and Agricultural Organization of the United Nations.

45. Chen C. and S. L. George, "The Bootstrap and Identification of Prognostic Factors via Cox's Proportional Hazards Regression Model,". Statistics in Medicine, Vol. 4, No. 1, 1985, pp. 39-46.

46. Chernoff, H. (1954). On the distribution of the likelihood ratio. Annals of Mathematical Statistics, 25, 573-578.

47. Coakley, C.W., Hettmansperger, T.P., 1993. A bounded-influence, high breakdown, efficient regression estimator. J. Amer. Statist. Assoc. 88, 872-880.

48. Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with Discussion). Journal of the Royal Statistical Society series B 49, 1-39.

49. Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman

50. Cruciani, G., Baroni, M., Clementi, S., Constantino, G. and Riganelli, D. (1992). "Predictive ability of regression models, Part II: Selection of the best predictive PLS model. Journal of Chemometrics, 6, 347-356

51. David B. Lobella, Marshall B. Burkeb, 2010. On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology. Agricultural and Forest Meteorology 150 (2010) 1443-1452. journal homepage: www.elsevier.com/locate/agrformet

52. Desmond A. F. and Chapman G. R. (1993). "Modelling Task Completion Data with Inverse Gaussian Mixtures". Journal of the Royal Statistical Society. Series C (Applied Statistics). Vol. 42, No. 4, pp. 603-613

53. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with Discussion). Journal of the Royal Statistical Society series B 39, 1-38.

54. Dejene, A. 1989. "The Training and Visit Agricultural Extension in Rainfed Agriculture: Lessons from Ethiopia." World Dev. 17:1647-1659.

55. Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81, 425-455.

56. Donoho, D. L. (2000). High dimensional data analysis: the curses and blessings of dimensionality. In Math Challenges of 21st Centuary (2000). American Mathematical Society. Plenary speaker. Available in: http://www.stat.stanford.edu/donoho/Lectures/AMS2000/.

57. Draper, N. R. and Smith, H. (1998). Applied regression analysis. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley and Sons, New York, 3rd ed.

58. Efron, B. and Morris, C. (1975) Data analysis using Stein's estimator and its generalizations, J. Am. Statist. Ass., 70, 311-319.

59. Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1-26.

60. Efron, B., 1982. The jackknife, the bootstrap and other resampling plans. In: CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 38. SIAM, Philadelphia, PA.

61. Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Amer. Statist. Assoc. 76, 312-319.

62. Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation (with discussions). Journal of the American Statistical Association, 99(467):619-642.

63. Evenson, R.E and G. Mwabu (2001). The effect of Agriculture Extension on Farm Yield in Kenya, (unpublished)

64. Evenson, R.E. 1997. "The economic contributions of agricultural extension to agricultural and rural development." In B.E. Swanson, R.P. Bentz, A.J. Sofranko. Improving agricultural extension: a reference manual. Rome, Italy: FAO, pp 27-36.

65. Fan, J. and Peng, H. (2004). Non-concave penalized likelihood with a diverging number of parameters. The Annals of Statistics, 32(3):928-961.

66. Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, vol 96, 1348-1360.

67. Fan, J. and Li, R. (2006) Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In Proc. of the Madrid Interna-

tional Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, European Mathematical Society, Zurich, 595-622.

68. Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. Journal of Royal Statistical Society. Series B 70, 849-911.

69. Fan, J. and Lv, J. (2010), "A selective overview of variable selection in high dimensional feature space," Statistica Sinica, 20, 101-148.

70. FAO (Food and Agriculture Organization of the United Nations). 2005. Fertilizer use by crop in Ghana. http://www.fao.org/docrep/008/a0013e/a0013e00.htm (retrieved June 10, 2012).

71. FAO (Food and Agriculture Organization of the United Nations). 2006. World Food Summit, Rome. http://www.fao.org/docrep/003/w3613e/w3613e00.htm (retrieved April 4, 2013).

72. Fisher, R.A. (1935). The Design of Experiments. Oliver and Boyd, Edinburgh.

73. Frank, I. and Friedman, J. (1993) A statistical view of some chemometrics regression tools. Technometrics, 35, 109-148.

74. FtM AGRA grant Narrative report, 2011. Linking Farmers to Markets (FTM) Project. Prepared for the Alliance for a Green Revolution in Africa (AGRA) by IFDC, 2011

75. Furnival, G. and Wilson, R. (1974). Regression by leaps and bounds. Technometrics 16, 499-511.

76. Gao H. M., Jiang J., Wilson B., Zhang W., Hong Jau-Shyoung and Lui Bin. (2002). "Microglial activation-mediated delayed and progressive degeneration of rat nigral dopaminergic neurons: relevance to Parkinson's disease. Journal of Neurochemistry, 81, 1285-1297.

77. Gautam, M. (1998), 'Returns to T and V Extension in Kenya: Some Alternative Findings', World Bank, Washington DC, mimeo.

78. Gautam, M. 2000. Agricultural extension: The Kenya experience: An impact evaluation. Operations Evaluation Studies. Washington, D.C.: World Bank.

79. Global Forum on Agriculture 2010. Policies for Agricultural Development, Poverty Reduction and Food Security. OECD Headquarters, Paris.November 2010. TAD/CA/APM/WP(2010)40.

80. Golbraikh A. and Tropsha A. (2002). "QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology". Journal of Chemical Information and Computational Science. 43 (1), pp 144-154.

81. Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). A distribution-free theory of nonparametric regression. Springer Series in Statistics. Springer-Verlag, New York.

82. Gui, J. and Li, H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics, 21, 3001-3008.

83. Ha, I. D. and Lee, Y. (2003), "Estimating frailty models via Poisson hierarchical generalized linear models," Journal of Computational and Graphical Statistics, 12, 663-681.

84. Ha, I.D. and Lee, Y. (2005a). Comparison of hierarchical likelihood versus orthodox BLUP approach for frailty models. Biometrika, 92, 717-723.

85. Ha, I.D. and Lee, Y. (2005b). Multilevel mixed linear models for survival data. Lifetime data analysis, 11, 131-142.

86. Ha, I.D., Lee, Y. and McKenzie, G. (2005). Model selection for multicomponent frailty models. Manuscript submitted for publication.

87. Ha, I.D., Lee, Y. and Song, J.K. (2001). Hierarchical likelihood approach for frailty models. Biometrika, 88, 233-243.

88. Ha, I.D., Lee, Y. and Song, J.K. (2002). Hierarchical likelihood approach for mixed linear models with censored data. Lifetime data analysis, 8, 163-176.

89. Ha, I. D., Sylvester, R., Legrand, C. and MacKenzie, G. (2011), "Frailty modelling for survival data from multi-centre clinical trials," Statistics in Medicine, 30, 2144-2159.

90. Hans C. van Houwelingen, Willi Sauerbrei, 2013. Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited. Open Journal of Statistics, 2013, 3, 79-102 (http://www.scirp.org/journal/ojs)

91. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

92. Harville D.A. (1976). Maximum likelihood approaches to variance component estimation and related problems (with discussion). Journal of the American Statistical Association 1977; 72:320-340.

93. Harrell, F. E., Lee K. L. and Mark D. B, (1996). "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors,". Statistics in Medicine, Vol. 15, No. 4, 1996, pp 361-387.

94. Hausman, J.A. (1978). Specification tests in econometrics. Econometrica, 46, 1251-1271.

95. Hertier, S., Ronchetti, E., 1994. Robust bounded-in5uence tests in general parametric models. J. Amer. Statist. Assoc. 89, 897-904.

96. Hocking, R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

97. Hurvich, C. and Tsai, C. L. (1989). Regression and time series model selection in small samples. Biometrika 76, 297-307.

98. H. Zou and T. Hastie, 2005. "Regularization and Variable Selection via the Elastic Net," Journal of the Royal Statistical Society Series B, Vol. 67, No. 2, 2005, pp 301-320

99. Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12(1):55-67.

100. Horst, W. J., and R. Hardter. 1994. Rotation of maize with cowpea improves yield and nutrient use of maize compared to maize mono-cropping in an alfisol in the northern Guinea Savanna of Ghana. Plant and Soil. 160:171-183.

101. Houwelingen H. C. and Cessie S. le, "Predictive Value of Statistical Models,". Statistics in Medicine, Vol. 9, No. 11, 1990, pp. 1303-1325.

102. Houwelingen H. C., Sauerbrei W.,(2013). "Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited". Open Journal of Statistics, 2013, 3, 79-102.

103. Hsiao, C. (1995). Analysis of panel data. Econometric Society Monograph, Cambridge: Cambridge University Press.

104. Huang, J., Horowitz, J. L. and Ma, S. G. (2006). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Technical Report no. 360, Department of Statistics and Actuarial Science, University of Iowa.

105. Huang, J., Horowitz, J. L. and Ma, S. G. (2006). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Technical

Report no. 360, Department of Statistics and Actuarial Science, University of Iowa.

106. Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. Ann. Statist. 33 1617-1642.

107. Huber, P. J. (1981). Robust Statistics. Wiley, New York.

108. IFAD (International Fund for Agricultural Development). 2012. Ghana: country programme evaluation. IFAD Publ. No. 84, May 2012. http://www.ifad.org/eval html/eksyst/doc/profile/pa/ghana2012.htm (retrieved February 19, 2013).

109. IFAD (International Fund for Agricultural Development). 2012. Ghana: country programme evaluation. IFAD Publ. No. 84, May 2012. http://www.ifad.org/eval html/eksyst/doc/profile/pa/ghana2012.htm (retrieved February 19, 2013).

110. IFAD (International Fund for Agricultural Development). 2012. Ghana: country programme evaluation. IFAD Publication. No. 84, May 2012. http://www.ifad.org/evaluation/public/html/eksyst/doc/profile/pa/ghana 2012.htm (retrieved February 19, 2013).

111. International Fund for Agricultural Development. 2003. IFAD in Ghana. International Fund for Agricultural Development, Rome.

112. Iizumi, T., Yokozawa, M., Nishimori, M., 2009. Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: application of a Bayesian approach. Agricultural and Forest Meteorology 149 (2), 333-348.

113. Izrailev, S., and Agrafiotis, D. K.,(2002) Variable selection for QSAR by artificial ant colony systems, SAR and QSAR in Environmental Research, Vol. 13, No. 3-4, pp. 417-423.

114. Jiao H, Wang S, Kamata A. (2005), Modelling local item dependence with the hierarchical generalized linear model. PubMed, J Appl Meas. 2005;6

(3):311-21.

115. Jorgensen, B. (1986). Some properties of exponential dispersion models. Scandinavian Journal of Statistics, 13, 187-198.

116. Jorgensen, B. (1987). Exponential dispersion models (with discussion). Journal of the Royal Statistical Society B, 49, 127-162.

117. Johnson, B. A., Lin, D. Y. and Zeng, D. (2008), "Penalized estimating functions and variable selection in semi-parametric regression models," Journal of the American Statistical Association, 103, 672-680.

118. Jouan-Rimbaud D., Walczak B., Poppi R. J., de Noord O. E., and Massart D. L. (1996). "Application of Wavelet Transform To Extract the Relevant Component from Spectral Data for Multivariate Calibration". Analytical Chemistry, 69 (21), pp 4317-4323

119. Jian Huang and Huiliang Xie (2007). "Asymptotic Oracle Properties of SCAD-Penalized Least Squares Estimators". Lecture Notes-Monograph Series. Vol. 55, pp. 149-166.

120. Kim, D., Lee, Y. and Oh, H.S. (2006). Hierarchical likelihood-based wavelet method for denoising signals with missing data. To appear in IEEE Signal Processing Letters.

121. Kim, J., Kim, Y., Kim, Y., 2008b. A gradient-based optimization algorithm for lasso. Journal of Computational and Graphical Statistics 17, 994-1009.

122. Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. Review of Economic Studies, 98, 361-393.

123. Kim, Y., Choi, H., Oh, H., 2008a. Smoothly clipped absolute deviation on high dimensions. Journal of the American Statistical Association 103, 1656-1673.

124. Knight, K., Fu, W.J., 2000. Asymptotics for lasso-type estimators. The Annals of Statistics 28, 1356-1378.

125. Kraybill, D., B. Bashaasha., and M. Betz. 2009. "Productivity and Marketed Surplus in Ugandan Agriculture, 1999-2007." Working paper, Department of Agric. Economics, Ohio State University.

126. Kubinyi H. (1996). "Evolutionary variable selection in regression and PLS analyses". Journal of Chemometrics. Volume 10, Issue 2, pages 119-133.

127. Kwon, S., Oh, S. and Lee Y. (2013), "The use of random-effect models for high-dimensional variable selection problems," Scandinavian Journal of Statistics 28, 56-78.

128. Leardi R., GonzÃ¡lezb A. L., (1998). "Genetic algorithms applied to feature selection in PLS regression: how and when to use them". Chemometrics and Intelligent Laboratory Systems. Volume 41, Issue 2, 195-207.

129. Lane, P.W. and Nelder, J.A. (1982). Analysis of covariance and standardization as instances of prediction. Biometrics, 38, 613-621.

130. Lauritzen, S.L. (1974). Sufficiency, prediction and extreme models. Scandinavian Journal of Statistics, 1, 128-134.

131. Lee, Y. and Oh, H. S. (2009), "Random-effect models for variable selection," Department of Statistics, Stanford University, Technical report No. 2009-4, 1-24.

132. Lee, Y. (1991). Jackknife variance estimators of the location estimator in the one-way random-effects model. Annals of the Institute of Statistical Mathematics, 43, 707-714.

133. Lee, Y. (2000). Discussion of Durbin Koopman's paper. Journal of the Royal Statistical Society B, 62, 47-48.

134. Lee, Y. (2001). Can we recover information from concordant pairs in binary matched paired? Journal of Applied Statistics, 28, 239-246.

135. Lee, Y. (2002a). Robust variance estimators for fixed-effect estimates with hierarchical-likelihood. Statistics and Computing, 12, 201-207.

136. Lee, Y. (2002b). Fixed-effect versus random-effect models for evaluating therapeutic preferences. Statistics in Medicine, 21, 2325-2330.

137. Lee, Y. (2004). Estimating intraclass correlation for binary data using extended quasi-likelihood. Statistical Modelling, 4, 113-126.

138. Lee, Y. and Nelder, J.A. and Pawitan, Y. (2006). Generalized Linear Models with Random Effects, Unified Analysis via H-likelihood. Boca Raton: Chapman and Hall/CRC.

139. Lee, D., Lee, W., Lee, Y. and Pawitan, Y. (2010), "Super sparse principal component analysis for high-throughput genomic data," BMC Bioinformatics, 11, 296.

140. Lee, D., Lee, W., Lee, Y. and Pawitan, Y. (2011a), "Sparse partial least-squares regression and its applications to high-throughput data analysis," Chemometrics and Intelligent Laboratory Systems, 109, 1-8.

141. Lee, W., Lee, D., Lee, Y. and Pawitan, Y. (2011b), "Sparse canonical covariance analysis for high-throughput data," Statistical Applications in Genetics and Molecular Biology, 10, 1-24.

142. Lee, Y. and Nelder, J.A. (1996). Hierarchical generalised linear models (with discussion). Journal of the Royal Statistical Society B, 58, 619-656.

143. Lee, Y. and Nelder, J.A. (1997). Extended quasi-likelihood and estimating equations approach. IMS Notes monograph series, edited by Basawa, Godambe and Tayler, 139-148.

144. Lee, Y. and Nelder, J.A. (1998). Generalized linear models for the analysis of quality-improvement experiments. Canadian journal of Statistics, 26, 95-105.

145. Lee, Y. and Nelder, J.A. (1999). The robustness of the quasi-likelihood estimator. Canadian Journal of Statistics, 27, 321-327.

146. Lee, Y. and Nelder, J.A. (2000a). The relationship between double exponential families and extended quasi-likelihood families. Applied Statistics, 49, 413-419.

147. Lee, Y. and Nelder, J.A. (2000b). Two ways of modelling overdispersion in non-normal data. Applied Statistics, 49, 591-598.

148. Lee, Y. and Nelder, J.A. (2001a). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect model and structured dispersion. Biometrika, 88, 987-1006.

149. Lee, Y. and Nelder, J.A. (2001b). Modelling and analysing correlated non-normal data. Statistical Modelling, 1, 7-16.

150. Lee, Y. and Nelder, J.A. (2002). Analysis of the ulcer data using hierarchical generalized linear models. Statistics in Medicine, 21, 191-202.

151. Lee, Y. and Nelder, J.A. (2003a). Robust Design via generalized linear models. Journal of Quality Technology, 35, 2-12.

152. Lee, Y. and Nelder, J.A. (2003b). False parsimony and its detection with GLMs. Journal of Applied Statistics, 30, 477-483.

153. Lee, Y. and Nelder, J.A. (2003c). Extended REML estimators. Journal of Applied Statistics, 30, 845-856.

154. Lee, Y, and Nelder, J.A. (2004). Conditional and marginal models: another view (with discussion). Statistical Science, 19, 219-238.

155. Lee, Y. and Nelder, J.A. (2005). Likelihood for random-effect models (with discussion). Statistical and Operational Research Transactions, 29, 141-182.

156. Lee, Y. and Nelder, J.A. (2006a). Double hierarchical generalized linear models (with discussion). Applied Statistics, 55, 139-185.

157. Lee, Y. and Nelder, J.A. (2006b). Fitting via alternative random effect models. Statistics and Computing, 16, 69-75.

158. Lee, Y, Nelder, J.A. and Noh, M. (2006). H-likelihood: problems and solutions. Statistics and Computing, revision.

159. Lee Y. Nelder J.A., 2002. "Analysis of ulcer data using hierarchical generalized linear models". Statist. Med. 2002; 21:191-202 (DOI: 10.1002/sim.978)

160. Lee, Y., Noh, M. and Ryu, K. (2005). HGLM modeling of dropout process using a frailty model. Computational Statistics, 20, 295-309.

161. Leng C. L., Y. Lin and G. Wahba,(2006) "A note on LASSO and Related Procedures in Model Selection," Statistica Sinica, Vol. 16, 2006, pp. 1273-1284.

162. Li, K.C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. The Annals of Statistics, 13(4):1352-1377.

163. Li, K.C. (1986). Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing. The Annals of Statistics, 14(3):1101-1112.

164. Li, K.C. (1987). Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics, 15(3):958-975.

165. Lobell, D.B., Ortiz-Monasterio, J.I., 2007. Impacts of day versus night temperatures on spring wheat yields: a comparison of empirical and CERES model predictions in three locations. Agronomy Journal 99 (2), 469-477.

166. Lu Xu, Wen-Jun Zhang, 2001. Comparison of different methods for variable selection. Analytica Chimica Acta 446 (2001) 477-483.

167. Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661-675.

168. Mallows, C. (1995). More comments on cp. Technometrics 37, 362-372.

169. Markatou, M., He, X., 1994. Bounded influence and high breakdown point testing procedures in linear models. J. Amer. Statist. Assoc. 89, 543-549.

170. Medium Term Agriculture Sector Investment Plan (METASIP) (2010). Ministry of Food and Agriculture (MOFA), Ghana.

171. McKay, A., and E. Aryeetey. 2004. Operationalizing pro-poor growth: A country case study on Ghana. A joint initiative of AFD, BMZ (GTZ, KfW Development Bank), DFID, and the World Bank. http://www.dfid.gov.uk/pubs/files/oppgghana.pdf. Accessed July 9, 2010.

172. McQuarrie, A. and Tsai, C. L. (1998). Regression and time series model selection. World Scientific.

173. McCullagh, P. (1980). Regression models for ordinal data (with discussion). Journal of the Royal Statistical Society B, 42, 109-142.

174. McCullagh, P. (1983). Quasi-likelihood functions. Annals of Statistics, 11, 59- 67.

175. McCullagh, P. (1984). Regression models for ordinal data (with discussion). Journal of the Royal Statistical Society B, 42, 109-142.

176. McCullagh, P. and Nelder, J.A. (1989). Generalized linear models, 2nd ed. Chapman and Hall, London.

177. McGilchrist, C. A. (1994), "REML Estimation for Survival Models with Frailty," Biometrics, 49, 221-225.

178. McCulloch, C.E. (1994). Maximum likelihood variance components estimation in binary data. Journal of the American Statistical Association, 89, 330-335.

179. McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association, 92, 162-170.

180. McCullagh, P. and Nelder, J.A. (1989). Generalized linear models - 2nd ed. London : Chapman and Hall, 1989.

181. McShane LM, Aamodt R, Cordon-Cardo C, Cote R, Faraggi D, Fradet Y, Grossman HB, Peng A, Taube SE, Waldman FM, and the National Cancer Institute Bladder Tumor Marker Network (1999). "Producibility of p53 immunohistochemistry in bladder tumors". Clinical Cancer Research, 6: 1854-1864

182. Meinshausen, N. and Buehlmann, P. (2010). "Stability Selection." Journal of the Royal Statistical Society, Series B, 72, 414-473.

183. Miller, A. (2002). Subset selection in Regression. New York: Chapman and Hill, 2nd ed.

184. MOFA. Ministry of Agriculture. 2007. Agriculture in Ghana 2006. Accra, Ghana: Statistics Research and Information Directorate.

185. MOFA (Ministry of Food and Agriculture). 2011. Agriculture in Ghana: facts and figures (2010). Statistics, Research and Information Directorate (SRID). http://mofa.gov.gh/site/wp-content/uploads/2011/10/AGRICULTURE-IN-GHANA-FF-2010.pdf (retrieved February 1, 2013).

186. Morris, M., V.A. Kelly, R.J. Kopicki and D. Byerlee. 2007. Fertilizer Use in African Agriculture: Lessons Learned and Good Practice Guidelines. Washington, D.C.: World Bank.

187. Moss, T., and S. Majerowicz. 2012. No longer poor: Ghana's new income status and implications of graduation from IDA. CGD Working Paper 300. Washington D.C.: Center for Global Development. http://www.cgdev.org/content/public (retrieved November 6, 2012).

188. Murphy, J.M., Sexton, D.M.H., Barnett, D.N., Jones, G.S., Webb, M.J., Collins, M., Stainforth, D.A., 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430 (7001), 768-772.

189. Nadaraya, E. A. (1964). On estimating regression. Theory of Probability and Its Applications, 10:186-190.

190. Noh, M., Ha, I. D. and Lee, Y. (2006) Dispersion frailty models and HGLMs. Statistics in Medicine,10:18-20.

191. Noh, M., Lee, Y. and Pawitan, Y. (2005). Robust ascertainment-adjusted parameter estimation. Genetic Epidemiology. 29, 68-75.

192. Norinder U. (1996). "Single and domain mode variable selection in 3D QSAR applications" Journal of Chemometrics. Volume 10, Issue 2, pages 95-105.

193. Nelder J.A. and Wedderburn R.W.M (1972) Generalized linear models. Journal of the Royal Statistical Society A, 135, 370-84.

194. Nelder, J.A. and Pregibon, D. (1987). An extended quasi-likelihood function. Biometrika, 74, 221-232.

195. Oehmke, J.F., Crawford, E.W., 1996. The impact of agricultural technology in sub-Saharan Africa. Journal of African Economies 5 (2), 271-292.

196. Owens, T., J. Hoddinott, and B. Kinsey. 2003. "The Impact of Agricultural Extension on Farm Production in Resettlement Areas of Zimbabwe." Economic Development and Cultural Change 51:337-357.

197. Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 33:1065-1076.

198. Park, M., Hastie, T., 2007. L_1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society, Series B 69, 659-667.

199. Pawitan, Y. (2001). In all likelihood : statistical modelling and inference using likelihood. Oxford: Clarendon Press.

200. Pawitan, Y. (2001) Estimating variance components in generalized linear mixed models using quasi-likelihood. Journal of Statistical Computation and Simulation, 69, 1-17.

201. Pawitan, Y., Reilly, M., Nilsson, E., Cnattingius, S. and Lichtenstein, P. (2004). Estimation of genetic and environmental factors for binary traits using family data. Statistics in Medicine, 23, 449-465.

202. Patterson H.D, Thompson R., (1997). Recovery of inter block information when block sizes are unequal. Biometrika 1971; 58:545-554.

203. Peng, S., Huang, J., Sheehy, J., Laza, R., Visperas, R., Zhong, X., Centeno, G., Khush, G., Cassman, K., 2004. Rice yields decline with higher night temperature from global warming. Proceedings of the National Academy of Sciences of the United States of America 101 (27), 9971-9975.

204. Pierce, D.A. and Schafer, D.W. (1986). Residuals in Generalized Linear Models. Journal of the American Statistical Association, 81, 977-986.

205. Policy, Planning, Monitoring and Evaluation Division (PPMED). 1991. Agriculture in Ghana. Facts and figures. Ministry of Food and Agriculture, Accra, Ghana.

206. Portnoy, S. (1984). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: I. Consistency. Ann. Statist. 12 1298-1309.

207. Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: II. Normal approximation. Ann. Statist. 13 1403-1417.

208. Potscher B.M, Leeb H.(2009) "On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding." Journal of Multivariate Analysis 2009; 100:2065-2082.

209. Purcell, D.L. and J.R. Anderson (1997), Agricultural Extension and Research: Achievements and Problems in National Systems, The World Bank, Washington, DC.

210. Radchenko, P. and James, G. (2008) Variable inclusion and shrinkage algorithms, J. Am. Statist. Ass., 103, 1304-1315.

211. Rissanen, J. (1984). Universal coding, information, prediction, and estimation. Institute of Electrical and Electronics Engineers. Transactions on Information Theory, 30(4):629-636.

212. Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). Statistical Science, 6, 15-51.

213. Roecker, E. (1991). Prediction error and its estimation for subset-selected models. Technometrics 20, 459-468.

214. Rocke, D.M., WoodruJ, D.L., 1996. Identification of outliers in multivariate data. J. Amer. Statist. Assoc. 91, 1047-1061.

215. Rosegrant, M.W., Paisner, M.S., Meijer, S. and Witcover, J. 2001. Global food projections to 2020: emerging trends and alternative futures. International Food Policy Research Institute, Washington, DC.

216. Rosenzweig, M.R. and H.P. Binswanger. 1993. "Wealth, Weather Risk and the Composition and Profitability of Agricultural Investments." Economics Journal, 103:56-78.

217. Ronchetti, E., Staudte, R.G., 1994. A robust version of Mallows's Cp. J. Amer. Statist. Assoc. 89, 550-559.

218. Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. The Annals of Mathematical Statistics, 27:832-837.

219. Rosset, S., Zhu, J., 2007. Piecewise linear regularized solution paths. The Annals of Statistics 35.

220. Sallah, P.Y.K., Twumasi-Afriyie, S., Badu-Apraku, B. and Dzah, B.D. 1993. Agronomic performance of different maturity groups of maize varieties in the interior savannah zone of Ghana. In: Proceedings of the Third Workshop on Improving Farming Systems in the Interior Savannah Zone of Ghana (eds H. Mercer-Quashie, K.O. Marfo, A.S. Langyintuo and R.K. Owusu). Nyankapala Agricultural Experimental Station, NAES, Tamale, Ghana.

221. Sanders, J.H., 1996. Measuring impacts of sorghum/millet technologies in sub-Saharan Africa. Paper presented at the INTSORMIL Principal Investigators Conference in Lubbock, Texas, September 21-22.

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

223. Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association, 82, 605-610.

224. Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88, 486-494.

225. Shao, J., 1996. Bootstrap model selection. J. Amer. Statist. Assoc. 91, 655-665.

226. Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88, 486-494.

227. Shao, J. (1997). An asymptotic theory for linear model selection (with discussions). Statistica Sinica, 7(2):221-264.

228. Sheehy, J.E., Mitchell, P.L., Ferrer, A.B., 2006. Decline in rice grain yields with temperature: models and correlations can give different estimates. Field Crops Research 98 (2-3), 151-156.

229. Shen, X., Pan, W., Zhu, Y. and Zhou, H. (2012). "On L_0 Regularization in High-dimensional Regression." Journal of the American Statistical Association, 91, 65-66

230. Shibata, R. (1981). An optimal selection of regression variables. Biometrika 68, 45-54.

231. Sharma, R. (2008) Systemic inflammatory response predicts prognosis in patients with advanced-stage colorectal cancer. Clinical Colorectal Cancer, 7, 331-337.

232. Shedden, K. (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat. Med., 14, 822-827.

233. Sen, A.K. 1962. "An Aspect of Indian Agriculture." Economics Weekly. Annual number: 243-66.

234. Simpson, J.R., Montgomery, D.C., 1998. The development and evaluation of alternative generalized-M estimation techniques. Comm. Statist. Simulation Comput. 27 1031-1049.

235. Soil Survey Staff 1994. Keys to soil taxonomy, 6th edn. USDA-SCS, Government Printing Office, Washington, DC.

236. Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society B, 64 , 583-640.

237. Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modelling (with discussions). The Annals of Statistics, 25(4):1371-1470.

238. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussions). Journal of the Royal Statistical Society. Series B, 36:111-147.

239. Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. Biometrics, 50, 1171-1177.

240. Sauerbrei W.,(1999). "The Use of Resampling Methods to Simplify Regression Models in Medical Statistics," Journal of the Royal Statical Society, Series C, Applied Statistics. Vol. 48, No. 3, pp 313-329.

241. Tao, F., Yokozawa, M., Zhang, Z., 2009. Modelling the impacts of weather and climate variability on crop productivity over a large area: a new process-based model development, optimization, and uncertainties analysis. Agricultural and Forest Meteorology 149 (5), 831-850.

242. Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 365 (1857), 2053-2075.

243. Tikhonov, A. N. (1943). On the stability of inverse problems. C. R. (Doklady) Acad. Sci. URSS (N.S.), 39:176-179.

244. Tibshirani, R. J. (1996) Regression shrinkage and selection via the LASSO, Journal of Royal Statist. Soc. B, 58, 267-288.

245. Tweedie, M.C.K. (1947). Functions of a statistical variate with given means, with special references to Laplacian distributions. Proceedings of the Cambridge Philosophical Society, 43, 41-49.

246. USAID (United States Agency for International Development). 2012. Introduction to Ghana. http://ghana.usaid.gov/content/introduction-ghana (retrieved February 3, 2013).

247. USAID (United States Agency for International Development). 2013. Agriculture and food security. http://www.usaid.gov/ghana/agriculture-and-food-security (retrieved February 3, 2013).

248. Van J. C., Houwelingen, 2001. "Shrinkage and Penalized Likelihood as Methods to Improve Predictive Accuracy," Statistica Neerlandica, Vol. 55, No. 1, 2001, pp. 17-34. doi:10.1111/1467-9574.00154

249. Van der Veen, M. "Analysis of Interfarm Variation in Rice Yields: An Economic Study of HYV Rice Production in Cavite Prov-ince, Philippines." Ph.D. dissertation, Pennsylvania State University.

250. Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components, Biometrics, 59, 254-262.

251. Vu, H.T.V., Segal, M.R., Knuiman, M.W. and James, I.R. (2001). Asymptotic and small sample statistical properties of random frailty variance estimates for shared gamma frailty models. Communications in Statistics: Simulation and Computation, 30, 581-595.

252. Vu, H.T.V. and Knuiman, M.W. (2002). Estimation in semiparametric marginal shared gamma frailty models. Australian and New Zealand Journal of Statistics, 44, 489-501.

253. Waldmann Patrick, Meszaros G., Gredler B., Fuerst C., SOlkner J. (2013). "Evaluation of the LASSO and the Elastic Net in genome-wide association studies." Frontiers in Genetics. Volume 4, Article 270, 1-11.

254. Walker, S.G. and Mallick, B.K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. Journal of the Royal Statistical Society B, 59, 845-860.

255. Wang, Y. (2004). Model selection. In Handbook of computational statistics, pp. 437-466. Springer-Verlag, Berlin.

256. Wang, H., Li, R. and Tsai, C. L. (2007), "Tuning parameter selectors for the smoothly clipped absolute deviation method," Biometrika, 94, 553-568.

257. Watson, G. S. (1964). Smooth regression analysis. Sankhya. Series A, 26:359-372.

258. Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika, 61, 439-447.

259. White, J.W., 2009. Comments on a report of regression-based evidence for impact of recent climate change on winter wheat yields. Agriculture, Ecosystems and Environment 129 (4), 547-548.

260. Wiggins, S., and H. Leturque. 2011. Ghana's sustained agricultural growth: putting underused resources to work. Overseas Development Institute. http://www.developmentprogress.org/sites/developmentprogress.org/files/resourc e-report/ghana-report-full.pdf (retrieved March 11, 2013).

261. Wilcox, R.R., 1997. Introduction to Robust Estimation and Hypothesis Testing. Academic Press, San Diego, CA.

262. Wilcox, R.R., 1998. The goals and strategies of robust methods. British Journal of Mathematics and Statistical Psychology. 51, 1-39.

263. Wisnowski, J.W., Simpson, J.R., Montgomery, D.C., Runger, G.C., 2002. An alternative prediction error criterion for regression model selection. Comm. Statist. Simulation Comput., to appear.

264. World Data Bank. 2013. World Development Indicators (WDI). http://databank.worldba and id=4 (retrieved January 10, 2013).

265. World Bank (1990), Staff Appraisal Report: Kenya, Second National Agricultural Extension Project, The World Bank, Washington, DC.

266. Wolfinger, R. (1993). Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistics, Computing and Simulations, 48, 233-243.

267. Wu, C.F.J., 1986. Jackknife, bootstrap, and other re-sampling methods in regression analysis. Ann. Statist. 14, 1261-1295.

268. Xinshen Diao, IFPRI,. 2010. Economic Importance of Agriculture for Sustainable Development and Poverty Reduction: Findings from a Case Study of Ghana. A paper presented at Global Forum on Agriculture 2010. Reference: TAD/CA/APM/WP(2010)40. pp 16, 19

269. Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika 92, 973-50.

270. Yun, S. and Lee, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. Computational Statistics and Data Analysis, 45, 639-650.

271. Yun, S. and Lee, Y. (2006). Robust estimation in mixed linear models with non-monotone missingness. Statistics in Medicine, in press.

272. Yun, S., Lee, Y. and Kenward, M.G. (2005). Using h-likelihood for missing observations. Manuscript prepared for publication.

273. Yun, S., Sohn, S.Y. and Lee, Y. (2006). Modelling and estimating heavy-tailed non-homogeneous correlated queues pareto-inverse gamma HGLMs with covariates. Journal of Applied Statistics, 33, 417-425.

274. Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. Journal of the American Statistical Association 86, 79-86.

275. Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R., and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. Journal of the American Statistical Association, 99(467):659-672.

276. Zhang, H. H. and Lu,W. (2007), "Adaptive Lasso for Cox's proportional hazards model," Biometrika, 94, 691-703.

277. Zhang, Y., Li, R. and Tsai, C. L. (2010), "Regularization parameter selections via generalized information criterion," Journal of the American Statistical Association, 105, 312-323.

278. Zhao, P., Yu, B., 2006. On model selection consistency of lasso. Journal of Machine Learning Research 7, 2541-2563.

279. Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67 301-320.

280. Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418-1429.

281. Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics 36, 1509-1533

282. 2010 Ghana millennium development goals report, 2012.United Nations Development Programme, Ghana and NDPC/GOG 2012

REFERENCES

Baidoo, A. (2013). The spread of hiv/aids in ghana. *International journal for the Spread HIV/AIDS*, 35:210–240.

Barnett, W. A., He, Y., and Yansi, F. (2002). Stabilization policy as bifurcation selection: Would stabilization policy work if the economy really were unstable. ,. *Journal of Science Maths*, 5:1025–1052.

Okyere, G. A. (2013a). *Introduction to LaTex*.

Okyere, G. A. (2013b). Mathematics is for great thinkers. *Ghana Mathematical Journal*, 34:23.

Appendix A

Table 5.1: Comparative Model Estimates for Gaussian GLM and Gaussian Joint-GLM's

		GAUSSIAN JOINT-GLM				GAUSSIAN GLM			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
$\log(\mu)$	(Intercept)	2789.82	630.85	4.4223	0.000645	5869.60	1104.53	5.3141	0.00017
	(Credit) 1	-694.45	368.33	-1.8854	0.044395				
	(Crop) 2	-13444.65	361.52	-3.7194	0.001991	-3489.1	627.32	-5.562	0.00012
	(Training) 1					-2598	710.71	-3.6555	0.002213
	(Tour) 1								
	(Practical) 1								
	(Networking) 1	831.65	358.33	2.3209	0.021354				
	(Equipment) 1	-983.56	339.32	-2.8986	0.007944				
	Farmers	-87.66	34.80	-2.5187	0.015246	-236.6	50.78	-4.6599	0.000448
	Plot size	541.36	22.68	23.8712	<0.00001	577.2	23.27	24.8103	<0.00001
$\log(\phi)$	(Intercept)	14.916823	0.208605	71.5075	<0.00001				
	(Credit) 1	0.444352	0.125076	3.55266	0.002623				
	(Crop) 2	-0.585502	0.11825	-4.9514	0.000289				
	(Training) 1	0.887435	0.134344	6.60569	0.00003				
	(Tour) 1	0.323429	0.121485	2.66229	0.011905				
	(Practical) 1	0.414736	0.115033	3.60536	0.002403				
	(Networking) 1	-0.470067	0.127219	-3.6949	0.002075				
	(Equipment) 1								
	Farmers								
	Plot size	0.069479	0.004379	15.8664	<0.00001				
$\log(\lambda)$	(Intercept)					17.96	0.08603	208.764	<0.00001
				Selection Criterion	Gaussian Joint-GLM	Gaussian GLM			
				-2ML(-2 h)	15894.87	16422.00			
				-2RL(-2$p_{beta}(h)$)	15772.61	16468.00			
				cAIC	15914.87	16442.00			

Table 5.2: Comparative Model Estimates for Gamma GLM and Gamma Joint-GLM's

		GAMMA GLM				GAMMA JOINT-GLM			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
log(μ)	(Intercept)	8.917414	0.07397	120.5543	<0.00001	8.91886	0.056282	158.467	<0.00001
	(Credit) 1								
	(Crop) 2	-0.193212	0.042012	-4.599	0.000491	-0.149376	0.036715	-4.0685	0.001129
	(Training) 1	-0.134143	0.047596	-2.8184	0.009112	-0.158761	0.035628	-4.4561	0.000612
	(Tour) 1					0.05197	0.037108	1.4005	0.09581
	(Practical) 1					-0.084918	0.033583	-2.5286	0.014987
	(Networking) 1								
	(Equipment) 1	-0.108639	0.043526	-2.496	0.015832	-0.129216	0.032611	-3.9624	0.001339
	Farmers	-0.006341	0.003401	-1.8648	0.045955				
	Plot size	0.032139	0.001558	20.627	<0.00001	0.031049	0.001343	23.1155	<0.00001
log(ϕ)	(Intercept)					-2.37071	0.235844	-10.052	<0.00001
	(Credit) 1					0.32474	0.141297	2.29828	0.022195
	(Crop) 2								
	(Training) 1					0.98346	0.15179	6.47908	0.000035
	(Tour) 1								
	(Practical) 1					0.509524	0.129886	3.92286	0.001427
	(Networking) 1					-0.401953	0.143677	-2.7976	0.009446
	(Equipment) 1					0.43765	0.13861	3.15742	0.005102
	Farmers								
	Plot size					0.009238	0.004967	1.85988	0.046273
log(λ)	(Intercept)	-1.008	0.06233	-16.1719	<0.00001				
				Selection Criterion	Gamma GLM	Gamma Joint-GLM			
				-2ML(-2 h)	16104.67	15984.99			
				-2RL(-2$p_{beta}(h)$)	16151.25	16045.60			
				cAIC	15914.87	16124.67			

Table 5.3: Comparative Model Estimates for Gaussian HGLM 1 and Gaussian HGLM 2

		GAUSSIAN HGLM 1				GAUSSIAN HGLM 2			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
log(μ)	(Intercept)	5869.6	1098.21	5.3447	0.000163	4753.8	697.2	6.8184	0.000023
	(Credit) 1								
	(Crop) 2	-3489.1	623.73	-5.594	0.000115	-2398.3	405.08	-5.9204	0.000074
	(Training) 1	-2598	706.64	-3.6765	0.002137	-1776.5	398.34	-4.4597	0.000609
	(Tour) 1					1164.8	416.4	2.7974	0.009439
	(Practical) 1					-726	367.79	-1.974	0.038317
	(Networking) 1								
	(Equipment) 1								
	Farmers	-236.6	50.49	-4.6867	0.00043	0.006634	0.00301	2.2039	0.019072
	Plot size	577.2	23.13	24.9531	<0.00001	0.028793	0.0015	19.20008	<0.00001
log(ϕ)	(Intercept)					15.32357	0.195893	78.224	<0.00001
	(Credit) 1					0.350278	0.117581	2.979	0.006916
	(Crop) 2					-0.46499	0.111278	-4.179	0.000945
	(Training) 1					0.458066	0.126042	3.6342	0.00229
	(Tour) 1					0.643937	0.11427	5.6352	0.000108
	(Practical) 1					0.303105	0.108174	2.802	0.009365
	(Networking) 1					-0.51501	0.119565	-4.307	0.000772
	(Equipment) 1					0.294979	0.115258	2.5593	0.014204
	Farmers								
	Plot size					0.056217	0.004127	13.622	<0.00001
log(λ)	(Intercept)	17.95	0.0855	209.9415	<0.00001				
	Province	-13.96	0.8563	-16.3027	<0.00001	-11.1	0.8563	-12.9627	<0.00001
	Community	-11.94	0.3922	-30.4436	<0.00001	-10.95	0.3922	-27.9194	<0.00001
				Selection Criterion	Gaussian HGLM-1	Gaussian HGLM-2			
				-2ML(-2 h)	16421.56	15982.81			
				-2RL(-2$p_{beta}(h)$)	16288.60	15858.20			
				cAIC	16441.60	16002.80			

Table 5.4: Comparative Model Estimates for Gamma HGLM 1 and Gamma HGLM 2

		GAMMA HGLM 2				GAMMA HGLM 1			
Model	covariates	Estimate	Std. Error	T-value	P-value	Estimate	Std. Error	T-value	P-value
log(μ)	(Intercept)	8.3421	0.461262	18.0854	<0.00001	8.425926	0.46619	18.07397	<0.00001
	(Credit) 1								
	(Crop) 2								
	(Training) 1								
	(Tour) 1								
	(Practical) 1	-0.05225	0.029128	-1.7939	0.043189				
	(Networking) 1								
	(Equipment) 1								
	Farmers	0.006923	0.002516	2.7514	0.005741	0.006634	0.00301	2.2039	0.019072
	Plot size	0.030736	0.001151	26.7071	<0.00001	0.028793	0.0015	19.20008	<0.00001
log(ϕ)	(Intercept)	-3.50947	0.45707	-7.6782	<0.00001				
	(Credit) 1	0.46636	0.20552	2.26917	0.016571				
	(Crop) 2								
	(Training) 1	1.16659	0.22775	5.12224	0.000018				
	(Tour) 1								
	(Practical) 1								
	(Networking) 1								
	(Equipment) 1	0.44948	0.20936	2.14692	0.021391				
	Farmers								
	Plot size								
log(λ)	(Intercept)					-1.624	0.09102	-17.8422	<0.00001
	Province	-3.627	0.8563	-4.2356	0.000162	-3.958	0.8563	-4.6222	0.000064
	Community	-1.641	0.3922	-4.1841	0.000184	-1.596	0.3922	-4.0694	0.000249
				Selection Criterion	Gamma HGLM-2	Gamma HGLM-1			
				-2ML(-2 h)	15509.20	15678.71			
				-2RL(-2$p_{beta}(h)$)	15564.50	15729.01			
				cAIC	15477.20	15649.61			

APPENDIX B

MODEL OUTPUTS FOR THE INTERACTION OF PHYSICAL CROP YIELD VARIABLES

MODEL 1: GAMMA/LOG GLM FOR FIXED AND INTERACTION TERMS

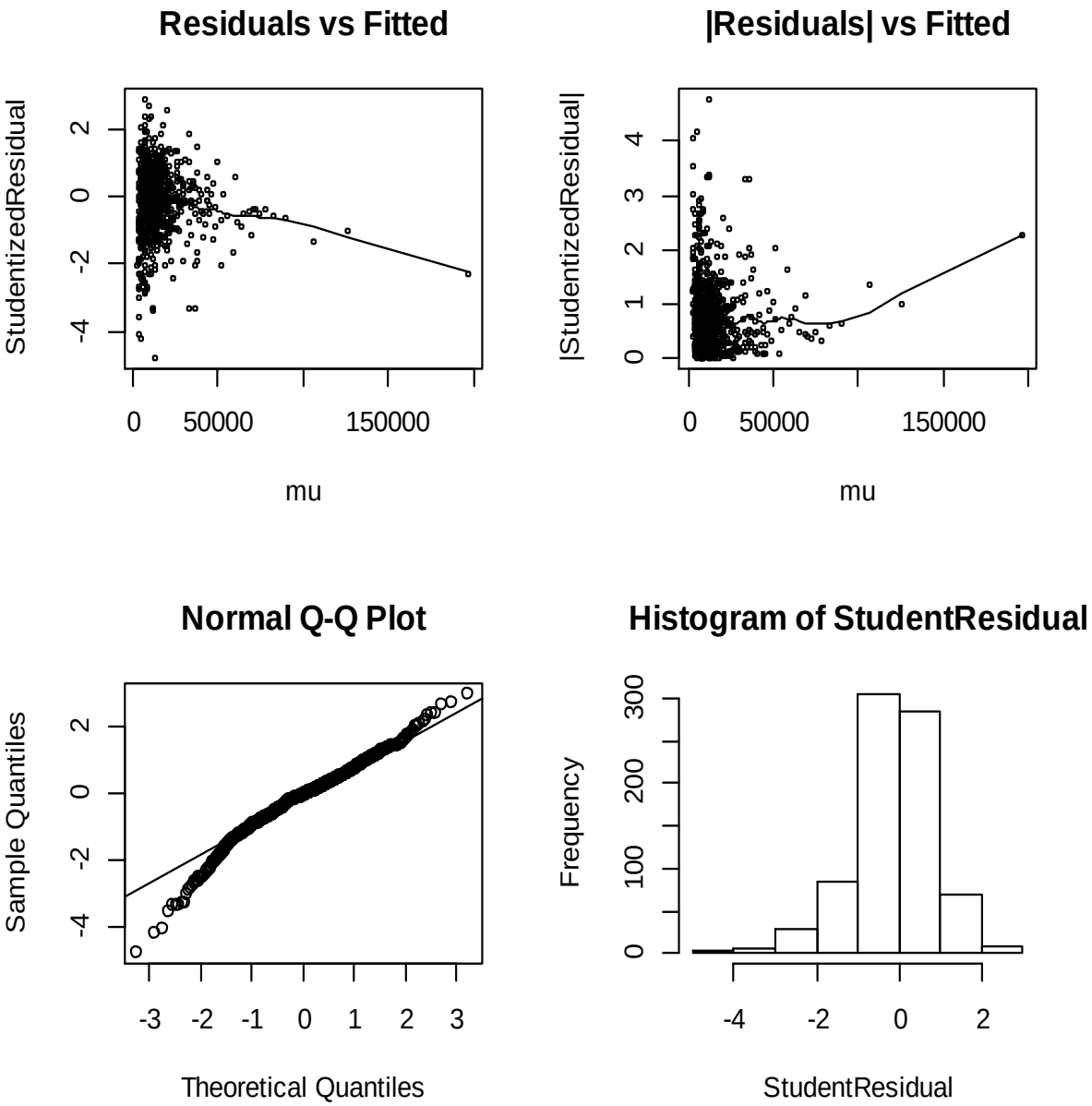

Fig. B1: Diagnostic plot of Gamma/log GLM for fixed and interaction variables

Table B1: Gamma/log GLM Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                          "Gamma"
[1] "Estimates from the model (mu)"
Yield ~ ...

[1] "Log"
                                                 Estimate Std. Error t-value
(Intercept)                                     8.3121825  1.876e-01 44.2989
as.factor(credit)1                              0.0762167  1.542e-01  0.4943
as.factor(crop)2                               -0.1535437  1.550e-01 -0.9909
as.factor(training)1                           -0.1427779  1.912e-01 -0.7467
as.factor(tour)1                                0.2576577  1.603e-01  1.6070
as.factor(practicals)1                          0.0256751  1.399e-01  0.1835
as.factor(networking)1                          0.2019217  1.674e-01  1.2065
as.factor(equipments)1                         -0.6577843  1.592e-01 -4.1317
farmers                                         0.0401614  1.142e-02  3.5168
plotsize                                        0.0446011  5.105e-03  8.7362
as.factor(credit)1:as.factor(crop)2            -0.2177506  9.934e-02 -2.1919
as.factor(credit)1:as.fa(training)1            -0.0394719  1.111e-01 -0.3552
as.factor(credit)1:as.factor(tour)1            -0.1575116  9.102e-02 -1.7305
as.factor(credit)1:as.f(practicals)1           -0.0293802  8.923e-02 -0.3293
as.factor(credit)1:as.fac(networking)1          0.1266971  1.113e-01  1.1379
as.factor(credit)1:as.fac(equipments)1         -0.0990281  1.033e-01 -0.9586
as.factor(credit)1:farmers                     -0.0082942  7.711e-03 -1.0756
as.factor(credit)1:plotsize                     0.0067039  3.076e-03  2.1794
as.factor(crop)2:as.factor(training)1          -0.2910814  1.368e-01 -2.1281
as.factor(crop)2:as.factor(tour)1               0.4197053  1.010e-01  4.1545
as.factor(crop)2:as.factor(practicals)1        -0.2230622  8.425e-02 -2.6477
as.factor(crop)2:as.factor(networking)1        -0.1465465  1.155e-01 -1.2692
as.factor(crop)2:as.factor(equipments)1         0.4806153  1.092e-01  4.4014
as.factor(crop)2:farmers                       -0.0093817  7.531e-03 -1.2458
as.factor(crop)2:plotsize                       0.0039852  3.521e-03  1.1319
as.factor(training)1:as.factor(tour)1           0.2653301  1.202e-01  2.2077
as.factor(training)1:as.factor(practicals)1     0.0380419  1.014e-01  0.3753
as.factor(training)1:as.factor(networking)1     0.1111546  1.189e-01  0.9348
as.factor(training)1:as.factor(equipments)1     0.1816572  1.148e-01  1.5825
as.factor(training)1:farmers                    0.0062735  8.686e-03  0.7223
as.factor(training)1:plotsize                  -0.0041749  3.744e-03 -1.1151
as.factor(tour)1:as.factor(practicals)1        -0.1565937  8.845e-02 -1.7704
as.factor(tour)1:as.factor(networking)1        -0.1372623  1.148e-01 -1.1959
as.factor(tour)1:as.factor(equipments)1        -0.3683106  9.104e-02 -4.0455
as.factor(tour)1:farmers                       -0.0083159  7.477e-03 -1.1122
as.factor(tour)1:plotsize                       0.0030447  3.483e-03  0.8740
as.factor(practicals)1:as.factor(networking)1 -0.0685746  8.985e-02 -0.7632
as.factor(practicals)1:as.factor(equipments)1  0.2109174  8.841e-02  2.3856
as.factor(practicals)1:farmers                 -0.0091309  7.275e-03 -1.2552
as.factor(practicals)1:plotsize                 0.0078753  3.499e-03  2.2507
as.factor(networking)1:as.factor(equipments)1 -0.0209805  1.143e-01 -0.1835
as.factor(networking)1:farmers                  0.0094410  8.082e-03  1.1681
as.factor(networking)1:plotsize                -0.0151176  3.921e-03 -3.8558
as.factor(equipments)1:farmers                 -0.0062294  8.096e-03 -0.7695
as.factor(equipments)1:plotsize                 0.0163462  3.665e-03  4.4598
farmers:plotsize                               -0.0008217  9.544e-05 -8.6096
```

```
[1] "Estimates from the model (phi)"

Phi ~ 1
<Environment: 0x078bfe88>

[1] "Log"
            Estimate Std. Error
(Intercept)   -1.296    0.07181

[1] "========== Likelihood Function Values and Condition AIC =========="
                              [, 1]
-2ML (-2 h)           :  15829.43
-2RL (-2 p_beta (h)) :  16029.60
cAIC                  :  15921.43

[1] "========== Degrees of freedom and Deviance =========="
                 [, 1]
DF:              46.0000
Deviance:        203.6405
```

MODEL 2: GAUSSIAN/IDENTITY GLM FOR FIXED AND INTERACTION TERMS

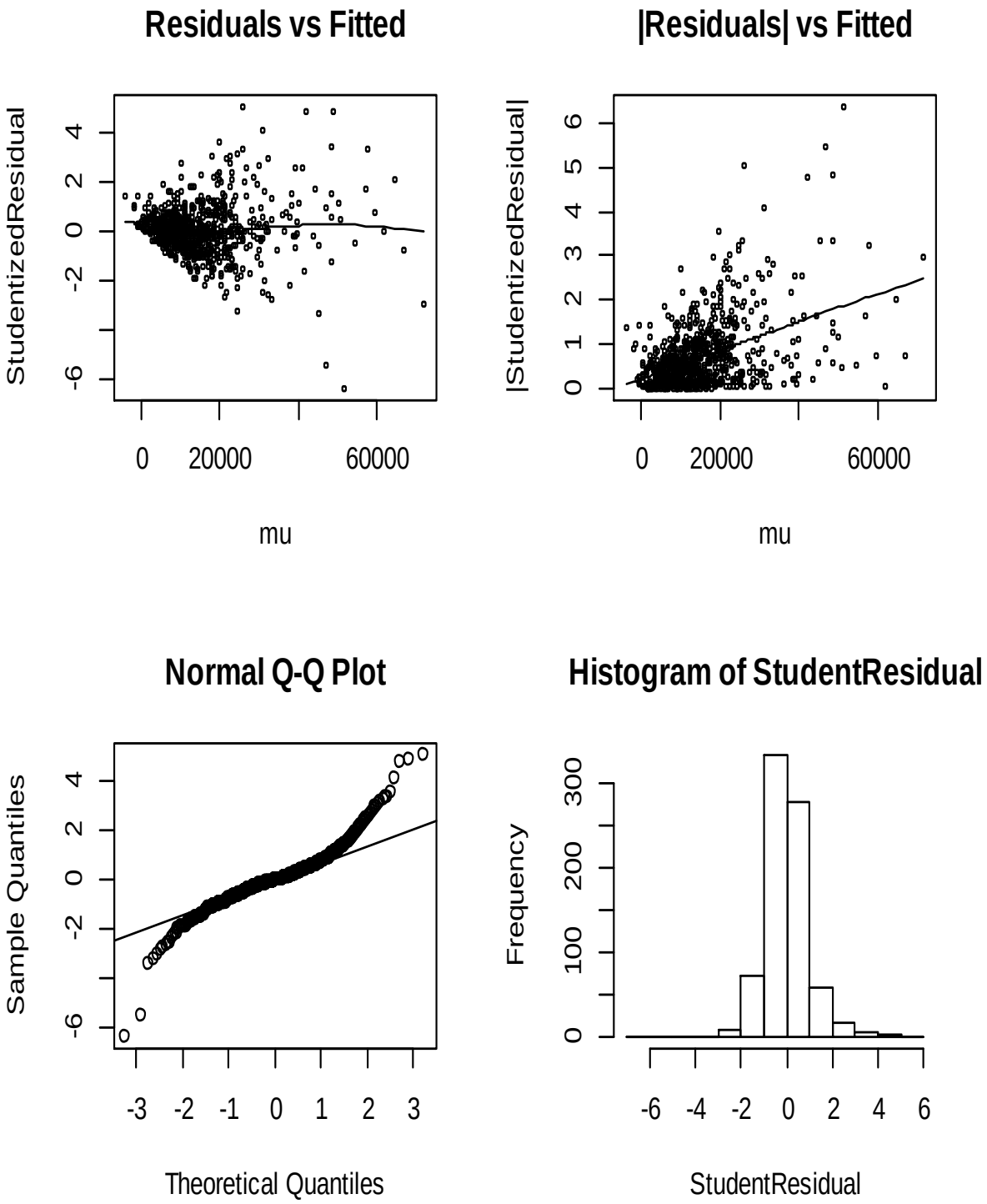

Fig. B2: Diagnostic plot of Gaussian/identity GLM for fixed and interaction variables

Table B2: Gaussian/identity GLM Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                    "Gaussian"
[1] "Estimates from the model (mu)"
Yield ~ ...

[1] "Identity"
                                                 Estimate Std. Error   t-value
(Intercept)                                      -464.486   2704.702 -0.171733
as.factor(credit)1                              -1973.942   2222.746 -0.888065
as.factor(crop)2                                -1814.873   2233.609 -0.812529
as.factor(training)1                             5723.936   2756.172  2.076771
as.factor(tour)1                                 2019.600   2311.106  0.873867
as.factor(practicals)1                           4978.451   2016.703  2.468609
as.factor(networking)1                           3436.551   2412.497  1.424479
as.factor(equipments)1                          -9591.122   2294.860 -4.179393
farmers                                            46.998    164.612  0.285506
plotsize                                          730.686     73.590  9.929129
as.factor(credit)1:as.factor(crop)2             -3810.396   1431.996 -2.660898
as.factor(credit)1:as.factor(training)1          -101.773   1601.859 -0.063534
as.factor(credit)1:as.factor(tour)1             -2228.289   1312.031 -1.698352
as.factor(credit)1:as.factor(practicals)1          -1.380   1286.238 -0.001073
as.factor(credit)1:as.factor(networking)1        1931.501   1604.926  1.203483
as.factor(credit)1:as.factor(equipments)1         597.436   1489.017  0.401229
as.factor(credit)1:farmers                         29.382    111.152  0.264344
as.factor(credit)1:plotsize                        96.774     44.338  2.182611
as.factor(crop)2:as.factor(training)1           -2156.874   1971.635 -1.093952
as.factor(crop)2:as.factor(tour)1                3579.936   1456.196  2.458416
as.factor(crop)2:as.factor(practicals)1         -2585.391   1214.395 -2.128953
as.factor(crop)2:as.factor(networking)1         -2524.078   1664.286 -1.516613
as.factor(crop)2:as.factor(equipments)1          7047.614   1573.997  4.477527
as.factor(crop)2:farmers                           43.693    108.549  0.402514
as.factor(crop)2:plotsize                         -69.281     50.749 -1.365177
as.factor(training)1:as.factor(tour)1            1769.046   1732.353  1.021181
as.factor(training)1:as.factor(practicals)1      -999.052   1460.912 -0.683855
as.factor(training)1:as.factor(networking)1      -831.800   1714.065 -0.485279
as.factor(training)1:as.factor(equipments)1      1281.474   1654.687  0.774451
as.factor(training)1:farmers                     -206.354    125.199 -1.648206
as.factor(training)1:plotsize                     -98.115     53.965 -1.818127
as.factor(tour)1:as.factor(practicals)1         -1377.250   1274.979 -1.080213
as.factor(tour)1:as.factor(networking)1         -1347.443   1654.502 -0.814410
as.factor(tour)1:as.factor(equipments)1         -2721.810   1312.324 -2.074038
as.factor(tour)1:farmers                          101.801    107.775  0.944576
as.factor(tour)1:plotsize                         -26.667     50.213 -0.531086
as.factor(practicals)1:as.factor(networking)1 -1071.089   1295.150 -0.827000
as.factor(practicals)1:as.factor(equipments)1  2879.793   1274.442  2.259651
as.factor(practicals)1:farmers                    -37.806    104.858 -0.360539
as.factor(practicals)1:plotsize                  -102.482     50.436 -2.031914
as.factor(networking)1:as.factor(equipments)1  -333.591   1648.128 -0.202406
as.factor(networking)1:farmers                    76.142    116.499  0.653580
as.factor(networking)1:plotsize                 -127.856     56.516 -2.262307
as.factor(equipments)1:farmers                  -121.517    116.695 -1.041318
as.factor(equipments)1:plotsize                  215.595     52.832  4.080769
farmers:plotsize                                  -3.061      1.376 -2.225056
```

```
[1] "Estimates from the model (phi)"

Phi ~ 1
<Environment: 0x08ed8c04>

[1] "Log"
            Estimate Std. Error
(Intercept)    17.67    0.09766


[1] "========== Likelihood Function Values and Condition AIC =========="
                              [, 1]
-2ML (-2 h)           :  16154.00
-2RL (-2 p_beta (h)) :  16354.17
cAIC                  :  16246.00

[1] "========== Degrees of freedom and Deviance =========="
                   [, 1]
DF :                 46
Deviance : 35065815280
```

MODEL 3: GAMMA/LOG JOINT-GLM FOR FIXED AND INTERACTION TERMS

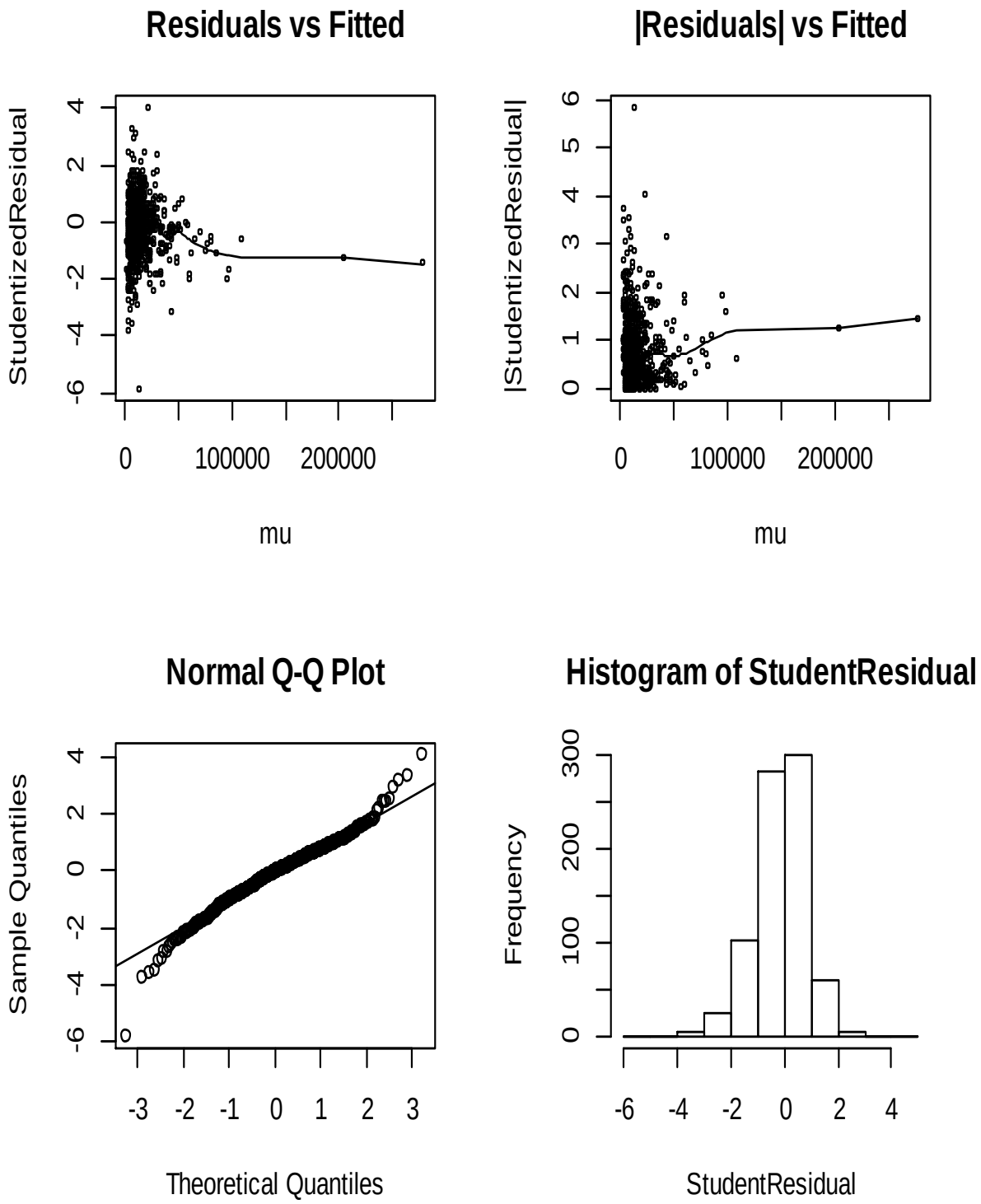

Fig. B3: Diagnostic plot of Gamma/log Joint-GLM for fixed and interaction variables

Table B3: Gamma/log Joint-GLM Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                          "Gamma"
[1] "Estimates from the model (mu)"
Yield ~ ...

[1] "Log"
                                                 Estimate Std. Error   t-value
(Intercept)                                     8.3652974  1.183e-01  70.72794
as.factor(credit)1                             -0.0403679  1.062e-01  -0.37999
as.factor(crop)2                               -0.2846343  1.015e-01  -2.80332
as.factor(training)1                           -0.1676278  1.401e-01  -1.19672
as.factor(tour)1                                0.2374078  1.177e-01   2.01788
as.factor(practicals)1                          0.0012257  9.760e-02   0.01256
as.factor(networking)1                         -0.0212389  1.089e-01  -0.19504
as.factor(equipments)1                         -0.5073023  1.095e-01  -4.63309
farmers                                         0.0412105  7.547e-03   5.46042
plotsize                                        0.0469243  3.794e-03  12.36684
as.factor(credit)1:as.factor(crop)2            -0.1562764  7.362e-02  -2.12284
as.factor(credit)1:as.factor(training)1        -0.0019077  7.654e-02  -0.02492
as.factor(credit)1:as.factor(tour)1            -0.2031009  6.892e-02  -2.94703
as.factor(credit)1:as.factor(practicals)1      -0.0757990  6.866e-02  -1.10393
as.factor(credit)1:as.factor(networking)1       0.1295895  8.060e-02   1.60778
as.factor(credit)1:as.factor(equipments)1       0.0286055  7.248e-02   0.39468
as.factor(credit)1:farmers                     -0.0006676  5.505e-03  -0.12128
as.factor(credit)1:plotsize                     0.0035738  2.480e-03   1.44094
as.factor(crop)2:as.factor(training)1          -0.3666257  8.519e-02  -4.30384
as.factor(crop)2:as.factor(tour)1               0.4684116  7.914e-02   5.91894
as.factor(crop)2:as.factor(practicals)1        -0.1677092  6.613e-02  -2.53597
as.factor(crop)2:as.factor(networking)1         0.0317786  8.734e-02   0.36384
as.factor(crop)2:as.factor(equipments)1         0.4042558  7.504e-02   5.38689
as.factor(crop)2:farmers                       -0.0080272  5.604e-03  -1.43247
as.factor(crop)2:plotsize                       0.0028877  2.588e-03   1.11563
as.factor(training)1:as.factor(tour)1           0.2456222  8.057e-02   3.04841
as.factor(training)1:as.factor(practicals)1    -0.0297179  7.146e-02  -0.41585
as.factor(training)1:as.factor(networking)1     0.1164308  1.019e-01   1.14302
as.factor(training)1:as.factor(equipments)1     0.2693948  7.583e-02   3.55278
as.factor(training)1:farmers                    0.0122279  5.773e-03   2.11805
as.factor(training)1:plotsize                  -0.0084444  2.644e-03  -3.19324
as.factor(tour)1:as.factor(practicals)1        -0.0887473  7.088e-02  -1.25205
as.factor(tour)1:as.factor(networking)1        -0.0770664  9.877e-02  -0.78030
as.factor(tour)1:as.factor(equipments)1        -0.4271944  6.796e-02  -6.28568
as.factor(tour)1:farmers                       -0.0223318  5.894e-03  -3.78903
as.factor(tour)1:plotsize                       0.0111725  2.871e-03   3.89180
as.factor(practicals)1:as.factor(networking)1 -0.0683870  6.954e-02  -0.98335
as.factor(practicals)1:as.factor(equipments)1  0.2017236  6.077e-02   3.31949
as.factor(practicals)1:farmers                 -0.0037174  5.162e-03  -0.72014
as.factor(practicals)1:plotsize                 0.0051250  2.574e-03   1.99126
as.factor(networking)1:as.factor(equipments)1 -0.0038870  8.399e-02  -0.04628
as.factor(networking)1:farmers                  0.0071381  5.901e-03   1.20962
as.factor(networking)1:plotsize                -0.0080811  3.165e-03  -2.55320
as.factor(equipments)1:farmers                 -0.0108459  5.172e-03  -2.09695
as.factor(equipments)1:plotsize                 0.0118724  2.493e-03   4.76209
farmers:plotsize                               -0.0008875  8.162e-05 -10.87372
```

[1] "Estimates from the model (phi)"

```
Phi ~ as.factor(credit) + as.factor(crop) + as.factor(training) +
    as.factor(tour) + as.factor(practicals) + as.factor(networking) +
    as.factor(equipments) + farmers + plotsize

[1] "Log"
                       Estimate Std. Error
(Intercept)            -3.05148   0.297486
as.factor(credit)1      0.58928   0.177384
as.factor(crop)2       -0.19228   0.166473
as.factor(training)1    1.37900   0.192221
as.factor(tour)1        0.44639   0.171572
as.factor(practicals)1  0.49830   0.161891
as.factor(networking)1 -0.87691   0.179945
as.factor(equipments)1  0.31264   0.173748
farmers                -0.01345   0.013715
plotsize                0.01246   0.006289

[1] "========== Likelihood Function Values and Condition AIC =========="
                              [, 1]
-2ML (-2 h)           :  15611.09
-2RL (-2 p_beta (h)) :  15901.23
cAIC                  :  15703.09

[1] "========== Degrees of freedom and Deviance =========="
               [, 1]
DF :        46.0000
Deviance : 210.2769
```

MODEL 4: GAUSSIAN/IDENTITY JOINT-GLM FOR FIXED AND INTERACTION TERMS

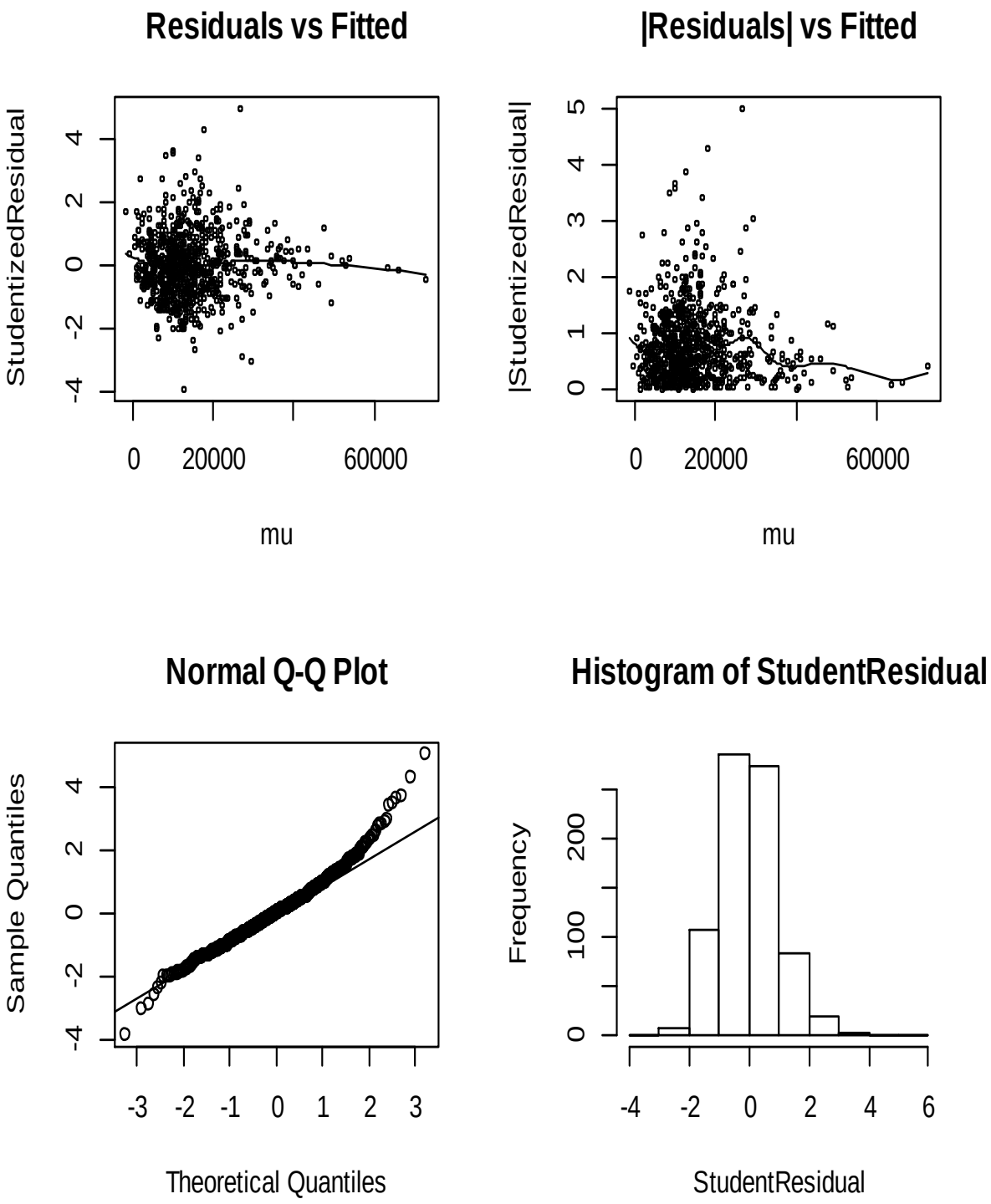

Fig. B4: Diagnostic plot of Gaussian/identity Joint-GLM for fixed and interaction variables

Table B4: Gaussian/identity Joint-GLM Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                    "Gaussian"
[1] "Estimates from the model (mu)"
Yield ~ ...

[1] "Identity"
                                                Estimate Std. Error  t-value
(Intercept)                                       88.320    1309.60  0.06744
as.factor(credit)1                             -3059.074    1221.60 -2.50416
as.factor(crop)2                                -610.509    1089.41 -0.56040
as.factor(training)1                            4028.669    1314.47  3.06485
as.factor(tour)1                                3380.635    1314.21  2.57237
as.factor(practicals)1                          -162.951     978.22 -0.16658
as.factor(networking)1                         -1392.572    1110.37 -1.25415
as.factor(equipments)1                         -5761.840    1106.06 -5.20935
farmers                                          107.367      94.63  1.13457
plotsize                                         788.652      62.16 12.68781
as.factor(credit)1:as.factor(crop)2            -1957.325     678.57 -2.88448
as.factor(credit)1:as.factor(training)1        -1337.102     775.74 -1.72364
as.factor(credit)1:as.factor(tour)1            -1223.396     665.96 -1.83703
as.factor(credit)1:as.factor(practicals)1        375.036     643.22  0.58306
as.factor(credit)1:as.factor(networking)1       2109.845     836.23  2.52303
as.factor(credit)1:as.factor(equipments)1       -205.507     702.97 -0.29234
as.factor(credit)1:farmers                        71.952      63.60  1.13124
as.factor(credit)1:plotsize                      122.302      40.33  3.03240
as.factor(crop)2:as.factor(training)1          -2897.284     840.67 -3.44638
as.factor(crop)2:as.factor(tour)1               2569.429     706.46  3.63705
as.factor(crop)2:as.factor(practicals)1        -2002.489     634.38 -3.15660
as.factor(crop)2:as.factor(networking)1         1382.154     833.69  1.65787
as.factor(crop)2:as.factor(equipments)1         5436.169     742.78  7.31868
as.factor(crop)2:farmers                         -68.656      63.86 -1.07502
as.factor(crop)2:plotsize                       -139.635      40.15 -3.47760
as.factor(training)1:as.factor(tour)1            648.014     818.63  0.79158
as.factor(training)1:as.factor(practicals)1     -265.074     632.50 -0.41909
as.factor(training)1:as.factor(networking)1      773.718     984.23  0.78612
as.factor(training)1:as.factor(equipments)1     3127.124     680.91  4.59254
as.factor(training)1:farmers                       2.468      64.65  0.03817
as.factor(training)1:plotsize                   -246.898      39.85 -6.19644
as.factor(tour)1:as.factor(practicals)1         -501.699     640.39 -0.78343
as.factor(tour)1:as.factor(networking)1          726.961    1025.52  0.70887
as.factor(tour)1:as.factor(equipments)1        -3935.135     630.59 -6.24038
as.factor(tour)1:farmers                        -176.020      66.99 -2.62746
as.factor(tour)1:plotsize                         46.099      41.33  1.11526
as.factor(practicals)1:as.factor(networking)1   -794.011     645.17 -1.23070
as.factor(practicals)1:as.factor(equipments)1   2363.379     542.38  4.35739
as.factor(practicals)1:farmers                    50.879      55.10  0.92332
as.factor(practicals)1:plotsize                   -8.583      38.12 -0.22517
as.factor(networking)1:as.factor(equipments)1   -161.822     840.24 -0.19259
as.factor(networking)1:farmers                  -14.359      73.80 -0.19457
as.factor(networking)1:plotsize                  -4.756      53.35 -0.08915
as.factor(equipments)1:farmers                    7.918      60.40  0.13110
as.factor(equipments)1:plotsize                   7.684      38.90  0.19752
farmers:plotsize                                 -4.191       1.58 -2.65252
```

[1] "Estimates from the model (phi)"

```
Phi ~ as.factor(credit) + as.factor(crop) + as.factor(training) +
    as.factor(tour) + as.factor(practicals) + as.factor(networking) +
    as.factor(equipments) + farmers + plotsize

[1] "Log"
                       Estimate Std. Error
(Intercept)            14.14330   0.272956
as.factor(credit)1      0.83998   0.162991
as.factor(crop)2       -0.59405   0.153182
as.factor(training)1    0.89055   0.175940
as.factor(tour)1        0.72271   0.157700
as.factor(practicals)1  0.33184   0.149172
as.factor(networking)1 -1.05164   0.165567
as.factor(equipments)1  0.20134   0.159770
farmers                -0.01298   0.012421
plotsize                0.07781   0.005654

[1] "========== Likelihood Function Values and Condition AIC =========="
                             [, 1]
-2ML (-2 h)           :  15496.28
-2RL (-2 p_beta (h)) :  14938.90
cAIC                  :  15588.28

[1] "========== Degrees of freedom and Deviance =========="
                  [, 1]
DF :                46
Deviance : 40447333610
```

MODEL 5: **HGLM 1- FIXED = GAMMA/LOG, RANDOM = INVERSE GAMMA / GAUSSIAN (FOR FIXED AND INTERACTION TERMS)**

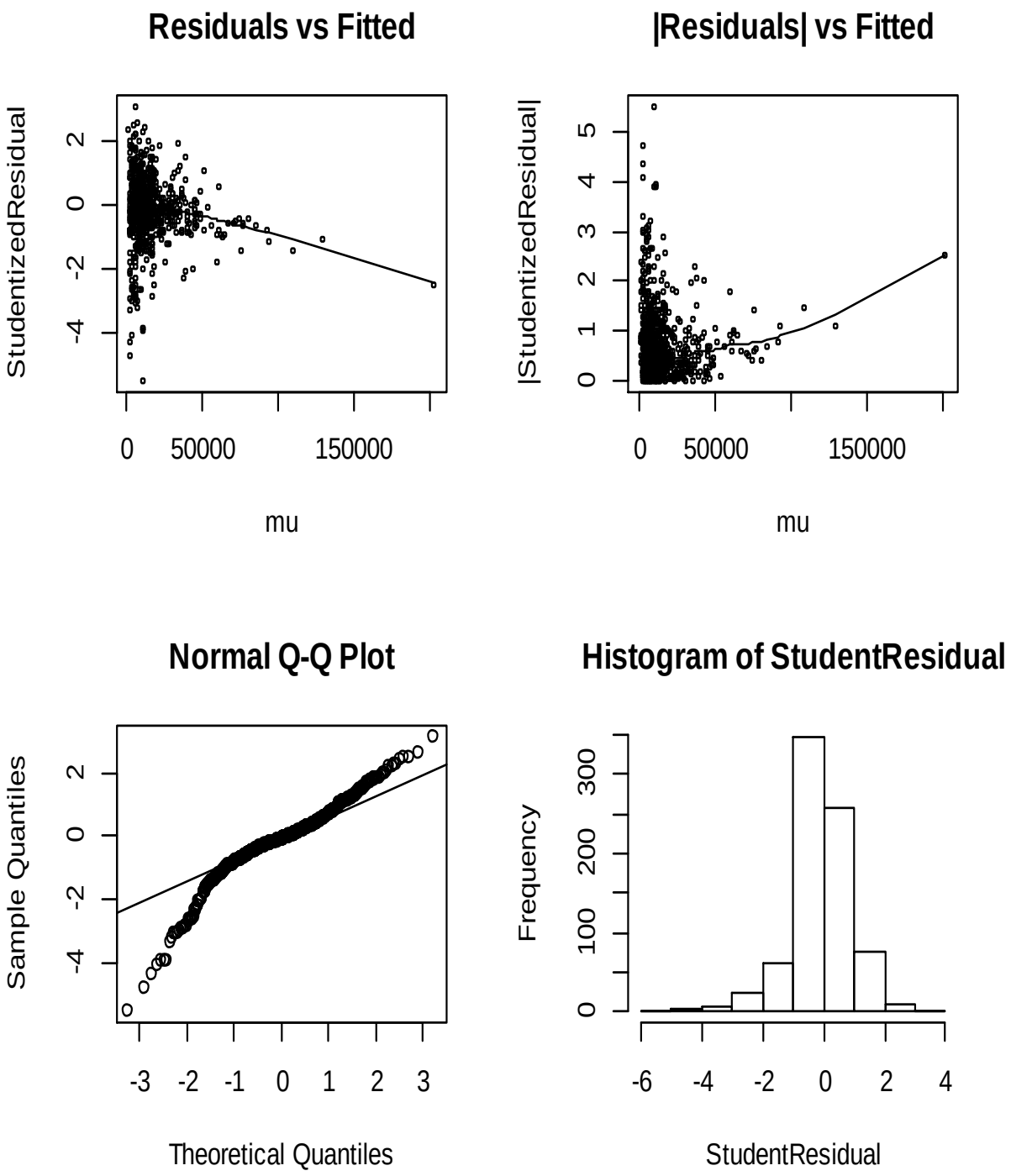

Fig. B5: Diagnostic plot of Gamma/log HGLM - 1 for fixed and interaction variables

Table B5: Gamma/log HGLM - 1 Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                          "Gamma"
[1] "Estimates from the model (mu)"
Yield ~ ...

[1] "Log"
                                                 Estimate Std. Error  t-value
(Intercept)                                     8.0818420  4.988e-01 16.20204
as.factor(credit)1                             -0.0195318  1.445e-01 -0.13520
as.factor(crop)2                               -0.2402774  2.200e-01 -1.09195
as.factor(training)1                           -0.0676464  1.763e-01 -0.38363
as.factor(tour)1                                0.0676873  1.526e-01  0.44345
as.factor(practicals)1                         -0.0728067  1.305e-01 -0.55782
as.factor(networking)1                          0.0512852  1.570e-01  0.32658
as.factor(equipments)1                         -0.0428338  1.636e-01 -0.26182
farmers                                         0.0316660  1.089e-02  2.90721
plotsize                                        0.0428802  4.928e-03  8.70053
as.factor(credit)1:as.factor(crop)2            -0.0667589  1.002e-01 -0.66598
as.factor(credit)1:as.factor(training)1        -0.0035933  1.071e-01 -0.03354
as.factor(credit)1:as.factor(tour)1            -0.0814992  8.730e-02 -0.93356
as.factor(credit)1:as.factor(practicals)1      -0.0575887  8.510e-02 -0.67669
as.factor(credit)1:as.factor(networking)1       0.0879100  1.081e-01  0.81311
as.factor(credit)1:as.factor(equipments)1       0.1185892  1.002e-01  1.18390
as.factor(credit)1:farmers                     -0.0056792  7.232e-03 -0.78530
as.factor(credit)1:plotsize                     0.0040425  2.919e-03  1.38498
as.factor(crop)2:as.factor(training)1          -0.0120610  1.512e-01 -0.07979
as.factor(crop)2:as.factor(tour)1               0.0575049  1.044e-01  0.55057
as.factor(crop)2:as.factor(practicals)1        -0.0266839  8.549e-02 -0.31213
as.factor(crop)2:as.factor(networking)1         0.0142990  1.276e-01  0.11209
as.factor(crop)2:as.factor(equipments)1         0.1090789  1.254e-01  0.86998
as.factor(crop)2:farmers                        0.0016016  7.208e-03  0.22220
as.factor(crop)2:plotsize                       0.0035752  3.444e-03  1.03823
as.factor(training)1:as.factor(tour)1           0.0858928  1.119e-01  0.76787
as.factor(training)1:as.factor(practicals)1     0.0532793  1.013e-01  0.52579
as.factor(training)1:as.factor(networking)1    -0.0146498  1.143e-01 -0.12817
as.factor(training)1:as.factor(equipments)1     0.0050763  1.096e-01  0.04633
as.factor(training)1:farmers                    0.0083454  8.026e-03  1.03986
as.factor(training)1:plotsize                  -0.0029158  3.450e-03 -0.84524
as.factor(tour)1:as.factor(practicals)1        -0.0497602  8.313e-02 -0.59860
as.factor(tour)1:as.factor(networking)1        -0.0694726  1.077e-01 -0.64505
as.factor(tour)1:as.factor(equipments)1        -0.1268631  8.889e-02 -1.42722
as.factor(tour)1:farmers                        0.0033291  7.182e-03  0.46351
as.factor(tour)1:plotsize                      -0.0004609  3.288e-03 -0.14017
as.factor(practicals)1:as.factor(networking)1  0.0782523  8.928e-02  0.87648
as.factor(practicals)1:as.factor(equipments)1 -0.0172940  8.657e-02 -0.19978
as.factor(practicals)1:farmers                  0.0015523  6.946e-03  0.22347
as.factor(practicals)1:plotsize                 0.0006440  3.388e-03  0.19010
as.factor(networking)1:as.factor(equipments)1 -0.1476357  1.109e-01 -1.33075
as.factor(networking)1:farmers                  0.0047054  7.500e-03  0.62739
as.factor(networking)1:plotsize                -0.0061620  3.703e-03 -1.66421
as.factor(equipments)1:farmers                 -0.0095119  7.586e-03 -1.25386
as.factor(equipments)1:plotsize                 0.0091950  3.450e-03  2.66526
farmers:plotsize                               -0.0007433  9.169e-05 -8.10724
```

[1] "Estimates for logarithm of lambda=var(u_mu)"

```
[1] "Gaussian"      "Inverse-gamma"

          Estimate Std. Error
Region      -5.372     0.8563
Community   -2.292     0.3922
```

[1] "Estimates from the model (phi)"

```
Phi ~ 1
<Environment: 0x127d48fc>

[1] "Log"
            Estimate Std. Error
(Intercept)   -1.669    0.08536

[1] "========== Likelihood Function Values and Condition AIC =========="
                                        [, 1]
-2ML (-2 p_v(mu) (h))           :  15639.85
-2RL (-2 p_beta(mu),v(mu) (h)) :  15897.72
cAIC                            :  15684.67

[1] "========== Degrees of freedom and Deviance =========="
                [, 1]
DF :        58.18551
Deviance : 148.66743
```

[1] "========== Random effect =========="

```
          [,1]
1  -0.12193473
2  -0.07712488
3  -0.04361149
1  -0.56131518
2   0.68579641
3  -0.24641596
4   0.35889849
5  -0.34983726
6   0.23289621
7   0.13190898
8  -0.12178861
9  -0.32456099
10 -0.13173636
11  0.04228939
12 -0.03104321
13  0.07223699
```

MODEL 6: **HGLM 1- FIXED = GAUSSIAN/IDENTITY, RANDOM = GAUSSIAN/GAMMA (FOR FIXED AND INTERACTION TERMS)**

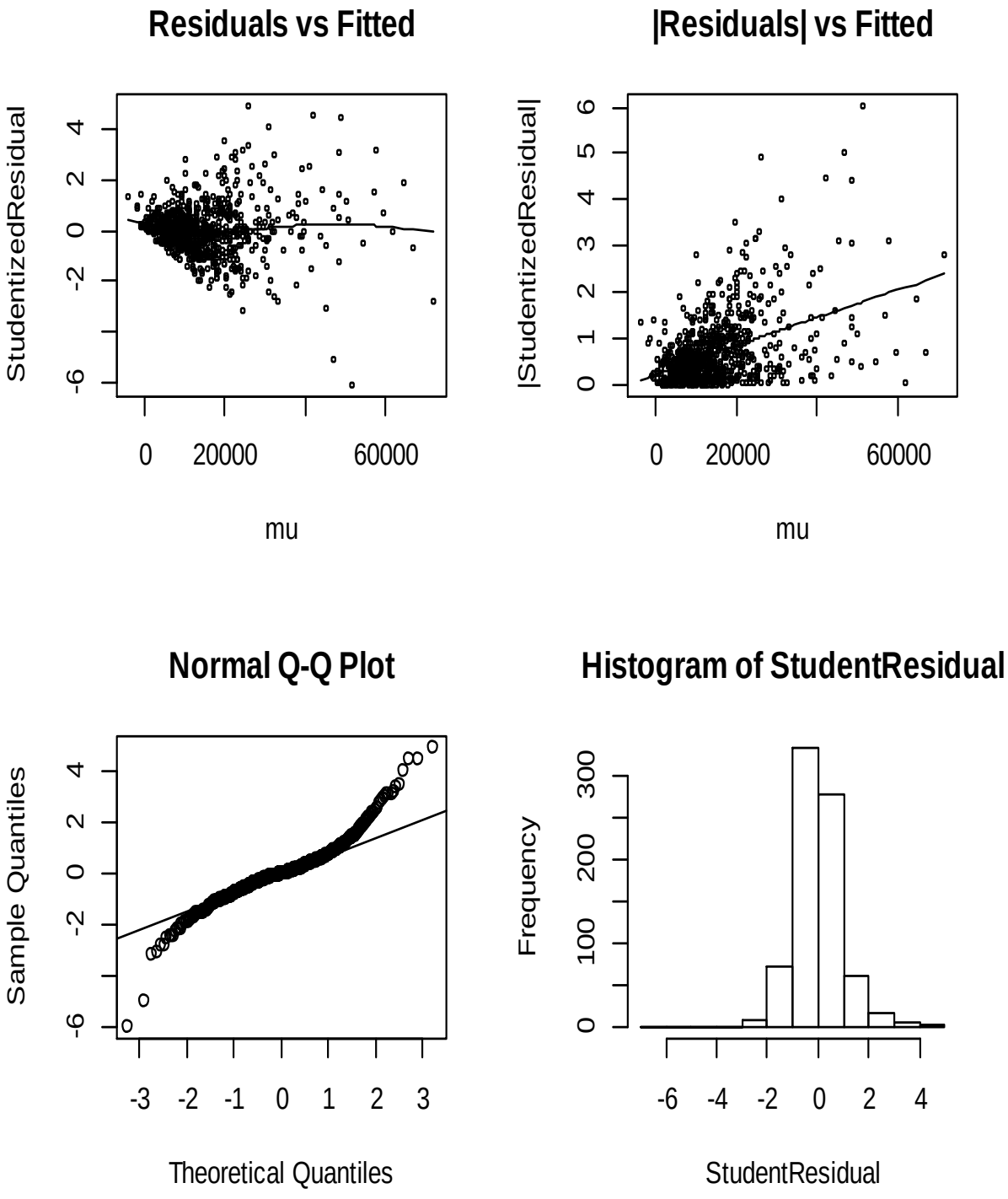

Fig. B6: Diagnostic plot of Gaussian/Identity HGLM - 1 for fixed and interaction variables

Table B6: Gaussian/Identity HGLM - 1 Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                      "Gaussian"
[1] "Estimates from the model (mu)"
Yield ~ ...

[1] "Identity"
                                                Estimate Std. Error   t-value
(Intercept)                                     -464.484   2626.439 -0.176849
as.factor(credit)1                             -1973.943   2158.429 -0.914528
as.factor(crop)2                               -1814.873   2168.978 -0.836741
as.factor(training)1                            5723.936   2676.419  2.138654
as.factor(tour)1                                2019.598   2244.232  0.899906
as.factor(practicals)1                          4978.451   1958.347  2.542169
as.factor(networking)1                          3436.549   2342.689  1.466925
as.factor(equipments)1                         -9591.120   2228.456 -4.303930
farmers                                           46.998    159.849  0.294013
plotsize                                         730.686     71.461 10.224999
as.factor(credit)1:as.factor(crop)2            -3810.395   1390.560 -2.740187
as.factor(credit)1:as.factor(training)1         -101.774   1555.508 -0.065428
as.factor(credit)1:as.factor(tour)1            -2228.289   1274.066 -1.748959
as.factor(credit)1:as.factor(practicals)1         -1.380   1249.019 -0.001105
as.factor(credit)1:as.factor(networking)1       1931.502   1558.486  1.239345
as.factor(credit)1:as.factor(equipments)1       597.437   1445.931  0.413185
as.factor(credit)1:farmers                       29.382    107.935  0.272221
as.factor(credit)1:plotsize                      96.774     43.055  2.247649
as.factor(crop)2:as.factor(training)1          -2156.874   1914.584 -1.126549
as.factor(crop)2:as.factor(tour)1               3579.934   1414.060  2.531671
as.factor(crop)2:as.factor(practicals)1        -2585.389   1179.256 -2.192390
as.factor(crop)2:as.factor(networking)1        -2524.077   1616.129 -1.561805
as.factor(crop)2:as.factor(equipments)1         7047.614   1528.452  4.610949
as.factor(crop)2:farmers                          43.693    105.408  0.414509
as.factor(crop)2:plotsize                        -69.281     49.281 -1.405857
as.factor(training)1:as.factor(tour)1           1769.045   1682.226  1.051610
as.factor(training)1:as.factor(practicals)1     -999.053   1418.639 -0.704233
as.factor(training)1:as.factor(networking)1     -831.801   1664.467 -0.499740
as.factor(training)1:as.factor(equipments)1     1281.473   1606.807  0.797528
as.factor(training)1:farmers                    -206.354    121.577 -1.697319
as.factor(training)1:plotsize                    -98.115     52.403 -1.872304
as.factor(tour)1:as.factor(practicals)1        -1377.249   1238.087 -1.112401
as.factor(tour)1:as.factor(networking)1        -1347.443   1606.627 -0.838678
as.factor(tour)1:as.factor(equipments)1        -2721.810   1274.351 -2.135840
as.factor(tour)1:farmers                         101.802    104.656  0.972724
as.factor(tour)1:plotsize                        -26.667     48.760 -0.546912
as.factor(practicals)1:as.factor(networking)1 -1071.089   1257.674 -0.851642
as.factor(practicals)1:as.factor(equipments)1  2879.792   1237.564  2.326983
as.factor(practicals)1:farmers                   -37.806    101.824 -0.371282
as.factor(practicals)1:plotsize                 -102.482     48.977 -2.092462
as.factor(networking)1:as.factor(equipments)1  -333.591   1600.438 -0.208438
as.factor(networking)1:farmers                   76.142    113.128  0.673055
as.factor(networking)1:plotsize                 -127.856     54.881 -2.329718
as.factor(equipments)1:farmers                  -121.517    113.319 -1.072348
as.factor(equipments)1:plotsize                  215.595     51.303  4.202368
farmers:plotsize                                  -3.061      1.336 -2.291359
```

[1] "Estimates for logarithm of lambda=var(u_mu)"

```
[1] "Gaussian" "gamma"

          Estimate Std. Error
Region      -14.29     0.8563
Community   -12.22     0.3922
```

[1] "Estimates from the model (phi)"

```
Phi ~ 1
<Environment: 0x11b9e730>

[1] "Log"
            Estimate Std. Error
(Intercept)    17.61    0.09319

[1] "========== Likelihood Function Values and Condition AIC =========="
                                        [, 1]
-2ML (-2 p_v(mu) (h))           :  16152.61
-2RL (-2 p_beta(mu),v(mu) (h)) :  15542.73
cAIC                            :  16244.61

[1] "========== Degrees of freedom and Deviance =========="
                  [, 1]
DF :                46
Deviance : 35065802655
```

[1] "========== Random effect =========="

```
             [, 1]
1   1.557672e-03
2   4.853383e-04
3  -2.043001e-03
1  -1.850068e-03
2   5.973378e-03
3  -2.565638e-03
4   1.972833e-03
5  -2.780214e-03
6   1.288779e-03
7   6.105267e-04
8  -1.723845e-03
9  -9.159768e-06
10  1.767677e-04
11 -1.249963e-04
12 -1.307343e-03
13  3.389884e-04
```

MODEL 7: HGLM 2- FIXED = GAMMA/LOG, RANDOM = GAUSSIAN/INVERSE GAMMA PHI = GAMMA/LOG (FOR FIXED AND INTERACTION TERMS)

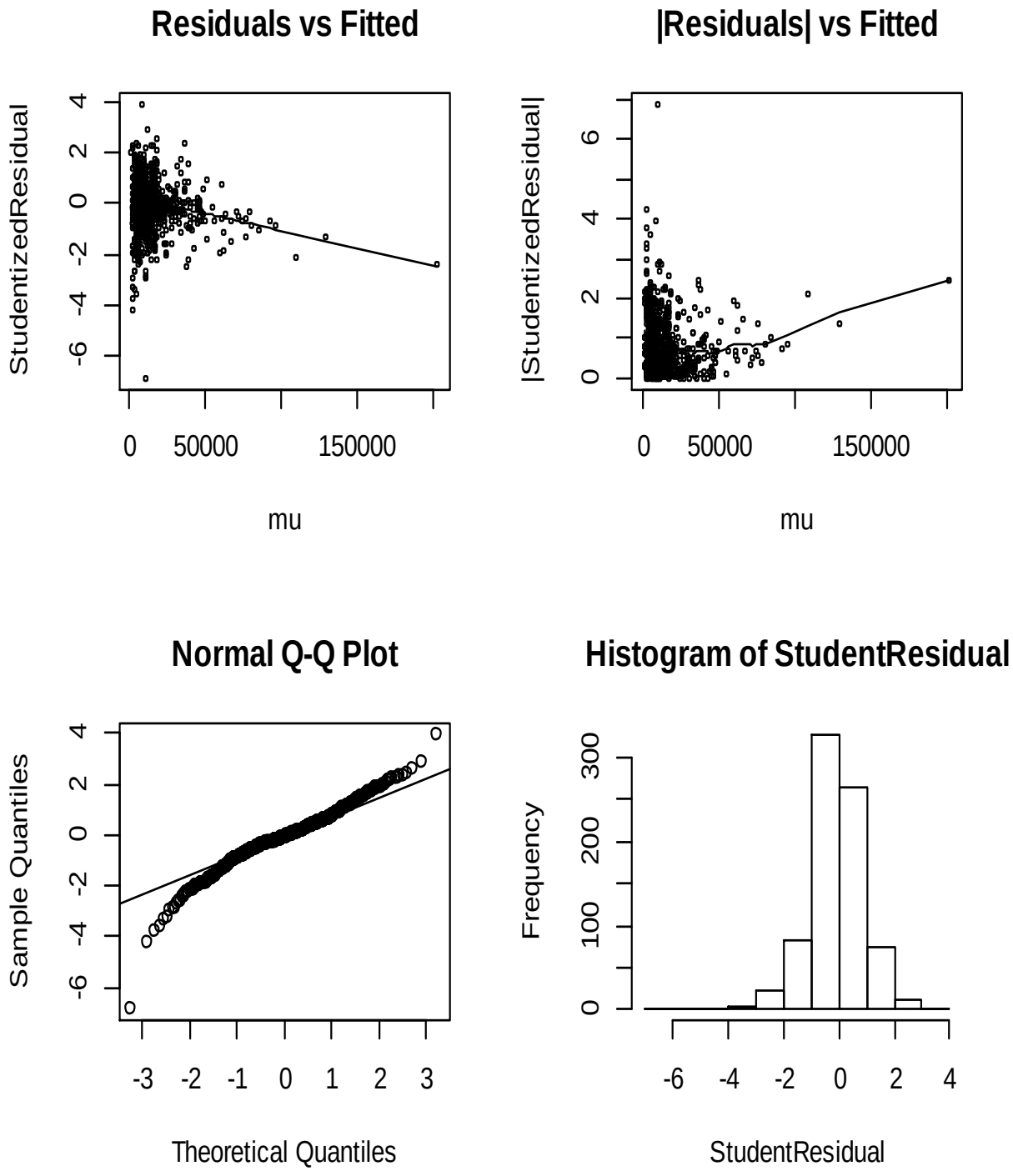

Fig. B7: Diagnostic plot of Gamma/log HGLM - 2 for fixed and interaction variables

Table B7: Gamma/log HGLM - 2 Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                        "Gamma"
[1] "Estimates from the model (mu)"
Yield ~ ...

[1] "Log"
                                                 Estimate Std. Error  t-value
(Intercept)                                     7.9451883  4.815e-01 16.50064
as.factor(credit)1                             -0.0464855  1.133e-01 -0.41037
as.factor(crop)2                               -0.0907664  1.711e-01 -0.53037
as.factor(training)1                           -0.0441746  1.461e-01 -0.30244
as.factor(tour)1                                0.0845560  1.209e-01  0.69911
as.factor(practicals)1                         -0.1076213  1.028e-01 -1.04697
as.factor(networking)1                          0.0418318  1.239e-01  0.33760
as.factor(equipments)1                          0.0470808  1.251e-01  0.37624
farmers                                         0.0364664  8.511e-03  4.28472
plotsize                                        0.0430731  3.940e-03 10.93345
as.factor(credit)1:as.factor(crop)2            -0.0439735  7.905e-02 -0.55627
as.factor(credit)1:as.factor(training)1         0.0029998  7.971e-02  0.03763
as.factor(credit)1:as.factor(tour)1            -0.0987920  7.077e-02 -1.39600
as.factor(credit)1:as.factor(practicals)1      -0.0572158  6.912e-02 -0.82781
as.factor(credit)1:as.factor(networking)1       0.0238834  8.681e-02  0.27511
as.factor(credit)1:as.factor(equipments)1       0.1349423  7.693e-02  1.75405
as.factor(credit)1:farmers                     -0.0003012  5.690e-03 -0.05294
as.factor(credit)1:plotsize                     0.0030714  2.390e-03  1.28500
as.factor(crop)2:as.factor(training)1          -0.0224889  1.052e-01 -0.21385
as.factor(crop)2:as.factor(tour)1               0.0735747  8.663e-02  0.84929
as.factor(crop)2:as.factor(practicals)1        -0.0115648  7.046e-02 -0.16412
as.factor(crop)2:as.factor(networking)1        -0.0293129  1.057e-01 -0.27730
as.factor(crop)2:as.factor(equipments)1         0.0595358  9.784e-02  0.60848
as.factor(crop)2:farmers                       -0.0022693  5.986e-03 -0.37912
as.factor(crop)2:plotsize                       0.0032342  2.678e-03  1.20761
as.factor(training)1:as.factor(tour)1           0.0981488  8.048e-02  1.21948
as.factor(training)1:as.factor(practicals)1     0.0430968  8.309e-02  0.51867
as.factor(training)1:as.factor(networking)1     0.0176581  1.007e-01  0.17532
as.factor(training)1:as.factor(equipments)1    -0.0182289  8.565e-02 -0.21284
as.factor(training)1:farmers                    0.0105546  5.934e-03  1.77865
as.factor(training)1:plotsize                  -0.0049157  2.579e-03 -1.90599
as.factor(tour)1:as.factor(practicals)1        -0.0396537  6.892e-02 -0.57539
as.factor(tour)1:as.factor(networking)1        -0.0665327  9.975e-02 -0.66700
as.factor(tour)1:as.factor(equipments)1        -0.1343733  7.164e-02 -1.87562
as.factor(tour)1:farmers                       -0.0041991  6.095e-03 -0.68896
as.factor(tour)1:plotsize                       0.0028956  2.711e-03  1.06798
as.factor(practicals)1:as.factor(networking)1   0.0415297  7.535e-02  0.55118
as.factor(practicals)1:as.factor(equipments)1   0.0062195  6.751e-02  0.09212
as.factor(practicals)1:farmers                  0.0018413  5.513e-03  0.33397
as.factor(practicals)1:plotsize                 0.0018499  2.600e-03  0.71163
as.factor(networking)1:as.factor(equipments)1 -0.1210462  9.274e-02 -1.30520
as.factor(networking)1:farmers                  0.0035345  6.465e-03  0.54673
as.factor(networking)1:plotsize                -0.0032418  3.238e-03 -1.00118
as.factor(equipments)1:farmers                 -0.0112985  5.641e-03 -2.00299
as.factor(equipments)1:plotsize                 0.0078423  2.546e-03  3.08067
farmers:plotsize                               -0.0007924  8.076e-05 -9.81200
```

[1] "Estimates for logarithm of lambda=var(u_mu)"

```
[1] "Gaussian"      "inverse-gamma"

          Estimate Std. Error
Region      -5.265     0.8563
Community   -2.254     0.3922
```

[1] "Estimates from the model (phi)"

```
Phi ~ as.factor(credit) + as.factor(crop) + as.factor(training) +
    as.factor(tour) + as.factor(practicals) + as.factor(networking) +
    as.factor(equipments) + farmers + plotsize

[1] "Log"
                          Estimate Std. Error
(Intercept)             -2.824893   0.338732
as.factor(credit)1       0.536420   0.203271
as.factor(crop)2        -0.090648   0.192497
as.factor(training)1     1.082872   0.217828
as.factor(tour)1        -0.021638   0.197526
as.factor(practicals)1   0.362818   0.187087
as.factor(networking)1 -0.861374   0.206733
as.factor(equipments)1 -0.025111   0.199204
farmers                  0.011444   0.015589
plotsize                 0.002609   0.007138

[1] "========== Likelihood Function Values and Condition AIC =========="
                                        [, 1]
-2ML (-2 p_v(mu) (h))           :  15470.24
-2RL (-2 p_beta(mu),v(mu) (h)) :  15748.33
cAIC                            :  15511.58

[1] "========== Degrees of freedom and Deviance =========="
                [, 1]
DF :         58.24963
Deviance : 149.05148
```

[1] "========== Random effect =========="

```
          [, 1]
1  -0.11636385
2  -0.09348388
3  -0.05251477
1  -0.61419772
2   0.69989334
3  -0.20205947
4   0.33460102
5  -0.33184174
6   0.23431035
7   0.14477254
8  -0.13254941
9  -0.32487451
10 -0.16650890
11  0.03983901
12 -0.02345832
13  0.07971131
```

MODEL 8: HGLM 2- FIXED = GAUSSIAN/IDENTITY, RANDOM = GAUSSIAN/GAMMA PHI = GAMMA/LOG (FOR FIXED AND INTERACTION TERMS)

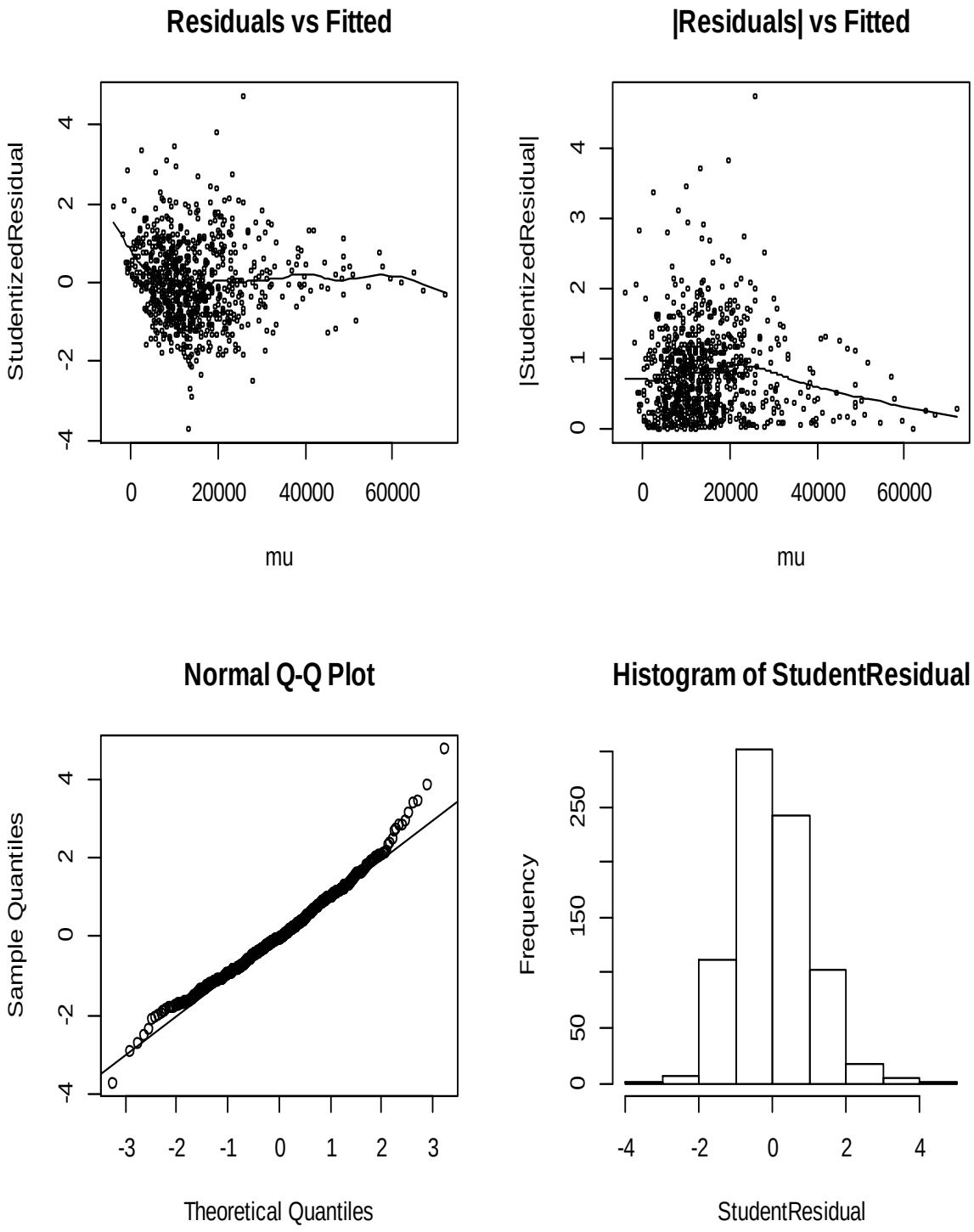

Fig. B8: Diagnostic plot of Gaussian/identity HGLM - 2 for fixed and interaction variables

Table B8: Gaussian/identity HGLM - 2 Model Estimates for fixed and interaction variables

```
Distribution of Main Response:
                     "Gaussian"
[1] "Estimates from the model (mu)"
Yield ~

[1] "Identity"
                                                 Estimate Std. Error t-value
(Intercept)                                      -448.923   1563.276 -0.2872
as.factor(credit)1                              -2682.593   1433.944 -1.8708
as.factor(crop)2                                  205.035   1311.649  0.1563
as.factor(training)1                             4484.912   1529.043  2.9331
as.factor(tour)1                                 3453.206   1498.904  2.3038
as.factor(practicals)1                            360.348   1194.162  0.3018
as.factor(networking)1                           -803.287   1348.171 -0.5958
as.factor(equipments)1                          -6475.975   1320.711 -4.9034
farmers                                           114.747    109.144  1.0513
plotsize                                          778.413     65.776 11.8343
as.factor(credit)1:as.factor(crop)2             -2223.702    816.884 -2.7222
as.factor(credit)1:as.factor(training)1         -1243.696    936.654 -1.3278
as.factor(credit)1:as.factor(tour)1             -1370.824    788.516 -1.7385
as.factor(credit)1:as.factor(practicals)1         430.557    757.639  0.5683
as.factor(credit)1:as.factor(networking)1        2202.608    982.068  2.2428
as.factor(credit)1:as.factor(equipments)1        -268.842    826.234 -0.3254
as.factor(credit)1:farmers                         57.348     75.406  0.7605
as.factor(credit)1:plotsize                       116.143     43.251  2.6853
as.factor(crop)2:as.factor(training)1           -3066.524   1028.328 -2.9820
as.factor(crop)2:as.factor(tour)1                2653.181    847.337  3.1312
as.factor(crop)2:as.factor(practicals)1         -2509.624    748.793 -3.3516
as.factor(crop)2:as.factor(networking)1          1253.758    999.522  1.2544
as.factor(crop)2:as.factor(equipments)1          5539.311    885.561  6.2551
as.factor(crop)2:farmers                          -88.466     76.279 -1.1598
as.factor(crop)2:plotsize                        -144.693     43.694 -3.3115
as.factor(training)1:as.factor(tour)1             558.955    985.860  0.5670
as.factor(training)1:as.factor(practicals)1      -266.119    774.868 -0.3434
as.factor(training)1:as.factor(networking)1       383.991   1113.331  0.3449
as.factor(training)1:as.factor(equipments)1      3019.865    836.578  3.6098
as.factor(training)1:farmers                      -38.035     77.643 -0.4899
as.factor(training)1:plotsize                    -213.346     44.625 -4.7808
as.factor(tour)1:as.factor(practicals)1          -641.078    754.920 -0.8492
as.factor(tour)1:as.factor(networking)1           634.957   1170.920  0.5423
as.factor(tour)1:as.factor(equipments)1         -3704.458    739.663 -5.0083
as.factor(tour)1:farmers                         -169.609     77.535 -2.1875
as.factor(tour)1:plotsize                          46.505     43.904  1.0592
as.factor(practicals)1:as.factor(networking)1    -708.796    776.805 -0.9125
as.factor(practicals)1:as.factor(equipments)1    2477.449    657.066  3.7705
as.factor(practicals)1:farmers                     62.067     67.045  0.9258
as.factor(practicals)1:plotsize                   -25.163     42.625 -0.5903
as.factor(networking)1:as.factor(equipments)1    -232.489    988.817 -0.2351
as.factor(networking)1:farmers                    -18.990     84.619 -0.2244
as.factor(networking)1:plotsize                   -24.096     56.182 -0.4289
as.factor(equipments)1:farmers                      8.558     71.105  0.1203
as.factor(equipments)1:plotsize                    36.957     42.875  0.8620
farmers:plotsize                                   -3.357      1.667 -2.0139
```

[1] "Estimates for logarithm of lambda=var(u_mu)"

```
[1] "Gaussian" "gamma"

          Estimate Std. Error
Region      -11.37     0.8563
Community   -11.57     0.3922
```

[1] "Estimates from the model (phi)"

```
Phi ~ as.factor(credit) + as.factor(crop) + as.factor(training) +
    as.factor(tour) + as.factor(practicals) + as.factor(networking) +
    as.factor(equipments) + farmers + plotsize

[1] "Log"
                         Estimate Std. Error
(Intercept)            14.825507    0.24043
as.factor(credit)1      0.722429    0.14429
as.factor(crop)2       -0.506487    0.13661
as.factor(training)1    0.558480    0.15462
as.factor(tour)1        0.528527    0.14021
as.factor(practicals)1  0.316325    0.13277
as.factor(networking)1 -0.794790    0.14672
as.factor(equipments)1  0.279803    0.14141
farmers                -0.003393    0.01106
plotsize                0.057569    0.00507

[1] "========== Likelihood Function Values and Condition AIC =========="
                                      [, 1]
-2ML (-2 p_v(mu) (h))          :  15654.07
-2RL (-2 p_beta(mu),v(mu) (h)) :  15081.15
cAIC                           :  15746.07

[1] "========== Degrees of freedom and Deviance =========="
                    [, 1]
DF :       4.600001e+01
Deviance : 3.506580e+10
```

[1] "========== Random effect =========="

```
            [, 1]
1   2.740153e-03
2   5.808270e-03
3  -7.457529e-03
1  -1.589856e-03
2   6.303971e-03
3  -1.973962e-03
4   5.004118e-03
5  -2.779775e-03
6   3.580008e-03
7  -2.264001e-03
8  -4.672704e-03
9   6.661968e-05
10  6.742703e-04
11  9.742384e-04
12 -1.102135e-03
13 -1.129898e-03
```