Analysis of stem borers and their parasitoids in maize and sorghum agro-ecosystems in eastern Ethiopia using generalized linear model

A thesis submitted to School of Graduate Studies, Addis Ababa University In Partial Fulfillment of the Requirement for the Degree of Masters of Science in Statistics

By

Bedanie Gemechu Bulty

June 2008

APPROVAL SHEET SCHOOL OF GRADUATE STUDIES

ADDIS ABABA UNIVERITY

As members of the Examining Board of the Final M. Sc. Open Defense, we certify that we have read and evaluated the thesis prepared by Bedanie Gemechu entitled "Analysis of stem borers and their parasitoids in maize and sorghum agro-ecosystems in eastern Ethiopia using generalized linear model" and recommended that it be accepted as fulfilling the thesis requirements for the Degree of Master of Science in Statistics.

Name of Chairman	Signature	Date
Girma Taye		22/07/08
Name of Major Advisor	Signature	Date
Emana Getu		22/07/08
Name of Co-Advisor	Signature	Date
Fentaw Abegaz		17/07/08
Name of Internal Examiner	Signature	Date
Emmanuel G. Yohannes		18/07/2008
Name of External Examiner	Signature	Date

Final approval and acceptance of the thesis is contingent upon the submission of the final copy of the thesis to the Council of Graduate Studies (CGS) through the Departmental Graduate Committee (DGC) of the candidate's major department.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the thesis requirement.

Name of Thesis Advisors	Signature	Date
Girma Taye		22/07/08
Emana Getu		22/07/08

Dedication

I dedicated this thesis manuscript to my lovely mother Temegnu Aga, for nursing me with affection and love from my childhood as mother and father and my younger brother Wesena Gemechu, who is courage and responsible and for their unlimited contribution to the success of my life.

Wesena, I never forget your dedication in maintaining the livelihood of our home together with our elder brother Faye, when I was at school.

Acknowledgement

Very special thanks go to the almighty 'Waqaa' for He does not leave me alone in all my careers.

I would like to express my heartfelt gratitude and indebtedness to my major advisor, Dr. Girma Taye, for his immense devotion, professional assistance to my thesis work and to bring me here from the very start. He was very kind and exceptionally helpful to me. I am very grateful to him.

Grateful appreciation is extended to my co- advisor, Dr Emana Getu, for all his resources, useful comments, unlimited professional assistance and encouragement during my study. Data for this thesis is a courtesy of Dr. Emana Getu from his PhD work. So I am very grateful to him as he found me from where I lost.

Finally, I need to extend my heart felt thanks to Jamjam Baptist church and Christian brothers and sisters for their unlimited support during my careers.

Table of contents

Contents Pages

List of Tables

List of Figures

Abstract

Maize and Sorghum are the most important food crops in Ethiopia. However, stem borers became the major problem resulting in yield losses. A study was conducted in the central Rift valley of Ethiopia to find out stem borers infestation, diversity and their parasitoids interaction, to determine the appropriate statistical model that should be used in the analysis of this sort of data. The analysis of stem borers and their parasitoids in maize and sorghum agro-ecosystems was made by generalized linear model using data recorded from eastern Ethiopia in 1999 and 2000. In previous studies, some researchers have used general linear Model (GLM) analysis on the stem borers' data.

But, general linear model is used for data satisfying assumption of normality where as there are situations in which non-normal data is treated. This includes using generalized linear model analysis that allows the use of different exponential distributions and non-linear link functions. Furthermore, generalized linear model was found to be relatively better than general linear model for the analysis of data which violate the assumption of linearity and normality. Moreover, the response variables: infestation and diversity of stem borers were found to be discrete and counts. The analysis of the data using this model and appropriate link functions is applied to identify the significant predictors which contribute positively or negatively to the diversity and infestation of stem borers among the explanatory variables. Accordingly, generalized linear model was found to be the best model in identifying the significant predictors. Consequently, infestation and diversity of stem borers were not significantly affected by the season, wild host and the predator species. It was also observed that the infestation was not affected by cropping systems where as the diversity is not affected by year, pest species and crop growth stages. Mean separation was made by using LSD and Duncan's multiple range tests to see the differences with in the explanatory variables. There fore, the effects of vegetation types across locations on the infestation and diversity of stem borers is not significantly different from district to districts at 95% confidence. Besides, there is no significant difference of

infestation of stem borers with or with out the wild hosts where as diversity does not differ significantly for the main crops (sorghum or maize), cropping systems (mixed or sole), for the parasitoid species (presence or absence) and for the nitrogen contents. But, there is significant difference in infestation of stem borers between the years, 1999 and 2000, between the main crops, more for sorghum (47.812) than maize (19.775), with in the pests species which high for the key species (44.867) than other species (18.639), with in the growth stages of the crops in which more infestation was measure o during the stubble stage (55.00) followed by maturity stage (43.463) the vegetative stage (29.683) at 95% level of significance. The presence of the parasitoids species resulted in more infestation than its absence. The interaction effects of year and location basically implies the difference of infestation and diversity of stem borers in the years 1999 and 2000 across the locations. This is due to the fact that different locations have specific determinant factors such as temperature, altitude and rain fall and etc. which are not included in this study.

Finally, it is recommended that generalized linear model is flexible, easy to use for any type of data, due attention should be given to the highly infested areas, control should be devised for the stem borers and great care should be given for the parasitoids during the use of chemicals.

Chapter One

Introduction

1.1 Background

Maize (Zea mays L) and sorghum (Sorghum bicolor L) are among the main cereals grown in Ethiopia. About 8.2 million hectares of land is under crop production in Ethiopia with more than 82% under cereal production. Among the cereals, maize and sorghum are grown on about 29% of the cultivated land (CSA, 2000). The total production was 1.2 million tons for maize and 0.3 million tons for sorghum in 1999 and 2000 (CSA, 2000). The average national yield of maize and sorghum is 2 tons and 1 ton ha^{-1}, respectively, which is very low when compared to the world average of 4 tons for maize and 2 tones ha^{-1} for sorghum.(Emana,2002).

A number of factors contributed to low yield of these crops. Herbivore by lepidopterous stem borers are among the most important contributing factors (Emana, 2002). Stem borers are pests that feed on the sorghum and maize stems. In Ethiopia, Assefa (1985) and Emana and Tsedeke (1999) reported 10 to 100% yield losses due to stem borers. Six stem borer species were reported on maize and sorghum in Ethiopia (Emana, 2001; Emana et al., 2001, 2002, 2003). These are the noctuids Busseola fusca (Fuller), Sesamia calamisitis (Hampson), and Sesamia nonagriodes botanephaga (Lefebvre), the crambid, Chilo partellus (Swinhoe), the Rhynchophorids, Rhynchaenus niger (Horn) and Pissodes dubius (Storm). However, B. fusca and C. partellus are the key species causing economic loss. These stem borers occur in mixed populations most of the time. Along stem borers a number of natural enemies consisting of parasitoids, predators and pathogens have been recorded since the start of stem borers' research in the country which is dated back to 1970s Emana et al., (2002). Emana et al., (2002) recorded 21 parasitoids, 14 predators and 7 pathogens in different stages (egg, larva and pupa) of stem borers.

Understanding the diversity of both the stem borers and their associated natural enemies are prerequisite for the successful management of insect pests in general and stem borers in particular (Emana, 2002). For example, natural enemies give certain level of stem borers' population suppression. Additional control measure will be applied based on the remaining pest population, i.e. the concept of economic threshold and economic injury level. Moreover, the control measure to be applied should be very safe for the natural enemies existing in the Agro-ecosystem. As is true for all living things, the diversity and effect of insects existing in certain ecosystem is dependent on many abiotic and biotic factors. The effect levels of these factors vary. These variations can be measured using different statistical models. Among the statistical models, generalized linear model is the most frequently used model because of its simplicity, ability to counterbalance orders, ability to analyze multiple factors with their interactions. Moreover, generalized linear model is used because (1) nonlinear as well as linear effects can be tested (2) it is applicable for categorical as well as for continuous predictor variables, (3) it can be used for any dependent variable whose distribution follows several special members of the exponential family of distributions (e.g., gamma, poisson, binomial, etc.), as well as any normally-distributed dependent variable.

In the general linear model a response variable Y is linearly associated with values on the X variables by:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k + e,$$

where Y is dependent (or response) variable, X's are a set of predictor (or explanatory) variables, b_0 is the regression coefficient for the intercept, the b_i values are the regression coefficients (for variables 1 through k) estimated from the data and e stands for the error variability that cannot be accounted for by the predictors. Note that the expected value of e is assumed to be 0. The relationship in the generalized linear model is assumed to be,

$$Y = g(b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k) + e$$

where e is the error, and g(...) is a function. Formally, the inverse function of g(...), say f(...), is called the link function; so that:

$$f(mu_y) = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

where mu_y stands for the expected value of y.

The generalized linear model differs from the general linear model (of which, for example, multiple regression is a special case) in two major respects: first, the distribution of the dependent or response variable can be (explicitly) non-normal, and does not have to be continuous, i.e., it can be binomial, multinomial, or ordinal multinomial (i.e., contain information on ranks only); second, the dependent variable values are predicted from a linear combination of predictor variables, which are connected to the dependent variable via a link function. The general linear model for a single dependent variable can be considered a special case of the generalized linear model: In the general linear model the dependent variable values are expected to follow the normal distribution, and the link function is a simple identity function (i.e., the linear combination of values for the predictor variables is not transformed).

Estimation of the values of the parameters (b_0 through b_k and the scale parameter) in the generalized linear model are obtained by maximum likelihood (ML) estimation, which requires iterative computational procedures. There are many iterative methods for ML estimation in the generalized linear model, of which the Newton-Raphson and Fisher-Scoring methods are among the most efficient and widely used. The Fisher-scoring (or iterative re-weighted least squares) method in particular provides a unified algorithm for all generalized linear models, as well as providing the expected variance-covariance matrix of parameter estimates as a byproduct of its computations Dobson, (1990).

Statistical tests for the significance of the effects in the model can be performed via the Wald statistic, the likelihood ratio (LR), or score statistic McCullagh and Nelder (1989). The Wald statistic (Dobson, 1990) is computed as the generalized inner product of the parameter estimates with the respective variance-covariance matrix and is an easily computed, efficient statistic for testing the significance of effects.

The score statistic is obtained from the generalized inner product of the score vector with the Hessian matrix (the matrix of the second-order partial derivatives of the maximum likelihood parameter estimates). The likelihood ratio (LR) test requires the greatest computational effort (another iterative estimation procedure) and is thus not as fast as the first two methods; however, the LR test provides the most asymptotically efficient test known (Agresti, et al., 1996)

Diagnostics in the generalized linear model can be made by using two basic types of residuals, Pearson residuals and deviance residuals. Pearson residuals are based on the difference between observed responses and the predicted values; deviance residuals are based on the contribution of the observed responses to the log-likelihood statistic. The deviance residual is computed as:

$r_D = \text{sign}(y-\mu)\ \text{sqrt}(d_i)$, Where $\Sigma\ ^{n}_{i=1}\ d_i = D$, and D is the overall deviance measure of discrepancy of a generalized linear model (McCullagh and Nelder, 1989). Thus, the deviance statistic for an observation reflects its contribution to the overall goodness of fit (deviance) of the model.

To evaluate the goodness of fit of a generalized linear model, a common statistic that is computed is the so-called Deviance statistic. It is defined as:

Deviance = -2 * (Lm - Ls) , where Lm denotes the maximized log-likelihood value for the model of interest, and Ls is the log-likelihood for the saturated model, i.e., the most complex model given the current distribution and link function (Agresti, 1996).

Hence, the current study was conducted to analyze factors affecting both the diversity of stem borers and their parasitoids in eastern Ethiopia where stem borers are the major problems in the production of maize and sorghum.

1.2 Statements of the problem

Studies have been carried out on maize and sorghum stem borers' diversity, infestation and their parasitoids in Eastern Ethiopia. The pests are the major problems resulting in tremendous grain yield losses in Africa in general and Ethiopia in particular. So, control methods should be devised to mitigate the problems of grain yield loss which contributes to the food scarcity in the country. The data consists of factors such as year, locations, wild and crop hosts, cropping systems, seasons, pest species, parasitoids species, nitrogen content and soil texture in relation to existence of parasitoids, extent of stem borers' infestation and diversity. But, appropriate models for analysis have not yet been used. So far, the analysis was based on the general linear model and only summary of statistics and percentage had been produced. In general linear model, some non linear and non normal responses have not been considered. Thus, general linear model may not be helpful for such studies as it might lead to inefficient conclusions and recommendations.

1.3 Objective of the study

General objectives:

To investigate the significant predictors or factors affecting stem borers' diversity/ infestation and their parasitoids interaction using generalized linear model.

To see the comparative advantage of generalized linear model to other models, especially general linear model.

Specific objectives:

1) To analyze the data of maize and sorghum stem borers and their parasitoids using the generalized linear models.

2) To employ some link functions to see the comparative results within the generalized linear model.

3) To investigate the comparative advantages of generalized linear model to other models.

4) To recommend the more efficient model among all other models used so far.

5) To devise effective strategies of stem borers' management.

Chapter Two

Literature Review

2.1 Maize and sorghum production in Ethiopia

Maize and sorghum are important food crops in Ethiopia. Both crops are produced in all regions of the country at elevations ranging from 500-2300 m above sea level (Benti and Ransom, 1993). The yield of these crops is much less than the average potential yields. This is due to various constraints among which poor soil fertility, moisture stress, lack of pure seed, backward cultural practices, diseases and insect pests are the major ones.(Emana, 2002)

2.2 Stem borers

In Ethiopia, Assefa (1985), Emana and Tsedeke (1999) reported three species of stem borers, namely Busseola fusca (Fuller), Chilo partellus (Swinho) and Sesamia calamistis (Hampson). In addition to maize and sorghum, stem borers attack cultivated cereals like rice, finger millet and sugarcane.They also attack numerous wild hosts (Ingram, 1958; Harris and Nwanze, 1992; Overholt et al., 1997; Van Den Berg et al., 1997). There are various controlling methods of stem borers which include chemical control, varietal resistance and biological control. The most prominent stem borers are C. Partellus and B. fusca.

C. Partellus moths are medium sized, straw or light brown in color, with numerous shiny brown spots on the fore wing margins. The hind wings are papery thin and white. The moth is nocturnal in habit and usually lives for approximately one week (Pats, 1992). Although adults are present throughout the year, their numbers are low between cropping seasons (Unnithan and Sexena, 1990). Mating generally takes place during the early hours of the day soon after emergence and on the two to three subsequent nights (Bonhof, 2000) and egg laying during the evening hours. The

female lays up to 500 eggs in batches of 10-80 near the midrib on the under surface of the leaves (Bonhof, 2000). Eggs can be found in the field from one week to three weeks after plant emergence (Seshu Reddy, 1983). Eggs hatch in 5-11 days early in the morning (Pats, 1992) and many larvae are carried away by the wind on silken threads and infest the neighboring plants (Berger, 1992).

B. fusca females lay eggs in batches of 30-150 on the inner surface of leaf Sheaths. A female lays about 400-500 eggs over a period of 5-6 days. Eggs hatch in 5-6 days, and the young larvae remain in clusters inside the leaf sheaths. The larvae disperse the following night and move to the leaf whorl for feeding. The larvae have buff to purple brown colored bodies. There are 6-7 larval instars, and larval development is completed in 24-36 days. Larvae pupate in the plant stem, and cut an exit hole before pupation. Adults emerge in 9-12 days, and exhibit a wide variation in color. Usually 2 to 3 generations are produced in a year. It is only the 1st generation that causes severe damage to the crop. The 2nd or 3rd generation larvae enter diapause with the on set of the dry season, and complete development in 6-7 months (Sharma, 1993). In Ethiopia similar phenology was reported by Assefa (1988).

2.3 The Parasitoids

Insect parasitoids have an immature life stage that develops on or within a single insect host, ultimately killing the host, hence the value of parasitoids as natural enemies. Adult parasitoids are free-living and may be predaceous. Parasitoids are often called parasites, but the term parasitoid is more technically correct. Most beneficial insect parasitoids are wasps and flies, although some rove beetles and other insects may have life stages that are parasitoids. Most insect parasitoids only attack a particular life stage of one or several related species. The immature parasitoid develops on or within a pest, feeding on body fluids and organs, eventually leaving the host to pupate or emerging as an adult. The life cycle of the pest and parasitoid can coincide, or that of the pest may be altered by the parasitoid to accommodate its development.

The life cycle and reproductive habits of beneficial parasitoids can be complex. In some species, only one parasitoid will develop in or on each pest while, in others, hundreds of young larvae may develop within the pest host. Over wintering habits may also vary. Female parasitoids may also kill many pests by direct feeding on the pest eggs.

Major characteristics of insect parasitoids:

- they are specialized in their choice of host
- they are smaller than their host
- only the female searches for host
- different parasitoid species can attack different life stages of host eggs or larvae are usually laid in, on, or near host

One of the most prominent parasitoids is Cotesia flavipes. Cameron flavipes adults start mating soon after emergence, especially in bright light and mating lasts for about one minute. Unmated females produce male progeny by parthenogenesis. Females start ovipositing on the day of emergence (Mohyuddin, 1971). C. flavipes female enters through holes in stems excavated by stem borers, traverses the tunnel to locate the stem borer and then directly inject eggs into the host. Oviposition lasts only few seconds. Eggs hatch within 3-4 days in its host larvae and first instars parasitoid larva begins feeding internally.

2.4 Generalized Linear Model

The study of Bailer and Oris (1996) showed that generalized linear models (GLiMs) provide a general model for data commonly encountered in aquatic toxicology. They proposed effective concentration (EC) estimator, labeled a relative inhibition or RI estimator was derived in the GLiM framework. This estimator represents the concentration, or more generally, the exposure level to some hazard, associated with a specified level of change in the response relative to the control response. Along with the construction of this estimator, standard errors and confidence intervals were presented. This RI estimator was then applied to dichotomous, count, and continuous responses to illustrate its use.

Kerr and Meador (1995) have also described a method to determine and model dose-response relationships from binomial response data using generalized linear models (GLM). The benefit in using this technique is that it allows LC_p or LD_p to be determined without an initial linear zing transformation. (LC_p and LD_p are the lethal concentration or dose that causes p proportion of test animals to die at a specified time period.) They found that the method of GLM is an appropriate way to analyze a dose-response relationship because it utilizes the inherent S-shaped feature of the toxicological response and incorporates the sample size of each trial in parameter estimation. This method is also much better behaved when the extremes of the response probability are considered because responses of 0% and 100% are included in the model. The advantageous feature of this method is confidence intervals (C.I.s) for both the dose estimate and response probabilities can be computed with GLM, which provides a more complete description of the estimates and their inherent uncertainty. Because C.I.s for both the dose estimate and response probabilities can be constructed, the lowest observed effect concentration (LOEC) can also be determined. Liang and Zeger (1986), on the other hand, proposed an extension of generalized linear models to the analysis of longitudinal data. They introduced a class of estimating equations that give consistent estimates of the regression parameters and of their variance under mild assumptions about the time dependence. The estimating

equations are derived without specifying the joint distribution of a subject's observations yet they reduce to the score equations for multivariate Gaussian outcomes. Asymptotic theory is presented for the general class of estimators. The approach was closely related to quasi-likelihood. They had also introduced a generalized estimating equations approach based on a 'working' correlation matrix to obtain consistent and efficient estimators of regression parameters in the class of generalized linear models for repeated measures data. They based the analysis on specifications for the means and variances of the observations, as usual for generalized linear models, but showed how specifications for the correlations between measurements made on the same unit could be avoided by using a 'working' correlation matrix. In some cases the parameters involved in this matrix are subject to an uncertainty of definition, which can lead to a breakdown of the asymptotic properties of the estimators.

As demonstrated by Crowder (1995), because of the uncertainty of definition of the working correlation matrix, the Lian-Zeger approach may in some cases lead to a complete breakdown of the estimation of the regression parameters. In his study he showed that, even though the Lian-Zeger approach in many situations yields consistent estimators for the regression parameters, these estimators are usually inefficient as compared to the regression estimators obtained by using the independence estimating equations approach.

Gareth & James M. presented a technique for extending generalized linear models (GLM) to the situation where some of the predictor variables are observations from a curve or function. The technique is particularly useful when only fragments of each curve have been observed. They demonstrated, on both simulated and real world data sets, how this approach could be used to perform linear, logistic and censored regression with functional predictors. In addition, they showed how functional principal components could be used to gain insight into the relationship between the response and functional predictors. Finally, they extended the methodology to apply GLM and principal components to standard missing data problems.

The class of generalized linear models is extended to allow for correlated observations, nonlinear models and error distributions not of the exponential family form. The extended class of models includes a number of important examples, particularly of the composite transformational type. Large-sample inference and maximum likelihood estimation for the extended class of generalized linear models are discussed, and the analysis of deviance is generalized to the extended class of models. Calculation of the maximum likelihood estimate for a general likelihood by Fisher's scoring method and a related method is considered, and the relation with the Gauss–Newton method is discussed.

Generalized linear models have unified regression methodology for a wide variety of discrete, continuous and censored response that can be assumed to be independent (Mc Cullagh and Nelder, 1989; Nelder and Wedderburn, 1972). Several authors have investigated the extension of random effects models to the generalized linear model (GLM) family. The beta-binomial (Williams, 1982) and Poisson gamma models (Breslow, 1984) were among the earlier. Here covariates cannot vary with in a cluster. Stiratelli, et al. (1984) and Anderson and Aitkin (1985) considered general covariates in logistic regression with a Gaussian intercept being expectation maximization (EM) and Newton algorisms, respectively. Ochi and Prentice (1984) and G-Mour, Anderson, and Rae (1985) discussed probit Gaussion models. Harville and Mee (1984) studied random effects models for ordered categorical data. With count data, Breslow (1984), Croweder (1985) and Tsutakawa (1988) have investigated log-linear models with random effects. Related models are used in Bayesian analysis of contingency table for example by Leonard (1975).

As Nelder (1985) has pointed out in his discussion of Lindlely and Smith's (1972) article on Bayesian methods in regression, there is a strong connection between the random effects and Bayesian regression models. Haberman and Renshaw (1988) had also attempted to demonstrate the versatility of generalized linear models and the statistical package GLIM for tackling a range of modeling problems. According to these authors three different types of problem based on data from different related

areas have been discussed. They have also indicated how to adapt the GLIM package to fit certain types of distribution, which are not immediately available within the system. Three examples are the Pareto, Burr and Weibull distributions. It is a trivial matter to fit other loss distributions such as the lognormal, gamma and log gamma to uncomplicated data. Modeling such large complex data sets may be viewed as a balancing act between model complexity and the need to encapsulate the salient underlying features present in the data. The simpler the model is the simpler the interpretation of the underlying data generating mechanism. Modeling does not necessarily have a unique solution, but a model may be deemed adequate only if it achieves this goal. One way of assessing the adequacy is through a thorough graphical analysis of model residuals, which ideally should be "pattern free". Additionally, what might be termed "fine tuning" might then be attempted, and its effects formally assessed.

The GLIM-based approach outlined here could pave the way for a completely new, scientifically sound approach to life insurance underwriting. It offers a more dynamic means of model building than has hitherto been attempted in this field, in which the effects of individual factors and their interactions on excess mortality may be assessed. It was highlighted the meager assumptions on which the models are based, the comparative ease with which they can be fitted and compared using GLIM and the appealing connection which these models have with the traditional actuarial standard mortality ratios.

Generalized linear models (GLMs) have been used by Khuri A.I., et al. (2006) quite effectively in the modeling of a mean response under nonstandard conditions, where discrete as well as continuous data distributions can be accommodated. The choice of design for a GLM is a very important task in the development and building of an adequate model. However, one major problem that handicaps the construction of a GLM design is its dependence on the unknown parameters of the fitted model. Several approaches have been proposed in the past 25 years to solve this problem. These approaches, however, have provided only partial solutions that apply in only

some special cases, and the problem, in general, remains largely unresolved. To focus attention on the aforementioned dependence problem, they provided a survey of various existing techniques dealing with the dependence problem. This survey includes discussions concerning locally optimal designs, sequential designs, Bayesian designs and the quantile dispersion graph approach for comparing designs for GLMs.

The research on designs for generalized linear models is still very much in its developmental stage. Not much work has been accomplished either in terms of theory or in terms of computational methods to evaluate the optimal design when the dimension of the design space is high. The situations where one has several covariates (control variables) or multiple responses corresponding to each subject demand extensive work to evaluate "optimal" or at least efficient designs. The curve fitting approach of Müller and Parmigiani (1995) may be one direction to pursue in higher-dimensional design problems. Finding robust and efficient designs in high-dimensional problems will involve formidable computational challenges and efficient search algorithms need to be developed. Ideas can be brought into case-control studies where the prime objective is to study the association between a disease (say, lung cancer) and some exposure variables (such as smoking, residence near a hazardous waste site, etc.). Classical case-control studies are carried out by sampling separately from the case (persons affected with the disease) and control (persons without the disease) populations, with the two sample sizes being fixed and often arbitrary. Chen (2000) proposed a sequential sampling procedure, which removes this arbitrariness. Specifically, he proposed a sampling rule based on all the accumulated data, which mandates whether the next observation (if any) should be drawn from a case or a control population. He showed also certain optimality of his proposed sampling rule. However, like much of the stochastic approximation literature, Chen touched very briefly on the choice of a stopping rule, but without any optimality properties associated with it. It appears that a Bayes stopping rule or some approximation thereof can be introduced along with Chen's sampling rule so that the issues of optimal stopping and choice of designs can be addressed simultaneously.

The use of the quantile dispersion graphs (of the mean-squared error of prediction) provides a convenient technique for evaluating and comparing designs for generalized linear models. The main advantages of these graphs are their applicability in experimental situations involving several control variables, their usefulness in assessing the quality of prediction associated with a given design throughout the experimental region, and their depiction of the design's dependence on the parameters of the fitted model. There are still several other issues that need to be resolved. For example, the effects of misspecification of the link function and/or the parent distribution of the data on the shape of the quantile plots of the quantile dispersion graph approach need to be investigated. In addition, it would be of interest to explore the design dependence problem in multi-response situations involving several response variables that may be correlated.

McCullagh and Nelder introduced the GLM for exponential family data with the following form:

$f_Y(y, \delta, \varphi) = \exp\{(y_\delta - b(\delta))/a(\varphi) + c(y, \varphi)\}$,............. (1)

where a (.), b (.) and c (.) are known functions, δ is the canonical parameter, and φ is the dispersion parameter. A sample of n independent observations $y = (y_1, \ldots, y_n)$_ are drawn, each with diversity given by (1), where in addition we assume:

$\mu_i = E[y_i] = g^{-1}(x_i'\beta)$ ....(2)

where g(.) is a given link function, xi is a p×1 vector of covariates for the ith subject, and β is a p×1 vector of regression parameters. The link function g (.) determines the function (2).

Writing a $(\varphi) = \varphi/mi$ for some known weights mi yields log-likelihood equations of the form:

$y(\beta, \varphi \mid X, y) = \sum_{i=1}^{n} (y \mid x(\beta, \varphi \mid x_i, y_i) = \sum_{i=1}^{n}\{m_i\, [y_i h(x_i'\beta) - b(h(x_i'\beta))] / \varphi + c(y_i, \varphi)\}$.

where X is the n × p matrix of covariates for all of the observations. In general, solution of the likelihood equations requires an iterative process, but this has been widely implemented in standard computing packages. One of the attractive properties of the GLM is that it allows for linear as well as non-linear models under a single framework.

It is possible to fit models where the underlying data are normal, Poisson, binomial or gamma (as well as others) by suitable choice of the functions a (.), b (.), c (.), g (.) for the given sample space.

The GLM presumes the response and all covariates are observed for each subject. The joint distribution can be partitioned such that f(Y,X) = f(Y |X)f(X). Given complete data the likelihood factors and MLE's can be found by maximizing f (Y |X), the conditional distribution of interest. The parameters governing the distribution of the covariates are ancillary to the parameters of interest (β) in the GLM and can be ignored. We now consider the setting where the vector of covariates xi may not be fully observed for all subjects.

Perry et al, (2000) has proposed the powerful general Pacala^Hassell host parasitoid model for a patchy environment, which allows host diversity-dependent heterogeneity (HDD) to be distinguished from between-patch, host diversity independent heterogeneity (HDI), is reformulated within the class of the generalized linear model (GLM) family. This improves accessibility through the provision of general software within well-known statistical systems, and allows a rich variety of models to be formulated. Covariates such as age class, host diversity and a biotic factor may be included easily. For the case where there is no HDI, the formulation is a simple GLM. When there is HDI in addition to HDD, the formulation is a hierarchical generalized linear model. Two forms of HDI model are considered, both with between-patch variability: one has binomial variation within patches and one has extra-binomial, over dispersed variation within patches. Examples are given demonstrating parameter estimation with standard errors, and hypothesis testing. For

one example given, the extra-binomial component of the HDI heterogeneity in parasitism is itself shown to be strongly diversity dependent.

Lamouroux N. and Jowett I. G. (2005) has used conventional in stream habitat models (e.g., the physical habitat simulation system) to predict the impact of regulation on the habitats of freshwater taxa. They link a hydraulic model with microhabitat-suitability models for taxa to predict habitat values at various discharge rates. Their use requires considerable field effort and experience. Recent analyses performed in France suggested that comparable results could be achieved using simplified hydraulic data. They tested this approach for 99 stream reaches and nine aquatic taxa in New Zealand. The resulting generalized habitat models predict habitat values similar to those predicted by conventional models from simplified hydraulic data (depth-discharge and width-discharge relationships, average particle size, and mean annual discharge). As in France, within reach changes in habitat values were linked to the specific discharge of reaches, while between-reach changes depended mainly on the Froude number at mean annual discharge. The generalized models perform well outside their calibration range. Models previously developed in France perform well in New Zealand. Such generalized models contribute to identifying the key hydraulic variables for freshwater taxa and should facilitate habitat studies with out replacement.

Emana Getu (2002) had used the general linear model (GLM) procedure in SAS program in order to analyze the data of ecological diversity of stem borers and its parasitoids in the eastern Ethiopia. Then the correlation matrix, stepwise regression and GIS were used to arrive at the result that stem borers result in 20-50% loss to the crops by the same researcher. Abiy Tilahun, (2005) had also used general linear model procedure in SAS Computer program for the data analysis of maize and sorghum stem borer species and their parasitoids. Prior to his analysis, he checked the data for normality and for data which lacked normality, he transformed using logarithmic and arcsine transformation. He did analysis of variance for each data and

correlation analysis to compare the production practices with maize and sorghum stem borer species and their parasitoids. He used Duncan's multiple range tests to separate means.

Chapter Three

Materials and Methods

3.1 The Study area and Sources of data

The data for this study was obtained from the study of "diversity of stem borers and its parasitoids interaction in eastern part of Ethiopia". Data were collected through the consecutive surveys of 1999 and 2000 in eastern Ethiopia which covered 148 sorghum and/or maize farmers' field of 46 districts categorized according to their vegetation types. During each survey five plants were sampled randomly and additionally five infested plants using purposive sampling were destructively sampled for the assessment from each field using zigzag pattern of sampling (Emana, 2000). The total number of size used for this study is 148, which is equal to the number of farmers' field. The data set consists of cell counts response variables.

Data Representation

Symbols	Description
year	Year (1999, 2000)
locbyveg	Location classified based on their vegetation type cropl=cropping land, other veg (dry land,grass land, savana land, shrub and others)
wh	wild host (present, absent)
ch	crop host (wch=with crop host, woch=with out crop host)
mcrop	main crops (maize, sorghum)
cstage	crops growth stage (stubble, maturity, vegitative)
cropping	cropping system (sole,mixed)
pessp	pest species: (keysp (Chilo partellus and Busseola fusca), othersp
cp	Chilo partellus
bf	Busseola fusca
sc	Sesamia calamisitis Hampson
new	unknown new pests
sb	Sesamia nonagrides botanephaga
diversity	number of pests (count data)
infesta	Stem borers infestation (discrete data)
parasp	parasitoids species (present, absent)
cf	Catesia flavipes
df	Dolochogendia fuscivora
predsp	predators species ((present, absent))
nc	Nitrogen contents (less than 0--0.2==0, greater than 0.2=1

3.2 Methods of Data Analysis

3.2.1 Generalized Linear Model

The class of generalized linear models is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an exponential family of distributions. Many widely used statistical models are generalized linear models. These include classical linear models with normal errors, logistic and probit models for binary data, and log-linear models for multinomial data. Many other useful statistical models can be formulated as generalized linear models by the selection of an appropriate link function and response probability distribution.

A traditional linear model is of the form:

$$y_i = x_i'\beta + \varepsilon_i$$

where y_i is the response variable for the i^{th} observation. The quantity x_i is a column vector of covariates, or explanatory variables for observation i, that is known from the experimental setting and is considered to be fixed, or non-random. The vector of unknown coefficients β is estimated by a least squares fit to the data y. The $\varepsilon_{i's}$ are assumed to be independent, normal random variables with zero mean and constant variance. i.e. $\varepsilon_i \sim NIID(0, \delta^2)$

The expected value of y_i, denoted by μ_i, is:

$$\mu_i = x_i'\beta$$

While traditional linear models are used extensively in statistical data analysis, there are types of problems for which they are not appropriate.

i) It may not be reasonable to assume that data are normally distributed. For example, the normal distribution (which is continuous) may not be adequate for modeling counts or measured proportions that are considered to be discrete.

ii) If the mean of the data is naturally restricted to a range of values, the traditional linear model may not be appropriate since the linear predictor $x_i'\beta$ can take on any value. For example, the mean of a measured proportion is between 0 and 1, but the linear predictor of the mean in a traditional linear model is not restricted to this range.

iii) It may not be realistic to assume that the variance of the data is constant for all observations. For example, it is not unusual to observe data where the variance increases with the mean of the data.

A generalized linear model extends the traditional linear model and is therefore applicable to a wider range of data analysis problems. A generalized linear model consists of the following components.

The linear component is defined just as it is for traditional linear models:

$$\eta_i = x_i'\beta$$

A monotonic differentiable link function g describes how the expected value of y_i is related to the linear predictor μ_i:

$$g(\mu_i) = x_i'\beta$$

The response variables y_i are independent for i = 1, 2, ... and have a probability distribution from an exponential family. This implies that the variance of the response depends on the mean μ through a variance function V:

$$\text{var}(y_i) = \phi V(\mu i)/ wi,$$

where ϕ is a constant and wi is a known weight for each observation. The dispersion parameter ϕ is either known, for example for the binomial distribution, or it must be estimated.

As in the case of traditional linear models, fitted generalized linear models can be summarized through statistics such as parameter estimates, their standard errors, and goodness-of-fit statistics. We can also make statistical inference about the parameters using confidence intervals and hypothesis tests. However, specific inference procedures are usually based on asymptotic considerations, since exact distribution theory is not available or is not practical for all generalized linear models.

3.2.2 Some Forms of Generalized Linear Models

We construct a generalized linear model by deciding on response and explanatory variables for our data and choosing an appropriate link function and response probability distribution. Some forms of generalized linear models follow depending on the behavior of response variables. Explanatory variables can be any combination of continuous variables, classification variables, and interactions.

Model	Response variable	Distribution	Link function
Traditional Linear Model	continuous	normal	identity, $\eta = \mu$
Logistic Regression	proportion	binomial	logit, $\mu = \log(\mu/1-\mu)$
Poisson Regression in Log Linear Model	count	poisson	log, $\eta = \log(\mu)$
Gamma Model with Log Link	positive, continuous	gamma	log, $\eta = \log(\mu)$

3.2 .3 Test of Goodness of Fit

Assessing the goodness of fit is to examine the extent to which the fitted values of the response variable under the model compare with observed values. The model attempts to generate the parameter estimates that make the result most likely (Hosmer and Lamshow, 2000).

Hypothesis: H_0: the hypothesized model fits the data versus H_A: not H_0

So, the Goodness of fit criteria of the generalized linear procedure (GENMOD) gives us either to accept the fitness or not. There are two approaches mentioned below to assess goodness of fit.

3.2.3.1 Log Likelihood Analysis

One of the methods of assessing the goodness of fit is log likelihood. We obtain the likelihood function as follows:

Let $x=(x_1,x_2,\ldots,x_n)$ be a random vector and $\{ f_x (x/ \theta : \theta \in \varphi)\}$ a statistical model parameterized by $\theta = (\theta_1, \theta_2,\ldots, \theta_k)$ be the parameter vector in the parameter space φ.The likelihood function is a map $L: \varphi \rightarrow \Re$ given by :

$$L(\theta|x) = f_x(x|\theta)$$

The parameter vector $\hat{\theta}$such that for all $L(\hat{\theta}) \geq L(\theta), \theta \in \varphi$is called a maximum likelihood estimate, or MLE, of θ.

Many of the diversity functions are exponential in nature; it is therefore easier to compute the MLE of a likelihood function L by finding the maximum of the natural log of L, known as the log-likelihood function:

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = \ln(L(\boldsymbol{\theta} \mid \boldsymbol{x}))$$

Suppose a sample of n data points X_i are collected. Assume that the $X_i \sim N(\mu, \delta^2)$ and the X_i's are independent of each other. Then, the joint probability density function (pdf) of the X_i, and hence the likelihood function is:

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{\Sigma(x_i - \mu)^2}{2\sigma^2}\right).$$

The log-likelihood function is:

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = -\frac{\Sigma(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi).$$

Taking the first derivative (gradient), we get:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \left(\frac{\Sigma(x_i - \mu)}{\sigma^2}, \frac{\Sigma(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2}\right).$$

Setting $\partial L / \partial \theta = 0$ and solve for $\theta = (\mu, \delta^2)$ we have

$$\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}^2) = \left(\overline{x}, \frac{n-1}{n} s^2\right),$$

Finally, we verify that $\hat{\theta}$ is indeed the MLE of θ by checking the negativity of the 2nd derivatives (for each parameter).

The test statistic for the above test was based on the likelihood function L.

$-2LL \sim \chi^2 (n-k)$, under H_0

For the model that fits perfectly, the likelihood is equal to 1 which is the same as -2LL=0

3.2.3.2 The Deviance Analysis

A likelihood-ratio test can be used under full maximum likelihood (ML). The use of such a test is a quite general principle for statistical testing. In hierarchical linear models, the deviance test is mostly used for multi-parameter tests and for tests about the random part of the model. This is based on the likelihood function of the observed predictors for the fitted model (hypothesized model) say L_h, and the likelihood function for the true distribution under the assumed perfect model (full or saturated model), say L_f.

The deviance denoted by D is given by:

$$D = -2\log(L_h/L_f) = -2(\log L_h - \log L_f)$$

$$D \sim \chi^2 (k-1)$$

Larger values of D are encountered when L_h is small relative to L_f, indicating that the current model is poor.

3.2.4 SAS Procedures

The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector β in SAS program. There is, in general, no closed form solution for the maximum likelihood estimates of the parameters. The GENMOD procedure estimates the parameters of the model numerically through an iterative fitting process. The dispersion parameter ϕ is also estimated by maximum likelihood, or optionally by the residual deviance or by Pearson's chi-square divided by the degrees of freedom. Co-variances, standard errors, and associated p-values are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators.

The proc GENMOD produce has five default outputs in SAS program. These are stated as below.

i) Model Information: It provides information about the specified model and the input data set.

ii) Class level information: It identifies the levels of the classification variables that are used in the model.

iii) Goodness of Fit Criteria: It contains statistics that summarize the fit of the specified model. These statistics are helpful in judging the adequacy of a model and comparing it with other models under consideration.

iii) Parameter Estimates: The "Analysis of Parameter Estimates" summarizes the results of the iterative parameter estimation process. For each parameter in the model, proc GENMOD prints columns with the parameter name; the degrees of freedom associated with the parameter, the estimated parameter value, the standard error of the parameter estimate, and a Wald chi-square statistic and associated p-value for testing the significance of the parameter to the model. If a column of the model matrix corresponding to a parameter is found to be linearly dependent, or aliased, with columns corresponding to parameters preceding it in the model, proc GENMOD assigns it zero degrees of freedom and prints a value of zero for both the parameter estimate and its standard error.

iv) Type1 and Type3 analysis: It is usually of interest to assess the importance of the main effects in the model. Type 1 and Type 3 analyses generate statistical tests for the significance of these effects. You can request these analyses with the type1 and type3 options in the model statement.

We have also other procedures of SAS program which are used for the analysis of different specific models. For instance, the GLM is a SAS procedure that uses the method of least squares to fit general linear model. Among the statistical methods available in PROC GLM are regression, analysis of Variance (ANOVA), analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA) and partial correlation.

PROC GLM analyzes data within the frame work of General linear model .It handles models relating one or several continuous dependent variable to one or several independent variables.

3.2.5 Generalized Additive Models

Generalized Additive Models are generalizations of generalized linear models. In generalized linear models, the transformed dependent variable values are predicted from (is linked to) a linear combination of predictor variables; the transformation is referred to as the link function; also, different distributions can be assumed for the dependent variable values. An example of a generalized linear model is the Logit Regression model, where the dependent variable is assumed to be binomial, and the link function is the logit transformation. In generalized additive models, the linear function of the predictor values is replaced by an unspecified (non-parametric) function, obtained by applying a scatter plot smoother to the scatter plot of partial residuals (for the transformed dependent variable values).

The Generalized Linear Model (GLZ) is a generalization of the general linear model. In its simplest form, a linear model specifies the (linear) relationship between a dependent (or response) variable Y, and a set of predictor variables, the X's, so that

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

In this equation b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables 1 through k) computed from the data.

3.2.6 Distribution of dependent variable

The dependent variable of interest may have a non-continuous distribution, and thus, the predicted values should also follow the respective distribution; any other predicted values are not logically possible. For example, a researcher may be interested in predicting one of three possible discrete outcomes (e.g., a consumer's choice of one of three alternative products). In that case, the dependent variable can only take on 3 distinct values, and the distribution of the dependent variable is said to

be multinomial. Or suppose we are trying to predict people's family planning choices, specifically, how many children families will have, as a function of income and various other socioeconomic indicators. The dependent variable -- number of children is discrete (i.e., a family may have 1, 2, or 3 children and so on, but cannot have 2.4 children), and most likely the distribution of that variable is highly skewed (i.e., most families have 1, 2, or 3 children, fewer will have 4 or 5, very few will have 6 or 7, and so on). In this case it would be reasonable to assume that the dependent variable follows a Poisson distribution.

3.2.7 Link function

A reason why the linear (multiple regression) model might be inadequate to describe a particular relationship is that the effect of the predictors on the dependent variable may not be linear in nature. For example, the relationship between a person's age and various indicators of health is most likely not linear in nature: During early adulthood, the (average) health status of people who are 30 years old as compared to the (average) health status of people who are 40 years old is not markedly different. However, the difference in health status of 60-year-old people and 70-year-old people is probably greater. Thus, the relationship between age and health status is likely non-linear in nature. Probably some kind of a power function would be adequate to describe the relationship between a person's age and health, so that each increment in years of age at older ages will have greater impact on health status, as compared to each increment in years of age during early adulthood. Put in other words, the link between age and health status is best described as non-linear, or as a power relationship in this particular example.

The generalized linear model can be used to predict responses both for dependent variables with discrete distributions and for dependent variables that are nonlinearly related to the predictors.

3.2.8 Link Function and Distribution Function

The link function in generalized linear models specifies a nonlinear transformation of the predicted values so that the distribution of predicted values is one of several special members of the exponential family of distributions (e.g., gamma, Poisson, binomial, etc.). The link function is therefore used to model responses when a dependent variable is assumed to be nonlinearly related to the predictors.

Various link functions are commonly used, depending on the assumed distribution of the dependent variable (y) values:

Distributions	Corresponding Link functions
Normal, Gamma, Inverse normal, and Poisson	Identity link: $f(z) = z$ Log link: $f(z) = \log(z)$ Power link: $f(z) = z^a$, for a given a
Binomial, and Ordinal Multinomial	Logit link: $f(z)=\log(z/(1-z))$ Probit link: $f(z)=\text{invnorm}(z)$, where invnorm is the inverse of the standard normal cumulative distribution function. Complementary log-log link: $f(z) =\log(-\log(1-z))$ Loglog link: $f(z) =-\log(-\log(z))$
Multinomial	Generalized logit link: $f(z_1 \mid z_2,\ldots, z_c)=\log(x_1/(1-z_1-\ldots-z_c))$, where the model has c+1 categories.

3.3 Assumptions

Quantitative models always rest on assumptions about the way the world works, and regression models are no exception. There are four principal assumptions: linearity (the relationship between dependent and independent variables), independence of the errors (no serial correlation), homoscedasticity (constant variance) of the errors versus time or the predictions (or versus any independent variable), normality of the error distribution and Zero residual mean which justify the use of linear regression models for purposes of prediction. If any of these assumptions is violated (i.e., if there is nonlinearity, serial correlation, heteroscedasticity, and/or non-normality), then the forecasts, confidence intervals, and economic insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading. The violations of these assumptions are treated so as to offset the problems encountered in analysis.

3.4 Violation of the assumptions and their treatments

Violation of one of these assumptions is a series problem in fitting a linear model and in prediction. Nonlinearity is usually most evident in a plot of the observed versus predicted values or a plot of residuals versus predicted values, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or a horizontal line in the latter plot. To fix this problem, we apply a nonlinear transformation to the dependent and/or independent variables. Violations of homoscedasticity or heteroscedasticity also makes difficult to gauge the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Some combination of logging and/or deflating will often stabilize the variance in this case. In the case of non normality, distribution may be skewed by the presence of a few large outliers; influences on parameter estimates, confidence intervals may be too wide or too narrow. This problem can be easily detected by normal probability plot (PP-plot) of the residuals or histogram. If

the distribution is normal, the points on this plot should fall close to the diagonal line. A bow-shaped pattern of deviations from the diagonal indicates that the residuals have excessive skewness (i.e., they are not symmetrically distributed, with too many large errors in the same direction). An S-shaped pattern of deviations indicates that the residuals have excessive kurtosis--i.e., there are either two many or two few large errors in both directions. Violations of normality often arise either because (a) the distributions of the dependent and/or independent variables are themselves significantly non-normal, and/or (b) the linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems.

3.5 Limitations

The major conceptual limitation of all regression techniques is that one can only ascertain relationships, but never be sure about underlying causal mechanism. For example, one would find a strong positive relationship (correlation) between the damage that a fire does and the number of firemen involved in fighting the blaze. Do we conclude that the firemen cause the damage? Of course, the most likely explanation of this correlation is that the size of the fire (an external variable that we forgot to include in our study) caused the damage as well as the involvement of a certain number of firemen (i.e., the bigger the fire, the more firemen are called to fight the blaze). Even though this example is fairly obvious, in real correlation research, alternative causal explanations are often not considered.

3.6 Choice of the Number of Variables and Multicollinearity

Multiple regressions are a seductive technique: "plug in" as many predictor variables as you can think of and usually at least a few of them will come out significant. This is because one is capitalizing on chance when simply including as many variables as one can think of as predictors of some other variable of interest. Multicollinearity is the common problem in many correlation analyses where two or more predictors are completely redundant. There are many statistical indicators of this type of

redundancy (tolerances, semi-partial R, etc., as well as some remedies (e.g., Ridge regression).

Chapter Four

Results and Discussion

4.1 Fitting General Linear Model

To study the effects of the explanatory variables on two response variables, namely infestation and diversity of stem borers, the general linear model was used prior to the model selections. The model was fitted by using GLM procedure in SAS program before and after making transformation and the results of model adequacy and analysis of variance was presented in the appendices and in the succeeding tables. The untransformed analysis of infestation and diversity of stem borers out put is given in the Tables A1, A2 and A3 of Appendix A and in the Tables B1, B2 and B3 of Appendix B, respectively. When diagnosis of the untransformed response variables are taken for the fulfilments of normality and linearity using normal probability plot and scatter plot matrix of residuals respectively, it is clearly observed in figures 1 and 3, and figures 9 and 11 that there exists slight problems of normality and linearity. To offset these problems, transformations should have been taken and the square root and logarithm transformation for infestation and diversity was taken respectively, which looked better solve the problems of normality, linearity and homogeneity among all other attempted transformations (like log, arcsine, cosine, sine etc.) as it can be seen in the diagnostic parts presented on the figures, 2, 4 for normality, figures 6 and 8 for linearity and figures 10 and 12 for homogeneity.

Therefore, the results of infestation and diversity of stem borers were dealt after the square root and logarithmic transformations respectively, and fitted general linear model analysis shown in the following tables.

Table 1: Tests of model adequacy for infestation of stem borers

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	604.437415	40.295828	8.19	<.0001
Error	132	649.152557	4.917822		
Corrected Total	147	1253.589972			

R-Square	Coeff Var	Root MSE	infestasqrt Mean
0.482165	39.11513	2.217616	5.669459

The model under consideration fits the data of infestation significantly well at both 99% and 95% level of significance (Table 1). But, the result of $R^2 = 0.482165$ can be interpreted as the infestation of the stem borers is explained by the explanatory variables with approximately 48% and the remaining 52% is the residual variability. This implies that the predictors are very poor in explaining the response variable. In the results shown in Table 2 below, by inserting one variable after the other, only the variables year, crop host, pest species, parasitoid species and the interaction effects of pest species and parasitoid species become important variables in explaining the infestation of stem borers which left out most of the explanatory variables insignificant. When the effect of the explanatory variables is treated separately, year and main crop are significant at 95% and 99% levels while pest species, parasitoids species and their interaction effect is significant at 95% level.

Table 2: Analysis of type I for infestation of Stem borers

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Year	1	324.1679137	324.1679137	65.92	<.0001
Locbyveg	1	0.3015550	0.3015550	0.06	0.8048
Wh	1	7.9870644	7.9870644	1.62	0.2048
ch	1	20.3612532	20.3612532	4.14	0.0439
mcrop	1	118.8599839	118.8599839	24.17	<.0001
Cstage	2	0.1954644	0.0977322	0.02	0.9803
cropping	1	1.5098670	1.5098670	0.31	0.5805
Season	0	0.0000000	.	.	.
pessp	1	48.3311197	48.3311197	9.83	0.0021
parasp	1	40.6172554	40.6172554	8.26	0.0047
predsp	1	0.9023389	0.9023389	0.18	0.6691
nc	1	1.2451123	1.2451123	0.25	0.6157
pessp*parasp	1	30.4670930	30.4670930	6.20	0.0141
Year*Loc	1	2.1559167	2.1559167	0.44	0.5091
Wh*ch	1	7.3354775	7.3354775	1.49	0.2241

Table 3: Analysis of type II for infestation of Stem borers

Source	DF	Type II SS	Mean Square	F Value	Pr > F
Year	1	158.0070254	158.0070254	32.13	<.0001**
Loc	1	5.7966905	5.7966905	1.18	0.2796
Wh	1	1.9849301	1.9849301	0.40	0.5263
ch	1	14.2706339	14.2706339	2.90	0.0908
mcrop	1	71.6243079	71.6243079	14.56	0.0002**
Cstage	1	6.5314865	6.5314865	1.33	0.2512
cropping	1	1.6256561	1.6256561	0.33	0.5663
Season	0	0.0000000	.	.	.
pessp	1	33.4386575	33.4386575	6.80	0.0102*
parasp	1	32.8171783	32.8171783	6.67	0.0109*
predsp	1	0.6634919	0.6634919	0.13	0.7140
nc	1	0.0325166	0.0325166	0.01	0.9353
pessp*parasp	1	28.5555809	28.5555809	5.81	0.0173*
Year*Loc	1	2.9030508	2.9030508	0.59	0.4437
Wh*ch	1	7.3354775	7.3354775	1.49	0.2241

Sum of Residuals	0.0000000
Sum of Squared Residuals	649.1525570
Sum of Squared Residuals - Error SS	-0.0000000
First Order Autocorrelation	0.1563143
Durbin-Watson D	1.6728561

In the same fashion as above, it can be seen that the model fits the data as it is shown on Table 4 below at 99% and 95% levels and the diversity of stem borers is approximately explained 45% by its explanatory variables which is again very poor and the left 55% is due to the errors from the value of R-square. But, only crop host, main crop and parasitoids species are found to be significant at 95% confidence levels by inserting each variables one after the other given in Table 5 and none of those variables found out significant when the effects of each variables dealt separately. This can be interpreted as the extent to which the explanatory variables explain the diversity is only 45% the majority is due to the error variability which leads to wrong information and conclusions.

The diagnostics in section 4.2 for both untransformed and transformed response variables showed the problems of one or more assumptions for both response variables. So, it obliged us to explore for another model that fits the data by solving the problems of assumptions and to obtain appropriate results with ample information of predictors depending on the data type in hand.

Table 4: Tests of model adequacy for diversity of stem borers

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	5.38907583	0.35927172	1.88	0.0307
Error	132	25.24025617	0.19121406		
Corrected Total	147	30.62933200			

R-Square	Coeff Var	Root MSE	divlog Mean
0.175945	44.98810	0.437280	0.971991

Table 5: Analysis of Type I for diversity of stem borers

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Year	1	0.02378877	0.02378877	0.12	0.7249
Locbyveg	1	0.15208673	0.15208673	0.80	0.3741
Wh	1	0.00921254	0.00921254	0.05	0.8266
ch	1	1.94395069	1.94395069	10.17	0.0018**
mcrop	1	1.00381971	1.00381971	5.25	0.0235*
Cstage	2	0.10863520	0.05431760	0.28	0.7532
cropping	1	0.20612004	0.20612004	1.08	0.3011
Season	0	0.00000000	.	.	.
pessp	1	0.14329091	0.14329091	0.75	0.3882
parasp	1	1.14863593	1.14863593	6.01	0.0156*
predsp	1	0.04633936	0.04633936	0.24	0.6233
nc	1	0.21567386	0.21567386	1.13	0.2902
pessp*parasp	1	0.33524628	0.33524628	1.75	0.1878
Year*Loc	1	0.05221845	0.05221845	0.27	0.6021
Wh*ch	1	0.00005736	0.00005736	0.00	0.9862

Table 6: Analysis of Type II for diversity of stem borers

Source	DF	Type II SS	Mean Square	F Value	Pr > F
Year	1	0.10062234	0.10062234	0.53	0.4695
Locbyveg	1	0.21620460	0.21620460	1.13	0.2896
Wh	1	0.04717469	0.04717469	0.25	0.6202
ch	1	0.46685262	0.46685262	2.44	0.1206
mcrop	1	0.59994473	0.59994473	3.14	0.0788
Cstage	1	0.00003355	0.00003355	0.00	0.9895
cropping	1	0.32101212	0.32101212	1.68	0.1973
Season	0	0.00000000	.	.	.
pessp	1	0.01936571	0.01936571	0.10	0.7508
parasp	1	0.89397734	0.89397734	4.68	0.0324
predsp	1	0.05096957	0.05096957	0.27	0.6065
nc	1	0.12616611	0.12616611	0.66	0.4181
pessp*parasp	1	0.33040271	0.33040271	1.73	0.1910
Year*Loc	1	0.05211143	0.05211143	0.27	0.6025
Wh*ch	1	0.00005736	0.00005736	0.00	0.9862

Sum of Residuals	0.00000000
Sum of Squared Residuals	25.24025617
Sum of Squared Residuals - Error SS	-0.00000000
First Order Autocorrelation	0.01710740
Durbin-Watson D	1.95894568

The following section will be discussed based the results presented after general linear model is fitted.

4.2 Diagnostics

In this section, data was checked for linearity, normality, homogeneity of errors variance and associations between variables and its general behavior before formal analysis of the data. The following sections will demonstrate how different assumptions of regression are examined.

4.2.1 Normality of the Residuals

It is important to check the normality test for the residuals rather than the raw scores. There is not a general agreement of the best way to test normality. Most normality tests have small statistical power (probability of detecting non-normal data) unless the sample size is large.

It is usually assumed that errors have a normal distribution with mean zero and variance σ^2. In this case the normal probability plot of residuals in STATA program is utilized to check the normality of data on infestation and diversity of stem borers before and after transformation.

In a normal probability plot, the normal distribution is represented by a straight line angled at 45 degrees. The standard residuals are compared against the diagonal line to show the departure. If the residuals follow along the straight line, it means that the departure from normality is slight. In the plots below (figure 1 and figure 2), it is quite clear that the deviation is slight which some how assures normality for stem borers' infestation before and after transformation. But, in the case of the diversity of stem borers (Figure 3 and 4) before and after transformation is taken using different functions like square root and logarithmic function, the log transformation seems fix problem of non-normality. Hence, the problem of normality might not be serious for this data even though, it does not mean that the normality is sufficient to use the normal distribution.

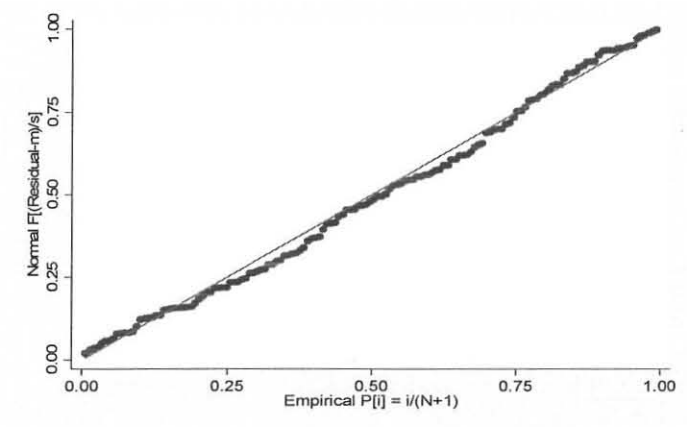

Figure 1. Normal probability plot of infestation of stem borers before transformation

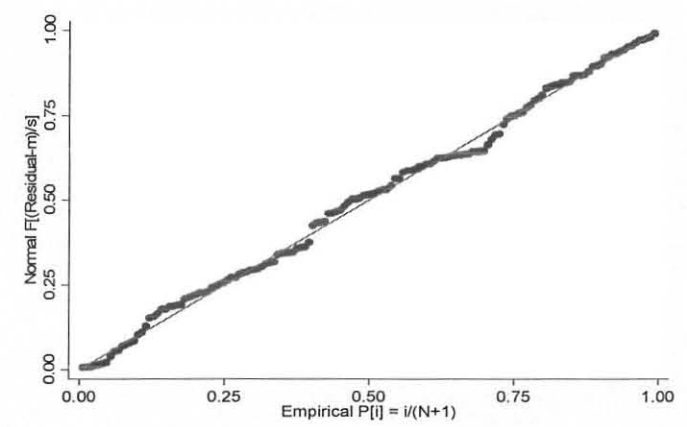

Figure 2. Normal probability plot of infestation of stem borers after square root transformation

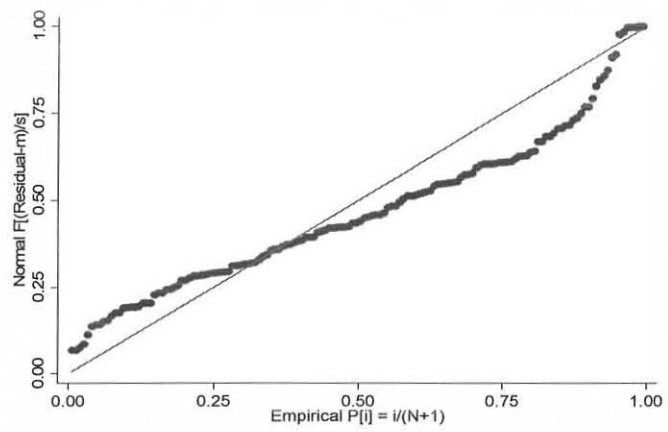

Figure 3. Normal probability plot of diversity of stem borers before transformation

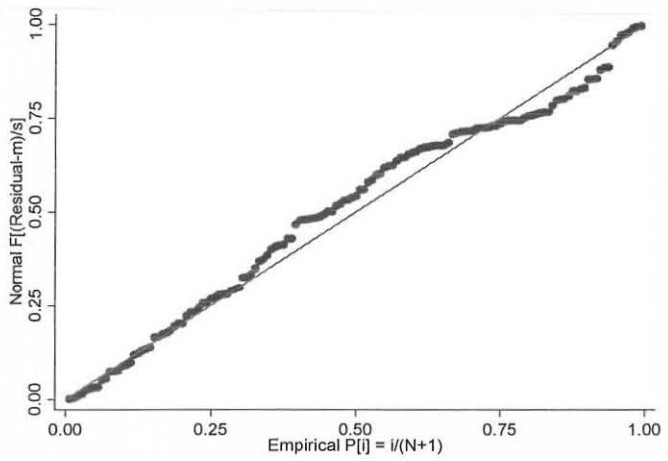

Figure 4. Normal probability plot of diversity of stem borers after log transformation

4.2.2 Linearity

To examine the assumption of linearity, one can apply a scatter plot matrix of STATA program in showing all predictors against the response variable in a pair wise manner. If the graph appears with some clusters within the subjects, a linear fit to all data points is not the best fit.

Although it is useful to plot the response variable against each predictor for the examination of linearity, these plots are inadequate because they only tell the partial

relationship between the response variable and each predictor, controlling for the other predictors. Therefore, it is desirable to use residual plots against the response variable. In the figures 5 and 6 below, there is some sort of clusters within the subjects before and after transformation in the diagnosis of infestation of stem borers. Moreover, linearity is also tested for the diversity of stem borers and figure 7 shows some clusters with in the subjects and clusters are also seen in figure 8 after transformation. So, we can not be quite sure for the linearity of the response variables which may in turn result in the violation of normality.

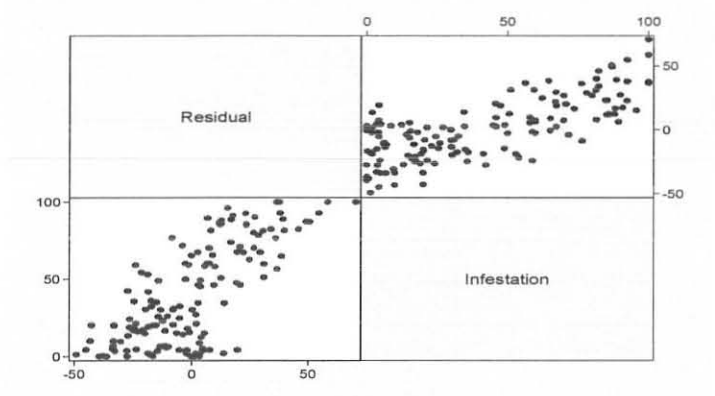

Figure 5. Scatter plot matrix of residuals vs. infestation of stem borers before transformation

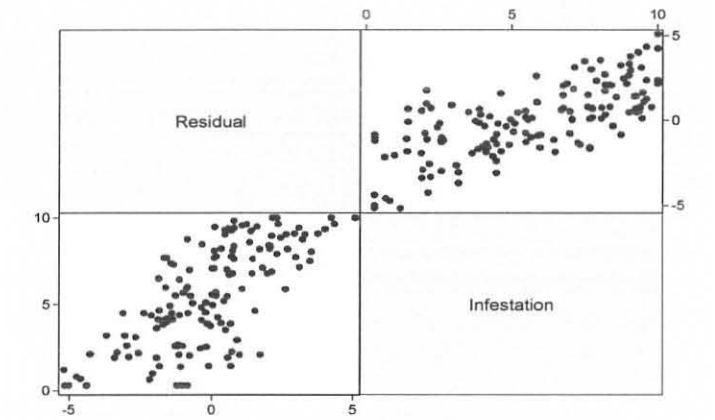

Figure 6. Scatter plot matrix of residuals vs. infestation of stem borers after square root transformation

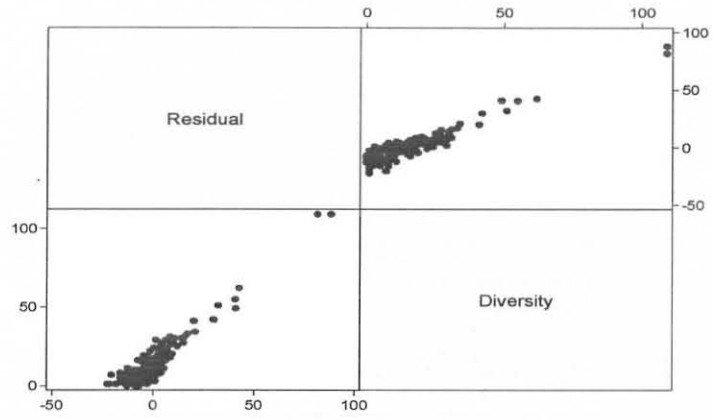

Figure 7. Scatter plot matrix of residuals vs. diversity of stem borers before transformation

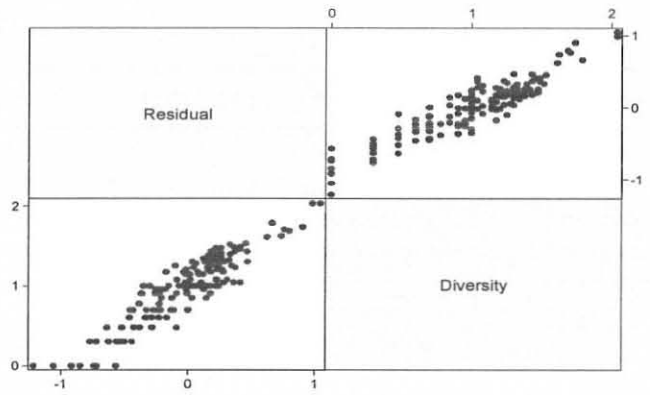

Figure 8. Scatter plot matrix of residuals vs. diversity of stem borers after log transformation

4.2.3 Test of homogeneity of Variance

The next step is to check for homogeneity of variance of residuals. This is done by plotting scatter plot of residual versus predictions of the response variables in the STATA program.

On the scatter plot the points form a "fun" shape about the line of best fit. In other words, the points are assumed to fall in a band of reasonably constant width on either side of the horizontal line. When the error term variance appears constant, the data are considered homoscedastic, otherwise, the data are said to be heteroscedastic. A scatter plot of residuals versus predicted values is produced using STATA program. Ideally, most of the residual autocorrelations should fall within the 95% confidence bands around zero, which are located at roughly plus-or-minus $2/\sqrt{n}$, where n is the sample size.

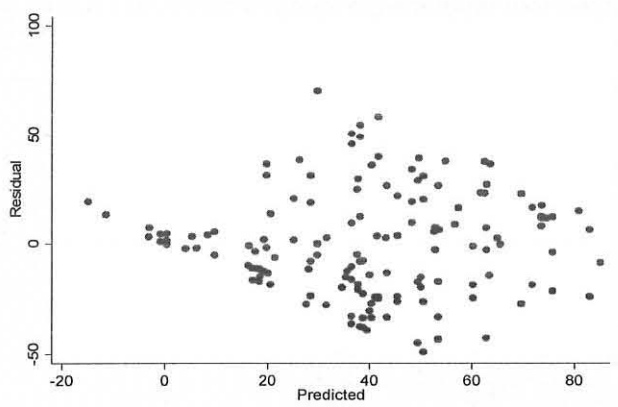

Figure 9. Scatter plot of infestation of stem borers before transformation

In Figures 9 and 11, it seems very difficult to examine the homoscedasticity in the given bands which is not constant and the plot is taking mega phonic pattern. Besides, the points are not equally distributed on both sides of the normal line. The pattern indicates that the variance of the error is not constant. So, transformations on the response variables are generally employed to stabilize variance for further analysis. After taking different transformations for infestation and diversity of stem borers,

square root and logarithmic function were found to be adequate respectively. Square root was chosen since it would adjust the data whose plot takes mega phonic pattern.

If the residuals variance is around zero, it implies that the assumption of homoscedasticity is not violated. There is a high concentration of residuals below zero in the figures 9 and 11 which confirms that the variance is not constant and thus a systematic error exists. After taking transformation, it can be seen that the variance of the residuals seems constant in the case of infestation of the stem with in the band ± 5 (Figure 10) and in the case of diversity in the bands of ±1.5 (Figure 12). Consequently, not only hetroscedasticity but also multicollinearity is not a major problem since it has been overcome by transformation.

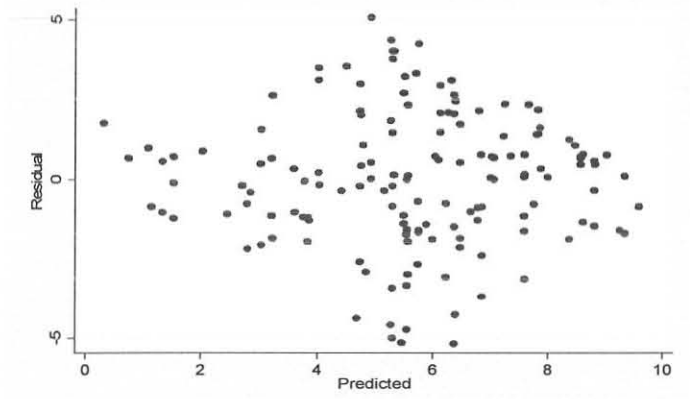

Figure 10. Scatter plot of stem borers' infestation after sqrt transformation

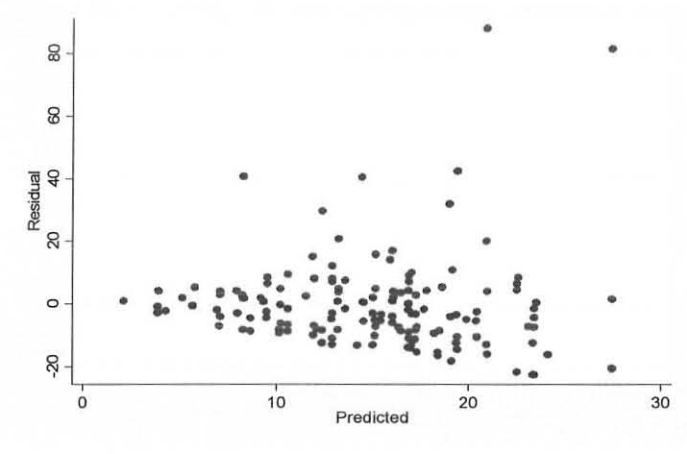

Figure 11. Scatter plot of diversity of stem borers before transformation

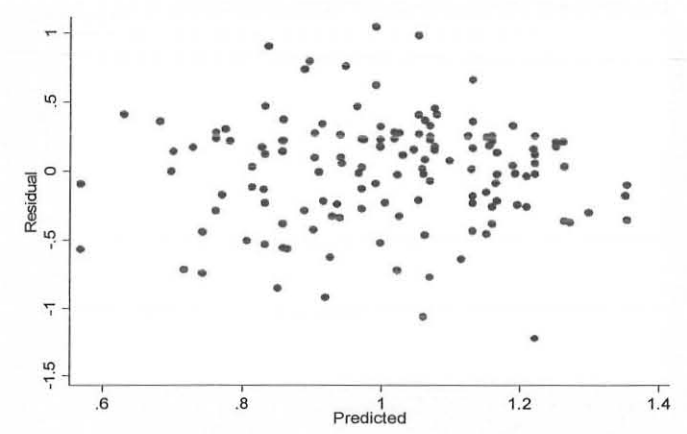

Figure 12. Scatter plot of diversity of stem borers after log transformation

4.2.4 Independence of Residuals

A general linear model requires independence of error terms. Again, a residuals plot can be used to check this assumption. Random, patternless residuals imply independent errors. So, it can be visualized that the plots in Figures 5 and 7 of residuals versus predicted values is random and patternless. Thus, the violation of independence of errors is unquestionable.

In general, most empirical data always arise in discrete form, but that in many cases it is at least possible to imagine a continuous distribution. Thus, the response variables; infestation and diversity of stem borers indicate the problems of linearity and homogeneity even after applying transformations (mostly square root and logarithmic functions). It has already been said that normality may fail to fulfill because of two reasons. First, the distributions of the dependent and/or independent variables may be themselves significantly non-normal and secondly, the violation of assumption of linearity. Moreover, the response variables may be discrete in behavior and in fact count in this case. So, the best model, non linear transformation and appropriate exponential family distributions should be utilized in order to generate the actual outputs. Hence, generalized linear model which can offset these problems was used in this study with Poisson distribution and logarithm link function which will be discussed in the following section.

4.3 Fitting Generalized Linear Model

Before applying generalized linear model analysis, data were subjected to the standard ANOVA analysis or general linear model analysis (section 4.1) and tested for the assumptions (section 4.2). The variables infestation and diversity of stem borers were not explained very well by the explanatory variables and some inadequacy exist in the assumptions even after the data is transformed. This might be because of the discrete behavior of the variables which are in fact cell counts in this study. So, this property of the data does not permit to use general linear model even though the test of the model seems adequate (Table 1 and 4).

In order to counterbalance such types of problems, the non linear transformation and appropriate distributions should be used to generate the generalized linear model results based on the behavior of the data. On applying generalized linear model, we benefits in two aspects: 1) it solves the problems of non normality and non linearity. 2) It helps us to use freely the exponential distributions other than normal distribution and non-linear link functions.

In this section, the results of the analysis from the generalized linear model, is presented. The software used is the SAS program. After exhaustive diagnostics and selection of models and link functions in the preceding sections, further analysis of the data was made based on the best fit model. The generalized linear model analysis is run to observe the effect of different factors or explanatory variables such as year, location classified by its vegetation type, wild and crop hosts, growth stage of crops, cropping systems, pest species, parasitoid species, predator species, nitrogen contents and some interaction effects related to infestation and diversity of stem borers. Generalized linear model was used since it enables us to use different exponential distributions and link functions that could solve the problem of violation of any one of the assumptions. It also allows us to use non-linear transformations and different exponential distributions and has become best fit model for each of the response variables to identify significant predictors.

The exponential family of Poisson distribution was the best for the analysis of infestation diversity stem borers with the link function of logarithm since the data of this study is discrete or counts. Moreover, the parameter estimates were presented with appropriate results in the following tables.

4.4 Distributions and link functions

In probability theory and statistics, the Poisson distribution is a two-parameter family of probability distributions used to analyze discrete or count data. For this specific analysis, Poisson distribution and logarithmic link function is selected as best method compared to the other exponential distributions such as gamma, normal etc. and link functions such as log, log-log and inverse power function among others. 148 observations of the dependent variables are related to 12 explanatory variables with different number of levels. The following tables 7 and 8 contains the model and class level information.

Table 7: Model Description

Data Set	WORK. stem borers
Distribution	Poisson
Link Function	Log
Dependent Variable	-infesta infesta -diversity diversity
Observations Used	148

Table 8: Class Level Description

Class	Levels	Values
Year	2	1999 2000
LocbyVeg	2	cropl otherveg
Wh	2	absent present
ch	2	wch woch
mcrop	2	maize sorg
Cstage	3	matu stuble veg
cropping	2	Mixed sole
Season	2	main off
pessp	2	Othersp keysp
parasp	2	absent present
predsp	2	absent present
nc	2	0 1

4.5 Analysis and Tests of Goodness of Fit

The main purpose of selecting generalized linear model is to see its relative importance with the general linear model. Therefore, this model is surely the best fit to the data under study. It can be tested by generating the smallest value of the deviance by trying different transformations and link functions. Thus, from the Tables 9 and 13, it can be concluded that the model was the best fit for the data. The test of Goodness of Fit in the Tables 9 and 13 contains statistics that summarize the fit of the specified model. These statistics are helpful in judging the adequacy of a model to fit to the data and in comparing it with other models under consideration. The value 17.6001 in the table 9 for the case of infestation of stem borers is much smaller when compared to its asymptotic chi-square with 132 degrees of freedom distribution which is approximately equal to 124.3 at 95% confidence, which is evidential to accept the null hypothesis and confirms that the specified model fits the data reasonably well. Similarly, the value 11.4639 in the Table 13 again is much smaller when compared to its asymptotic chi-square with 132 degrees of freedom distribution which is approximately equal to 124.3 at 95% confidence, which is evidential to accept the null hypothesis and confirms that the specified model fits the data as well.

The analysis of parameter estimates summarizes the result of iterative parameter estimation process. If in the process, a column of the model matrix corresponding to a parameter is found to be linearly dependent, or aliased, with columns corresponding to parameters preceding it in the model, PROC GENMOD assigns it zero degrees of freedom and prints a value for both the parameter estimate and its standard errors.

For Type 1 analysis in tables 11 and 15, each entry in the deviance column represents the deviance for the model containing the effect for that row and all effects preceding it in the table. For instance, the deviance corresponding to YEAR in the Table 11 is the deviance of the model containing an INTERCEPT and YEAR. As more terms are included in the model, the deviance decreases. Entries in the chi-square column are likelihood ratio-statistics for testing the significance of the effect added to the model containing all the preceding effects. For example, the chi-square value of 827.26 for the

YEAR in the Table 11 represents twice the difference in log likelihood between fitting a model with only an INTERCEPT term and a model with an INTERCEPT and YEAR. Since the scale parameter is set to 1 in this analysis, the value of the chi-square is equal to the difference in deviances.

The Type 3 analysis showed in tables 12 and 16 has resulted in the same conclusions as the Type 1 analysis. The Type 3 chi-square value for YEAR, for example, is twice the difference between the log likelihood for the model with INTERCEPT, YEAR, and all succeeding variables included and the log likelihood for the model with YEAR excluded. The hypothesis tested in this case is the significance of YEAR in the model with the others variables already included.

Table 9: Tests of Goodness of Fit for infestation of stem borers

Criterion	DF	Value	Value/DF
Deviance	132	2323.2097	17.6001
Scaled Deviance	132	2323.2097	17.6001
Pearson Chi-Square	132	2193.0567	16.6141
Scaled Pearson X2	132	2193.0567	16.6141
Log Likelihood		17223.7896	

Table 10: Analysis of Parameter Estimates of infestation of stem borers

Parameter	levels		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept			1	3.9501	0.1162	3.7223,4.1779	1155.07	<.0001
Year	1999		1	-0.8473	0.0459	-0.9372, -0.7574	340.91	<.0001
Year	2000		0	0.0000	0.0000	0.0000, 0.0000	.	.
LocbyVeg	cropl		1	0.1873	0.0356	0.1175, 0.2571	27.64	<.0001
LocbyVeg	other veg		0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh	absent		1	0.2007	0.0631	0.0771, 0.3243	10.12	0.0015
Wh	present		0	0.0000	0.0000	0.0000, 0.0000	.	.
ch	wch		1	0.7073	0.0840	0.5426, 0.8719	70.87	<.0001
ch	woch		0	0.0000	0.0000	0.0000, 0.0000	.	.
mcrop	maize		1	-0.6685	0.0406	-0.7481, -0.5888	270.43	<.0001
mcrop	sorg		0	0.0000	0.0000	0.0000, 0.0000	.	.
Cstage	matu		1	-0.4094	0.0563	-0.5198, -0.2991	52.88	<.0001
Cstage	stuble		1	-0.0893	0.0861	-0.2580, 0.0794	1.08	0.2996
Cstage	veg		0	0.0000	0.0000	0.0000, 0.0000	.	.
cropping	Mixed		1	0.0374	0.0297	-0.0209, 0.0957	1.58	0.2086
cropping	sole		0	0.0000	0.0000	0.0000, 0.0000	.	.
Season	main		0	0.0000	0.0000	0.0000, 0.0000	.	.
Season	off		0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp	Otherssp		1	0.0784	0.0583	-0.0358, 0.1927	1.81	0.1785
pessp	keysp		0	0.0000	0.0000	0.0000, 0.0000	.	.
parasp	absent		1	-0.2327	0.0325	-0.2964, -0.1691	51.35	<.0001
parasp	present		0	0.0000	0.0000	0.0000, 0.0000	.	.
predsp	absent		1	0.0984	0.0593	-0.0178, 0.2145	2.75	0.0970
predsp	present		0	0.0000	0.0000	0.0000, 0.0000	.	.

nc	0		1	-0.0158	0.0483	-0.1105, 0.0788	0.11	0.7429
nc	1		0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp*parasp	Others p	absent	1	-1.6671	0.1197	-1.9017, -1.4325	193.99	<.0001
pessp*parasp	Others p	present	0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp*parasp	keysp	absent	0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp*parasp	keysp	present	0	0.0000	0.0000	0.0000, 0.0000	.	.
Year*LocbyVeg	1999	cropl	1	-0.0991	0.0551	-0.2071, 0.0089	3.24	0.0720
Year*LocbyVeg	1999	other veg	0	0.0000	0.0000	0.0000, 0.0000	.	.
Year*LocbyVeg	2000	cropl	0	0.0000	0.0000	0.0000, 0.0000	.	.
Year*LocbyVeg	2000	other veg	0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh*ch	absent	wch	1	-0.4250	0.0775	-0.5769, -0.2731	30.06	<.0001
Wh*ch	absent	woch	0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh*ch	present	wch	0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh*ch	present	woch	0	0.0000	0.0000	0.0000, 0.0000	.	.
Scale			0	1.0000	0.0000	1.0000, 1.0000		

Table 11: LR Statistics for Type 1 Analysis of infestation of stem borers

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	4263.3182			
Year	3436.0586	1	827.26	<.0001**
LocbyVeg	3435.2564	1	0.80	0.3704
Wh	3413.8238	1	21.43	<.0001**
ch	3358.5411	1	55.28	<.0001**
mcrop	2892.2565	1	466.28	<.0001**
Cstage	2882.0454	2	10.21	0.0061**
cropping	2880.6424	1	1.40	0.2362
Season	2880.6424	0	0.00	.
pessp	2740.5730	1	140.07	<.0001**
parasp	2603.9308	1	136.64	<.0001**
predsp	2599.5602	1	4.37	0.0366*
nc	2596.3327	1	3.23	0.0724
pessp*parasp	2355.5599	1	240.77	<.0001**
Year*LocbyVeg	2353.6478	1	1.91	0.1667
Wh*ch	2323.2097	1	30.44	<.0001**

Table 12: LR Statistics for Type 3 Analysis of infestation of stem borers

Source	DF	Chi-Square	Pr > ChiSq
Year	1	470.08	<.0001**
LocbyVeg	1	23.14	<.0001**
Wh	1	0.10	0.7577
ch	1	62.72	<.0001**
mcrop	1	308.99	<.0001**
Cstage	1	51.75	<.0001**
cropping	1	1.58	0.2087
Season	0	0	0.00
pessp	1	195.06	<.0001**
parasp	1	337.89	<.0001**
predsp	1	2.81	0.0935
nc	1	0.11	0.7430
pessp*parasp	1	231.52	<.0001**
Year*LocbyVeg	1	3.24	0.0717
Wh*ch	1	30.44	<.0001**

Table 13: Tests of Goodness of Fit for diversity of stem borers

Criterion	DF	Value	Value/DF
Deviance	132	1513.2381	11.4639
Scaled Deviance	132	1513.2381	11.4639
Pearson Chi-Square	132	1823.0607	13.8111
Scaled Pearson X2	132	1823.0607	13.8111
Log Likelihood		3886.6630	

Table 14: Analysis of Parameter Estimates of diversity of stem borers

Parameter	levels		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept			1	2.2735	0.2349	1.8131, 2.7339	93.67	<.0001
Year	1999		1	-0.0270	0.0808	-0.1852, 0.1313	0.11	0.7385
Year	2000		0	0.0000	0.0000	0.0000, 0.0000	.	.
LocbyVeg	cropl		1	0.0925	0.0776	-0.0596, 0.2445	1.42	0.2333
LocbyVeg	other veg		0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh	absent		1	0.2328	0.1646	-0.0898, 0.5555	2.00	0.1572
Wh	present		0	0.0000	0.0000	0.0000, 0.0000	.	.
ch	wch		1	0.5116	0.1975	0.1246, 0.8986	6.71	0.0096
ch	woch		0	0.0000	0.0000	0.0000, 0.0000	.	.
mcrop	maize		1	-0.3444	0.0552	-0.4526, -0.2361	38.88	<.0001
mcrop	sorg		0	0.0000	0.0000	0.0000, 0.0000	.	.
Cstage	matu		1	-0.1100	0.0821	-0.2710, 0.0510	1.79	0.1805
Cstage	stuble		1	-0.7039	0.1735	-1.0440, -0.3637	16.45	<.0001
Cstage	veg		0	0.0000	0.0000	0.0000, 0.0000	.	.
cropping	Mixed		1	0.2477	0.0473	0.1550, 0.3404	27.43	<.0001
cropping	sole		0	0.0000	0.0000	0.0000, 0.0000	.	.
Season	main		0	0.0000	0.0000	0.0000, 0.0000	.	.
Season	off		0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp	Othersp		1	0.1838	0.0944	-0.0013, 0.3689	3.79	0.0516

pessp	keysp		0	0.0000	0.0000	0.0000, 0.0000	.	.
parasp	absent		1	-0.0199	0.0512	-0.1202, 0.0805	0.15	0.6978
parasp	present		0	0.0000	0.0000	0.0000, 0.0000	.	.
predsp	absent		1	-0.0918	0.1017	-0.2910, 0.1075	0.81	0.3667
predsp	present		0	0.0000	0.0000	0.0000, 0.0000	.	.
nc	0		1	0.2795	0.0774	0.1278, 0.4311	13.05	0.0003
nc	1		0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp*parasp	Othersp	absent	1	-0.5726	0.1292	-0.8259, -0.3194	19.64	<.0001
pessp*parasp	Othersp	present	0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp*parasp	keysp	absent	0	0.0000	0.0000	0.0000, 0.0000	.	.
pessp*parasp	keysp	present	0	0.0000	0.0000	0.0000, 0.0000	.	.
Year*LocbyVeg	1999	cropl	1	0.1750	0.0935	-0.0083, 0.3583	3.50	0.0613
Year*LocbyVeg	1999	other veg	0	0.0000	0.0000	0.0000, 0.0000	.	.
Year*LocbyVeg	2000	cropl	0	0.0000	0.0000	0.0000, 0.0000	.	.
Year*LocbyVeg	2000	other veg	0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh*ch	absent	wch	1	-0.1063	0.1884	-0.4755, 0.2628	0.32	0.5724
Wh*ch	absent	woch	0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh*ch	present	wch	0	0.0000	0.0000	0.0000, 0.0000	.	.
Wh*ch	present	woch	0	0.0000	0.0000	0.0000, 0.0000	.	.
Scale			0	1.0000	0.0000	1.0000,1.0000		

Table 15: LR Statistics for Type 1 Analysis of diversity of stem borers

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	1805.5641			
Year	1766.6187	1	38.95	<.0001**
LocbyVeg	1750.0841	1	16.53	<.0001**
Wh	1749.9914	1	0.09	0.7607
ch	1678.4127	1	71.58	<.0001**
mcrop	1623.6295	1	54.78	<.0001**
Cstage	1606.2338	2	17.40	0.0002**
cropping	1578.0704	1	28.16	<.0001**
Season	1578.0704	0	0.00	.
pessp	1565.1240	1	12.95	0.0003**
parasp	1557.8677	1	7.26	0.0071**
predsp	1557.5222	1	0.35	0.5566
nc	1536.4163	1	21.11	<.0001**
pessp*parasp	1517.3499	1	19.07	<.0001**
Year*LocbyVeg	1513.5585	1	3.79	0.0515
Wh*ch	1513.2381	1	0.32	0.5714

Table 16: LR Statistics for Type 3 Analysis of stem borers' diversity

Source	DF	Chi-Square	Pr > ChiSq
Year	1	0.68	0.4083
LocbyVeg	1	13.62	0.0002**
Wh	1	3.78	0.0518
ch	1	11.51	0.0007**
mcrop	1	40.97	<.0001**
Cstage	1	1.78	0.1824
cropping	1	27.58	<.0001**
Season	0	0.00	.
pessp	1	2.54	0.1110
parasp	1	18.62	<.0001**
predsp	1	0.80	0.3715
nc	1	13.33	0.0003**
pessp*parasp	1	19.15	<.0001**
Year*LocbyVeg	1	3.51	0.0610
Wh*ch	1	0.32	0.5714

4.6 Parameter Estimates and Test of Association

The Analysis of Parameter Estimates in Table 10 and Table 14 summarize the results of the iterative parameter estimation process by the likelihood ratio statistics (LR statistics). For each parameter in the model, PROC GENMOD prints columns with the parameter name; the degrees of freedom associated with the parameter, the estimated parameter value, the standard error of the parameter estimate, and a Wald chi-square statistic and associated p-value for testing the significance of the parameter to the model. If a column of the model matrix corresponding to a parameter is found to be linearly dependent, or aliased, with columns corresponding to parameters preceding it in the model, PROC GENMOD assigns it zero degrees of freedom and prints a value of zero for both the parameter estimate and its standard error.

The type 1 generalized linear model analysis of the infestation of stem borers on the explanatory variables is run using GENMOD procedure of SAS program. This revealed that the explanatory variables year, wild host, crop host, main crop, crops growth stage, pest species, parasitoids species, the interaction effects of pest and parasitoids species, the interaction effects of wild and crop hosts are significant predictors of stem borers' infestation at both 99% and 95% confidence levels while predators species is significant at 95% level when the variables are introduced forward in to the model (Table 11). On the other hand, the effects of the predictors such as location by vegetation type, cropping systems, nitrogen contents and the interaction effect of year and location by vegetation are insignificant. Season is linearly dependent variable with the rest of the predictors and hence its effect is aliased.

The type 3 generalized linear model analysis of the infestation of stem borers was dealt to see the independent effects of the predictor variables and hence, year, location by vegetation, crop host, main crop, crop growth stage, pest species, parasitoids species, the interaction effects of pest and parasitoids species, the interaction effects of wild and crop hosts are significant predictors of stem borers' infestation at both 99% and 95% confidence levels. But, the effects of the predictors such as wild host, cropping systems, predator species, nitrogen contents and the interaction effect of year and location by vegetation are insignificant where as season has aliased effect with the other variables. Of course, other explanatory variables such as temperature, altitude, rainfall, soil texture and etc that are not included in this study might have determinant effects.

Similarly, type 1 analysis in the table 15 provides that the diversity of stem borers is explained very well at 99% and 95% by the variables year, location by vegetation, crop hosts, main crop, crop growth stage, cropping system, pest and parasitoids species, nitrogen contents, and the interaction effects of pest and parasitoid species are significant predictors after the insertion of the variables one after the other. But, the effects of wild host, predator species and the interaction effects of year and location by

vegetation and wild and crop host are insignificant where as season has an aliased effect.

Type 3 generalized linear model analysis in the Table 16 also gives that location by vegetation, crop host, main crop, cropping system, parasitoids species, nitrogen contents and pest and parasitoids species interaction effect are found to be significant. But, the effects of year, wild host, crop growth stage, pest and predator species and the interaction effects of year and location by vegetation and wild and crop host are insignificant where as season has an aliased effect.

The estimate of the parameters given in the tables 10 and 14 represents the extents (negatively or positively) to which the weighted effects for each variables is measured.

Pair wise comparison was made between means of the explanatory variables using LSD and Duncan's multiple range tests at 95% confidence level. Consequently, there is significance difference in infestation of stem borers between the years. Accordingly, high infestation was observed in 2000 with mean 59.020 than the year 1999 with mean 28.084. There is no significant difference between the locations classified according to its vegetation type. However, slightly larger infestation was observed for the cropping lands (41.233) than other vegetation types (40.203) such as dry land, grass land, savanna land, and shrub land. There is also no significant difference in infestation of stem borers when the farmers' land is with crop hosts (38.814) or with out crop host (49.465). Significant difference in infestation was observed between the two main crops. As a result, high infestation was measured for sorghum (47.812) than maize (19.775). There are significant differences in infestation for the crop growth stages, between the key species (Chilo partellus and Busseola fusca) and other pest species (Sesamia clamistis,new species and Sesamia nonagrides botanephaga) and between the absence or presence of parasitoid species. Larger infestation was measured in the stubble stage (55.00) followed by maturity stage (43.463) than the vegetative stage (29.683). The key species resulted in high infestation (44.867) than other pest species (18.639) and more infestation was observed in the presence (49.316) parasitoids species than in its absence (30.649).

The results of mean separation using LSD and Duncan's multiple ratio tests at 95% level of confidence revealed that the diversity of stem borers does not differ significantly across locations based on their vegetation type except that slightly more stem borers was measured from cropping land (16) than location of other vegetation types (13). There are no significant differences of density of stem borers was measured for the main crops, cropping systems, parasitoid species and nitrogen contents. But slightly more stem borers were measured for sorghum (16) than maize (12), more for mixed cropping systems (16) than sole cropping system (14), more in the presence (16) of parasitoids than in their absence (14) and more in the ranges of 0-0.2 (18) than for the ranges greater than 0.2 (14) of nitrogen contents.

Chapter five

Conclusions and Recommendations

5.1 Conclusions

The study was made on the stem borers' diversity, infestation and status of their parasitoids interaction on farmers' field in the eastern Ethiopia on maize and sorghum cereal crops during the consecutive years in 1999 and 2000. The purpose of this study was basically concerned with the selection of statistical models which appropriately helps to identify the true predictors of the infestation and diversity of stem borers and their parasitoids status among all other predictors or factors when one or more assumptions fail to satisfy. The data is subjected to general linear model before and after transformations by logarithmic and square root and assumptions were also checked (section 4.2). For this study, the generalized linear model (PROC GENMOD) analysis has become the best model to be applied for the analysis. Another hard work was selection of appropriate distribution and link functions. Obviously, the data violates some assumptions like linearity, homogeneity and hence normal distribution should not be applied. Moreover, the behavior of the data for this study is discrete. Therefore, there are many other exponential distributions and Poisson is one of them. Poisson distribution is used because the data is discrete and counts in its behavior. Different non-linear link functions were also tried and logarithm function was used for our purposes since it minimizes the values of deviances that assure the adequacy of the model. Consequently, the generalized linear model analysis procedure in SAS statistical program was found to be best together with Poisson distribution and logarithm link function to identify the predictors of the response variables.

The results indicated that the infestation of stem borers do not differ significantly from vegetation to vegetation types across the locations. How ever, districts with crop land vegetation type have slightly more infestation than that with other vegetation types. In addition, the infestation is not significantly different with or with out the crop hosts around the farms. But, very significant difference of infestation of stem borers is observed from year to year, for the main crops sorghum and maize, for the pest species, for the crop growth stages and in the presence or absence of parasitoid species. Consequently, more infestation was measure in the year 2000 than 1999, for sorghum farm than maize, for stubble stage followed by maturity stage than vegetative stage. More over, the key pest species resulted in more infestation than the other species and also infestation was high in the presence of the parasitoid species.

Similarly, diversity of stem borers was not significantly different from vegetation to vegetation across the locations, between the main crops and cropping systems, in the presence and absence of parasitoid species and for the nitrogen contents of the soil. One important finding was that season has no association with infestation and diversity of stem borers in these districts. The crops or plants in the boundary of the main crops also share the effects of stem borers and the interaction of the pest species and the parasitoid species is significant. The extent to which the parasitoids affect the stem borers differ based on the parasitoids species.

5.2 Recommendations

- The Researchers should use generalized linear model than general linear model since the former enables to use different exponential family distribution depending on the behaviour of data to be used and even for data violating basic assumptions for further study.
- Instead of the normal distributions, the exponential family distribution, Poisson distribution should be used for discrete data or count data.
- Appropriate non link functions should be used to offset the problems of non linearity
- Due attention should be given to the key pest species, stubble and maturity stages of the crops and the sorghum cereal crop than maize, identified in the analysis during the stem borers' management.
- The parasitoids are found to be good enemy of stem borers and thus attention should be given for conserving them particularly when pesticides are used.

References

Agresti A. (1996). An Introduction to Categorical Data Analysis.

Abiy Tilahun (2005). Parasitoids species diversity and rates of parasitism on maize and sorghum stem borers in central Rift valley of Ethiopia. M.Sc. thesis, Haromaya University).

Anderson, D.A. and Aitkin, M. (1985). A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55: 218-234.

Assefa Gebreamlak. (1985). Survey of lepidopterous stem borers attacking maize and sorghum in Ethiopia. Ethiopia journal of Agricultural Science 4:55-59.

Assefa, G/ Amlak. 1988. Ecology and management of maize stalk borer, Busseola fusca (Fuller) (Lepidoptera: Noctuidae), in Southern Ethiopia. Ph. D. dissertation. Swedish University of Agricultural Sciences, Uppsala, Sweden.

Bailer, A.J. and Oris, J.T. (1996). Estimating Inhibition Concentrations for Different Response Scales Using Generalized Linear Models. Environmental toxicology and chemistry 16 (7): 1554–1559.

Benti,T. and Ransom,J.K (eds).1993.Proceeding of the first national maize workshop of Ethiopia.5-7, May, 1992, IAR/CIMMYT, Addis Ababa, Ethiopia.73 pp.

Berger, A. 1992. Larval movements of Chilo partellus (Lepidoptera: Pyralidae) within and between plants: timing, diversity responses and survival. Bulletin of Entmological Research 82: 441-448.

Bon hof, M.J .2000. The impact of predators on maize stem borers in coastal Kenya. Ph.D thesis Wageningen Agricultural University, the Netherlands, 181 PP.

Breslow, N. (1984) Extra-Poisson variation in log-linear models. Appl. Statist., 33, 38-44. Francis B.and Whittaker J., eds, Generalized Linear Models.

Chen, et al (2000). Sequential selection procedures. Using sample means to improve efficiency.

Crowder, B.1995. Accuracy of efficiency of estimating approach.Biometrika 82,407-410.

Crowder (1985) and Tsutakawa (1988). Extensions of generalized linear models to include random effects on log-linear models .

CSA (Central Statistical Authority). 2000. Agricultural Sampling Survey 1999/2000.Report on area and production for major crops (private peasant holdings, Meher season), Statistical Bulletin No.171.CSA, Addis Ababa, Ethiopia. 73 pp.

Culham ,J. Advantages of Generalized Linear Model. http:/www. Dummies website (adopted Nov, 21/2006).

Dobson (1990). Generalized Linear Models (GLZ).

Emana, G. and Tsedeke, A.1999. Management of maize stem borers using sowing date at Arsi Negele. Pest Management journal of Ethiopia.3:47-52.

Emana,G.2001. Ecological analysis of stem borers and their natural enemies under maize and sorghum based agro-ecosystems in Ethiopia. Ph.D.thesis.Pp 123-128,Kenyatta University.

Emana, G., Overholt,W.A and Kairu,E.2001.Distribution and species composition of stem borers and their natural enemies under maize and sorghum based agro-ecosystems in Ethiopia. Insect Science and Its Application 21. Pp.353-359.

Emana, G., Overholt,W.A and Kairu,E.2002. Predicting the distribution of C.artellus (Swinhoe) and Cotesia flavipes Cameron in Ethiopia using Geographic Information System and stepwise regressions. Insect Science and Its Application 22:523-539.

Emana, G., Overholt,W.A and Kairu,E.2003.Evidence of the establishment of Cotesia flavipes (Hymenoptera: Braconidae), a parasitoid of cereal stem borers, and its host range expansion in Ethiopia.Bulletin of Entomological Research 93:125-129.

Gilmour, A. R., Anderson, R. D. & RAE, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. Biometrika 72:593-9.

Haberman, S. and Renshaw, A. E. (1988). Generalized Linear Models in actuarial work. Presented to the Staple Inn Actuarial Society on 2nd February.

Harris, K. M. and Nwanze, K.F. 1992. Busseola fusca (Fuller), the African maize stalk borers: A handbook of information. Information Bulletin No. 33. ICRISAT, Patancheru, A.P. India and CAB International, Wallingford, UK. 92 pp.

Harville and Mee (1984). Extensions of generalized linear models to include random effects.

Horton, N.J. and Laird, N.M. (1998). Maximum Likelihood Analysis of Generalized Linear Models with Missing Covariates. Statistical Methods in Medical Research 8: 37 - 50.

Hosmer,W. and Lemshow,S. (2000). Applied Logistic regression(2nd Ed.). New York, John Wiley and Sons, Inc., New York.

Ingram, W.R.1958.The Lepidopterous stalk borers associated with graminae in Uganda.Bull.Entomol.Res. 49 :367-383.

Jørgensen, B. (). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models.

Kerr D.R. and Meador J. P. (1995). Modeling Dose Response using Generalized Linear Models.

Khan Z. R , Pickett J. A. (2000). Exploiting chemical ecology and species diversity: stem borer and striga control for maize and sorghum in Africa.

Khuri A.I., Mukherjee B., Sinha B.K. and Ghosh M. (2006). Design Issues for Generalized Linear Models 21(3): 376–399.

Lamouroux N. and Jowett I. G. (2005). Generalized in stream habitat models.

Leonard T. K. (1975). The Epidemiology of Epilepsy in Rochester, Minnesota, 1935 through 1967 Epilepsia 16 (1): 1–66.

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

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. Journal of the Royal Statistical Society 34: 1-41.

McCullagh, P. & Nelder, J. A. (1989). Generalized linear models, 2nd edn. London: Chapman & Hall.

Mohyuddin A.I. (1971). Comparative biology and ecology of Apanteles flavipes (Cam) and A. Sesamia Cam. as parasites of graminaceous borers. Bulletin of Entomological Research 61: 33-39.

Mueller, P. and Parmigiani, G. (1995). ``Optimal design via curve fitting of ``Monte Carlo integration in general dynamic models.

Nelder and Wed-deburn (1972). In Generalized Linear Model theory.

Ochi and Prentice (1984). Maximum Likelihood Variance Components Estimation for Binary Data.

Overholt, W. A. A.J. Ngi-Song, C.O. Omwega. S.W. Kimani- Njougu, J Mbapila, M.N. Sallam and V. Ofomata. (1997). A review of the introduction and establishment of Cotesia flavipes Cameron in east Africa for biological control of cereal stem borers. Insect Science and its Application. 17: 79-88.

Pats, P. 1992. Reproductive biology of the cereal stem borer Chilo partellus. Ph.D. thesis submitted to Swedish University of Agricultural Sciences. Department of Plant and Forest Protection. Uppsala, 97 PP.

Perry J.N., Noh M. S., Lee Y.(2000). Fitting host-parasitoid models with CV241 using hierarchical generalized linear models.

Potting, R. P. J. 1997. Evolutionary and applied aspects of the behaverioal ecology of the stem borer parasitoid Cotesia flavipes. Insect Science and its application. 17: 109-118.

Seshu Reddy, K.V. 1983. Sorghum stem borers in eastern Africa. Insect Sciences and Its Application 4: 33-39.

Smith.S; Widenmann, R.N. and Overholt, W.A. (1993). Parasites of Lepidopteran Stem borers of tropical Gramminaceous plant. ICIPE Science press, Nairobi 89 PP.

Stiralli (1984). Random effects models for serial observation with binary responses.

Sutradhar B.C. and K Das. On the efficiency of regression estimators in generalised linear models for longitudinal data.

Tsutakawa R. K. and Johnson J.C (1988). The effect of uncertainty of item parameter estimation on ability estimates. Department of Statistics, University of

Missouri, 222 Math Sciences, 65211 Columbia, MO.

Van den Berg, J., Van Rensburg, G.D.J. and Van der Westhuizen, M.C. (1997). Economic threshold levels for Chilo partellus (Lepidoptera: Pyralidae). Control on resistant and susceptible sorghum plants. Bulletin of Entomological Research 87: 89-93.

Unnithan, G.C and Saxena, K.N. (1990). Population Monitoring of Chilo partellus (Swinhoe) (Lepidoptera: Pyralidae) using pheromone traps. Insect Science and its Application 11: 795- 805.

Williams, D. A. (1982).Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika 61: 439-47.

Appendices

Appendix A. SAS code and some out puts of Infestation of Stem borers

A1. General Linear Model (GLM) analysis for untransformed data

```
Data Stem borers;

Set work. Stem borers;

Proc GLM data= Stem borers;

Title 'General Linear model analysis';

Class Year LocbyVeg Wh ch mcrop Cstagecropping Season
      pessp parasp predsp nc ;

Model infesta=Year LocbyVeg Wh ch mcrop Cstagecropping Season
      pessp parasp predsp nc pessp*parasp
      Year*LocbyVeg/ wh*ch/p ss1 ss2;

Run;
```

Table A1: Tests of model adequacy for infestation of stem borers

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	68074.8056	4538.3204	7.13	<.0001
Error	132	84055.0083	636.7804		
Corrected Total	147	152129.8139			

R-Square	Coeff Var	Root MSE	infesta Mean
0.447478	62.13349	25.23451	40.61338

Table A2. Analysis for type I of infestation of Stem borers

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Year	1	34424.67145	34424.67145	54.06	<.0001**
Locbyveg	1	32.64356	32.64356	0.05	0.8212
Wh	1	1262.01305	1262.01305	1.98	0.1615
ch	1	3201.29714	3201.29714	5.03	0.0266**
mcrop	1	15454.65526	15454.65526	24.27	<.0001**
Cstage	2	374.17926	187.08963	0.29	0.7459
cropping	1	97.41816	97.41816	0.15	0.6963
Season	0	0	0.00000	.	.
pessp	1	3905.02358	3905.02358	6.13	0.0145**
parasp	1	5119.52523	5119.52523	8.04	0.0053**
predsp	1	18.41686	18.41686	0.03	0.8652
nc	1	55.27960	55.27960	0.09	0.7687
pessp*parasp	1	2108.04124	2108.04124	3.31	0.0711
Year*Loc	1	383.17719	383.17719	0.60	0.4393
Wh*ch	1	1638.46406	1638.46406	2.57	0.1111

Table A3: Analysis for type II of infestation of Stem borers

Source	DF	Type II SS	Mean Square	F Value	Pr > F
Year	1	20882.13448	20882.13448	32.79	<.0001**
Loc	1	658.70084	658.70084	1.03	0.3110
Wh	1	263.47160	263.47160	0.41	0.5212
ch	1	2710.12495	2710.12495	4.26	0.0411**
mcrop	1	9957.60039	9957.60039	15.64	0.0001**
Cstage	1	1536.85163	1536.85163	2.41	0.1227
cropping	1	132.81499	132.81499	0.21	0.6486
Season	0	0	0.00000	.	.
pessp	1	2600.27720	2600.27720	4.08	0.0453**
parasp	1	4245.79308	4245.79308	6.67	0.0109**
predsp	1	26.43471	26.43471	0.04	0.8389
nc	1	0.03548	0.03548	0.00	.
pessp*parasp	1	1881.08334	1881.08334	2.95	0.0880
Year*Loc	1	533.81355	533.81355	0.84	0.3616
Wh*ch	1	1638.46406	1638.46406	2.57	0.1111

Sum of Residuals	0.00000
Sum of Squared Residuals	84055.00828
Sum of Squared Residuals - Error SS	0.00000
First Order Autocorrelation	0.09020
Durbin-Watson D	1.79944

A2. General Linear Model (GLM) analysis for transformed data

```
Data Stem borers;

Set work. Stem borers;

Proc GLM data= Stem borers;

Title 'General Linear model analysis';

Title 'General Linear model analysis';

Class Year Locbyveg Wh ch mcrop Cstagecropping Season
      pessp parasp predsp nc ;

Model infestasqrt=Year Locbyveg Wh ch mcrop Cstagecropping
      Season pessp parasp predsp nc pessp*parasp
      Year*Locbyveg wh*ch/p ss1 ss2;

Run;
```

A3. Generalized Linear Model (GENMOD) analysis with dist= Poisson and link= log

```
Proc data Stem borers;

Set work. Stem borers;

Proc GENMOD data= Stem borers descending;

Title 'Generalized Linear model analysis';

Class Year Locbyveg Wh ch mcrop Cstage cropping Season
      pessp parasp predsp nc ;

Model infesta=Year Locbyveg Wh ch mcrop Cstage cropping
      Season pessp parasp predsp nc pessp*parasp Year*Loc
wh*ch/dist=poisson link=log type1 type3;

Run;
```

Appendix B. SAS code and some out puts of Diversity of Stem borers

B1. General Linear Model (GLM) analysis for untransformed data

```
Data Stem borers (untransformed);

Set work. Stem borers;

Proc GLM data= Stem borers descending;

Title 'General Linear model analysis';

Class Year LocbyVeg Wh ch mcrop Cstagecropping Season
      pessp parasp predsp nc ;

Model diversity=Year LocbyVeg Wh ch mcrop Cstagecropping
      Season pessp parasp predsp nc pessp*parasp
    Year*LocbyVeg wh*ch/p ss1 ss2;

Run;
```

Table B1: Tests of model adequacy for diversity of stem borers

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	4230.38472	282.02565	1.14	0.3293
Error	132	32714.17609	247.83467		
Corrected Total	147	36944.56081			

R-Square	Coeff Var	Root MSE	diversity Mean
0.114506	105.8578	15.74277	14.87162

Table B2: Analysis of Type I for diversity of stem borers

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Year	1	567.5721744	567.5721744	2.29	0.1326
Locbyveg	1	248.5032735	248.5032735	1.00	0.3185
Wh	1	1.0221191	1.0221191	0.00	0.9489
ch	1	848.6705689	848.6705689	3.42	0.0665
mcrop	1	810.5321943	810.5321943	3.27	0.0728
Cstage	2	281.3948798	140.6974399	0.57	0.5682
cropping	1	450.8691202	450.8691202	1.82	0.1797
Season	0	0	0.0000000	.	.
pessp	1	210.5625070	210.5625070	0.85	0.3583
parasp	1	93.5542063	93.5542063	0.38	0.5400
predsp	1	6.2055076	1 6.2055076	0.03	0.8745
nc	1	342.1971352	342.1971352	1.38	0.2421
pessp*parasp	1	303.5701190	303.5701190	1.22	0.2704
Year*Loc	1	65.7308623	65.7308623	0.27	0.6074
Wh*ch	1	0.0000520	0.0000520	0.00	0.9996

Table B3: Analysis of Type II for diversity of stem borers

Source	DF	Type II SS	Mean Square	F Value	Pr > F
Year	1	6.1349267	6.1349267	0.02	0.8752
Locbyveg	1	279.2655972	279.2655972	1.13	0.2904
Wh	1	49.1198941	49.1198941	0.20	0.6569
ch	1	121.9615885	121.9615885	0.49	0.4842
mcrop	1	595.7664182	595.7664182	2.40	0.1234
Cstage	1	50.4896646	50.4896646	0.20	0.6525
cropping	1	446.9839779	446.9839779	1.80	0.1816
Season	0	0	0.0000000	.	.
pessp	1	96.2134123	96.2134123	0.39	0.5343
parasp	1	29.2845514	29.2845514	0.12	0.7316
predsp	1	11.6583047	11.6583047	0.05	0.8286
nc	1	238.9239331	238.9239331	0.96	0.3280
pessp*parasp	1	306.1351588	306.1351588	1.24	0.2684
Year*Loc	1	65.1984242	65.1984242	0.26	0.6089
Wh*ch	1	0.0000520	0.0000520	0.00	0.9996

Sum of Residuals	-0.00000
Sum of Squared Residuals	32714.17609
Sum of Squared Residuals - Error SS	-0.00000
First Order Autocorrelation	0.00772
Durbin-Watson D	1.98088

B2. General Linear Model (GLM) analysis for transformed data by logarithm

```
Data Stem borers;

Set work. Stem borers;

Proc GLM data= Stem borers descending;

Title 'General Linear model analysis';

Class Year  Locbyveg  Wh  ch  mcrop  Cstage cropping  Season
      pessp  parasp  predsp  nc  ;

Model divlog=Year Locbyveg  Wh  ch  mcrop Cstage  cropping
      Season  pessp  parasp  predsp  nc  pessp*parasp
   Year*Locbyveg  wh*ch/p ss1 ss2;

Run;
```

B3. Generalized Linear Model (GENMOD) analysis with dist=poisson and link=log

```
Data Stem borers;

Set work. Stem borers;

Proc GENMOD data= Stem borers descending;

Title 'Generalized Linear model analysis';

Title 'Generalized Linear model analysis';
Class Year LocbyVeg Wh ch mcrop Cstage cropping Season
      pessp parasp predsp nc;
Model diversity=Year LocbyVeg Wh ch mcrop Cstage cropping
      Season pessp parasp predsp nc pessp*parasp
    Year*LocbyVeg wh*ch/dist= poisson link=log type1 type3;
Run;
```