ADDIS ABEBA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
MASTER'S THESIS

STUDIES ON :

Analysis of data from crop protection experiments using Generalized Linear Model: the Case of parthenium

A Thesis Submitted
To

The school of Gradute Studies
Addis Abeba University

In Partial fulfillment of the requirements
for the Degree of Master of
science in Statistics

By
BEGIZEW YAREGAL

JULY,2008
ADDIS ABEBA

ADDIS ABEBA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
MASTER'S THESIS PROSAL

STUDIES ON :

Analysis of data from crop protection experiments using Generalized Linear Model: the Case of parthenium

By:

BEGIZEW YAREGAL
Department of Statistics
Addis Ababa University

Approved by the Board of Examiners:

Name

SELSHI FANTA(Ms.c.)	----------
Chairman, Dept.Graduate Committee	Signature
GIRMA TAYE, (Ph.D)	----------
Research Advisor	Signature
.....( )	----------
Internal Examiner	Signature
…….. ( )	----------
External Examiner	Signature

List of Tables

page

Table1: Link functions----------18

Table2: Overall Dominance power of Parthenium and other species in 100 plot areas----------46

Table 3: Level of parthenium infestation----------50

Table 4: Standard ANOVA table----------51

Table 5-SAS Logistic Procedure output summary----------56

Table 6- SAS probit Procedure output summary----------59

Table7- SAS Poisson Procedure output summary.----------61

Table8- Criteria For Assessing Goodness Of Fit for Poisson regression (SAS output)----------62

Table9- SAS Poisson Procedure output summary(APPENDIX I)----------63

Table10- Criteria For Assessing Goodness Of Fit for poisson regression(SAS output) ----------63

Table11- Criteria For Assessing Goodness Of Fit for Negative Binomial regression(SAS output)--64

Table12-Summary of SAS output for the parameter estimates and confidence interval.----------65

Table13- Quadrant to quadrant Contrast Estimate Results----------66

Table 14-Parameter estimates for Multinomial logit model----------67

Table 15- A Summary of Goodness of fit criteria for each Model----------69

List of Figures

Figure1: The logistic function with z on the horizontal axis and f(z) on the vertical axis----------21

Figure2: The cumulative normal probability distribution.----------25

Figure 3: The data collection layout----------44

Figure 4: A simple Barchart of Number of all species per quadrant ----------45

Figure 5: Level of Parthenium abundance----------47

Figure 6: Graph of parthinium data(before transformation)----------48

Figure 7: Bar gaph for a square root transformed Parthenium data----------49

Figure 8: Boxplot of Number of partienium plant per plot----------53

Figure 9: Histogram for Log transformed Data----------54

Figure 10:Histogram for SQRT transformed Data----------55

Table of content

page

Acknowledgment----------i

Acronym----------ii

Absract----------iii

1. Chapter one –INTRODUCTION----------1

1.1. Background ----------1

1.1.1 Naming of Parthenium Weed in Ethiopia----------3

1.1.2 Morphology of Parthenium Weed in Ethiopia ----------3

1.1.3. Infestation of Parthenium weed ----------4

1.2 Introduction to Methods ----------6

1.3 Statement of the Problems ----------6

1.4 Objective ----------7

1.5 Organization of the thesis----------7

2 Chapter Two-Litrature Review----------8

3 Chapter Three- Materials and Methods----------13

3.1 Data and source----------13

3.2 Generalized Linear Models (GLM) ----------13

3.2.1. Distribution of the dependent variable----------17

3.2.2 Link function. ----------17

3.3 Processes in Model Fitting----------19

3.4.Some principles to Guide the modeler----------19

3.5.Common Generalized Linear Models----------20

3.5.1. Logistic regression----------20

3.5.2. Probit model----------24

3.5.3. Multinomial model----------28

3.5.4. Poisson regression----------29

3.5.5.Negative Binomial Regression----------32

3.6. Parameter Estimation in the generalized linear model----------33

3.7. Diagnostics in the generalized linear model. ----------36

3.8. Methods of Analyses ----------36

3.9. Data Transformation----------38

3.10. Software----------40

3.10.1 Estimation in SAS and SPSS----------40

4 Chapter Four- Results and Discussion ---------- 44

4.1. Data type and source---------- 44

4.2 Descriptive Statistics---------- 45

4.3. Analysis of Variance----------50

4.4 Diagnostics ----------52

4.5. Model Fitting----------56

4.6 Model checking and Selection----------67

5. Chapter Five- Conclusion and Recommendations ---------- 69

6. References----------71

7. Annexes----------74

Acknowledgment

I am very grateful to **Dr. Girma Taye**, my thesis advisor and instructor for his invaluable comments and suggestions that contributed to the successful realization of the study.

I am highly indebted to the Department of Statistics for the provision of computer facilities during office hours as well as weekends.

My deep appreciation and thanks also goes to Dr. Taye Tessema, from Ethiopian Institute of Agricultural Research, in providing the data collected and other related materials to the study. I would like to express my gratitude to the Ethiopian Institute of Agricultural Research Library staff members for their cooperation and provision in refering materials.

I am very much indebted to my family members, for helping me in giving their unreserved and constructive ideas. I am particularly grateful to my friend Zewdu Ayalew who informed me to join masters program from the very beginning. I am also indebted to my very good friend, Zelalem Bekele, who provided me internet accesses especially at the beginning of this study. I want to thank my batch colleagues as a whole especially those who wishes and pushes me to complete this study.

Finally, my deep appreciation extended to all instructors (all dedicated) in giving us their lectures quiet amazingly in this special batch, especially Dr. Bute Gutu, for his patience to give us two different course lectures at the beginning of the program.

Acronym

GLMs	Generalized linear models
AIC	Akaike Information Criterion
ASE	Asymptotic standard error
ML	Maximum Likelihood
Cdf	Cumulative density function
LR	Likelihood Ratio
NegBino	Negative Binomial
EIAR	Ethiopian Institute of Agricultural Research
RGCC	Relative gradient convergence criterion
SC	Schwarz Criterion
DF	Degrees of freedom
DGM	Data generating mechanism
CDVM	Categorical dependent variable model

Abstract

Among many weeds that cause crop loss, parthenium was found to be the most terrible one according to some exploratory studies. The problem of parthenium is not only that it cause very sever crop loss, but also it cause health problems to human and animal beings. Control of Parthenium by Farmers ,cultural and labour intensive, caused farmers to suffer from skin allergy, itching, fever, and asthma.

This study tried to popularize different Generalized linear models for modeling agricultural data which is used for describing the data sufficiently well and then identify the natural relationship between different variables for further analayis as well as applications. Generalized linear models (GLMs) are used to do regression modeling for non-normal data with a minimum of extra complication compared with normal linear regression. One of the available programs that is important in current statistical practice is the GLM procedure in the SAS software package**.**

The study is based on the result of a parthenium and other species count data, secondary data,obtained from Ethiopian instisute of Agricultre research. Descriptive statistics supported by graphical presentations have been discussed to show the dominance of parthenium on other species per plot area. Furthermore, to evaluate the probability of a plot or a quadrant to be free of parthenium, models form GLM family are applied to the data using SAS software.

Based on the parameter estimates, fitted models were formulated, parameters are interpreted and comparison of fitted models conducted. In this model fitting process, an attempt was made to alleviate a confusion of which model to which data. The logit and probit model fitting gives similar results for the same data as expected and the choice of one model cannot be made based on AIC, because the AIC for both models is the same.

The poisson regression model fit is found to be inadequate for two different variables, as its Deviance value is far from one. The Negative Binomial Model gives a better fit and its Deviance shows the model is adequate for the same data used for poisson regression. The multinomial logit model for parthenium infestation in five categories as dependent variable and the sum of all other species gives a better result, as infestation level increase i.e as the severity of parthenium infestation increase, the number of the sum of other species gets low which in turn means that the probability of getting other species gets very low.

CHAPTER ONE

1. INTRODUCTION

1.1 Background

Crop protection is a discipline which studies how to protect crop productivity from loss due to pest, weed and other diseases. For example, striga is one of the major biological constraints of subsistence agricultural in sub-saharan Africa. It is a major problem in 59 countries accounting for an average crop loss of 40%.

Parthenium is a highly competitive, adaptable and allergenic weed. It is an invasive annual weed believed to be introduced to Ethiopia in 1970s. The weed has currently spread to the most part of the country. It was observed growing in different habitats from hot arid and semi-arid low altitude (912 m) to high-mid-altitude (2500 m) of the study areas. It grows on roadsides, vacant sites, towns, villages, gardens, waterways, grasslands and in different crop fields both during the crop season and after harvest. Parthenium was detected as a major weed of crops in the north and eastern regions of Ethiopia with infestation of greater than 20 plants per m^2 in some locations. Infestation of parthenium in the crop field varied from field to field depending on the time of parthenium introduction into the area and the efforts made by the farmers to control the weed. It grows on different types of soil in different habitats.

Parthenium infestation was found very low at high altitudes and was not observed on mountains. However, it has the ability to grow in any temperature and rainfall regions.

Control of Parthenium by Farmers is entirely based on cultural and labour intensive practices such as hand weeding, mowing, hoeing and slashing. This caused farmers to suffer from skin allergy, itching, fever, and asthma.

Parthenium is a herbaceous invasive weed believed to be originated in tropical America, now occurs widely in India, Australia, and east and south Africa. It is a procumbent, diffused leafy herb, 0.5 - 2 m tall, bearing alternate, pinnatified leaves, belonging to the family Compositae. In Ethiopia, it is believed to have been introduced in 1976/77 with army vehicles from Somalia and has become a serious weed both in arable and grazing lands (Taye, 2002).

Other than direct competition with crops, Parthenium poses allelopathic effect on different crops and other plants and health hazard to humans and animals (Chippendale and Panneta, 1994). Parthenium can also cause severe crop yield losses. In India, a yield reduction of 40% in agricultural crops (Khosla and Sobti, 1981) and 90% reduction in forage production in grass lands (Nath, 1988) were reported. In eastern Ethiopia, Tamado (2001) has reported that Parthenium is the second most frequent weed (54 %) after *Digitaria abyssinica* (63%) and that sorghum grain yield was reduced from 40 - 97 % depending on the year and the location. In the Caribbean, Parthenium is the fourth most serious weed (Hammerton, 1981) while in Kenya, it was reported as one of the important weed in coffee plantations (Njorge, 1991).

In Ethiopia, the distribution and ecology of Parthenium infestation were not yet determined though this information is important in order to apply various control measures against the weed. Therefore, a study that aims to draw an exhaustive inventory of Parthenium infestation, distribution and its

ecology in Ethiopia was conducted. This paper used the secondary data (count data type) collected by this exploratory study.

1.1.1 Naming of Parthenium Weed in Ethiopia

The name Parthenium is commonly accepted for *P. hysterophorus* in Ethiopia though it has many different vernacular names used in different regions and in different locations within a region. It is known as “Qinche” in Tigray region; “Qinche Arem” or “Chebchabe” in Amhara region; “Arama Sorgo”, “Arama Kuba”, “Biyabassa”, “Faramssiissa” or “Dayyeessa” in Oromiya region; “Arama Kuba” or “Terekabi” in Afar region; and “Kalignole” in Somalia region. The names imply its introduction and/or invasiveness or morphology. For example, “Arama Sorgo” implies that the weed was introduced from Somalia during the Ethio-Somalia war in 1976/77. “Biyabassa” means leave the place or a region, “Faramssiissa” means sign to leave the land, “Qinche“ is a kind of food made of coarse grinned barley or wheat that looks like the white flowers of Parthenium, and “Kalignole” means living alone indicating its allelopathic and strong competitive nature.

1.1.2 Morphology of Parthenium Weed in Ethiopia

All Parthenium populations that were surveyed in different parts of Ethiopia: Southern Tigray, North and South Wollo, East Wollega, central farmlands of Shewa, Rift Valley, and West and East Hararghe have a rosette leaves that did not elongate until flowering and form white flowers. Hence, they were classified under the white-flowered Parthenium group. Based on the flower colour and its rosette stage, it is similar to plants growing in Kenya, Australia, India, Mexico and USA.

1.1.3. Infestation of Parthenium weed

Parthenium was observed to grow on roadsides, vacant sites, towns, villages, gardens, waterways, grasslands and in crop fields both during the crop season and after harvest so long as enough moisture is available. Field crops [maize , sorghum], finger millet, cotton, haricot bean, tef, vegetables [potato, tomato, onion, cabbage, and carrot], and orchards [citrus, mango, papaya and banana] were found to be infested by Parthenium.

In the surveyed areas, infestation of Parthenium in the crop field varies from field to field depending on the time of its introduction into the area and the efforts made by the farmers to control the weed. It became a major weed of crops in the northern and eastern regions of Ethiopia. In eastern Ethiopia, although Parthenium grows as dense stands in every plot of land, farmers were aware of Parthenium problem. They keep their farmlands almost free of Parthenium through intensive cultivation, hoeing and hand weeding and use of inter-cropping.

In the central farmlands of East Shewa: Dukem, Debre Zeit, Mojo, and Koka areas heavy and widespread infestation occurs mostly on roadsides, wastelands, towns, villages and gardens.

High infestation of Parthenium (> 20 plants per m^2) was observed in sorghum fields around Kobo, and in sorghum, maize and teff fields around Robit, Gobie, Woldiya, and Kombolcha both during the growing period (September – November, 2000) and after harvesting time (January - March, 2000). Similarly, in East Shewa (Wolenchitti, Wonji and Methara), Afar region (Awash, Anano, and Miesso), and West and East Hararghe, heavy infestation of sorghum and maize was observed both during fallow and cropping seasons. The scale of Parthenium infestation in these areas is 5, i.e. severe infestation of Parthenium (>20 plants per m2). Scale of importance of Parthenium was also 4,

showing very serious or heavy yield reduction due to Parthenium. In Ataye, Shewa Robit, Ambo, and Nazareth area, Parthenium has entered crop fields having a scale of infestation and importance 4 and 3, respectively.

In highly infested areas from Woldiya to Alamata, the original grass and shrub vegetation had been very open and the disturbance allowed a dense stand of Parthenium to cover thousands of hectares of grazing and cultivated lands. From Sirinka to Mersa and then to Dessie, Parthenium was present on the narrow strip along the main road for several kilometres. Occasional dense stands were observed around farm buildings, crop fields and grazing lands especially where overgrazing had taken place. Similarly, from Kamisse down to Shewa Robit, dense infestation was observed along the main road, villages and waterways. Infestation of crop fields (maize, sorghum, tef) around Gubalafto, Ataye and Shewa Robit was also observed.

In many districts of West Shewa: Shoboka, Tibe, Guder, and Wolliso, only localized infestation of Parthenium was observed on roadsides and rarely in crop fields. The introduction in these areas is very recent, probably since 1997 for there had been no Parthenium observed in West Shewa region from 1995 – 1996 (Taye et al., 1998) during which intensive qualitative and quantitative determination of weeds occurring in these areas took place.

Parthenium is also known to affect animal and human health. Based on the field surveys it was known that Parthenium is replacing native grass species. In highly infested grasslands, it was observed to cause dermatitis, bloating and diarrhoea on animals. Farmers reported that the milk and meat of animals grazing on Parthenium and the honey obtained from bees visiting Parthenium flowers are bitter and untasteful. Individuals making hand weeding or hoeing in Parthenium infested

crops suffer from skin allergy, itching, fever, and asthma. In some areas, Parthenium also serve as a niche for pests like rodent and monkeys.

1.2. Intoduction to Methods

The statistical methods used to analyze the data is both descriptive and inferential statistics in fitting the data to different types of Generalized linear models(GLMs). Families of generalized linear models are to be used for this study and the whole process comprises comparing, selecting and applying the best model.

The softwares used includes SAS, STATA, and SPSS. The dependent variable(number of parthenium plant per quadrant) which is to be related with the independent variables quadrants and loss in crop productivity, and area of land infested, health status of farmers, death of animals due to parthenium and so on.

1.3. Statement of the problem

This thesis deals with modeling data from crop protection using generalized linear model(GLM). This includes checking the data for its normality, transformation, fitting the data for different models for parameter estimation, and test of significance of each parameter. The data used is a count data of parthenium weed and other plant species collected from a one kilometer-by-one kilometer area. Different models from among the class of generalized linear models(GLMs) fitted to the data and the one with best fit appreciated for prediction purpose.

1.4. OBJECTIVE

The general objective of the study is to explore the benefit of Generalized linear models in analyzing data from crop protection experiment.

The specific objectives of this study are:-

1. To explore models and analysis methods used in crop protection.

2. To show the advantage of generalized linear model as a whole.

3. To Recommend more precise statistical methods and models.

1.5. Organization of the thesis

The thesis includes five chapters and 11 appendixes. The first chapter, introduction, deals with much on parthenium origin, naming, morphology, infestation, and damages caused by parthenium. Chapter two is on literature review. The next chapter, chapter 3. is materials and methods. In this chapter, the data and its source, GLM, and software applications are dealt. Chapter four, results and discussions, deals with ANOVA, diagnostics, model fitting and discussing results as well as model checking. The last but not least chapter is chapter 5, which describes conclusions and recommendations.

CHAPTER TWO

2. LITRATURE REVIEW

According to the exploratory field surveys conducted in major Parthenium infested areas of Ethiopia: the central farm land and rift valley, South and North Wollo, West and East Hararghe, and East Wollega the major crops growing are teff (Eragrostis tef) and wheat (Triticum aestivum L.) in central farmlands and in some parts of Wollo; sorghum (Sorghum bicolor L.) and maize (Zea mays L.) in Hararghe and in Wollo, and maize in East Wollega. Though most of the study areas are arid or semi-arid characterized by low and erratic rainfall, the central farmlands, East Wollega and parts of Hararghe and Wollo are humid mid altitude areas suitable for crop production. Assessment was conducted in different habitats: cultivated lands, vacant lands and in grasslands to draw exhaustive inventory of Parthenium infestation.

Assessment was done at a regular interval (5 – 10 km) along the main roadsides after crop harvest (January – March 2000) and before harvest (August – November 2000). Fields were selected regardless of the size and on the basis of accessibility to the main road and representation of the growing conditions found in each location. The infestation of Parthenium was estimated by counting the number of plants per m^2 following the methods used by Yohannes et al. (1999) with some modifications. Five counts were taken per field and 3 – 5 fields were assessed per location and then converted into the scale of Parthenium infestation developed (Taye, 2002) as 0 – 5: where 0 = no Parthenium plants observed in the field, 1 = beginning or presence of Parthenium infestation only on road sides, 2 = presence of Parthenium infestation on roadsides and non-crop lands (grass land, gardens, waste land, etc.), 3 = infestation on road sides, non-agricultural lands, and beginning

infestation in crop fields, 4 = infestation in crop fields up to 20 plants per m^2, 5 = severe infestation of Parthenium (>20 plants per m^2).

Multivariate method of clustering is used to group locations based on temperature, rainfall and altitude to clarify the pattern of Parthenium infestation.

In addition to the above simple count study, a ph.D. research was conducted under the title "Investigation of Pathogens for Biological Control of Parthenium (Parthenium hysterophorus L.) in Ethiopia" by Taye Tessma which tried to use statistics in different forms about parthinium. The research paper didn't display statistical model for parameter estimation or for prediction of future values or for any other statistical purpose. Some researchers in the Ethiopian Agricultural Research institute used a software called 'CANOCA' to analyze their data in its very simplest standard, which they feel that it is very fascinating and/or easy to interact with.

Taye(Ph.D.) et al under "occurrence and distribution of parthenium phyllody"(2004) showed incidence of parthenium as percent of parthenium plants over total plants in 4mx4m plots($16m^2$) of the count data. Five counts per field and 3-5 fields were assessed at random at an interval of 2-3 km per location and then scaled as < 1 %, 1-5 %, 6-20 %, 21-50 % and > 50 %. Diseased plant samples were collected, tagged and pressed for later examinations in the laboratory.

Tebkew(2004), showed two graphs of fitted linear model for the data of randomly selected sample of 20 chickpea plants in an interval of 5-10 km in each field along a crossed diagonal line. The first fitted graph shows the effect of altitude on the magnitude of pod damage in chickpea by Helicoverpa armigera. The second graph depict the relation between mean percent pod damage and estimated weight loss using 100 seed weight.

Assfa, Tanner and Bennie(2005) under Effect of straw management, Tillage and cropping sequence on weed population Dynamics in south Eastern Highlands of Ethiopia" used the weed count data and transform it by square root transformation, scale it for analysis, used percentages and graphs to show data analysis**.**

Unpublished M.sc. thesis, Samuel Ashebir(2005) under Effect of Entomopathogenic Nematodes and Fungi on Barley chafer Grub(coptognatus curtipennis; Coleoptra: Dynastidae) used a linear regression analysis to show the relationship between concentration and percent Mortality in three different entomophathogenic nematodes concentrations and grub mortality. Similarly, Million Alemayhu(2003) in unpublished M.sc. thesis carried out Characterization of indigenous stone Bunding(Kab) and its effect on crop yield and soil productivity at Mesobit-Gedeba, North shewa zone of Amhara region. This fellow used simple linear correlation and regression analysis for the different parameters of the study. The different data of the terraced sites were analyzed according to analysis of variance (ANOVA) applied to split-split plot design using a software called MSTAC-C.

Sewalem A., Mebrate and B.M. Cooke(2001) under Aggressiveness of septoria Tritici isolates on Detached and Intact leaves of wheat cultivars used correlation between detached and intact leaf tests to show cultivar effects on the components of partial disease resistance of the septoria tritici isolates/mixtures. They also used the correlation coefficients to show the components of partial disease resistance tested. In addition to this different bar graphs are applied to compare the isolates/mixtures tested in terms of their NLS,DS and IPG.

Tamiru Hirpa (2005) under Effect of plant population and harvesting Time on Growth and dry Matter production of potato used a split plot design to fit into randomized complete blocks with three replications. Two verities were assigned to the main plot, while the twelve combination treatments representing planting densities and harvesting time to the sub plots were assigned. The statistical method that they used to analyze the data collected from this design were the mean number of stems, total number of leaves and leaf area index(LAI) as recorded on per plant basis; a table display and line graphs.

All the aforementioned research results didn't used GLM in modeling data; especially in the area of crop protection i.e GLMs did not appear in the agricultural research area literatures as seen in the Ethiopian agricultural research agency. Whereas if one browse internet, different research studies can be found that used GLM for data analysis and modeling in a lots of walk of life. In these studies, researchers used experimental data that they generate and/or secondary data collected for some other reason. Some, actually two, that are somehow related are tried to be discussed below.

Timothy H.(Ph.D.) under Optimal Experimental Design for Nonlinear and Generalised Linear Models used GLM to show the Optimal Experimental Design performance and parameter estimation in the area of chemistry and pharmacology instead of using linear models. It examines existing criteria for model discrimination for both nonlinear and generalised linear models, and combines them with criteria for parameter estimation. The thesis shows that these designs can be quite efficient in terms of each of the criteria.

Ruth Salway Jon Wakefield under 'Gamma Generalized Linear Models' for Pharmacokinetic Data (Pharmacokinetics is the study of the time course of a drug and its metabolites after introduction into the body) presented models based on generalized linear models for pharmacokinetic data as an alternative or as compared to compartmental models that have been used to analyze such data. For the generalized linear models desirable statistical properties exist, with a logarithmic link and gamma distribution.

Finally, previous studies in the area of crop protection used Multivariate method of clustering to group locations based on temperature, rainfall and altitude to clarify the pattern of parthenium infestation. Other researchers used contingency table and chi-square testes, and percentages. These statistics are more descriptive in nature and the precision of the result from such analysis would be very small, because other properties of the data like its normality, variance homogeneity and so on

are not investigated. In most cases, the application of statistical models to fit to the data and see the nature of estimated parameters, interpret it in its context, and hypothesis testing are not widely observed and not at their satisfactory level. Hence, in this paper dealing with one class of statistical model, fitting it to the data, popularize it to researchers, and show the way how they can deal with it to improve precision of their result is the ultimate goal. The proposed model to be used is a generalized linear model. In applying this model, the first step is to identify a distribution that is appropriate for the data in question. Then follows fitting different models of the generalized linear model for the data and then compare, select and apply the best one.

CHAPTER 3 - Materials and Methods

3.1. Data and source

The data used in this study are mainly secondary data. The data were collected by Ethiopian institute of Agricultural research for an exploratory study. Since it is a count of parthenium and all other species per plot in locality of welenchti. There are 100 plots and within each plot there are four quadrants and within each quadrant there are two plants. Totally we have 800 series of data. The count of each species was made in each 800 locations and registered accordingly.

3.2 Generalized Linear Models (GLM)

Generalized linear models (GLMs) are used to do regression modeling for non-normal data with a minimum of extra complication compared with normal linear regression. GLMs are flexible enough to include a wide range of common situations, but at the same time allow most of the familiar ideas of normal linear regression to carry over.

The GLMs are an extension of the linear modeling process that allows models to be fitted to data that follows other probability distributions other than the Normal distribution, such as the Poisson, Binomial, Multinomial, and others. Generalized Linear Models also relax the requirement of equality or constancy of variances that is required for hypothesis tests in traditional linear models.

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).

To illustrate, 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 ***e*** stands for the error variability that cannot be accounted for by the predictors; and the expected value of ***e*** is assumed to be 0. In this equation $\boldsymbol{b_0}$ is the intercept and the $\boldsymbol{b_i}$ values are the regression coefficients (for variables 1 through *k*) computed from the data. While 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 $\boldsymbol{mu_y}$ stands for the expected value of *y*.

Generalized linear models encompass a large class of models, from simple linear regression models to models for quantal responses to models for survival data. They can be studied as a single class, and are all defined by three characteristics:

1. The distribution of the **n×1** vector of independent responses, $\mathbf{Y = (Y_1, \ldots, Y_n)'}$, with means $\mathbf{E(Y_i) = \pi_i}$ and variances $\mathbf{Var(Y_i) = a(\Phi)V(\pi_i)}$, where a(Φ) is a scale factor which doesn't depend on $\boldsymbol{\pi_i}$ **and** $\boldsymbol{\pi = (\pi_1, \ldots, \pi_n)'}$.

2. The linear predictor $\boldsymbol{\eta = (\eta_1, \ldots, \eta_n)' = X\theta}$, where X is the n × p matrix of known functions of the **m** explanatory variables, and $\boldsymbol{\theta = (\theta_1, \ldots, \theta_p)'}$ is the p × 1 vector of model parameters.

3. The link function **g**, providing the link between the mean vector $\boldsymbol{\pi}$ and the linear Predictor $\boldsymbol{\eta}$: $\mathbf{g(\pi_i) = \eta_i}$.

A traditional linear model is of the form $\mathbf{y_i} = \mathbf{x'_i\beta} + \boldsymbol{\varepsilon_i}$ where $\mathbf{y_i}$ is the response variable for the i^{th} observation. The quantity $\mathbf{x_i}$ 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 $\boldsymbol{\beta}$ is estimated by a least squares fit to the data **y**. The $\boldsymbol{\varepsilon_i}$ are assumed to be independent, normal random variables with zero mean and constant variance. The expected value of $\mathbf{y_i}$, denoted by $\boldsymbol{\mu_i}$, is $\boldsymbol{\mu_i} = \mathbf{x'_i\beta}$. While traditional linear models are used extensively in statistical data analysis, there are types of problems for which they are not appropriate.

1. 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.

2. 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 $\mathbf{x'_i\beta}$ 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.

3. 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.

The Generalized Linear Model 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

$$\mathbf{Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k}$$

So for example, one could estimate (i.e., predict) a person's weight as a function of the person's height and gender. You could use linear regression to estimate the respective regression coefficients from a sample of data, measuring height, weight, and observing the subjects' gender. For many data analysis

problems, estimates of the linear relationships between variables are adequate to describe the observed data, and to make reasonable predictions for new observations.

There are many relationships that cannot adequately be summarized by a simple linear equation, for two major reasons:

1. the dependent variable of interest may have a non-continuous distribution, and thus, the predicted values should also follow the respective distribution.
2. the effect of the predictors on the dependent variable may not be linear in nature.

These two reasons are briefed in the next two consecutive sections in connection to practical examples.

One of the available programs that is important in current statistical practice is the GLM procedure in the SAS software package**.** GLM stands for "general linear model." A great many standard statistical methods fall under the general linear model, including bivariate and multivariate linear regression, fixed-effects analysis of variance, and analysis of covariance. While the general linear model may be viewed as a special case of the generalized linear model with identity link, most results of interest are obtained exactly only for the general linear model. Thus, the development of the general linear model has undergone a somewhat longer historical development. Results for the generalized linear model with non-identity link and fitted variance parameters are largely asymptotic (tending to work well with large samples).

A simple, very important example of a generalized linear model (also an example of a general linear model) is linear regression. Here the distribution function is the normal distribution with constant variance and the link function is the identity.

3.2.1. Distribution of the dependent variable.

The first reason why a simple linear equation is inadequate to describe a particular relationship is that a 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 farmer's choice of one of three alternative seeds of maize). 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 you 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.

3.2.2 Link function.

The link function provides the relationship between the linear predictor and the distribution function (through its mean). There are many commonly used link functions, and during the choice it is important to match the domain of the link function to the range of the distribution function's mean.

Table1 below shows some common link functions and their inverses (sometimes referred to as the mean function) used for several distributions in the exponential family.

Table1: Link functions

Distribution	Name	Link Function	Mean Function
Normal	Identity	$X\beta = \mu$	$\mu = X\beta$
Exponential	Inverse	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
Gamma	Inverse	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
Poisson	Log	$X\beta = \ln(\mu)$	$\mu = \exp(X\beta)$
Binomial	Logit	$X\beta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)}$
Multinomial	Logit	$X\beta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)}$

In the case of exponential and gamma distributions, the domain of the link function (that is, the range of the mean function) is not the same as the permitted range of the mean. In particular, the linear predictor may be negative, which would give an illegal negative mean.

The second 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.

The generalized linear model can be used to predict responses both for dependent variables with discrete distributions and for dependent variables which are nonlinearly related to the predictors. Pridictor is the function used to predict future values of the dependent variable using the fitted model.

The generalized linear model is a generalization of the linear regression model such that (1) nonlinear, as well as linear, effects can be tested (2) for categorical predictor variables, as well as for continuous predictor variables, using (3) any dependent variable whose distribution follows several special members of the exponential family of distributions (e.g., gamma, Possion, binomial, etc.), as well as for any normally-distributed dependent variable.

Categorical Predictor Variable. A categorical predictor variable is a variable, measured on a nominal scale, whose categories identify class or group membership, which is used to predict responses on one or more dependent variables. Gender would be an example of a categorical predictor variable, with the two classes or groups Male and Female. To site an example in this study ; a plot can be free of parthenium or infected by parthenium.

Categorical Dependent Variable. A categorical dependent variable is a variable of interest (a researcher wants to predict), measured on a nominal scale, whose values identify class or group membership (e.g., Gender, with classes Male and Female; or Education, with classes No High School Degree, High School Degree, Some College, College Degree, Some Graduate School, Graduate Degree). A researcher may be interested in predicting the group membership of observations based on the values of some independent or predictor variables. For example, Credit Risk, with values Good and Bad, would be a categorical dependent variable that one might want to predict based on measured independent variables that are possibly related to credit risk.

The assumptions in the generalized linear models are:-

1. Observations are independent (or at least uncorrelated)-excludes time series & spatial processes.

2. There is a single error term in the model. This assumption excludes models for the analysis of experiments having more than one explicit error term. Eg. Models for Split-plot design, which has two error terms, one for between-whole-plot variance and one for within-whole-plot variance.

3.3 Processes in Model Fitting

In general model fitting is to get an aid for interpreting data. There are three processes in model fitting:

1. model selection or identification
2. parameter estimation
3. prediction of future values

Model Selection

Selection of models to fit to data are usually taken from a particular class and, if the model-fitting process is to be useful, this class must be broadly relevant to the kind of data under study.

Parameter estimation

In case of generalized linear models, estimation proceeds by defining a measure of goodness of fit between the observed data and the fitted values generated by the model. The parameter estimates are the values that minimize the goodness-of-fit criterion.

3.4.Some principles to Guide the modeler

According to McCullagh and Nelder(1983), it is advisable to follow the following modeling principles:-

1. All models are wrong; some, though, are better than others and we can search for the better ones. At the same time we must recognize that eternal truth is not within our grasp.
2. Not to fall in love with one model, to the exclusion of alternatives. Data will often point with almost equal emphasis at several possible models and it is important that the analyst accept this.
3. Checking thoroughly the fit of a model to the data, for example by using residuals and other quantities derived from the fit to look for outlying observations, and so on. Since such procedures are not yet fully formalized, imagination is required of the analyst.

3.5. Common Generalized Linear Models

3.5.1. Logistic regression

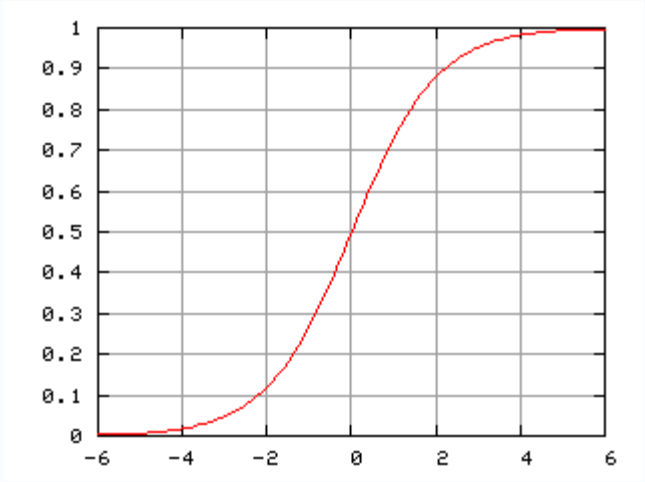
Logistic regression is a statistical regression model for Bernoulli-distributed dependent variables. It is a generalized linear model that uses the logit as its link function. Logistic regression is used extensively in the agricultural, biological, medical and social sciences. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription.

An explanation of logistic regression begins with an explanation of the logistic function:

$$f(z) = \frac{1}{1+e^{-z}}$$

The logistic function is useful because it can take as an input, any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1. The variable ***z*** represents the exposure to some set of risk factors, while *f*(*z*) represents the probability of a particular outcome, given that set of risk factors. The variable *z* is a measure of the total contribution of all the risk factors used in the model and is known as the logit.

Figure1: The logistic function with z on the horizontal axis and f(z) on the vertical axis.



The variable *z* is usually defined as

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots + \beta_k x_k,$$

where β_0 is called the "intercept" and β_1, β_2, β_3,…. B_k, are called the "regression coefficients" of x_1, x_2, x_3 ,…..,x_k respectively. The intercept is the value of z when the value of all the other risk factors is zero (i.e., the value of *z* in someone with no risk factors). Each of the regression coefficients describes the size of the contribution of that risk factor. A positive regression coefficient means that the risk factor increases the probability of the outcome, while a negative regression coefficient means that the risk factor decreases the probability of that outcome; a large regression coefficient means that risk

factor strongly influences the probability of that outcome; while a near-zero regression coefficient means that the risk factor has little influence on the probability of that outcome.

Logistic regression is a useful way of describing the relationship between one or more risk factors (e.g., age, sex, etc.) and an outcome such as infestation (which only takes two possible values: infested or not infested).

Logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure. Discriminant analysis is also used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables. Thus, in instances where the independent variables are a categorical, or a mix of continuous and categorical, logistic regression is preferred.

The logistic model takes the form

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i},$$

$$i = 1, \ldots, n,$$

where there are n units with covariates X and

$$p_i = E(Y/X_i) = \Pr(Y_i = 1)$$

The logarithm of the odds (the probability divided by one minus the probability of the outcome) is modeled as a linear function of the explanatory variables, X_i. This can be written equivalently as

The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success θ, or the value 0 with probability of failure 1-θ. This type of variable is called a Bernoulli (or binary) variable. Although not as common, applications of logistic regression have also been extended to cases where the dependent variable is of more than two cases, known as multinomial or polytomous [Tabachnick and Fidell (1996) use the term polychotomous].

The independent or predictor variables in logistic regression can take any form. That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and response variables is not a linear function in logistic regression, instead, the logistic regression function is used, which is the logit transformation of θ:

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}$$

Where α = the constant of the equation and, β = the coefficient of the predictor variables.

An alternative form of the logistic regression equation is:

$$\text{logit}\left[\theta(x)\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. Several different options are available during model creation.

There are two main uses of logistic regression. The first is the prediction of group membership. Since logistic regression calculates the probability of success over the probability of failure, the results of the analysis are in the form of an odds ratio. For example, logistic regression is often used in

epidemiological studies where the result of the analysis is the probability of developing cancer after controlling for other associated risks. Logistic regression also provides knowledge of the relationships and strengths among the variables (e.g., smoking 10 packs a day puts you at a higher risk for developing cancer than working in an asbestos mine).

3.5.2. Probit model

Probit model is a popular specification of a generalized linear model, using the probit link function. The term "probit' was coined in the 1930's by Chester Ittner Bliss and stands for probability unit. Because the response is a series of binomial results, the likelihood is often assumed to follow the binomial distribution. Let *Y* be a binary outcome variable, and let *X* be a vector of regressors. The probit model assumes that

where Φ is the cumulative distribution function of the standard normal distribution. The parameters β are typically estimated by maximum likelihood.

While easily motivated without it, the probit model can be generated by a simple latent variable model. Suppose that

$$Y^* = x'\beta + \varepsilon,$$

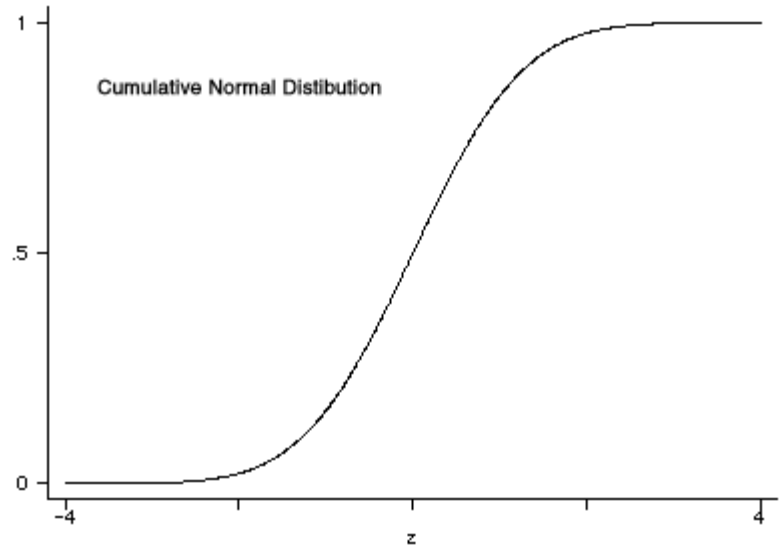
where $\varepsilon|x \sim \mathcal{N}(0,1)$, and suppose that *Y* is an indicator for whether the latent variable Y^* is positive:

Then it is easy to show that

$$\Pr(Y = 1|X = x) = \Phi(x'\beta).$$

The logit and probit analyses are very similar to one another. Logit analysis is based on log odds while probit uses the cumulative normal probability distribution. Here is what a cumulative normal distribution looks like.

Figure2: The cumulative normal probability distribution.



The probit model is defined as

$$\mathbf{Pr(y=1|x) = \Phi(b_o + xb)}$$

where **Φ** is the standard cumulative normal probability distribution and **xb** is called the probit score or index.

Since **xb** has a normal distribution, interpreting probit coefficients requires thinking in the **Z** (normal quantile) metric. The interpretation of a probit coefficient, **b**, is that a one-unit increase in the predictor leads to increasing the probit score by **b** standard deviations. Leaning to think and communicate in the **Z** metric takes practice and can be confusing to others. One can make use of a number of tools developed by Long and Freese to aid in the interpretation of the results. Below is a brief description of it.

The log-likelihood function for probit is

$$\ln L = \sum w_j \ln \Phi(x_j b) + \sum w_j \ln\left(1 - \Phi\left(x_j b\right)\right)$$

where $\mathbf{w_j}$ denotes optional weights.

Currently, logit models are more popular than probit models due to two reasons; 1) the exponentiated logistic coefficients can be interpreted as odds ratios, and 2) there are more diagnostic tools available in logistic regression. Although, this last reason can be a chicken-egg issue, that is, there might be more diagnostic tools because it is being used more often.

Cumulative normal distribution:- The S-shaped curve which results when you add up the bell-shaped normal curve, moving from z = -infinity to z=+infinity. The cumulative normal distribution is used by probit and not by logit.

Probit coefficients correspond to the **b** coefficients in regression or the logit coefficients in logit or logistic regression. All are effect coefficients. Logit and probit analysis generally arrive at the same conclusions for the same data, but the logit and probit coefficients differ in magnitude.

In the case of probit coefficients, the coefficient is how much difference a unit change in the independent makes in terms of the cumulative normal probability of the dependent variable. This means the probit coefficient measures the effect of the independent on the Z scores of the dependent. Note that the probability of the dependent is not a linear function of Z, but rather is a cumulative normal function of Z. This means that the effect of a unit change in the independent on the probability of the dependent depends on the level of the independents. Therefore to assess the effect of probit coefficients it is necessary to choose some level of the independents as a reference point and in particular the standard reference point is when all independents are at their sample means.

One substitutes the sample means of the independents into the probit regression equation to get the estimated Z score, then one looks in a table of the standard normal distribution to find the corresponding probability level. This gives us the baseline statement that when all variables are at their sample means, the probability that the dependent will have a value of 1 is that probability. This baseline can then be used to understand the effects of one unit change in an independent, given its

probit coefficient value. The answer will be the probability when the independent is at its sample mean plus one unit, minus the baseline probability, with all other independents held constant at their sample means. That is, one takes the baseline equation and simply substitutes the mean of the independent plus 1, then calculates the probability. Then one subtracts the baseline probability from this probability.

This calculated probability difference is called the **elasticity** of P(Y) with respect to the independent, when variables are held at their sample means, where Y is the dependent. The elasticity is the effect of a unit increase in the independent variable on the probability that the dependent=1, when all other independents are held constant at their mean values.

Probit regression is an alternative log-linear approach to handling categorical dependent variables. Its assumptions are consistent with having a categorical dependent variable assumed to be a proxy for a true underlying continuous normal distribution. A typical use of probit is to analyze dose-response data in medical studies. Like logit or logistic regression, the researcher focuses on a transformation of the probability that Y, the dependent, equals 1. Where the logit transformation is the natural log of the odds ratio, the function used in probit is the inverse of the standard normal cumulative distribution function.

Where logistic regression is based on the assumption that the categorical dependent reflects an underlying qualitative variable and uses the binomial distribution, probit regression assumes the categorical dependent reflects an underlying quantitative variable and it uses the cumulative normal distribution. As with logit regression, there are oprobit (ordinal probit) and mprobit (multinomial probit) options. Discussing and applying these models is left for the upcoming fellows.An extended discussion of probit is found in Pampel (2000: 54-68).

Probit may be the more appropriate choice when the categories are assumed to reflect an underlying normal distribution of the dependent variable, even if there are just two categories.

3.5.3. **Multinomial model**

One may apply regression models to the categorical dependent variables. However, due to the non-linearities of these models the statistical analysis and interpretation of these models is not an easy task. The most promising approach is via the method of maximum likelihood estimation in developing the logit and probit models for binary and ordinal data. The multinomial logit model is often used for nominal data.

Multinomial models refer to situations in which there can be multiple causes for a single event and allow you to estimate in the independent contribution of each of those causes.

This type of model applies to cases where an observation can fall into one of k categories. Binary data occurs in the special case where k=2. If there are m_i observations in a subpopulation i, then the probability distribution of the number falling into the k categories $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots y_{ik})$ can be modeled by the multinomial distribution. The multinomial model is an *ordinal* model if the categories have a natural order.
In multinomial logit models, the dependent variable is nominal. That is, there is no inherent order to the levels of the categories of the response variable.

The GENMOD procedure of SAS software orders the response categories for ordinal multinomial models from lowest to highest by default. This is different from the binomial distribution, where the response probability for the highest of the two categories is modeled. The way GENMOD orders the response levels can be changed with the ORDER option in the PROC GENMOD statement.

The GENMOD procedure supports only the ordinal multinomial model. If $(p_{i1}, p_{i2}, \ldots p_{ik})$ are the category probabilities, the cumulative category probabilities are modeled with the same link functions used for binomial data. Let $p_{ir} = \sum_{j=1}^{r} = p_{ij}$, r = 1, 2, ... , k-1 be the cumulative category probabilities (note that P_{ik} = 1). The ordinal model is

$$g(P_{ir}) = \mu_r + \mathbf{x}_i'\beta \quad \text{for} \quad r = 1, 2, \cdots k-1$$

where $\boldsymbol{\mu_1, \mu_2, \ldots, \mu_{k-1}}$ are intercept terms that depend only on the categories and $\mathbf{x}_i$ is a vector of covariates that does not include an intercept term. The logit, probit, and complementary log-log link functions *g* are available. These are obtained by specifying the MODEL statement options DIST=MULTINOMIAL and LINK=CUMLOGIT (cumulative logit), LINK=CUMPROBIT (cumulative probit), or LINK=CUMCLL (cumulative complementary log-log). Alternatively,

$$P_{ir} = \mathrm{F}(\mu_r + \mathbf{x}_i'\boldsymbol{\beta}) \quad \text{for} \quad r = 1, 2, \cdots k-1$$

where F = g^{-1} is a cumulative distribution function for the logistic, normal, or extreme value distribution.

PROC GENMOD estimates the intercept parameters $\boldsymbol{\mu_1, \mu_2, \ldots, \mu_{k-1}}$ and regression parameters $\boldsymbol{\beta}$ by maximum likelihood.

The subpopulations *i* are defined by constant values of the AGGREGATE= variable. This has no effect on the parameter estimates, but it does affect the deviance and Pearson chi-square statistics; it also affects parameter estimate standard errors if you specify the SCALE=DEVIANCE or SCALE=PEARSON options.

3.5.4. Poisson regression

Poisson regression models are generalized linear models with the logarithm as the (canonical) link function, and the Poisson distribution function.

In statistics, **Poisson regression** is a form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable *Y* has a Poisson distribution, and assumes the logarithm of its expected value can be modelled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a **log-linear model**, especially when used to model contingency tables. The log mean is the natural parameter for Poisson distributions, and the log link is the canonical link for GLM with Poisson random component. A Poisson loglinear model is a GLM that assumes a Poisson distribution for Y and uses the log link.

Poisson regression is often used to analyze count data. Poisson regression can be used to model the number of occurrences of an event of interest or the rate of occurrence of an event of interest, as a function of some independent variables. For example, rate of insurance claims, number of doctor visits, incidence of diseases, crime incidence, number of days a child is absent from school, colony counts for bacteria, can be modeled using Poisson regression.

The poisson regression is usually estimated with the log link function, and is sometimes called the "exponential poisson regression." The Poisson distribution for the dependent variable is limited to positive values, and has a variance equal to it's mean. That is, in populations where events are very rare, the distribution is highly right skewed; as the mean of events rises, the distribution increasingly resembles the normal.

In Poisson regression the dependent variable Y has a Poisson distribution given the independent variables X_1, X_2, ..., X_m,

$$P(Y=k|\ x1,\ x2,\ ...,\ xm) = e^{-\mu}\ \mu^k\ /\ k!, \qquad k=0,\ 1,\ 2,\ .....,$$

where the log of the mean μ is assumed to be a linear function of the independent variables. That is,

$$\log(\mu) = \text{intercept} + b_1*X_1 + b_2*X_2 + ...+ b_m*X_m,$$

which implies that μ is the exponential function of independent variables,

$$\mu = \exp(\text{intercept} + b_1*X_1 + b_2*X_2 + ...+ b_m*X_m).$$

In many situations the rate or incidence of an event needs to be modeled instead of the number of occurrences. For example, suppose that we know the number of occurrences of certain disease by county and we want to find out if frequency of occurrence depends on certain demographic variables and health policy programs also recorded by county. Since more at risk subjects result in more occurrences of the disease, we need to adjust for the number of subjects at risk in each county. For such data, we can write a Poisson regression model in the following form:

$$\log(\mu) = \log(N) + \text{intercept} + b_1*X_1 + b_2*X_2 + ...+ b_m*X_m,$$

where N is the total number of subjects at risk by county. The logarithm of variable N is used as an offset, that is, a regression variable with a constant coefficient of 1 for each observation. The log of the incidence, log (μ / N), is modeled now as a linear function of independent variables. The maximum likelihood method is used to estimate the parameters of Poisson regression models. In SAS, the GENMOD procedure can fit Poisson regression models.

Poisson regression is appropriate when the dependent variable is a count, for instance of events such as the arrival of a telephone call at a call centre. The events must be independent in the sense that the arrival of one call will not make another more or less likely, but the probability per unit time of events is understood to be related to covariates such as time of day.

Poisson regression is also appropriate for rate data, where the rate is a count of events occurring to a particular unit of observation, divided by some measure of that unit's *exposure*. For example, biologists may count the number of tree species in a forest, and the rate would be the number of species per square kilometre. Demographers may model death rates in geographic areas as the count of deaths divided by person-years. More generally, event rates can be calcula ted as events per unit time, which allows the observation window to vary for each unit. In these examples, exposure is respectively unit area, person-years and unit time. In Poisson regression this is handled as an **offset**, where the exposure variable enters on the right-hand side of the equation, but with a parameter estimate constrained to 1.

$$\text{Log}(E(Y))= \log(\text{exposure}) + a + bx.$$

which implies

$$\text{Log}(E(Y)) - \log(\text{exposure}) = \log(E(Y)/ \text{exposure}) = a + bx.$$

It is equivalent to modeling the dependent variable as a rate model (count / exposure) = a + bx, or model count = offset(exposure) + a + bx. One way to check if this assumption is valid is to explicitly model using the offset variable. One confusing matter in using an offset term is that printed GENMOD output does not directly show how the offset was used. The offset option in Proc GENMOD does provide a very flexible means for testing hypotheses about model parameters - one can easily fix

parameters to desired levels using the offset term and reestimate remaining model parameters. This is one method that can be used to fit combined additive and multiplicative model structures.

A characteristic of the Poisson distribution is that its mean is equal to its variance. In certain circumstances, it will be found that the observed variance is greater than the mean; this is known as overdispersion and indicates that the model is not appropriate. A common reason is the omission of relevant explanatory variables.

Another common problem with Poisson regression is excess zeros: if there are two processes at work, one determining whether there are zero events or any events, and a Poisson process determining how many events there are, there will be more zeros than a Poisson regression would predict. An example would be the distribution of cigarettes smoked in an hour by members of a group where some individuals are non-smokers. When the Poisson model may not fit adequately, other generalized linear models such as the negative binomial model may function better in these cases.

3.5.5.Negative Binomial Regression

Fitting negative binomial regression is very similar to fitting Poisson regression, assuming that the model is the same as the one described in Poisson Regression , that is, the log of the mean, μ, is a linear function of independent variables,

$$\log(\mu) = \text{intercept} + b_1*X_1 + b_2*X_2 + ...+ b_m*X_m,$$

which implies that μ is the exponential function of independent variables,

$$\mu = \exp(\text{intercept} + b_1*X_1 + b_2*X_2 + ...+ b_m*X_m).$$

Instead of assuming as before that the distribution of Y, number of occurrences of an event, is Poisson, now it is assumed that Y has a negative binomial distribution. That means, in particular, relaxing the assumption about equality of mean and variance (Poisson distribution property), since the variance of negative binomial is equal to $\mu + k\mu^2$, where $k >= 0$ is a dispersion parameter. The maximum likelihood method is used to estimate k as well as the parameters of the regression model for $\log(\mu)$.

The null hypothesis is:

H_0 : k=0

and the alternative hypothesis is:

H_a : k>0.

To carry out the test, follow the steps below:

(i) Run the regression model using negative binomial distribution, record LL (log likelihood) value.

(ii) Record LL for the Poisson model.

(iii) Use the LR (likelihood ratio) test, that is, compute LR statistic, -2(LL (Poisson) – LL (negative binomial)). The asymptotic distribution of the LR statistic has probability mass of one half at zero and one half – Chi-sq distribution with 1 d.f. To test the null hypothesis at the significance level α, use the critical value of Chi-sq distribution corresponding to significance level 2α, that is reject H_0 if LR statistic > $\chi^2_{(1-2\alpha\ ,\ 1\ df)}$.

The negative binomial regression model adds an "overdispersion" parameter to estimate the possible deviation of the variance from that expected under the Poisson. This has the consequence of generating an (usually more conservative) estimate of standard errors and may (in a practical sense, though not necessarily in theory) modify parameter estimates.

3.6 Parameter Estimation in the generalized linear model.

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 (see Dobson,1990). 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 by-product of its computations.

3.6.1.Statistical significance testing

Tests for the significance of the effects in the model can be performed via the Wald statistic, the likelihood ratio (LR), or score statistic. Detailed descriptions of these tests can be found in McCullagh and Nelder (1989). The Wald statistic (e.g., see Dobson,1990), which is computed as the generalized inner product of the parameter estimates with the respective variance-covariance matrix, 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. For details concerning these different test statistics, see Agresti(1996), McCullagh and Nelder(1989), and Dobson(1990).

3.6.2. Estimation and Testing

The parameters in a generalized linear model can be estimated by the maximum likelihood method. For a given probability distribution specified by $f(y_i ; \boldsymbol{ß}, F)$ and observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, the log-likelihood function for $\boldsymbol{ß}$ and F, expressed as a function of mean values $\mu = (\mu_1,\ldots, \mu_n)$ of the responses $\{Y_1, Y_2, \ldots, Y_n\}$, has the form

$$l(\boldsymbol{\mu};\mathbf{y}) = \sum_{i=1}^{n} \log f(y_i;\boldsymbol{\beta},\phi)$$

.

The maximum likelihood estimates of the parameters $\boldsymbol{ß}$ can be obtained by iterative re-weighted least squares (IRLS). Detailed information about the iterative algorithm and asymptotic properties of the parameter estimates can be found in McCullagh and Nelder (1989).

Analogous to the residual sum of squares in linear regression, the goodness-of-fit of a generalized linear model can be measured by the scaled deviance

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2[l(\mathbf{y};\mathbf{y}) - l(\hat{\boldsymbol{\mu}};\mathbf{y})]$$

,

where $l(\mathbf{y};\mathbf{y})$ is the maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed values, and $l(\hat{\boldsymbol{\mu}};\mathbf{y})$ is the log-likelihood function calculated at the estimated

parameters **ß**. The deviance function is very useful for comparing two models when one model has parameters that are a subset of the second model. The deviance is additive for such nested models if maximum likelihood estimates are used (McCullagh and Nelder 1989). Consider two nested models with the second having some covariates omitted and denote the maximum likelihood estimates in the two models by $\hat{\mu}_1$ and $\hat{\mu}_2$, respectively. Then the deviance difference $\{D(\mathbf{y};\hat{\boldsymbol{\mu}}_2) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_1)\}$ is identical to the likelihood-ratio statistic and has an approximate χ^2 distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models. For probability distributions in the exponential family the χ^2 approximation is usually quite accurate for differences of deviance even though it may be inaccurate for the deviances themselves (McCullagh and Nelder 1989).

AIC (Akaike Information Criterion)- It is calculated as AIC = -2 Log L + 2((*k*-1) + *s*), where *k* is the number of levels of the outcome variable and *s* is the number of predictors in the model. **AIC** is used for the comparison of models from different samples or nonnested models. Ultimately, the model with the smallest **AIC** is considered the best.

However, there are limited ways in which we can interpret the individual regression coefficients. A positive coefficient mean that an increase in the predictor leads to an increase in the predicted probability. A negative coefficient means that an increase in the predictor leads to a decrease in the predicted probability.

3.7. Diagnostics in the generalized linear model.

The two basic types of residuals are the so-called 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. In addition, leverage scores, studentized residuals, generalized Cook's D, and other observational statistics (statistics based on individual observations) can be computed. For a description and discussion of these statistics, see Hosmer and Lemeshow (1989).

Compare the deviance values for two models to determine if a squared term would improve the fit significantly.

To check the goodness of fit, also look at a probability plot of the Pearson residuals. These are normalized so that when the model is a reasonable fit to the data, they have roughly a standard normal distribution. (Without this standardization, the residuals would have different variances.)

For each of the five distributions that the SAS glmfit supports, there is a canonical (default) link function. For the binomial distribution, the canonical link is the logit. However, there are also three other links that are sensible for binomial models. All four maintain the mean response in the interval [0, 1]. It's often difficult for the analyst to distinguish between these four link functions, and a choice is often made on theoretical grounds.

3.8 Methods of Analyses

The design for an analysis can include effects for continuous as well as categorical predictor variables. Designs may include polynomials for continuous predictors (e.g., squared or cubic terms) as well as interaction effects (i.e., product terms) for continuous predictors. For categorical predictor variables, one can fit ANOVA-like designs, including full factorial, nested, and fractional factorial designs, etc. Designs can be incomplete (i.e., involve missing cells), and effects for categorical predictor variables can be represented using either the sigma-restricted parameterization or the overparameterized (i.e., indicator variable) representation of effects.

In addition to fitting the whole model for the specified type of analysis, different methods for automatic model building can be employed in analyses using the generalized linear model. Specifically, forward entry, backward removal, forward stepwise, and backward stepwise procedures can be performed, as well as best-subset search procedures. In forward methods of selection of effects to include in the model (i.e., forward entry and forward stepwise methods), score statistics are compared to select new (significant) effects. The Wald statistic can be used for backward removal methods (i.e., backward removal and backward stepwise, when effects are selected for removal from the model).

The best subsets search method can be based on three different test statistics: the score statistic, the model likelihood, and the AIC (Akaike Information Criterion, see Akaike, 1973). Note that, since the score statistic does not require iterative computations, best subset selection based on the score statistic is computationally fastest, while selection based on the other two statistics usually provides more accurate results.

Several forms of the Generalized Linear Model are now commonly used and implemented in many statistical software packages. Logistic Regression, Multiway Frequency Analysis (Log-Linear Models), Logit Models, and Poisson Regression are all forms of the Generalized Linear Model. In Logistic Regression, the binary response variable is modeled as a Binomial random variable with the logit link function. For Multiway Frequency Analysis (Log-Linear Models), the response variable is usually modeled as a Poisson random variable with the log link function. However, one could assume that the response variable is Binomial or Multinomial, but the results would not differ from those obtained assuming the response variable to be Poisson distributed (Agresti 1996). For logit models, binary response variables are modeled as Binomial random variables, while polychotomous response variables are modeled as Multinomial random variables, but in both instances the link function is the logit function. In Poisson regression, the response variable is modeled as a Poisson random variable with the log link function.

3.9. Data Transformation

Data transformation is the appropriate remedial measure for variance heterogeneity where the variance & the mean are functionally related. With this technique, the original data are converted into a new scale resulting in a new data set that is expected to satisfy the condition of homogeneity of variance. Because a common transformation scale is applied to all observations, the comparative values between treatments are not altered and comparisons between them remain valid.

The appropriate data transformation to be used depends on the specific type of relationship between the variance and the mean. Three of the most commonly used transformations for data in agricultural research are explained below.

3.9.1.Logarithmic Transformation

The logarithmic transformation is most appropriate for data where the standard deviation is proportional to the mean or where the effects are multiplicative. These conditions are generally found in data that are whole numbers and cover a wide range of values. Data on the number of insects per plot or the number of egg masses per plant(or per unit area) are typical examples.

To transform a data set into the logarithmic scale, simply take the logarithm of each and every component of the data set.

Logarithmic transformation is effective in converting multiplicative effect to additive effect.

If the data set involves small values(e.g. less than 10), log(**x**+1) should be used instead of log**x**, where **x** is the original data.

3.9.2.Square Root Transformation

Square -Root transformation is appropriate for data consisting of small whole numbers, for example, data obtained in counting rare events, such as the number of infested plants in a plot, the number of insects caught in traps, or the number of weeds per plot. For these data, the variance tends to be proportional to the mean.

The Square -Root transformation is also appropriate for percentage data where the range is between 0 and 30% or between 70 and 100%. For other ranges of percentage one can use arc sine transformation.

If most of the values in the data set are small(e.g. less than 10), especially with zeroes percent, $(x + 0.5)^{1/2}$ should be used instead of $x^{1/2}$, where x is the original data.

3.9.3.Arc sine Transformation

An arc sine or angular transformation is appropriate for data on proportions, data obtained from a count, and data expressed as decimal fractions or percentages. Note that percentage data that are derived from count data, such as percentage barren tillers (which is derived from the ratio of the number of nonbearing tillers to the total number of tillers), should be clearly distinguished from other types of percentage data, such as percentage protein or percentage carbohydrates, which are not derived from count data.

The mechanics of data transformation are greatly facilitated by using a table of the arc sine transformation. The values of 0% should be substituted by (1/4n) and the values of 100% by (100-1/4n), where n is the number of units upon which the percentage data was based (i.e. the denominator used in computing the percentage).

Not all percentage data need to be transformed and, even if they do, arc sine transformation is not the only transformation possible. The Square-Root transformation is occasionally used for percentage data.

3.10. Software

GLM's can be fitted and evaluated using S-PLUS, SAS, SPSS, and a number of other statistical packages. Of the major packages, S-PLUS and SAS provide greater flexibility in fitting and evaluating GLM's. A brief comparion of SAS and SPSS is discussed in the next section.

3.10.1. Estimation in SAS and SPSS

3.10.1.1. Fitting the Probit Model using SAS

For most GLMs, the equations that determine the ML parameter estimates are non-linear, and the estimates do not have a closed- form expression. Software calculates the estimates using an iterative algorithm for solving nonlinear equations. The algorithm requires an initial guess(but softwares for GLMs doesn't require the user to provide an initial guess) for the parameter values that maximize the likelihood function. Successive approximations produced by the algorithm tend to fall closer to the ML estimates. A popular algorithm for doing this, called Fisher scoring, was first proposed by R.A.Fisher for fitting probit models. For binomial logistic regression and poisson loglinear models, Fisher scoring simplifies to a general-purpose method called the *Newton- Raphson* algorithm.

The Newton- Raphson algorithm approximates the log-liklihood function in a neighborhood of the initial guess by a simpler polynomial function that has the shape of a concave (mound-shaped) parabola. It has the same slop and curvature at the initial guess as does the log-likelihood function. It is simple to determine the location of the maximum of this approximating polynomial. That location comprises the second guess for the ML estimates.

The Newton- Raphson method utilizes a matrix, called the information matrix, that provides asymptotic standared error(ASE) values for the parameter estimates. The matrix is based on the curvature of the log-liklihood function at the ML estimate. The greater the curvature, the greater the information about the parameter values. The standard errors are the square roots of the diagonal elements for the inverse of the information matrix. The greater the curvature of the log likelihood, the smaller the standard errors. This is reasonable, since large curvature implies that the log liklihood

drops quickly as β moves away from $\hat{\beta}$; hence, the data would have been much more liklely to occur if β took value $\hat{\beta}$ than if it took some value not close to $\hat{\beta}$. Software for GLMs routinely calculates the information matrix and the associated standard errors.

An informal description of the method used to fit a probit model in the SAS PROBIT procedure according to G.J. Johnston,(1993) is described as follows**.**

The SAS PROBIT procedure uses a modified Newton-Raphson algorithm to fit the probit model. Suppose, for simplicity, that the data are binary (success-failure) and that there is a single covariate x. Then the probit model for the probability of success is

$$\mathbf{p_i = \Phi(a + bx_i)}$$

for observation *i* where a and b are unknown parameters to be estimated and Φ is the standard normal cumulative distribution function. PROC PROBIT fits more complicated models (multinomial response, threshold parameter, and more covariates), but this simpler model extends directly to more complicated ones and should give you an idea of how SAS fits the probit model.

The Newton-Raphson step for updating the parameter vector $\theta^{(j)}$ at the j^{th} step in the iterative fitting process is

$$\mathbf{\theta^{(j+1)} = \theta^{(j)} - H^{-1}g}$$

where **H** is the matrix of second derivatives of the log likelihood function with respect to the parameters and **g** is the vector of first derivatives. The likelihood function is

$$L = \prod_i \binom{n_i}{r_i} P_i^{r_i} \left(1 - P_i\right)^{n_i - r_i}$$

(where n_i is the number of trials, r_i is the number of successes on observation i, and, p_i is the probability of success of the i^{th} observation) and after dropping the binomial coefficient which has no effect on parameter estimates or covariances, the log likelihood, l , is

$$l = \sum_i [r_i \log(p_i) + (n_i - r_i)\log(1 - p_i)]$$

where $\mathbf{n_i}$ is the number of trials and $\mathbf{r_i}$ is the number of successes on observation i, and the product and sum are over observations.

You need to compute the first and second derivatives of the log likelihood function to implement the Newton-Raphson algorithm. For any parameter, the derivatives are

$$\frac{\partial l}{\partial \theta} = \sum_i \left[\frac{r_i}{p_i} - \frac{n_i - r_i}{1 - \mathrm{P}_i} \right] \frac{\partial \mathrm{P}_i}{\partial \theta}$$

and

$$\frac{\partial^2}{\partial \theta_1 \partial \theta_2} = \sum_i \left[-\left(\frac{r_i}{\mathrm{P}_i^2} + \frac{n_i - r_i}{(1 - p_i)^2} \right) \frac{\partial p_i}{\partial \theta_1} \frac{\partial p_i}{\partial \theta_2} + \left(\frac{r_i}{p_i} - \frac{n_i - r_i}{1 - p_i} \right) \frac{\partial^2 p_i}{\partial \theta_1 \partial \theta_2} \right]$$

Note that, from the definition of the probit model,

$$\frac{\partial p_i}{\partial a} = \phi(a + bx_i)$$

$$\frac{\partial p_i}{\partial b} = x_i \phi(a + bx_i)$$

$$\frac{\partial^2 p_i}{\partial^2 a} = -(a + bx)\phi(a + bx)$$

and so on, where ϕ is the standard normal probability density function. Inserting these into the expressions for the derivatives of the log likelihood function and replacing **a** and **b** with their current estimates gives the formulas to use in the Newton-Raphson step. This is the basic iteration step of the algorithm that SAS uses to fit the probit model.

SAS estimates the covariance matrix of the parameter estimates with the inverse of the negative of the matrix of second derivatives, or the inverse of the *observed* information matrix, evaluated at the last iteration. Notice that the expected value of the second derivative is

$$\mathrm{E}\frac{\partial^2 \iota}{\partial\theta_1\partial\theta_2} = \sum_i \left[\frac{n_i}{p_i(1-p_i)} \frac{\partial p_i}{\partial\theta_1} \frac{\partial p_i}{\partial\theta_2} \right]$$

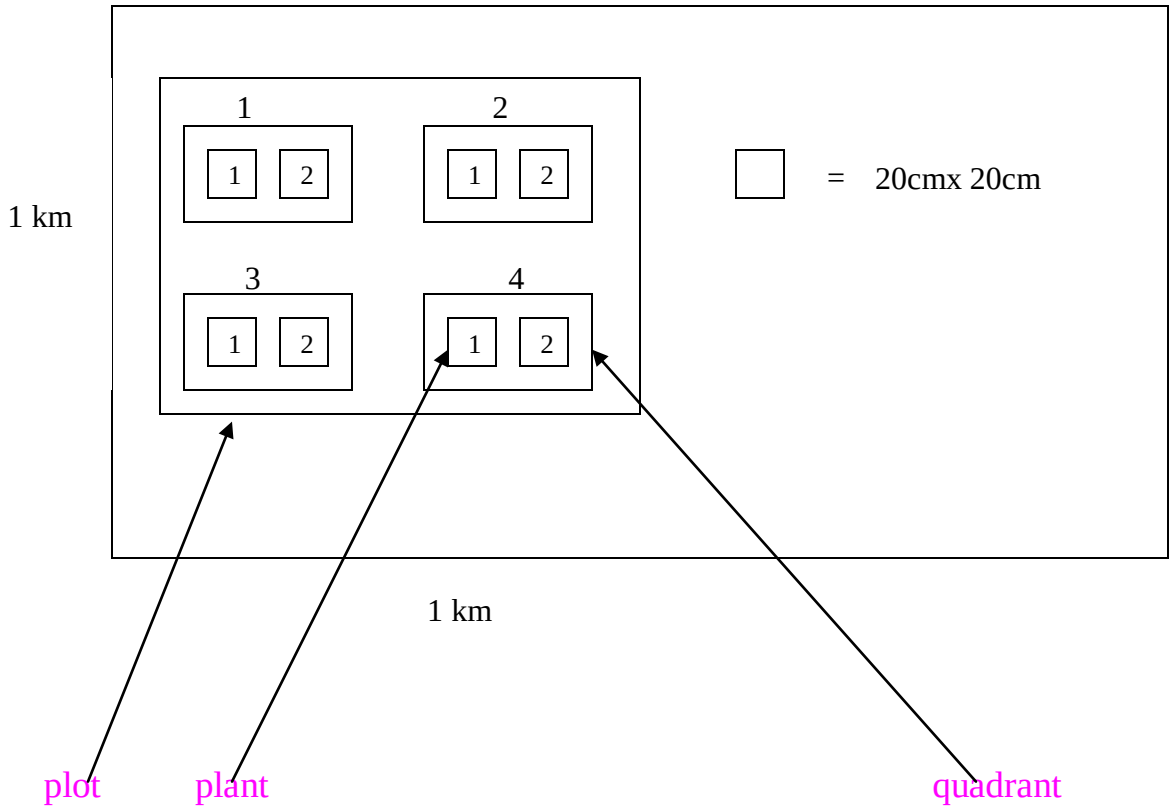
You can replace the matrix of second derivatives with the matrix of expected values of second derivatives and use this in the Newton-Raphson algorithm. The covariance matrix is then estimated by the inverse of the *expected* information matrix. Finney uses this method, and some statistical computing packages also use it. The two methods will generally converge to the same parameter estimates, but the estimates of the covariance matrix will be slightly different.

Chapter Four- Results and Discussion

4.1. Data type and Source

The data type for this research paper is a parthenium count data(the number of parthenium and other plant species) on a one square kilo meter area at welenchti. The data source is Ethiopian Institute of Agricultural Research. The process of counting species had conducted as follows. The one square kilo meter area divided into 10 equal places or columns, and each column is 100 meters wide. Then in each column a one meter by one meter area called ' plot ' were created and a total of 100 plots formulated. Within each plot 4 quadrants are available; each quadrant is a 20cm by 20 cm area. Within each quadrant there are two plants, plant1 and plant2, there are a total of 800 plants in the data set. Plant in each quadrant is the smallest fractional area in the whole experiment. The count of the number of species (including parthenium plant) in each quadrant is conducted and registered sequentially. The graphical explanation is presented below:-

Figure 3-the data collection layout



The data used in this way is presented in **APPENDIX** H.

4.2 Descriptive Statistics

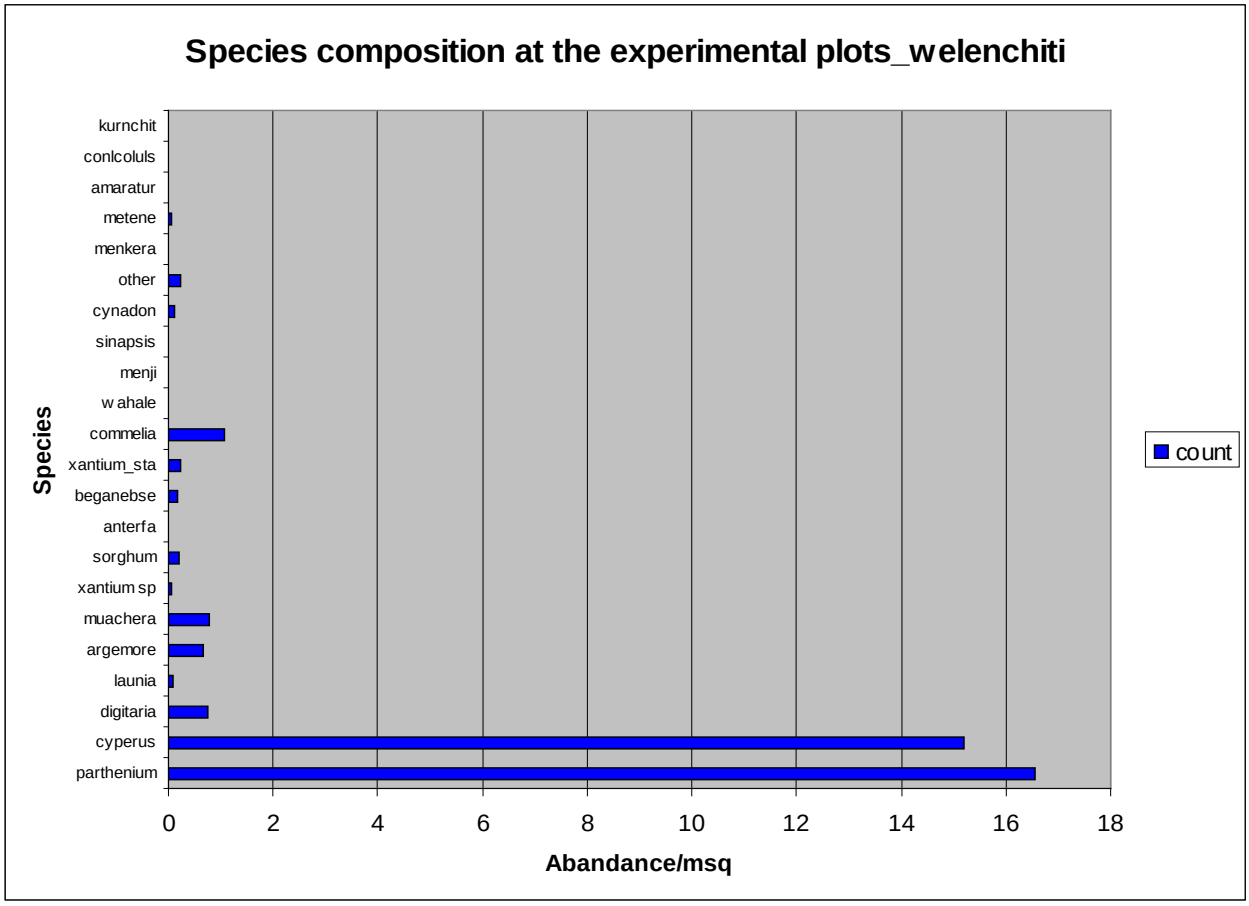
The dependent variable is-

1. Severity of parthenium- explained interms of the parthenium count in each quadrant, or in the Multinomial model as categorical variable categorized by infestation level.
2. absence/presence of parthenium- mainly used in logistic and probit models.

The independent variables are – plot, quadrant, plant, and sum of all other species per quadrant. The last independent variable is used in multinomial model.

The following histogram shows the aboundence of species including parthenium in each plot and the dominance power of parthenium and the rare occurrence of many other species.
The following Histograph shows the count data of species per quadrant overall area.

Figure 4. A Barchart of Number of all species per quadrant.



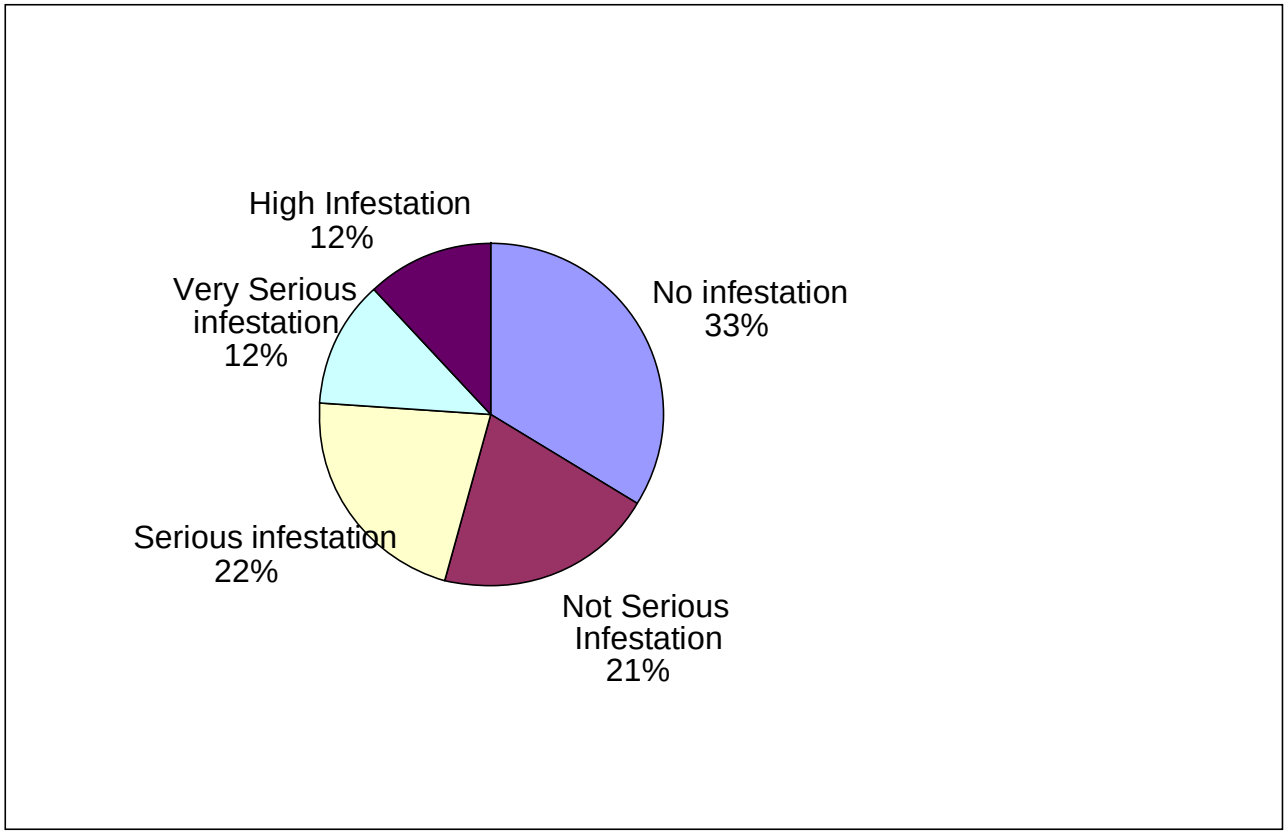
As the graph shows the abundance of parthenium per meter square is more than 16 plant per meter square followed by cyperus which is around 15 per meter square. The rest species are very rarely observed. This indicates that parthenium is the most dominant species in the area. This is because of the high competitive nature of parthenium weed. This can also be seen from the following table.

Table 2- Overall Dominance power of Parthenium and other species in 100 plot areas

S.no.	Scientific name	Number of Occurrence in total area of plots A	Total number of species in 100 plots B	Dominance Index (B/171,96)*100) 171.96 = sum of 71 species(all species) dominance index
1	*Echinochloa crus-galli L.*	68	3368	1958.60
2	*Euphorbia hirta*	21	225	130.84
3	*Euphorbia peplus*	17	321	186.67
4	*Leucas martinicensis*	4	13	7.56
5	*Oplismenus hirtellus*	1	2	1.16
6	*Oxygonum sinuatum*	1	1	0.58
7	*Parthenium hysterophorus L.*	95	8413	4892.42
8	*Phylanthus amure*	4	13	7.56
9	*Sample 2*	1	8	4.65
10	*Sample 3*	25	524	304.72
11	*Sample 6*	2	19	11.05
12	*Scorpiurus muricatus*	2	13	7.56
13	*Setaria pumila*	12	37	21.52
14	*Setaria verticillata*	10	50	29.08
15	*Trichodesma zeylanicum*	1	1	0.58
16	*Vicia spp*	2	5	2.91
17	*Viginia fisheries*	1	1	0.58
18	*Withania somnifera*	1	52	30.24
19	*Xanthium strumarium*	2	2	1.16

As shown in the table above, the dominance as well as the abundance of parthenium species is much higher than other species. Its occurrence is 95% which is greater by 27 % from the 2nd most abounden specie and higher by 94 % from the least abundant species. Parthenium dominance index also revealed its dominance power is unreachable by other species creating a minimum gap of 2892.42% . In relation to this, the following graph displays the abundance of parthenium overall the area**:**

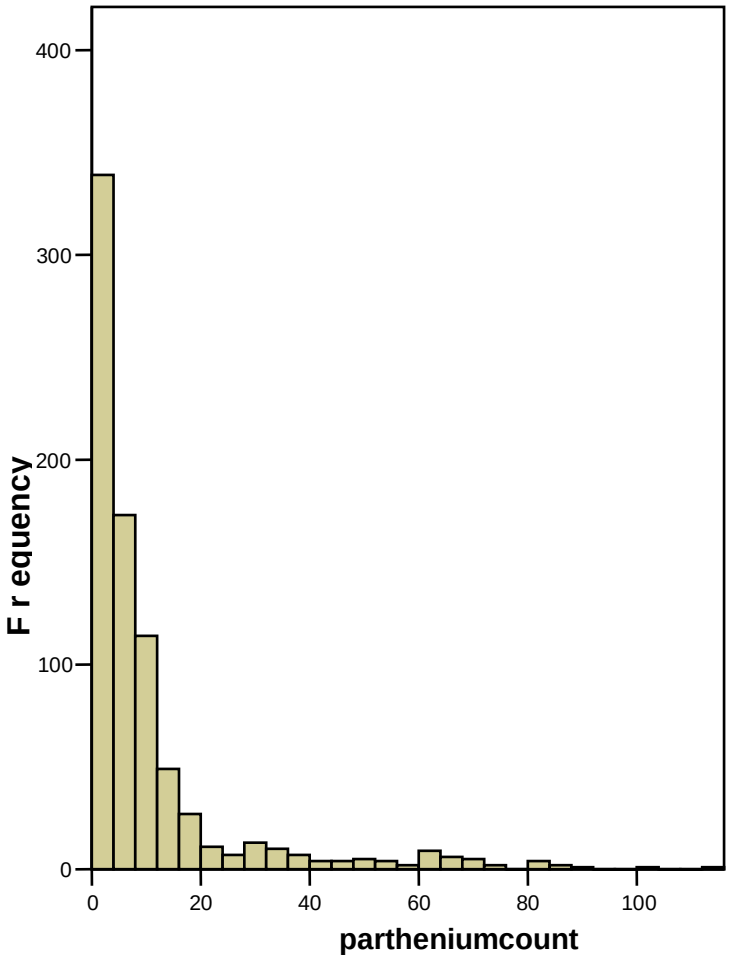
Figure 5. Level of Parthenium abundance



As the pie chart indicate around 70 % of the area are infested by parthenium. Serious infestation which is 22% gets the lead and the rest infestation levels are not that simple to dominate other species especially crops. These descriptive statistics show the probability of getting parthenium plant in each plot is much higher than other species, which illustrate the greater dominance of parthenium.

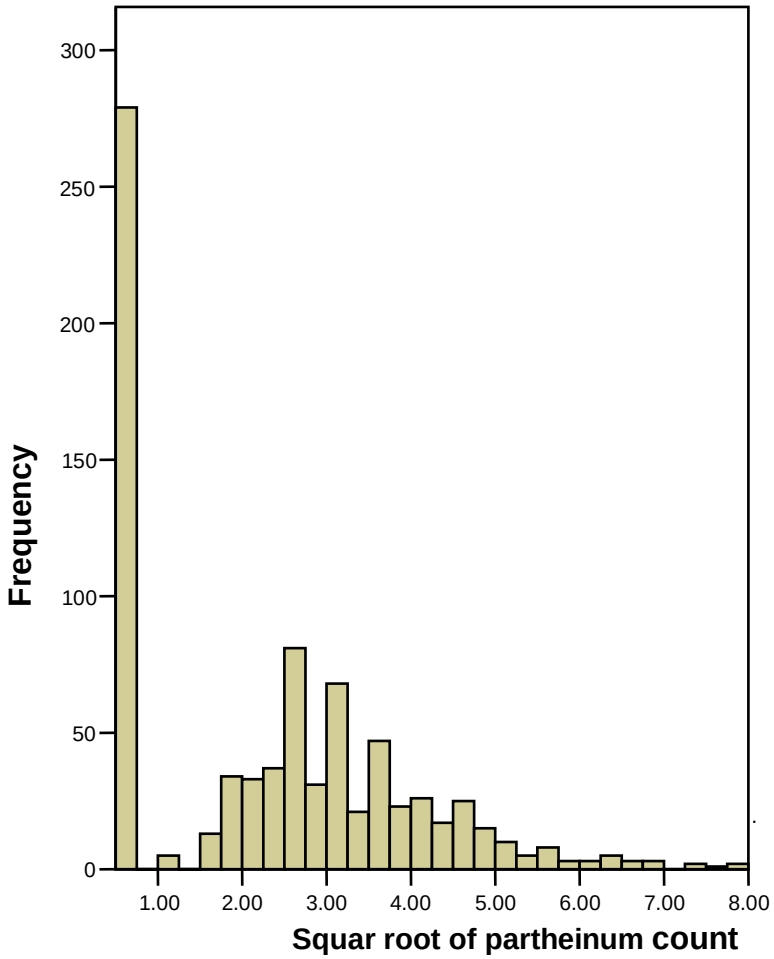
In the above discussion, it is mentioned that how parthenium widely spread and affect other species. The following discussion concentrates on parthenium count per quadrant in relation to the variable plant. The figure below presents the histogram of original parthenium count data.

Figure 6. Histogram of parthinium data(before transformation)



The graph revealed the data is extremely left skewed. In statistical terms this is an indication of the fact that the data is not normal, so we go for transformation. The appropriate transformation is square root transformation. The bar graph of the data obtained by square root transformation is displayed figure 8.

Figure 7. Histogram for a square root transformed Parthenium data



The parthenium variable clearly does not appear to follow a normal distribution, and we would not expect normality, because the outcome variable represents a count (with several zeroes). It is an indication to see an option to fit the data using generalized linear model. We therefore consider a generalized linear model for the parthenium outcome, with a distribution for the outcome variable that is appropriate for count data. Two commonly used distributions for count outcomes are the Poisson distribution and the Negative Binomial distribution among all other families of GLMs.

Suppose scaling number of parthenium in each quadrant as follows:-

Table 3- Level of parthenium infestation.

Level	Range of no. of parthenium in each quadrant	Category of Infestation	count	percentage
1	0	No infestation	269	33.6
2	1 to 5	Not Serious Infestation	164	20.5
3	6 to 10	Serious infestation	176	22
4	11 to 20	Very Serious infestation	96	12
5	> 20	High Infestation	95	11.9

As infestation gets stronger, the area to be covered also gets wider. When parthenium gets a chance to continue, the consequence will be total infestation and total crop loss.

4.3 Analysis of variance

The purpose of analysis of variance (ANOVA) is to test for significant differences between means by comparing (i.e., analyzing) variances. More specifically, by partitioning the total variation into different sources (associated with the different effects in the design), we are able to compare the variance due to the between-groups (or treatments) variability with that due to the within-group (treatment) variability. Under the null hypothesis (that there are no mean differences between groups or treatments in the population), the variance estimated from the within-group (treatment) variability should be about the same as the variance estimated from between-groups (treatments) variability.

As explained in the above sections, the dependent variable is parthenium count and the independent variables are plot, quadrant and plant. Then the model structure which shows dependent and independent variables is:

Y(partheniumcount) = plot

Y(partheniumcount) = plot + Quadrant(plot)

These alternative models will be fitted using SAS to investigate the contribution of each independent variable.

Since, each plant is nested within each quadrant and each quadrant is within each plot, the approach should follow a Nested model approach.

Using these two models, the result of the standard ANOVA obtained from SAS(APPENDIX G) is given in table 4.

Table 4- Standared ANOVA table

Source	DF	Anova SS	Mean Square	FValue	Pr>F
Plot	99	65983.92375	666.50428	4.02	<0.0001
Quadrant(Plot)	300	75575.87500	251.91958	1.52	<0.0001
Error	400	66237.50	165.5938		
Total	799	207797.298			

Based on the ANOVA in table 4, there is significant difference between the 100 plots and the four quadrants per plot. According to this result the model is also significant. But since the data is a count data, this conclusion might not be correct and there needs to be strong diagnostic supportive statistical results to reach at the right conclusion.

4.4 Diagnostics

Before fitting the data using GLMs to show the functional relationship between the dependent and the independent variable the data was examined as follows using different diagnostic methods.

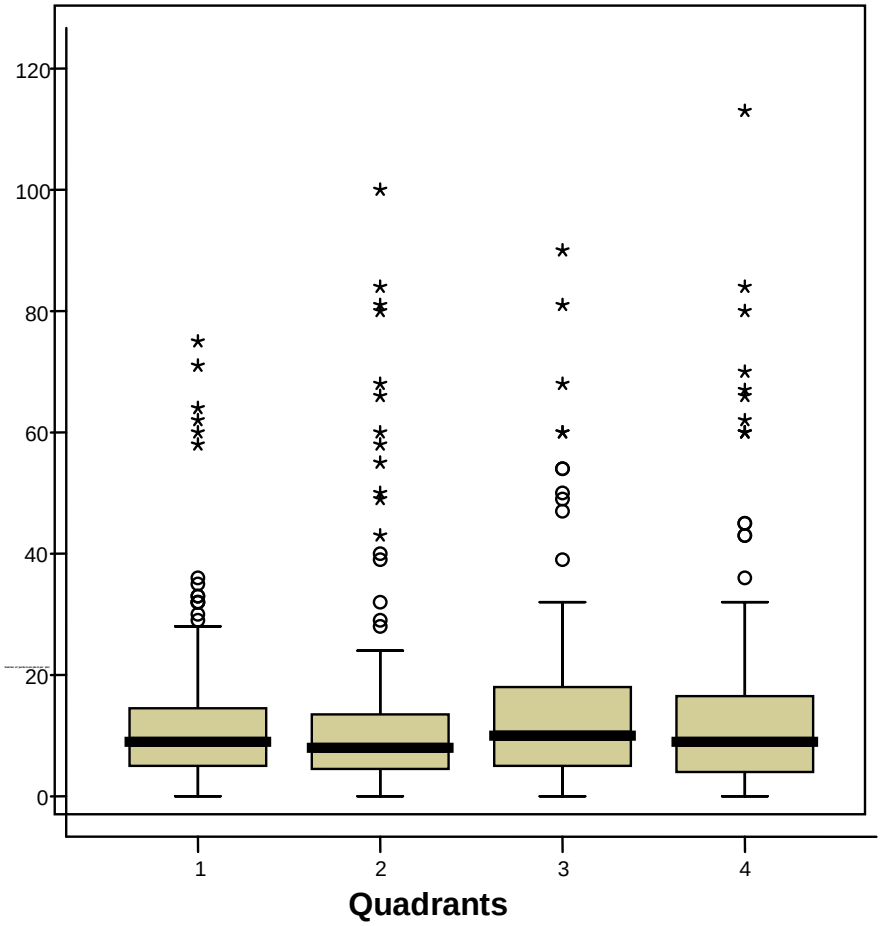
4.4.1 Normal Test

Normal distributions take the form of a symmetric bell-shaped curve. Normality can be visually assessed by looking at a histogram of frequencies, or by looking at a normal probability plot output by most computer programs.

Box plots are a graphical way of testing for lack of homogeneity of variances, normality, and presence of outliers. One requests side-by-side boxplots of each group, such that samples form the x axis. The more the width of the boxes varies markedly by sample, the more the assumption of homogeneity of variances is violated.

Boxplot tests of the normality assumption: Box plot produces charts in which the Y axis is the interval dependent and categories of the independent are arrayed on the X axis. The graph shows that, for each variable, there is a rectangle indicating the spread of the values for that category. If these rectangles are roughly at the same Y elevation for all categories, this indicates little difference among groups. Within each rectangle is a horizontal dark line, indicating the mean. If most of the rectangle is on one side of the mean or the other of the mean line, this indicates that the number of parthenium plant per plot is skewed (not normal) for that quadrant (category). Below is Boxplot of the data.

Figure 8. Boxplot of Number of partienium plant per plot



Based on previous discussion regarding the above box plot, there is evidence that the data (number of parthenium plant per plot in each quadrant) may not be normally distributed and that corrective measure is necessary.

Here are histograms for natural logarithem transformations and square root transformations. One can easily compare and clearly see which transformation stablized the variance of the variable of interest.

Figure 9-Histogram for Log transformed Data

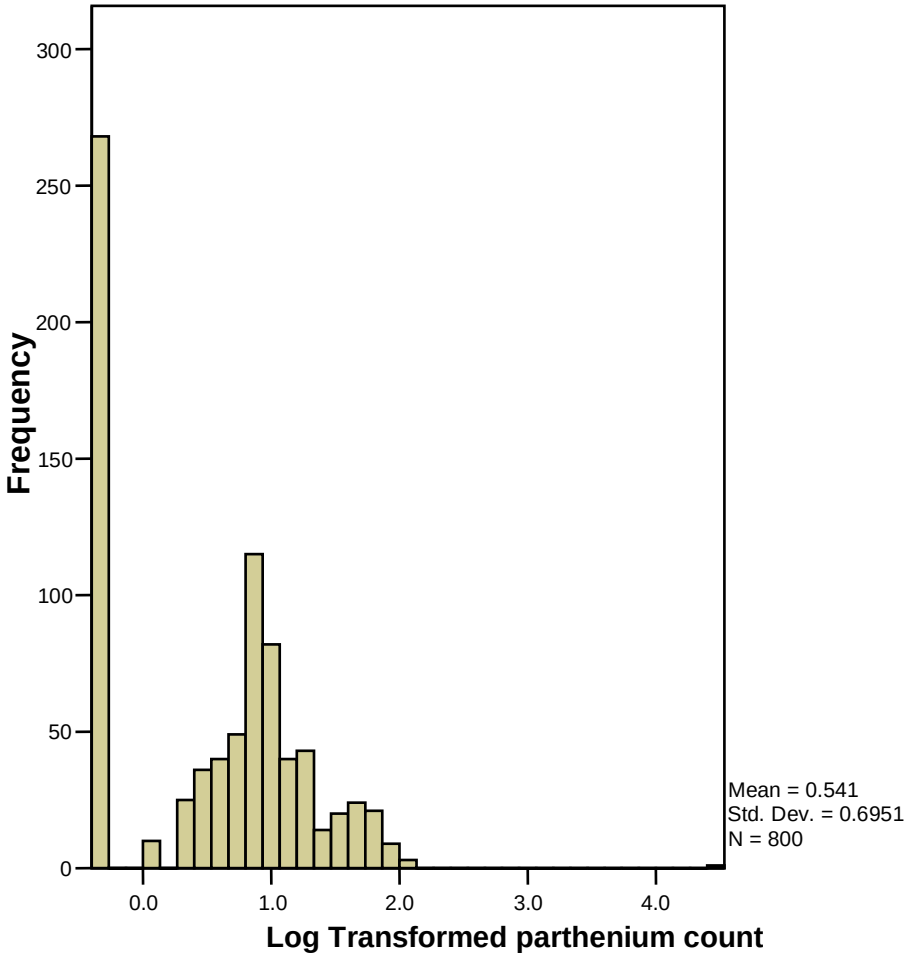
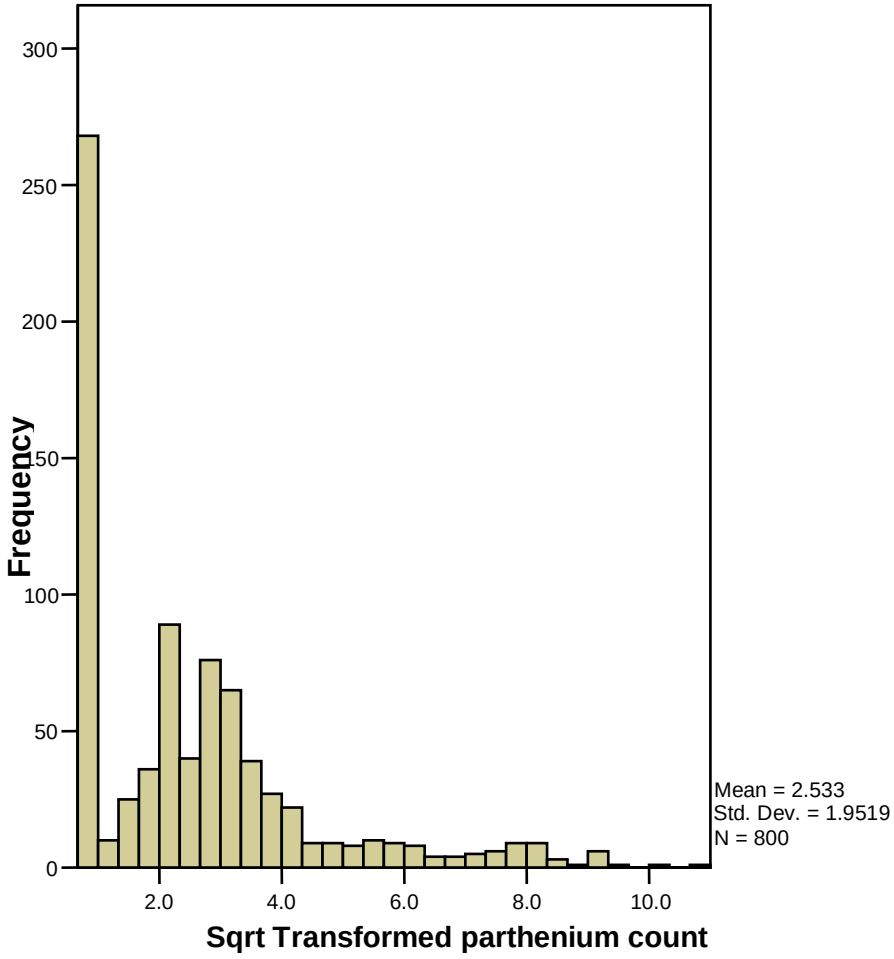


Figure 10-Histogram for SQRT transformed Data



From these two consecutive graphs, (Figure 9 and 10), the standard deviation(0.6951) obtained by log transformation is smaller than the one obtained by SQRT (1.9519). Hence, log transformation seems to stablize the variance better than the square root. Hence log data is used for this analysis.

4.5. Model Fitting

In this section, we deal with generalized linear models for the data under consideration. The fitting process used presence of parthenium as dependant variable, and plant and quadrant variables as independent variables. The complete SAS outputs used for this purpose are attached (annexes A to G) only the summaries are presented in this section.

4.5.1.Fitting Logistic model

Logistic regression is a statistical regression model for Bernoulli-distributed dependent variables. Hence, presence or absence of the dependent variable Parthenium in each quadrant can be modeled by logistic model.

The Logistic model is

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i},$$

$i = 1, \ldots, n,$ and the paramter estimates obtained using SAS software output results summary are displayed below.

Table 5-SAS Logistic Procedure output summary(see APPENDIX A)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4693	0.2982	24.2847	<.0001
Quadrant	1	0.0432	0.0675	0.4101	0.5219
Plant	1	0.4411	0.1512	8.5078	0.0035
Odds Ratio Estimates					
Effect		Point Estimate	95% Wald Confidence Limits		
Quadrant		1.044	0.915	1.192	
Plant		1.554	1.156	2.091	

From the result above, intercept and plant parameter estimates are significantly different from zero and can be included in the model i.e. Plant has a significant rising effect on parthenium, and the Quadrant effect fails to achieve the p<.05 level. On the other hand, since the coefficient of quadrant is insignificant, the occurrence of parthenium has no association with quadrants, the variable can't be included in the model.

The resultant model looks like the following:-

$$\text{model: } \text{logit}(p_i) = -1.4693 + 0.4411x$$

where x represent plant. Finally, the probability of a positive response is

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)} = \frac{\exp(-1.4693 + 0.4411x)}{1 + \exp(-1.4693 + 0.4411x)}$$

The odds of an event is defined as the ratio of the probability that an event occurs to the probability that it fails to occur.

The positive large coefficients indicate that the occurrence of parthenium together with the variable plant is strong or highly associated with the existence of parthenium. Even though it doesn't give much sense to assign the value zero for x for this particular data, one can see some mathematical interpretations as follows. In this logistic model the predicted probability for the presence of parthenium free plot is $\{e^{-1.4693}/(1+ e^{-1.4693})\}= 0.1871$, when the value of plant(or x) is zero. Using this predicted probability the ratio of probability of no parhenium(success) which is 0.1872 over the probability of there is parthenium(failure) which is 1-0.1871= 0.8129 is 0.2302 i.e. {(0.1872)/(1-0.1872)}= 0.2302. This is the odds of response 1(i.e. the odds of a "success") and is equal to $e^{\alpha} = e^{-1.4693} = 0.2302$ for zero value of x. This probability which is success probability divided by failure is very small. This shows that the probability of getting parthenium free plot is very small and parthenium is highly dominant over the area. In other words the probability of getting parthenium infested plot is 0.8129. The odds ratio for quadrants is 1.044, which is almost equal to 1. It implies that the odds for the presence of parthenium are independent of quadrant. This shows that there is no

strong associations between parthenium infestation and quadrants. The odds ratio for plants is 1.554, which shows that the odds for plant1 to be free of parthenium is 1.554 times that of the odds for plant2.

Since the variable quadrant is dropped from the fitted model, the model needs to be re-estimated, i.e. we need to fit a simple logistic regression model. In doing so, the SAS outputs are displayed below, (from appendix A1).

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3606	0.2442	31.0311	<.0001
plant	1	0.4408	0.1512	8.5036	0.0035

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
plant	1.554	1.156	2.090

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x)} = \frac{\exp(-1.3606 + 0.4408x)}{1+\exp(-1.3606 + 0.4408x)}$$

As the re-estimated model results just above reveals, parameter estimates are somehow different from the pervious estimated model. But the odds ratios are equal.

Using this re-estimated model, the predicted probability for the presence of parthenium free plot is $\{e^{-1.3606}/(1+ e^{-1.3606})\}= 0.2041$,when the value of plant(or x) is zero. The odds of a response 1 (i.e. the odds of a "success") which is the ratio of probability of no parhenium(success), 0.2041, over the probability of there is parthenium(failure), which is 1- 0.2041=0.7959 will be 0.2564 i.e {(0.2041)/(1-0.2041)}=0.2564. This probability which is success probability divided by failure is very small, but there is little improvement in increasing the ratio in this re-estimated model.

An alternative model to fit the data is the probit. The probit model is a GLM with binomial random component and probit link. Parameter estimates differ for the two models(logistic and probit), since their links have different scales. When both models fit well, slope estimates in logistic regression models are roughly about 1.6 to 2.0 times those in probit models (Internet) . To see the difference between the two the following sections present the fit of a probit model for the same data.

4.5.2.Fitting Probit model

Using SAS software PROBIT command, estimates of the linear probit model can be obtained, where the dependent variable takes on only two values. PROBIT, uses analytic first and second derivatives to obtain maximum likelihood estimates via the Newton-Raphson algorithm. The numerical implementation involves evaluating the normal density and cumulative normal distribution functions.

The cumulative normal distribution function is computed from an asymptotic expansion, since it has no closed form.

The probit model is defined as

$$\mathbf{Pr(y=1|x) = \Phi(xb)} = \beta_0 + \beta x$$

where **Φ** is the standard cumulative normal probability distribution and **xb** is called the probit score or index. To obtain parameter estimates for the probit model a SAS software output results are displayed below-

Table6- SAS probit Procedure output summary(APPENDIX B)

Probit model analysis					
The LOGISTIC Procedure					
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9038	0.1803	25.1321	<.0001
Quadrant	1	0.0264	0.0411	0.4127	0.5206
Plant	1	0.2693	0.0921	8.5554	0.0034

From Table6, it is only plant that is statistically significant at 0.05 alpha level of significance, meaning that the regression coefficient for plant is significantly different from zero and to be included in the model. Hence, using this result the model fitted presented as follows:-

Probit model: $\mathbf{Pr(y=1|x) = \Phi(bx)}= -0.9038 + 0.2693x$

Even though it doesn't give much sense to assign the value zero for x for this particular data, one can see some analogues mathematical interpretations as follows as it was mentioned in logistic model. The fitted probit for this model equals -0.9038. The fitted probability **Φ(bx)** is to the left-tail probability for the standard normal distribution at 0.9038, which equals 0.1736(0.5- 0.3264(z value at 0.9038)). The mean for the normal cdf is 0.9038/0.2694=3.4 and the standard deviation is 1/(/0.2694/)= 3.7. The predicted probability of infestation equals half at plant 3.4; that is, x=3.4 has a fitted probit of 0.9038 – 0.2694(3.4)=0, which is the z-score corresponding to a left-tail probability of 0.5. The fitted probit value of 0.9038 means that the mean is 0.9038 standard deviations below the mean of a normal distribution with mean 3.4 and standard deviation 3.7.

Similarly, Since the variable quadrant is dropped from the fitted model, the model needs to be re-estimated, i.e. we need to fit a simple probit model. In doing so, the SAS outputs are displayed below, (from appendix B1).

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8372	0.1474	32.2435	<.0001
plant	1	0.2691	0.0921	8.5444	0.0035

Using this result the model fitted presented as follows:-

Probit model: $\mathbf{Pr(y=1|x) = \Phi(bx)}= -0.8372 + 0.2691x$

The refitted probit model (just above) parameter estimates are different from the previous fitted probit model parameter estimates. Similar interpretations with limitations that it doesn't give that much sense can be done.

The fitted probit for this model equals -0.8372. The fitted probability **Φ(bx)** is to the left-tail probability for the standard normal distribution at 0.8372, which equals 0.1736{(0.5- 0.3133(z value at 0.8372))}. The mean for the normal cdf is 0.8372/0.3133=2.7 and the standard deviation is {1/(/0.3133/)}= 3.2. The predicted probability of infestation equals half at plant 2.7; that is, x=2.7 has a fitted probit of {0.8372 – 0.3133(2.7)}= 0, which is the z-score corresponding to a left-tail probability of 0.5. The fitted probit value of 0.8372 means that the mean is 0.8372 standard deviations below the mean of a normal distribution with mean 2.7 and standard deviation 3.2.

4.5.3.Fitting Poisson Loglinear Model

Poisson regression is often used to fit rare occurrence using count data. The number of parthenium is a count data and Poisson regression can be applied to the data. Hence, the following discussion shows how the Poisson regression applied to this count data.

The Poisson loglinear model has a form

$$\log(\mu) = \text{intercept} + b_1*X_1 + b_2*X_2 + ...+ b_m*X_m,$$

or equivalently, in terms of exponential relationship, $\mu = \exp(\beta_0 + \beta x) = e^{\beta_0}(e^{\beta})^x$ where μ is mean of parthenium count(Y), X_1 is for variable plant and X_2 is for variable quadrant; β_0 is an intercept; b_i's and β are regression coefficients.

Using SAS software, parameter estimates outputs (APPENDIX C) are displayed below;

Table7- SAS Poisson Procedure output summary(APPENDIX C)

Analysis of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.7780	0.0283	1.7225	1.8335	3939.22	<.0001
Quadrant 1	1	-0.0698	0.0318	-0.1322	-0.0074	4.81	0.0283
Quadrant 2	1	-0.0620	0.0318	-0.1242	0.0003	3.81	0.0510
Quadrant 3	1	-0.0480	0.0316	-0.1100	0.0140	2.30	0.1290
Quadrant 4	--	0.0000	0.0000	0.0000	0.0000	.	.
Plant1	1	0.8997	0.0249	0.8509	0.9486	1302.86	<.0001
Plant2	--	0.0000	0.0000	0.0000	0.0000	.	.
Scale	--	1.0000	0.0000	1.0000	1.0000		

From table 7, the intercept, quadrant1, quadrant2 and plant1 parameters are significant. The likelihood ratio test statistics allow one to test the null hypothesis that including a given predictor in a model is not improving the fit of the model. In the LR section of the SAS output(see APPENDIX C), plant is significant and quadrant is insignificant. Hence, only plant and the intercept are to be included in the model. But before using these parameters in the model, assessing goodness of fit (using the outputs from SAS) is advisable. The criteria output looks the following.

Table8- Criteria For Assessing Goodness Of Fit for poisson regression(SAS output)

Criterion	DF	Value	Value/DF
Deviance	705	11945.8639	16.9445
Scaled Deviance	705	11945.8639	16.9445
Pearson Chi-Square	705	20799.7760	29.5032
Scaled Pearson X^2	705	20799.7760	29.5032
Log Likelihood	10760.9766		

A good diagnostic check of whether or not the Poisson distribution is a good fit for this count outcome is to investigate the Value/DF for the model Deviance (Which is similar to the sum of squares for non-normal outcomes). Basically, a model with a good fit to the data will have a Value/DF close to 1 or below 1, but this value in this case is nearly 17, suggesting a very poor fit. This can be noticed in the initial data summary that the variance of the count of parthenium was larger than the mean, and these types of variables are often referred to as “over-dispersed” count variables as a result. A distribution often used for count outcomes with this characteristic is the negative binomial distribution.

In situations where the model fit is adequate, estimated parameters can be interpreted as follows. A one-unit increase in number of plant has a multiplicative impact of e^{β} on μ. The mean of partheniumcount(Y) at plant+1 equals the mean of partheniumcount(Y) at plant multiplied by e^{β}. If β=0, then $e^{\beta} = e^{0} =$ 1 and the multiplicative factor is 1. This means that the mean of partheniumcount(Y) does not change as plant changes. If β is positive, then e^{β} is greater than one and the mean of partheniumcount(Y) increases as plant(x) increases. If β is negative, then the mean decreases, by e^{β} times μ, for a unit change in plant.

Before proceeding to the next section, we will fit and discusses Poisson regression for another species known as chifirg as dependent variable. The Poisson regression, as explained earlier, is used to fit a rare event. In this data set there are chifirg species considered as occuring rarely in the field.

Hence, one can apply Poisson regression for these particular variable. Using SAS software, parameter estimates outputs (APPENDIX I) are displayed below.

Table9- SAS Poisson Procedure output summary(APPENDIX I)

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.3139	0.0601	0.1960	0.4317	27.24	<.0001
Quadrant1	1	-0.6225	0.0797	-0.7787	-0.4664	61.03	<.0001
Quadrant2	1	-0.4697	0.0761	-0.6188	-0.3206	38.11	<.0001
Quadrant3	1	0.3898	0.0610	0.2703	0.5093	40.86	<.0001
Quadrant4	--	0.0000	0.0000	0.0000	0.0000	.	.
Plant1	1	0.8308	0.0537	0.7255	0.9360	239.28	<.0001
Plant2	--	0.0000	0.0000	0.0000	0.0000	.	.

From Table 9, the intercept, quadrant1, quadrant2, quadrant3, quadrant4, plant1 and plant2 parameters are significant; meaning that there is significant difference in the occurrence of the species chifirg between quadrants and plants. Hence, all variables and the intercept are to be included in the model. But before using these parameters in the model, assessing goodness Of fit (using the outputs from SAS) is advisable. The criteria output is shown in Table 10.

Table10- Criteria For Assessing Goodness Of Fit for poisson regression(SAS output)

Criterion	DF	Value	Value/DF
Deviance	794	8978.3933	11.3078
Scaled Deviance	794	8978.3933	11.3078
Pearson Chi-Square	794	38682.3777	48.7184
Scaled Pearson X2	794	38682.3777	48.7184
Log Likelihood		-199.2097	

From Table 10, a Value/DF is close to 11, suggesting once again a very poor fit. Hence, Poisson regression is not a good fit for assessing the association of the occurrence of the chifirg species with quadrants and plants. As mentioned before, the distribution often used with this kind of situation is the negative binomial distribution. In situations where the model fit is adequate, estimated parameters can be interpreted as mentioned above. The next section discusses how to fit and interpret a negative binomial model.

4.5.4.Fitting Negative Binomial Model

The negative binomial regression model adds an "overdispersion" parameter to estimate the possible deviation of the variance from that expected under the Poisson. This has the consequence of generating an estimate of standard errors and may modify parameter estimates.

The first step in this case is to assess the goodness of fit criteria (obtained using SAS).

Table11- Criteria For Assessing Goodness Of Fit for Negative Binomial regression.

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	795	880.4206	1.1074
Scaled Deviance	795	880.4206	1.1074
Pearson Chi-Square	795	692.9427	0.8716
Scaled Pearson X2	795	692.9427	0.8716
Log Likelihood	15795.7852		

There is a marked improvement in the Value/DF criterion for the model Deviance, from 17 in Poisson regression to 1 in negative binomial regression. A model with deviance value 1 or very close to 1 is a better fit. This condition of adequacy is attained for this model. Now there is enough reason to believe that this model fits the data well (or at least better than the Poisson distribution).

Table12-Summary of SAS output for the parameter estimates and confidence interval;

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.7508	0.1296	1.4968	2.0047	182.59	<.0001
Quadrant1	1	-0.0200	0.1612	-0.3360	0.2960	0.02	0.9012
Quadrant2	1	-0.0366	0.1611	-0.3524	0.2791	0.05	0.8201
Quadrant3	1	-0.0121	0.1611	-0.3279	0.3037	0.01	0.9402
Quadrant4	--	0.0000	0.0000	0.0000	0.0000	.	.
Plant1	1	0.8992	0.1141	0.6756	1.1227	62.15	<.0001
Plant2	--	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion	1	2.4726	0.1432	2.2072	2.7698		

From Table 12, only the intercept and the plant variable estimators are significant and quadrant having no contribution. The theoretical and the practical model using these estimators takes the following form**:**

NegBino Model: $\log(\mu) = \text{intercept} + b_1*X_1 + b_2*X_2 + ...+ b_m*X_m$,

$$\log(\hat{\mu}_i) = 1.7508 + 0.8992*X_{i1},$$

Equivalently, as a multiplicative model

$$\hat{\mu}_i = e^{1.7508}e^{0.8992x_{i1}}$$

For a 1 unit increase in log(plant), the estimated count increases by a factor of $e^{0.8992} = 2.46$.

When fitting regression models to count data with log links, it is helpful to exponentiate the estimated coefficients. The resulting values represent ratios of expected counts on the dependent variable associated with a one-unit increase in a given predictor. In other words, the question 'what is the expected multiplicative increase in the mean of the count response that is associated with a one-unit increase in a given predictor?' can be answered. This interpretation, using SAS outputs (more in APPENDIX D) are displayed below. Although quadrant is not significant, to show how to compare them, quadrant to quadrant contrast results are presented for the benfit of the researchers.

Table13- Quadrant to quadrant and plant to plant Contrast Estimate Results

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
plant1 vs. plant2	0.8992	0.1141	0.05	0.6756	1.1227	62.15	<.0001
Exp(plant1 vs. plant2)	2.4576	0.2803	0.05	1.9653	3.0732		
1 vs. 2	0.0166	0.1612	0.05	-0.2994	0.3326	0.01	0.9179
Exp(1 vs. 2)	1.0168	0.1639	0.05	0.7413	1.3946		
1 vs. 3	-0.0079	0.1612	0.05	-0.3238	0.3080	0.00	0.9608
Exp(1 vs. 3)	0.9921	0.1599	0.05	0.7234	1.3607		
1 vs. 4	-0.0200	0.1612	0.05	-0.3360	0.2960	0.02	0.9012
Exp(1 vs. 4)	0.9802	0.1580	0.05	0.7146	1.3444		
2 vs. 3	-0.0245	0.1612	0.05	-0.3404	0.2913	0.02	0.8790
Exp(2 vs. 3)	0.9758	0.1573	0.05	0.7115	1.3382		
2 vs. 4	-0.0366	0.1611	0.05	-0.3524	0.2791	0.05	0.8201
Exp(2 vs. 4)	0.9640	0.1553	0.05	0.7030	1.3220		
3 vs. 4	-0.0121	0.1611	0.05	-0.3279	0.3037	0.01	0.9402
Exp(3 vs. 4)	0.9880	0.1592	0.05	0.7205	1.3548		

Here only the significant contrast output is for variable plant. As the result in Table13 shows, there is significant difference between plant1 and plant2. The estimate of the difference between the two is included in 95% confidence intervals for the ratios of counts, based on the estimates and their standard errors.

It is easily observed that the predicted mean count of parthenium for plant1 is about 2.5 times that for plant2 (controlling for the other variables in the model). The predicted count decreases by about 1% from quadrant1 to quadrant3(which is insignificant statistically). Similar insignificant decreases are observed between quadrant 1 and 4, 2 and 3, 2 and 4, 3 and 4. The only count increment seen from quadrant1 to 2, which is 1.02 times that of quadrant 2; but it is statistically not significant. In general only plant variable is estimated to have a significant relationship with the parthenium outcome and is taken to construct the model.

4.5.5.Multinomial Logit Model Fitting

In the multinomial logit model we assume that the log-odds of each response follow a linear model

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + x_i' \beta_j;$$

Where α_j is a constant and β_j is a vector of regression coefficients, for j =1, 2,……. J - 1. and i for number of independent variables.

This model is analogous to a logistic regression model, except that the probability distribution of the response is multinomial instead of binomial and we have J -1 equations instead of one. The J -1 multinomial logit equations contrast each of categories 1, 2,…,J -1 with category J, whereas the single logistic regression equation is a contrast between successes and failures. If J = 2 the multinomial logit model reduces to the usual logistic regression model.

The goal of fitting multinomial model in this thesis is to examine the effect of the infestation level on the number of other species. The independent variable is sum of the number of other species observed per quadrant (Nospecies) and the dependent variables are five infestation levels(InfestationL), which are categorical into five as mentioned earlier in Table3. The SAS output for multinomial logit model fitting displayed below:-

Table 14-Parameter estimates for Multinomial logit model(APPENDIX E)

Analysis of Maximum Likelihood Estimates					
Parameter	Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	1.8666	0.1715	118.40	<.0001
	2	1.2295	0.1806	46.35	<.0001
	3	1.2034	0.1766	46.44	<.0001
	4	0.3724	0.1948	3.66	0.0559
NOspecies	1	-0.0342	0.00489	48.73	<.0001
	2	-0.0252	0.00477	27.93	<.0001
	3	-0.0200	0.00413	23.51	<.0001
	4	-0.0104	0.00385	7.25	0.0071

From the outputs it can easily be seen that the intercepts 1.8666, 1.2295, 1.2034, 0.3724 are the alphas' and the β's(β_1, β_2, β_3, β_4) for each category except category 5, are -0.0342, -0.0252, -0.0200, -0.0104, respectively. All the intercepts and the coefficients (β's) are significant at the 5% significance level except for the fourth intercept (we can be sure of 94% or at 6% level of significance). Then the estimated multinomial model is as follow;-

$\eta_{i1} = 1.8666 - 0.0342\, x'_i$ $\qquad$ $\eta_{i3} = 1.2034 - 0.0200 x'_i$

$\eta_{i2} = 1.2295 - 0.0252 x'_i$ $\qquad$ $\eta_{i4} = 0.3724 - 0.0104\, x'_i$

Here, we see that we can be reasonably confident that the one or the other, or both of the log odds of no infestation versus infestation or infestation versus serious infestation or serious infestation versus very serious infestation or serious infestation versus high infestation are different from zero. That is, Nospecies does have affected the infestation level.

The liklihood ratio statistic for the overall model has 408 degrees of freedom (392 scores or levels of X are observed for each of two logits, yielding 90 pieces of information, four of which are consumed by the estimation of the four free parameters of the model). One cannot readily reject the hypothesis that the model fits the data (that is, the differences between the predicted frequencies and the actual frequencies (APPENDIX F) are not so large that they could not have reasonably occurred by chance).

The output in appendix 'Multinomial LOGIT Catmode' shows, for each "poplation" (i.e. unique score on X) the predicted log odds on the "first function" (i.e. the log odds of no infestation versus other) and the predicted log odds on the "second function" (i.e. the log odds of infestation versus high infestation). Perhaps more usefully, the ouput shows the frequency distribution of cases in each population on the dependent variable (i.e. how many other species are available in no infestation, infestation, serious infestation, very serious infestation and high infestation.), and the probabilities of each outcome predicted by the regression function. For the last population, for example, the one case was actually a no infestation and the model predicted a `0.001535` probability of that outcome (generating a residual of `-0.00153`).

The figure in appendix E(plotted using SAS) display the association of the level of parthenium infestation probability and the sum of the number of other species. From this figure, one can see as the number of other species increase the probability of getting high infestation level decreases.

4.6.Model checking and Selection

Goodness of fit criteria such as deviance, scaled deviance, pearson chisquare, scaled pearson chi-square are given as an output in fitting GLMs using SAS software . All of these refer to the size of the residual variance, and test whether the residuals are correlated or not. Large values indicate lack of fit, and can be tested against a chi-square distribution for the df given. In addition the AIC, SC, -2log L, R-square,adjusted R-square etc values are given to investigate model adequacy condition. Detail description of each tool is given in APPENDIX H. The following table summarizes the goodness of fit criteria for each fitted model.

Table15 - A Summary of Goodness of fit criteria for each Model.

S.N.	Criteria For Assessing Goodness Of Fit	GLMs value/Df				Multinomial logit
		Logistic	Probit	Poisson	Negative Binomial	
1	Deviance			16	1.1074	
2	scaled deviance			16	1.1074	
3	pearson chisquare			21	0.8716	
4	Scaled Pearson X^2			21	0.8716	
5	AIC	1020.884	1020.884			
6	Liklihood Ratio					451.51

The Deviance value greater than 1 indicate overdispersion, that is, the true variance is bigger than the mean, values smaller than 1 indicate underdispersion, the true variance is smaller than the mean. Evidence of underdispersion or overdispersion indicates inadequate fit of the Poisson model. Having

said this, one can conclude that negative binomial is a better fit than poisson. Where as in comparing logit and probit models, AIC is the same for both cases. If both models have similar level of fit to the data and if it is not possible to select one over the other, the choice may be based on the simplicity of parameter estimation and interpretation. Since the variables used to fit the multinomial logit model aren't the same as used in fitting logistic and probit, there cannot be a comparison between them . But the multinomial logit model used to model in situations where the dependent variables are a categorical one. In such cases one can identify how the independents affect or not the dependents or the relation of the two at different categories.

Chapter Five- Conclusion and Recommendations

5.1.Conclusion

The following conclusion can be drawn from this study:-

- As infestation of parthenium weed gets stronger, the area to be covered by parthenium also gets wider. When parthenium gets a chance to continue, the consequence will be total infestation and total crop loss. This will cause a negative implication in economic, health and social condition of the affected area in particular and the country in general.
- The spread of parthenium infestation have a strong association with plant and plot. parthenium don't have an association with quadrant.
- The probability of getting parthenium weed per quadrant is very high; 81 percent. This expresses that parthenium highly dominate other species.
- The logistic, probit and multinomial models fit for the variables in the data set, and the poisson regression model cannot fit for two different species in the data set. This shows that poisson regression is not adequate to explain for such type of data.
- In fitting Poisson regression model, if the result shows overdispersion, the Negative Binomial model gives a better fit.
- The coefficients as well as the probability estimates obtained from the probit model by the maximum likelihood method have satisfactory asymptotic properties as compared to linear models. After a review of the disadvantages of linear models for estimating the probability of success from the independents, it is better to use a probit model. The variables used in selection have a significant impact on the probability of success, and each variable seems to be associated with a specific aspect of the phenomenon.

- The multinomial logit model is commonly used as compared to the multinomial probit model. The multinomial probit model is not often used mainly due to the practical difficulty in estimation and interpretation. In the multinomial logit model, the independent variables contain characteristics of individuals.

- In applaying GLMs, the role of statistical softwares is immense. Without the help of these softwares, it is very difficult to use GLMs. But the knowledge of statistics and the application of which software for which purpose matters a lot.

5.2.Recommendations

- Controlling the infestation of parthenium must be strengthened over all areas by creating awareness among the community through media, distribution of brouchers and posters.

- The disease caused by parthenium, crop loss amount due to it, and other related data must be collected and documented overtime to provide opportunity for meta analysis. This will enable to give overall recommendation about the weed.

- In situations where the research data is not normal even after transformations, one can easily choose one of the models from the family of the Genralized linear models.

- It is better if statisticians give due attention and work better on standardizing model selection criteria in the generalized linear models.

- Ethiopian Agricultural Research Institute(EARI) has to adapt and familiarize families of generalized linear models against the traditional ANOVA approach.

References

1 Alen Agresti (1996) An introduction to Categorical Data Analysis.

2 Crop Protection society of Ethiopia(2001). Pest Management Journal of Ethiopia, Volume 5.

3 Crop Protection society of Ethiopia(2004). Pest Management Journal of Ethiopia. Volume8.

4 Crop Protection society of Ethiopia(2005). Pest Management Journal of Ethiopia, Volume9.

5 Million Alemayhu(2003). Characterization of indigenous stone bunding(kab) and its effect on crop yield and soil productivity: A thesis submitted to the school of graduate studies Alemaya university.

6 Samuel Ashiber(2005).Effect of Entomopathogenic nematodes and fungi on barly chafer grub: A thesis submitted to the school of graduate studies of Addis Ababa university.

7 P.McCullagh and J.A.Nelder (1989) Generalized linear Models. chapman and Hall.

8 P.McCullagh and J.A.Nelder (1983). Generalized linear Models.chapman and Hall.

9 Taye Tessema(2006).Investigation of pathogens for Biological control of parthenium in Ethiopia.Ph.d thesis.

10 Taye et al (2004). Ooccurrence and distribution of parthenium phyllody. An Exploratory study: *EIAR, Plant Protection Research Centre.*

11 The Biological Society of Ethiopia(2005) . Ethiopian Journal of Biological sciences, volume 4.

12 KWANCHAI A.Gomez(1984, 2nd Ed.).Statistical Procedures for Agricultural Research.John wiley & sons, Inc.

13 INTERNET

13.1 http://www2.sas.com/proceedings/sugi30/213-30.pdf

13.2 http://www.statsoft.com/textbook/stglz.html

13.3http://www.google.com.et/search?q=methodologies+in+fitting+a+generalized+linear+model+for+count+data&hl=en&start=60&sa=N

13.4 http://www.gseis.ucla.edu/courses/ed231c/notes3/probit1.html

13.5 http://www2.chass.ncsu.edu/garson/pa765/structur.htm#uni

13.6 http://www2.chass.ncsu.edu/garson/pa765/logit.htm

13.7 http://www.gseis.ucla.edu/courses/ed231c/notes3/probit1.html

13.8 http://www2.chass.ncsu.edu/garson/pa765/signif.htm

13.9 http://www2.chass.ncsu.edu/garson/pa765/normal.htm

13.10 http://www2.chass.ncsu.edu/garson/pa765/anova.htm#glm

13.11 http://www.uoregon.edu/~robinh/gnmd03_basics.txt

13.12 http://www.ics.uci.edu/~dgillen/Stat211/Handouts/lecture6.pdf

13.13 http://www.ats.ucla.edu/STAT/sas/dae/probit.htm(SAS Data Analysis Examples& Probit Regression

13.14 http://www.spss.com/advanced_models/data_analysis.htm

13.15 http://support.sas.com/ctx/samples/index.jsp?sid=493

13.16 http://www.ats.ucla.edu/STAT/sas/examples/ara/arasas15.htm

13.17 http://www.tspintl.com/products/tsphelp/probit.htm

13.18 http://shazam.econ.ubc.ca/intro/logit3.htm

13.19 http://econ.la.psu.edu/~hbierens/ML_LOGIT.PDF

13.20 http://roso.epfl.ch/mbi/papers/discretechoice/node13.html

13.21 http://www.math.yorku.ca/SCS/Courses/grcat/grc7.html

13.22 http://www.uc.edu/sashtml/stat/chap29/sect37.htm

13.23 http://www.tspintl.com/products/tsphelp/nonlinear.htm

13.24 http://shazam.econ.ubc.ca/intro/olslog.htm

13.25 http://www.gseis.ucla.edu/courses/ed231c/notes2/clog.html

13.26 http://userwww.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.htm

13.27 http://www.statsoft.com/textbook/glosfra.html?glosc.html&1

13.28 http://support.sas.com/techsup/technote/ts279.pdf

13.29 http://en.wikipedia.org/wiki/Logistic_regression

13.30 http://www.tspintl.com/products/tsphelp/probit.htm

13.31 http://en.wikipedia.org/wiki/Multicollinear

13.32 http://en.wikipedia.org/wiki/Poisson_regression

13.33 http://en.wikipedia.org/wiki/Binomial_regression

13.34 http://faculty.ucr.edu/~hanneman/soc271/count.html

13.35 http://www.uky.edu/ComputingCenter/SSTARS/P_NB_3.htm

13.36 http://www.stat.ufl.edu/~presnell/Courses/sta4504-2000sp/SAS/

13.37 http://faculty.ucr.edu/~hanneman/soc271/mlogit.html

13.38 http://www.uark.edu/misc/lampinen/tutorials/multinomial.htm

13.39 http://www.personal.umich.edu/~kwelch/workshops/financial_engineering /2007/finan_binary.doc

13.40 http://data.princeton.edu/wws509/notes/c6.pdf

13.41 http://home.ubalt.edu/ntsbarsh/stat-data/Topics.htm#rsckta

13.42 http://www.unece.org/stats/publications/metadatamodeling.pdf

13.43 http://www.indiana.edu/~statmath/stat/all/cdvm/cdvm.pdf

APPENDIX A: LOGISTIC MODEL

```
proc logistic data=PARTHENIUMANALYSIS descending;
  model partheniumBINO = quadrant plant/ rsquare;
 class quadrant plant;
  output out=pdat dfbetas= _all_
                  difchisq = d_chisq
                  difdev = d_dev
                  reschi = res_chisq
                  resdev = res_dev;
run;
```

```
                    The SAS System

                  The LOGISTIC Procedure

                     Model Information

Data Set                          WORK.PARTHENIUMANALYSIS
Response Variable                 PartheniumBINO
Number of Response Levels         2
Number of Observations            800
Model                             binary logit
Optimization Technique            Fisher's scoring


                   Response Profile

          Ordered     Parthenium         Total
            Value     BINO           Frequency

                1           1             267
                2           0             533

       Probability modeled is PartheniumBINO=1.


                Model Convergence Status

    Convergence criterion (GCONV=1E-8) satisfied.


                  Model Fit Statistics

                                        Intercept
                          Intercept         and
       Criterion            Only       Covariates

       AIC                1020.884       1015.901
       SC                 1025.569       1029.954
       -2 Log L           1018.884       1009.901


 R-Square    0.0112    Max-rescaled R-Square    0.0155


                Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio         8.9836        2         0.0112
Score                    8.9562        2         0.0114
```

```
Wald                    8.9005        2         0.0117


                The SAS System

                    The LOGISTIC Procedure

              Analysis of Maximum Likelihood Estimates

                                 Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -1.4693      0.2982       24.2847        <.0001
Quadrant      1      0.0432      0.0675        0.4101        0.5219
Plant         1      0.4411      0.1512        8.5078        0.0035


                    Odds Ratio Estimates

                       Point          95% Wald
            Effect     Estimate      Confidence Limits

            Quadrant      1.044       0.915       1.192
            Plant         1.554       1.156       2.091


   Association of Predicted Probabilities and Observed Responses

        Percent Concordant      50.0    Somers' D    0.123
        Percent Discordant      37.7    Gamma        0.140
        Percent Tied            12.3    Tau-a        0.055
        Pairs                 142311    c            0.561
```

APPENDIX A1: LOGISTIC MODEL- Refitted

```
DATA PARTHENIUMANALYSIS;
INPUT plot  quadrant    plant Nospecies partheniumcount partheniumBINO
      InfestationL;
CARDS;
1     1     1     15    6     0     2

proc logistic data=PARTHENIUMANALYSIS descending;
  model partheniumBINO = plant / rsquare;
 class plant quadrant;
  output out=pdat dfbetas= _all_
                  difchisq = d_chisq
                  difdev = d_dev
                  reschi = res_chisq
```

```
                    resdev = res_dev;
run;
```

The SAS System

The LOGISTIC Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Response Variable	partheniumBINO
Number of Response Levels	2
Number of Observations	800
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	parthenium BIN	Total Frequency
1	1	267
2	0	533

Probability modeled is partheniumBINO=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1020.884	1014.311
SC	1025.569	1023.680
-2 Log L	1018.884	1010.311

R-Square 0.0107 Max-rescaled R-Square 0.0148

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.5733	1	0.0034
Score	8.5503	1	0.0035
Wald	8.5036	1	0.0035

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3606	0.2442	31.0311	<.0001
plant	1	0.4408	0.1512	8.5036	0.0035

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
plant	1.554	1.156	2.090

Association of Predicted Probabilities and Observed Responses

Percent Concordant	30.7	Somers' D	0.110
Percent Discordant	19.8	Gamma	0.217
Percent Tied	49.5	Tau-a	0.049
Pairs	142311	c	0.555

APPENDIX B: PROBIT MODEL

```
PROC sort data=PARTHENIUMANALYSIS;
by descending partheniumBINO;
run;

PROC logistic data=PARTHENIUMANALYSIS;
title 'probit model analysis';
class partheniumBINO;
model partheniumBINO(event='1')=quadrant plant/ link=probit;
run;
```

probit model analysis

The LOGISTIC Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Response Variable	PartheniumBINO
Number of Response Levels	2
Number of Observations	800
Model	binary probit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Parthenium BINO	Total Frequency
1	0	533

```
                          2           1            267

          Probability modeled is PartheniumBINO=1.

                  Model Convergence Status

        Convergence criterion (GCONV=1E-8) satisfied.

                     Model Fit Statistics

                                          Intercept
                               Intercept        and
          Criterion              Only       Covariates

          AIC                 1020.884       1015.896
          SC                  1025.569       1029.950
          -2 Log L            1018.884       1009.896

                  Testing Global Null Hypothesis: BETA=0

     Test                 Chi-Square       DF     Pr > ChiSq

     Likelihood Ratio         8.9883        2         0.0112
     Score                    8.9562        2         0.0114
     Wald                     8.9571        2         0.0114


                          probit model analysis

                        The LOGISTIC Procedure

                  Analysis of Maximum Likelihood Estimates

                                  Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -0.9038      0.1803       25.1321        <.0001
Quadrant      1      0.0264      0.0411        0.4127        0.5206
Plant         1      0.2693      0.0921        8.5554        0.0034


     Association of Predicted Probabilities and Observed Responses

          Percent Concordant      50.0    Somers' D    0.123
          Percent Discordant      37.7    Gamma        0.140
          Percent Tied            12.3    Tau-a        0.055
          Pairs                 142311    c            0.561
```

OR

```
PROC sort data=PARTHENIUMANALYSIS;
by descending partheniumBINO;
run;

PROC PROBIT data=PARTHENIUMANALYSIS order= data;
title 'probit model analysis';
class partheniumBINO;
model partheniumBINO =quadrant plant;
run;
COMMENT-The privious is better in specifaying the predicted outcome.
```

```
                          probit model analysis

                            Probit Procedure
```

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Dependent Variable	PartheniumBINO
Number of Observations	800
Name of Distribution	Normal
Log Likelihood	-504.9479237

Class Level Information

Name	Levels	Values
PartheniumBINO	2	1 0

Response Profile

Ordered Value	Parthenium BINO	Total Frequency
1	1	267
2	0	533

PROC PROBIT is modeling the probabilities of levels of PartheniumBINO having LOWER Ordered Values in the response profile table.

Algorithm converged.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Quadrant	1	0.4149	0.5195
Plant	1	8.5557	0.0034

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.9038	0.1803	-1.2572	-0.5505	25.13	<.0001
Quadrant	1	0.0264	0.0410	-0.0540	0.1069	0.41	0.5195
Plant	1	0.2694	0.0921	0.0889	0.4498	8.56	0.0034

APPENDIX B1: PROBIT MODEL- Refitted

```
DATA PARTHENIUMANALYSIS;
INPUT plot  quadrant    plant Nospecies partheniumcount partheniumBINO
      InfestationL;
CARDS;
1     1     1     15    6     0     2
```

```
PROC logistic data=PARTHENIUMANALYSIS;
title 'probit model analysis';
class partheniumBINO;
model partheniumBINO(event='1')=plant/ link=probit;
run;
```

probit model analysis
The LOGISTIC Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Response Variable	partheniumBINO
Number of Response Levels	2
Number of Observations	800
Model	binary probit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	parthenium BINO	Total Frequency
1	0	533
2	1	267

Probability modeled is partheniumBINO=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1020.884	1014.311
SC	1025.569	1023.680
-2 Log L	1018.884	1010.311

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.5733	1	0.0034
Score	8.5503	1	0.0035
Wald	8.5444	1	0.0035

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8372	0.1474	32.2435	<.0001
plant	1	0.2691	0.0921	8.5444	0.0035

Association of Predicted Probabilities and Observed Responses

Percent Concordant	30.7	Somers' D	0.110
Percent Discordant	19.8	Gamma	0.217
Percent Tied	49.5	Tau-a	0.049
Pairs	142311	c	0.555

APPENDIX C: POISSON REGRESSION

```
    proc means data=PARTHENIUMANALYSIS n nmiss mean var;
var partheniumcount;
run;
proc genmod data=PARTHENIUMANALYSIS;
class quadrant plant;
model partheniumcount=quadrant plant/dist=poisson link=log type3 ;
      output out=proba p= pred xbeta=z;
run;
proc print data=proba;
run;
```

The SAS System

The MEANS Procedure

Analysis Variable : Partheniumcount

N	Miss	Mean	Variance
800	0	9.7887500	260.0717131

The SAS System

The GENMOD Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Distribution	Poisson
Link Function	Log
Dependent Variable	Partheniumcount
Observations Used	800

Class Level Information

Class	Levels	Values
Quadrant	4	1 2 3 4
Plant	2	1 2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	795	12831.3515	16.1401
Scaled Deviance	795	12831.3515	16.1401
Pearson Chi-Square	795	17111.2606	21.5236
Scaled Pearson X2	795	17111.2606	21.5236
Log Likelihood		10755.1217	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.7780	0.0283	1.7225	1.8335	3939.22	<.0001
Quadrant	1	1	-0.0698	0.0318	-0.1322	-0.0074	4.81	0.0283
Quadrant	2	1	-0.0620	0.0318	-0.1242	0.0003	3.81	0.0510
Quadrant	3	1	-0.0480	0.0316	-0.1100	0.0140	2.30	0.1290

```
      Quadrant     4     -      0.0000      0.0000      0.0000      0.0000        .           .
      Plant        1     1      0.8997      0.0249      0.8509      0.9486    1302.86        <.0001
 Plant        2     -      0.0000      0.0000      0.0000      0.0000        .           .
 Scale              -      1.0000      0.0000      1.0000      1.0000

NOTE: The scale parameter was held fixed.

                                    The SAS System

                                The GENMOD Procedure

                          LR Statistics For Type 3 Analysis

                                                Chi-
                      Source           DF     Square    Pr > ChiSq

                      Quadrant          3       5.84        0.1196
                      Plant             1    1437.72        <.0001
```

APPENDIX C1: POISSON MODEL- Refitted

```
DATA PARTHENIUMANALYSIS;
INPUT plot  quadrant    plant Nospecies partheniumcount partheniumBINO
      InfestationL;
CARDS;
1     1     1     15    6     0     2

proc genmod data=PARTHENIUMANALYSIS;
class quadrant plant;
model partheniumcount=plant/dist=poisson link=log type3 ;
      output out=proba p= pred xbeta=z;
run;
proc print data=proba;
run;
```

probit model analysis
The GENMOD Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Distribution	Poisson
Link Function	Log
Dependent Variable	partheniumcount
Observations Used	800

Class Level Information

Class	Levels	Values
quadrant	4	1 2 3 4
plant	2	1 2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	798	12837.1924	16.0867
Scaled Deviance	798	12837.1924	16.0867
Pearson Chi-Square	798	17161.8010	21.5060
Scaled Pearson X2	798	17161.8010	21.5060
Log Likelihood		10752.2013	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.7334	0.0210	1.6922	1.7746	6802.78	<.0001
plant	1	1	0.8997	0.0249	0.8509	0.9486	1302.86	<.0001
plant	2	-	0.0000	0.0000	0.0000	0.0000	.	.
Scale		-	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
plant	1	1437.72	<.0001

APPENDIX D: NEGATIVE BINOMIAL

```
proc genmod data=PARTHENIUMANALYSIS;
class quadrant plant;
model partheniumcount=quadrant plant/dist=negbin link=log type3;
estimate "plant1 vs. plant2" plant 1 -1/exp;
estimate "1 vs. 2" quadrant 1  -1 0 0/exp;
estimate "1 vs. 3" quadrant 1  0 -1 0/exp;
estimate "1 vs. 4" quadrant 1  0 0 -1/exp;
estimate '2 vs. 3' quadrant 0  1 -1 0/exp;
estimate '2 vs. 4' quadrant 0  1 0 -1/exp;
estimate '3 vs. 4' quadrant 0  0 1 -1/exp;
run;
```

The SAS System

The GENMOD Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Distribution	Negative Binomial
Link Function	Log

Dependent Variable	Partheniumcount
Observations Used	800

Class Level Information

Class	Levels	Values
Quadrant	4	1 2 3 4
Plant	2	1 2

Parameter Information

Parameter	Effect	Quadrant	Plant
Prm1	Intercept		
Prm2	Quadrant	1	
Prm3	Quadrant	2	
Prm4	Quadrant	3	
Prm5	Quadrant	4	
Prm6	Plant		1
Prm7	Plant		2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	795	880.4206	1.1074
Scaled Deviance	795	880.4206	1.1074
Pearson Chi-Square	795	692.9427	0.8716
Scaled Pearson X2	795	692.9427	0.8716
Log Likelihood		15795.7852	

Algorithm converged.

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.7508	0.1296	1.4968	2.0047	182.59	<.0001
Quadrant	1	1	-0.0200	0.1612	-0.3360	0.2960	0.02	0.9012
Quadrant	2	1	-0.0366	0.1611	-0.3524	0.2791	0.05	0.8201
Quadrant	3	1	-0.0121	0.1611	-0.3279	0.3037	0.01	0.9402
Quadrant	4	0	0.0000	0.0000	0.0000	0.0000	.	.
Plant	1	1	0.8992	0.1141	0.6756	1.1227	62.15	<.0001
Plant	2	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	2.4726	0.1432	2.2072	2.7698		

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Quadrant	3	0.05	0.9967
Plant	1	57.83	<.0001

Contrast Estimate Results

	Standard		Chi-

Label	Estimate	Error	Alpha	Confidence Limits		Square	Pr > ChiSq
plant1 vs. plant2	0.8992	0.1141	0.05	0.6756	1.1227	62.15	<.0001
Exp(plant1 vs. plant2)	2.4576	0.2803	0.05	1.9653	3.0732		
1 vs. 2	0.0166	0.1612	0.05	-0.2994	0.3326	0.01	0.9179
Exp(1 vs. 2)	1.0168	0.1639	0.05	0.7413	1.3946		
1 vs. 3	-0.0079	0.1612	0.05	-0.3238	0.3080	0.00	0.9608
Exp(1 vs. 3)	0.9921	0.1599	0.05	0.7234	1.3607		
1 vs. 4	-0.0200	0.1612	0.05	-0.3360	0.2960	0.02	0.9012
Exp(1 vs. 4)	0.9802	0.1580	0.05	0.7146	1.3444		
2 vs. 3	-0.0245	0.1612	0.05	-0.3404	0.2913	0.02	0.8790
Exp(2 vs. 3)	0.9758	0.1573	0.05	0.7115	1.3382		
2 vs. 4	-0.0366	0.1611	0.05	-0.3524	0.2791	0.05	0.8201
Exp(2 vs. 4)	0.9640	0.1553	0.05	0.7030	1.3220		
3 vs. 4	-0.0121	0.1611	0.05	-0.3279	0.3037	0.01	0.9402
Exp(3 vs. 4)	0.9880	0.1592	0.05	0.7205	1.3548		

APPENDIX D: NEGATIVE BINOMIAL-REFITTED

```
proc genmod data=PARTHENIUMANALYSIS;
class plant;
model partheniumcount=plant/dist=negbin link=log type3;
estimate "plant1 vs. plant2" plant 1 -1/exp;
run;
```

The SAS System

The GENMOD Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	partheniumcount
Observations Used	800

Class Level Information

Class	Levels	Values
plant	2	1 2

Parameter Information

Parameter	Effect	plant
Prm1	Intercept	
Prm2	plant	1
Prm3	plant	2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	798	880.4133	1.1033
Scaled Deviance	798	880.4133	1.1033
Pearson Chi-Square	798	692.6635	0.8680
Scaled Pearson X2	798	692.6635	0.8680
Log Likelihood		15795.7579	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.7334	0.0814	1.5739	1.8929	453.64	<.0001
plant	1	1	0.8997	0.1140	0.6764	1.1231	62.34	<.0001
plant	2	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	2.4728	0.1432	2.2075	2.7701		

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

The SAS System
The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
plant	1	57.99	<.0001

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
plant1 vs. plant2	0.8997	0.1140	0.05	0.6764	1.1231	62.34	<.0001
Exp(plant1 vs. plant2)	2.4589	0.2802	0.05	1.9667	3.0743		

APPENDIX E: The MULTINOMIAL LOGIT MODEL SAS CODE AND OUTPUTS

DATA ENTRY CODE

```
PROC FORMAT;
VALUE fd  1='no infestation' 2='infestation' 3='seinfestation' 4='vseinfestation'
5='high infestation';
* 8.1 LOGIT MODELS FOR NOMINAL RESPONSES;
DATA PARTHENIUMANALYSIS;
FORMAT InfestationL IL. ;
INPUT Nospecies InfestationL;
CARDS;
```

CODES

```
PROC SORT; BY InfestationL Nospecies;
PROC CATMOD DATA=PARTHENIUMANALYSIS;
RESPONSE logits / OUT=prob OUTEST=param;
DIRECT Nospecies;
MODEL InfestationL = Nospecies / noprofile; * pred=prob;
run;
DATA prob; SET prob;
KEEP Nospecies InfestationL obsvd pred resid;
obsvd=_obs_ ;
pred = _pred_ ;
resid=_resid_ ;
IF _type_ = 'PROB';

PROC PLOT DATA=prob;
PLOT pred*Nospecies=InfestationL / vaxis=0 to 1 by .2;
TITLE3 'Predicted Probabilities for Primary InfestationL';
```

(CATMOD, being a program for the analysis of categorical data, tends to assume that all variables are CLASS, unless told otherwise, DIRECT Nospecies statement above told catmod that this variable is a continous one). The "RESPONSE logits;" statement tells CATMOD to model generalized logits. That is, CATMOD calculates the log odds of each category of the dependent variable relative to the last category of the dependent variable.

```
                                The SAS System

          Predicted Probabilities for Primary InfestationL

                            The CATMOD Procedure

                                Data Summary

Response           InfestationL           Response Levels    5
Weight Variable    None                   Populations      104
Data Set           PARTHENIUMANALYSIS     Total Frequency  800
Frequency Missing  0                      Observations     800



                          Maximum Likelihood Analysis

 Sub         -2 Log  Convergence              Parameter Estimates
```

Iteration	Iteration	Likelihood	Criterion	1	2	3	4
0	0	2575.1007	1.0000	0	0	0	0
1	0	2386.1898	0.0734	1.7692	0.9096	0.9194	0.2782
2	0	2376.1168	0.004221	1.8538	1.2251	1.1966	0.3545
3	0	2376.0957	8.8786E-6	1.8666	1.2295	1.2034	0.3723
4	0	2376.0957	6.921E-10	1.8666	1.2295	1.2034	0.3724

Maximum Likelihood Analysis

Parameter Estimates

Iteration	5	6	7	8
0	0	0	0	0
1	-0.0318	-0.0223	-0.0193	-0.0127
2	-0.0338	-0.0251	-0.0199	-0.009879
3	-0.0342	-0.0252	-0.0200	-0.0104
4	-0.0342	-0.0252	-0.0200	-0.0104

Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	4	158.38	<.0001
NOspecies	4	59.94	<.0001
Likelihood Ratio	408	451.54	0.0673

The SAS System

Predicted Probabilities for Primary InfestationL

The CATMOD Procedure

Analysis of Maximum Likelihood Estimates

Parameter	Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	1.8666	0.1715	118.40	<.0001
	2	1.2295	0.1806	46.35	<.0001
	3	1.2034	0.1766	46.44	<.0001
	4	0.3724	0.1948	3.66	0.0559
NOspecies	1	-0.0342	0.00489	48.73	<.0001
	2	-0.0252	0.00477	27.93	<.0001
	3	-0.0200	0.00413	23.51	<.0001
	4	-0.0104	0.00385	7.25	0.0071

The SAS System

Predicted Probabilities for Primary InfestationL

The CATMOD Procedure

Data Summary

Response	InfestationL	Response Levels	5
Weight Variable	None	Populations	104
Data Set	PARTHENIUMANALYSIS	Total Frequency	800
Frequency Missing	0	Observations	800

Maximum Likelihood Analysis

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates 1	2	3	4
0	0	2575.1007	1.0000	0	0	0	0
1	0	2386.1898	0.0734	1.7692	0.9096	0.9194	0.2782
2	0	2376.1168	0.004221	1.8538	1.2251	1.1966	0.3545
3	0	2376.0957	8.8786E-6	1.8666	1.2295	1.2034	0.3723
4	0	2376.0957	6.921E-10	1.8666	1.2295	1.2034	0.3724

Maximum Likelihood Analysis

Iteration	Parameter Estimates 5	6	7	8
0	0	0	0	0
1	-0.0318	-0.0223	-0.0193	-0.0127
2	-0.0338	-0.0251	-0.0199	-0.009879
3	-0.0342	-0.0252	-0.0200	-0.0104
4	-0.0342	-0.0252	-0.0200	-0.0104

Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	4	158.38	<.0001
NOspecies	4	59.94	<.0001
Likelihood Ratio	408	451.54	0.0673

The SAS System

Predicted Probabilities for Primary InfestationL

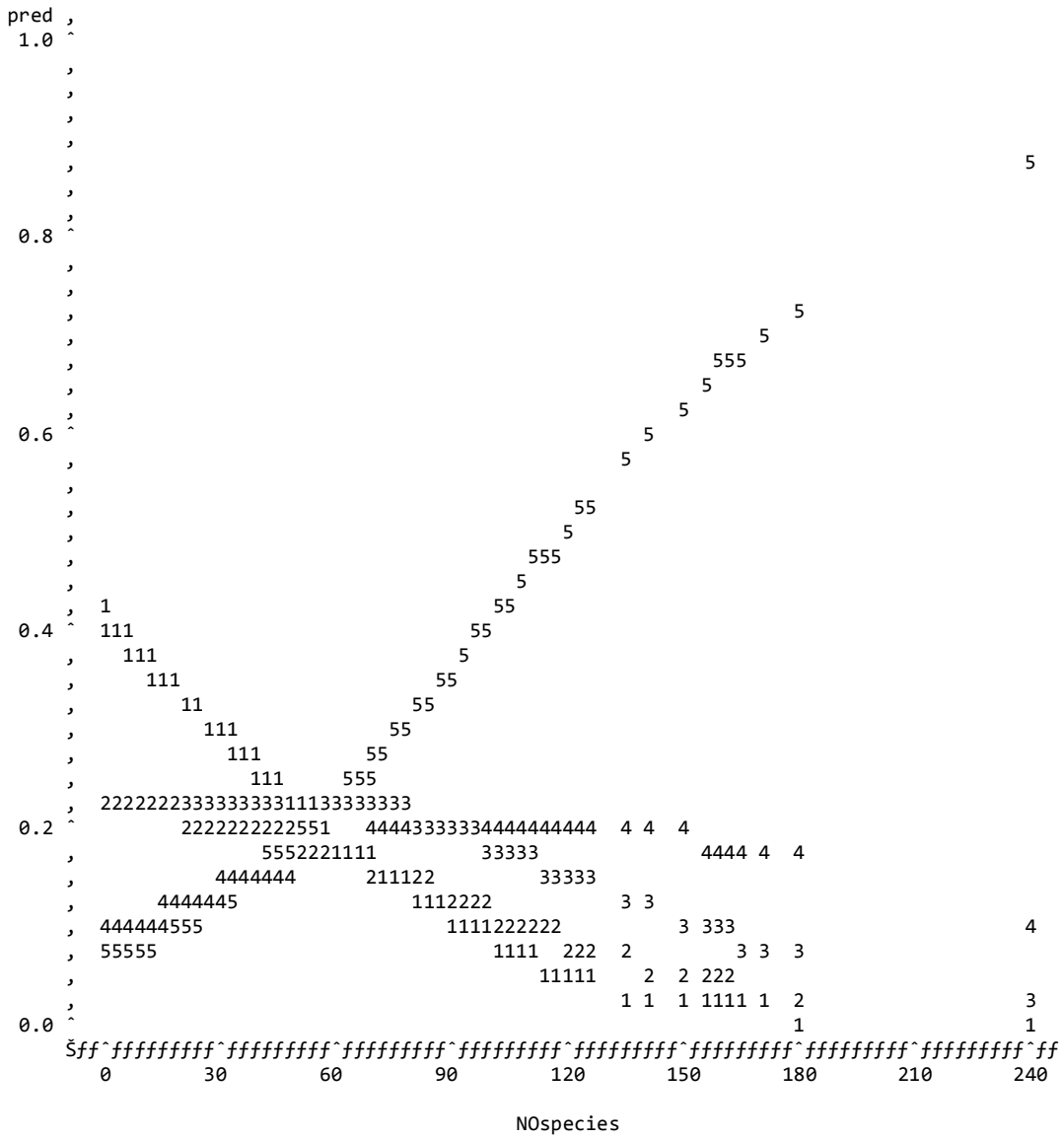
The CATMOD Procedure

Analysis of Maximum Likelihood Estimates

Parameter	Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	1.8666	0.1715	118.40	<.0001
	2	1.2295	0.1806	46.35	<.0001
	3	1.2034	0.1766	46.44	<.0001
	4	0.3724	0.1948	3.66	0.0559
NOspecies	1	-0.0342	0.00489	48.73	<.0001
	2	-0.0252	0.00477	27.93	<.0001
	3	-0.0200	0.00413	23.51	<.0001
	4	-0.0104	0.00385	7.25	0.0071

Predicted Probabilities for Primary InfestationL

Plot of pred*NOspecies. Symbol is value of InfestationL.



NOTE: 286 obs hidden.
As the nuber of other species increase the probability of the number of getting high infestation level decreases i.e. the number of getting many '4' and '5' gets decreasing.

APPENDIX G: ANOVA FOR NESTED DESIGNS

```
DATA PARTHENIUMANALYSIS;
INPUT plot quadrant plant partheniumcount;
CARDS;
.............
proc anova data=PARTHENIUMANALYSIS;/*nested anova for balanced design*/
class plot quadrant plant;
```

```
    model partheniumcount= plot quadrant(plot);
test h=plot e=quadrant(plot);
run;
```

```
                                        The SAS System

                                    The ANOVA Procedure

                                  Class Level Information

Class       Levels  Values

plot           100  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
                    29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
                    54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
                    79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

quadrant         4  1 2 3 4

plant            2  1 2

                                  Number of observations    800

                                    The ANOVA Procedure

Dependent Variable: partheniumcount

      Source                       DF         Sum of Squares     Mean Square    F Value    Pr > F

      Model                       399     141559.7988        354.7865       2.14    <.0001

      Error                       400      66237.5000        165.5938

      Corrected Total             799     207797.2988

                  R-Square     Coeff Var      Root MSE    partheniumcount Mean

                  0.681240      131.4603      12.86832                9.788750

      Source                       DF        Anova SS     Mean Square    F Value    Pr > F

      plot                         99     65983.92375       666.50428       4.02    <.0001
      quadrant(plot)              300     75575.87500       251.91958       1.52    <.0001

          Tests of Hypotheses Using the Anova MS for quadrant(plot) as an Error Term

      Source                       DF        Anova SS     Mean Square    F Value    Pr > F

      plot                         99     65983.92375       666.50428       2.65    <.0001
```

OR

```
proc glm data=PARTHENIUMANALYSIS; /*nested anova- MOre general form*/
class plot quadrant plant;
model partheniumcount= plot quadrant(plot);
test h=plot e=quadrant(plot);
 random quadrant(plot) test;
 means plot plot*quadrant/stderr pdiff;
 means plot/stderr pdiff e=quadrant(plot);
run;
```

```
                                        The SAS System
```

The GLM Procedure

Class Level Information

Class	Levels	Values
plot	100	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
quadrant	4	1 2 3 4
plant	2	1 2

Number of observations 800

The SAS System

The GLM Procedure

Dependent Variable: Partheniumcount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	399	141559.7987	354.7865	2.14	<.0001
Error	400	66237.5000	165.5938		
Corrected Total	799	207797.2988			

R-Square	Coeff Var	Root MSE	Partheniumcount Mean
0.681240	131.4603	12.86832	9.788750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
plot	99	65983.92375	666.50428	4.02	<.0001
quadrant(plot)	300	75575.87500	251.91958	1.52	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
plot	99	65983.92375	666.50428	4.02	<.0001
quadrant(plot)	300	75575.87500	251.91958	1.52	<.0001

Tests of Hypotheses Using the Type III MS for quadrant(plot) as an Error Term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
plot	99	65983.92375	666.50428	2.65	<.0001

APPENDIX H: PARTHENIUM COUNT DATA

S.N.	plot	quadrant	plant	SUM OF OTHER SPECIES AS Nospecies	partheniumcount	partheniumBIN	InfestationL
1	1	1	1	15	6	0	2
2	1	1	2	10	8	0	2

3	1	2	1	7	0	1	1
4	1	2	2	9	0	1	1
5	1	3	1	13	30	0	5
6	1	3	2	12	10	0	3
7	1	4	1	21	5	0	2
8	1	4	2	20	4	0	4
9	2	1	1	9	6	0	3
10	2	1	2	12	4	0	2
11	2	2	1	7	3	0	2
12	2	2	2	8	3	0	2
13	2	3	1	22	32	0	5
14	2	3	2	12	16	0	4
15	2	4	1	14	6	0	2
16	2	4	2	3	4	0	2
17	3	1	1	21	12	0	4
18	3	1	2	10	6	0	2
19	3	2	1	4	9	0	3
20	3	2	2	9	5	0	2
21	3	3	1	68	14	0	4
22	3	3	2	35	18	0	4
23	3	4	1	83	15	0	4
24	3	4	2	26	12	0	4
25	4	1	1	52	13	0	4
26	4	1	2	6	9	0	3
27	4	2	1	90	10	0	3
28	4	2	2	12	14	0	4
29	4	3	1	127	25	0	5
30	4	3	2	24	16	0	4
31	4	4	1	64	6	0	2
32	4	4	2	27	10	0	3
33	5	1	1	49	75	0	5
34	5	1	2	13	28	0	5
35	5	2	1	92	19	0	4
36	5	2	2	25	16	0	4
37	5	3	1	18	23	0	5
38	5	3	2	22	19	0	4
39	5	4	1	19	18	0	4
40	5	4	2	21	12	0	4
41	6	1	1	10	18	0	4
42	6	1	2	7	10	0	3

APPENDIX I:- POISSON FOR SPECIES CHIFIRG

```
data PARTHENIUMANALYSIS;
Input plot Quadrant Plant NOspecies Partheniumcount PartheniumBINO  InfestationL
chifirg digitra;
```

```
    cards;
1     1     1     83    6     0     2     0     7
1     1     2     68    8     0     2     0     10
1     2     1     28    0     1     1     0     7
1     2     2     35    0     1     1     0     9
1     3     1     174   30    0     5     0     13
.................................................
```

```
proc genmod data=PARTHENIUMANALYSIS descending;
class quadrant plant;
model chifirg=quadrant plant/dist=poisson link=log type3 wald;
run;
```

probit model analysis
The GENMOD Procedure

Model Information

Data Set	WORK.PARTHENIUMANALYSIS
Distribution	Poisson
Link Function	Log
Dependent Variable	chifirg
Observations Used	799

Class Level Information

Class	Levels	Values
Quadrant	4	1 2 3 4
Plant	2	1 2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	794	8978.3933	11.3078
Scaled Deviance	794	8978.3933	11.3078
Pearson Chi-Square	794	38682.3777	48.7184
Scaled Pearson X2	794	38682.3777	48.7184
Log Likelihood		-199.2097	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	0.3139	0.0601	0.1960	0.4317	27.24	<.0001
Quadrant	1	1	-0.6225	0.0797	-0.7787	-0.4664	61.03	<.0001
Quadrant	2	1	-0.4697	0.0761	-0.6188	-0.3206	38.11	<.0001
Quadrant	3	1	0.3898	0.0610	0.2703	0.5093	40.86	<.0001
Quadrant	4	--	0.0000	0.0000	0.0000	0.0000	.	.
Plant	1	1	0.8308	0.0537	0.7255	0.9360	239.28	<.0001
Plant	2	--	0.0000	0.0000	0.0000	0.0000	.	.
Scale		--	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

The GENMOD Procedure

Wald Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Quadrant	3	256.45	<.0001
Plant	1	239.28	<.0001

APPENDIX J: SAS Annotated Output

This page shows regression analysis with footnotes explaining the output in SAS.

- **Data Set** - This is the SAS dataset analyzed with probit or whatever regression.
- **Response Variable** - This is the outcome (dependent) variable in the probit regression.
- **Number of Response Levels** - This is the number of levels of the dependent variable.
- **Model** - This is the model that SAS is fitting. Like, binary refers to the outcome variable (the two levels of **the** response) and probit refers to the distribution used in fitting the model.
- **Optimization Technique** - This refers to the iterative method of estimating the regression parameters. In SAS, the default method is Fisher's scoring method, whereas in Stata, it is the

Newton-Raphson algorithm. Both techniques yield the same estimate for the regression coefficient; however, the standard errors differ between the two methods.

- **Model Convergence Status** - This describes whether or not the maximum-likelihood algorithm has converged and what kind of convergence criterion is used for convergence. The default convergence criterion is the relative gradient convergence criterion **(GCONV)**, and the default precision is 10^{-8}.
- **Criterion** - These are various measurements used to assess the model fit. The first two, Akaike Information Criterion (**AIC**) and Schwarz Criterion (**SC**) are variants of negative two times the Log-Likelihood (**-2 Log L**). **AIC** and **SC** penalize the Log-Likelihood by the number of predictors in the model.
- **Intercept Only** - This column refers to the respective **Criterion** statistics with no predictors.
- **Intercept and Covariates** - This column corresponds to the respective **Criterion** statistics for the fitted model. A fitted model includes all predictors and the intercept. We can compare the values in this column with the criteria corresponding **Intercept Only** value to assess model fit/significance.
- **AIC** - This is the Akaike Information Criterion. It is calculated as AIC = -2 Log L + 2((*k*-1) + *s*), where *k* is the number of levels of the outcome variable and *s* is the number of predictors in the model. **AIC** is used for the comparison of models from different samples or nonnested models. Ultimately, the model with the smallest **AIC** is considered the best.
-

- **SC** - This is the Schwarz Criterion. It is defined as - 2 Log L + ((*k*-1) + *s*)*log(Σf_i), where f_i's are the frequency values of the i^{th} observation, and *k* and *s* were defined previously. Like **AIC**, **SC** penalizes for the number of predictors in the model and the smallest **SC** is most desirable.
- **-2 Log L** - This is negative two times the log likelihood. The **-2 Log L** is used in hypothesis tests for nested models.
- **Deviance**- something like the sum of squares for non-normal outcomes.
- **Test** - These are three asymptotically equivalent Chi-Square tests. They test against the null hypothesis that at least one of the predictors' regression coefficient is not equal to zero in the

model. The differences between the three tests can be attributed to evaluating the log-likelihood function at different points.

- **Chi-Square** - This is the **Chi-Square** test statistic corresponding to the specific **test** that all of the predictors are simultaneously equal to zero.
- **DF** - This is the number of degrees of freedom. It determines the distribution of the Chi-Square test statistics and is defined by the number of predictors in the model. Our model includes three predictors, so **DF** = 3.
- **Pr > ChiSq** - This is the probability the **Chi-Square** test statistic (or a more extreme test statistic) would be observed under the null hypothesis that a particular predictor's regression coefficient is zero, given that the rest of the predictors are in the model. For a given alpha level, **Pr > ChiSq** determines whether or not the null hypothesis can be rejected. If **Pr > ChiSq** is less than alpha, then the null hypothesis can be rejected and the parameter estimate is considered statistically significant at that alpha level.
- **Likelihood Ratio** - This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model. The LR Chi-Square statistic can be calculated by -2 Log L(null model) - 2 Log L(fitted model), where L(null model) refers to the **Intercept Only** model and L(fitted model) refers to the **Intercept and Covariates** model.
- **Score** - This is the Score Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.
- **Wald** - This is the Wald Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.

- **Parameter** - These refer to the independent variables in the model as well as intercepts (constants) for the adjacent levels of the dependent variable.
- **DF** - This column gives the degrees of freedom corresponding to the **Parameter**. For each **Parameter** estimated in the model, one **DF** is required, and the **DF** defines the Chi-Square distribution to test whether the individual regression coefficient is zero given the other variables are in the model.

- **Estimate** - These are the regression coefficients. There are limited ways in which we can interpret the individual regression coefficients. A positive coefficient mean that an increase in the predictor leads to an increase in the predicted probability. A negative coefficient means that an increase in the predictor leads to a decrease in the predicted probability.

- **Intercept** - The constant term. This means that if all of the predictors are evaluated at zero, the predicted probability of X is F(Intercept) = a .
- **Standard Error - These are the standard errors of the individual regression coefficients. They are used in the calculation of the Wald Chi-Square test statistic.**
- AIC = weighs goodness-of-fit & model complexity (smaller is better) \
- Wald = $(\text{parameter}_{(\text{estimated})}/\ \text{standard error}_{(\text{estimated})})^2$.
- **Wald Chi-Square** - This is the Wald test statistic for the hypothesis test that an individual predictor's regression coefficient is zero given the rest of the predictors are in the model. The **Wald Chi-Square** test statistic is the squared ratio of the **Estimate** to the **Standard Error** of the respective predictor. The probability that a particular **Wald Chi-Square** test statistic is as extreme as, or more so, than what has been observed under the null hypothesis is given by **Pr > ChiSq**.
- **Pr > ChiSq** - This is the p-value corresponding to the **Wald Chi-Square** test statistic that all of the predictors are simultaneously equal to zero. We are testing the probability (**Pr > ChiSq**) of observing a **Chi-Square** statistic as extreme as, or more so, than the observed one under the null hypothesis; the null hypothesis is that all of the regression coefficients in the model are equal to zero. Typically, **Pr > ChiSq** is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01. The small p-value from the all three **tests** would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero

  The **Wald Chi-Square** test statistic for the **Intercept** is 18.6630 with an associated p-value <.0001. If we set our alpha level to 0.05, we would reject the null hypothesis and conclude that the model intercept has been found to be statistically different from zero given **gre, topnotch** and **gpa** are in the model.

- **Percent Concordant** - A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value.

- **Percent Discordant** - If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant.

- **Percent Tied** - If a pair of observations with different responses is neither concordant nor discordant, it is a tie.
- **Pairs** - This is the total number of distinct pairs with one case having a positive response (**response** = 1) and the other having a negative response (**response** = 0). The total number of ways the 400 observations (for instance)can be paired up (excluding be matched up with themselves) is 400(399)/2 = 79,800. Of the 79,800 possible pairings, 34,671 have different values on the response variable and 79,800 - 34,671 = 45,129 have the same value on the response variable.
- **Somers' D** - Somer's D is used to determine the strength and direction of relation between pairs of variables. Its values range from -1.0 (all pairs disagree) to 1.0 (all pairs agree). It is defined as $(n_c-n_d)/t$ where n_c is the number of pairs that are concordant, and n_d the number of pairs that are discordant, and t is the number of total number of pairs with different responses. In our example, it equals the difference between the percent concordant and the percent discordant divided by 100.
- **Gamma** - The Goodman-Kruskal Gamma method does not penalize for ties on either variable. Its values range from -1.0 (no association) to 1.0 (perfect association). Because it does not penalize for ties, its value will generally be greater than the values for Somer's D.
- **Tau-a** - Kendall's Tau-a is a modification of Somer's D to take into the account the difference between the number of possible paired observations and the number of paired observations with different response. It is defined to be the ratio of the difference between the number of concordant pairs and the number of discordant pairs to the number of possible pairs $(2(n_c-n_d)/(N(N-1))$. Usually Tau-a is much smaller than Somer's D since there would be many paired observations with the same response.
- **c** - Another measure of rank correlation of ordinal variables. It ranges from 0 to (no association) to 1 (perfect association). It is a variant of Somer's D index.

The scale parameters are related to the dispersion parameter as shown previously with the probability distribution definitions. Thus, the scale parameter output in the "Analysis of Parameter Estimates" table is related to the exponential family dispersion parameter. If you specify a constant

scale parameter with the SCALE= option in the MODEL statement, it is also related to the exponential family dispersion parameter in the same way.

$$X^2 = \sum_i \frac{w_i(y_i - \mu_i)^2}{V(\mu_i)}$$

- Pearson's chi-square statistic is defined as and the scaled Pearson's chi-square is X^2/ϕ, Where ϕ is a dispersion parameter.

APPENDIX K-Multinomial logit model BY CATMODE

All data for ppendix F.(predictions for Response and for Predictor variables)

```
proc catmod; response logits; direct Nospecies;
   model InfestationL=Nospecies /pred=prob pred=freq;
run;
```

```
                                The SAS System

                              The CATMOD Procedure

                                  Data Summary

Response           InfestationL           Response Levels    5
Weight Variable    None                   Populations      104
Data Set           PARTHENIUMANALYSIS     Total Frequency  800
Frequency Missing  0                      Observations     800
```

The SAS System

The CATMOD Procedure

Population Profiles

Sample	NOspecies	Sample Size
1	0	90
2	1	5
3	2	14
4	3	39
5	4	35
6	5	30
7	6	30
8	7	37
9	8	27
10	9	20
11	10	39
12	11	20
13	12	25
14	13	19
15	14	13
16	15	13
17	16	15
18	17	22
19	18	15
20	19	15
21	20	20
22	21	11
23	22	10
24	23	8
25	24	20
26	25	10
27	26	9
28	27	5
29	28	7
30	29	8
31	30	11
32	31	7
33	32	3
34	33	5
35	34	5
36	35	2
37	36	4
38	37	10
39	38	7
40	39	3
41	40	5
42	41	1
43	42	3
44	43	1
45	44	1
46	45	2

The SAS System

The CATMOD Procedure
Population Profiles

Sample	NOspecies	Sample Size
47	46	5
48	47	2
49	48	4
50	49	4
51	50	3
52	51	1

	53	52	6
	54	53	2
55	54		1
56	55		1
57	56		5
58	57		2
59	58		3
60	61		2
61	64		2
62	66		2
63	68		2
64	69		1
65	72		2
66	73		1
67	74		4
68	75		1
69	76		1
70	77		1
71	80		2
72	81		3
73	82		1
74	83		2
75	84		2
76	85		1
77	86		1
78	90		1
79	92		1
80	93		1
81	96		1
82	98		1
83	103		1
84	105		1
85	106		1
86	109		1
87	112		1
88	114		1
89	116		1
90	120		2
91	123		1
92	124		1

The SAS System

The CATMOD Procedure
Population Profiles

Sample	NOspecies	Sample Size
93	126	1
94	127	1
95	134	1
96	141	1
97	150	1
98	157	1
99	159	1
100	162	1
101	165	1
102	170	2
103	179	1
104	240	1

Response Profiles

Response	Infestation L
1	1

2	2
3	3
4	4
5	5

Maximum Likelihood Analysis

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates 1	2	3	4
0	0	2575.1007	1.0000	0	0	0	0
1	0	2386.1898	0.0734	1.7692	0.9096	0.9194	0.2782
2	0	2376.1168	0.004221	1.8538	1.2251	1.1966	0.3545
3	0	2376.0957	8.8786E-6	1.8666	1.2295	1.2034	0.3723

Maximum Likelihood Analysis

Iteration	Parameter Estimates 5	6	7	8
0	0	0	0	0
1	-0.0318	-0.0223	-0.0193	-0.0127
2	-0.0338	-0.0251	-0.0199	-0.009879
3	-0.0342	-0.0252	-0.0200	-0.0104

The SAS System

The CATMOD Procedure

Maximum Likelihood Analysis

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates 1	2	3	4
4	0	2376.0957	6.921E-10	1.8666	1.2295	1.2034	0.3724

Maximum Likelihood Analysis

Iteration	Parameter Estimates 5	6	7	8
4	-0.0342	-0.0252	-0.0200	-0.0104

Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	4	158.38	<.0001
NOspecies	4	59.94	<.0001
Likelihood Ratio	408	451.54	0.0673

Analysis of Maximum Likelihood Estimates

Parameter	Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	1.8666	0.1715	118.40	<.0001
	2	1.2295	0.1806	46.35	<.0001
	3	1.2034	0.1766	46.44	<.0001
	4	0.3724	0.1948	3.66	0.0559
NOspecies	1	-0.0342	0.00489	48.73	<.0001
	2	-0.0252	0.00477	27.93	<.0001
	3	-0.0200	0.00413	23.51	<.0001
	4	-0.0104	0.00385	7.25	0.0071

The CATMOD Procedure

Maximum Likelihood Predicted Values for Frequencies

		-------Observed------		------Predicted------		
	Infestation		Standard		Standard	
NOspecies	L	Frequency	Error	Frequency	Error	Residual
0	1	37	4.667857	37.14293	2.233548	-0.14293
	2	16	3.627059	19.64145	1.791233	-3.64145
	3	31	4.508018	19.13581	1.705832	11.86419
	4	5	2.173067	8.335892	1.091961	-3.33589
	5	1	0.994429	5.743916	0.841297	-4.74392
1	1	4	0.894427	2.044197	0.120279	1.955803
	2	0	0	1.090714	0.097019	-1.09071
	3	1	0.894427	1.06816	0.092947	-0.06816
	4	0	0	0.469821	0.060386	-0.46982
	5	0	0	0.327109	0.047121	-0.32711
2	1	10	1.690309	5.669646	0.326522	4.330354
	2	1	0.963624	3.052353	0.26493	-2.05235
	3	3	1.535299	3.004779	0.255321	-0.00478
	4	0	0	1.334444	0.168324	-1.33444
	5	0	0	0.938778	0.133015	-0.93878
3	1	20	3.121472	15.64313	0.882226	4.356869
	2	11	2.810238	8.497542	0.720071	2.502458
	3	5	2.087816	8.40859	0.698042	-3.40859
	4	2	1.377474	3.77053	0.46687	-1.77053
	5	1	0.987096	2.680206	0.373565	-1.68021
4	1	22	2.858571	13.90317	0.768342	8.096831
	2	5	2.070197	7.620338	0.630853	-2.62034
	3	5	2.070197	7.579774	0.615101	-2.57977
	4	3	1.656157	3.431838	0.417246	-0.43184
	5	0	0	2.46488	0.337994	-2.46488
5	1	16	2.73252	11.80073	0.639593	4.199272
	2	5	2.041241	6.526195	0.528238	-1.52619
	3	5	2.041241	6.525205	0.517974	-1.5252
	4	2	1.36626	2.98302	0.356235	-0.98302
	5	2	1.36626	2.164853	0.292092	-0.16485
6	1	13	2.71416	11.68437	0.621752	1.315631
	2	9	2.50998	6.519995	0.516464	2.480005
	3	7	2.316607	6.5529	0.509223	0.4471
	4	0	0	3.024732	0.354932	-3.02473
	5	1	0.983192	2.218004	0.294512	-1.218
7	1	18	3.04027	14.26712	0.746325	3.732875
	2	7	2.382368	8.032843	0.623387	-1.03284

The SAS System

The CATMOD Procedure

Maximum Likelihood Predicted Values for Frequencies

		-------Observed------		------Predicted------		
	Infestation		Standard		Standard	
NOspecies	L	Frequency	Error	Frequency	Error	Residual
	3	8	2.504051	8.115358	0.617916	-0.11536
	4	4	1.888801	3.782268	0.436286	0.217732
	5	0	0	2.802406	0.36627	-2.80241
8	1	10	2.509242	10.30631	0.530818	-0.30631
	2	12	2.581989	5.854997	0.44571	6.145003
	3	4	1.845916	5.945895	0.444042	-1.9459
	4	0	0	2.798036	0.317424	-2.79804
	5	1	0.981307	2.094764	0.269543	-1.09476
9	1	5	1.936492	7.556616	0.383886	-2.55662
	2	6	2.04939	4.331534	0.323905	1.668466
	3	6	2.04939	4.421651	0.324238	1.578349
	4	2	1.341641	2.100931	0.234527	-0.10093
	5	1	0.974679	1.589267	0.201379	-0.58927
10	1	11	2.810238	14.58388	0.732273	-3.58388
	2	9	2.631174	8.434867	0.620565	0.565133
	3	8	2.521701	8.65512	0.623973	-0.65512
	4	5	2.087816	4.152329	0.45637	0.847671
	5	6	2.253203	3.173805	0.396129	2.826195
11	1	8	2.19089	7.4012	0.368155	0.5988
	2	4	1.788854	4.319149	0.31318	-0.31915
	3	3	1.596872	4.454973	0.316186	-1.45497
	4	3	1.596872	2.158019	0.233668	0.841981
	5	2	1.341641	1.666659	0.204958	0.333341
12	1	8	2.332381	9.154362	0.452261	-1.15436
	2	6	2.135416	5.390324	0.385937	0.609676
	3	6	2.135416	5.588741	0.391071	0.411259
	4	3	1.624808	2.733477	0.291793	0.266523
	5	2	1.356466	2.133096	0.25854	-0.1331
13	1	6	2.026145	6.883502	0.338687	-0.8835
	2	3	1.589439	4.089658	0.289718	-1.08966
	3	4	1.777047	4.262244	0.294518	-0.26224
	4	3	1.589439	2.104898	0.221679	0.895102
	5	3	1.589439	1.659698	0.198333	1.340302
14	1	4	1.664101	4.659275	0.228983	-0.65928
	2	5	1.754116	2.793101	0.196198	2.206899
	3	1	0.960769	2.926106	0.199996	-1.92611
	4	2	1.300887	1.459064	0.151721	0.540936

The CATMOD Procedure

Maximum Likelihood Predicted Values for Frequencies

		Observed		Predicted		
NOspecies	Infestation L	Frequency	Standard Error	Frequency	Standard Error	Residual
	5	1	0.960769	1.162454	0.137009	-0.16245
15	1	3	1.519109	4.608807	0.22693	-1.60881
	2	4	1.664101	2.78771	0.194604	1.21229
	3	5	1.754116	2.935643	0.198818	2.064357
	4	1	0.960769	1.478014	0.151877	-0.47801
	5	0	0	1.189826	0.138368	-1.18983
16	1	7	1.932184	5.259654	0.260273	1.740346
	2	5	1.825742	3.210015	0.223211	1.789985
	3	1	0.966092	3.397933	0.228445	-2.39793
	4	1	0.966092	1.727355	0.175559	-0.72736
	5	1	0.966092	1.405043	0.161289	-0.40504
17	1	6	2.088932	7.628858	0.380589	-1.62886
	2	6	2.088932	4.697862	0.326166	1.302138
	3	5	1.965613	4.998735	0.334235	0.001265
	4	5	1.965613	2.565773	0.258161	2.434227
	5	0	0	2.108771	0.23906	-2.10877
18	1	1	0.966092	5.143383	0.259481	-4.14338
	2	3	1.549193	3.195806	0.222065	-0.19581
	3	7	1.932184	3.41816	0.227738	3.58184
	4	2	1.316561	1.771503	0.176631	0.228497
	5	2	1.316561	1.471148	0.16478	0.528852
19	1	4	1.712698	5.085329	0.260209	-1.08533
	2	4	1.712698	3.188169	0.222242	0.811831
	3	4	1.712698	3.427721	0.227997	0.572279
	4	2	1.316561	1.793686	0.177403	0.206314
	5	1	0.966092	1.505095	0.166648	-0.5051
20	1	4	1.788854	6.703119	0.348857	-2.70312
	2	5	1.936492	4.240236	0.297206	0.759764
	3	6	2.04939	4.58254	0.304882	1.41746
	4	3	1.596872	2.421244	0.237791	0.578756
	5	2	1.341641	2.05286	0.224811	-0.05286
21	1	4	1.595448	3.644244	0.19341	0.355756
	2	2	1.279204	2.326009	0.164296	-0.32601
	3	0	0	2.526851	0.168467	-2.52685
	4	4	1.595448	1.348042	0.131601	2.651958
	5	1	0.953463	1.154855	0.125155	-0.15485

The CATMOD Procedure

Maximum Likelihood Predicted Values for Frequencies

NOspecies	Infestation L	Observed Frequency	Observed Standard Error	Predicted Frequency	Predicted Standard Error	Residual
22	1	4	1.549193	3.274392	0.17764	0.725608
	2	1	0.948683	2.108752	0.15042	-1.10875
	3	1	0.948683	2.302746	0.154127	-1.30275
	4	2	1.264911	1.240398	0.1205	0.759602
	5	2	1.264911	1.073713	0.115219	0.926287
23	1	4	1.414214	2.588717	0.143868	1.411283
	2	0	0	1.682171	0.121416	-1.68217
	3	1	0.935414	1.846472	0.124287	-0.84647
	4	2	1.224745	1.004267	0.097189	0.995733
	5	1	0.935414	0.878374	0.093389	0.121626
24	1	4	1.788854	6.394931	0.364754	-2.39493
	2	5	1.936492	4.192877	0.306794	0.807123
	3	6	2.04939	4.626334	0.313683	1.373666
	4	3	1.596872	2.54059	0.245198	0.45941
	5	2	1.341641	2.245269	0.236663	-0.24527
25	1	2	1.264911	3.159106	0.185233	-1.15911
	2	1	0.948683	2.089928	0.155283	-1.08993
	3	2	1.264911	2.317972	0.158566	-0.31797
	4	3	1.449138	1.285278	0.123841	1.714722
	5	2	1.264911	1.147716	0.120011	0.852284
26	1	1	0.942809	2.808741	0.169533	-1.80874
	2	2	1.247219	1.874863	0.141672	0.125137
	3	3	1.414214	2.090252	0.14447	0.909748
	4	1	0.942809	1.170251	0.112694	-0.17025
	5	2	1.247219	1.055892	0.109601	0.944108
27	1	0	0	1.541312	0.095877	-1.54131
	2	2	1.095445	1.0381	0.079885	0.9619
	3	2	1.095445	1.163377	0.081351	0.836623
	4	1	0.894427	0.657646	0.063362	0.342354
	5	0	0	0.599564	0.06182	-0.59956
28	1	2	1.195229	2.13116	0.136746	-0.13116
	2	2	1.195229	1.44829	0.113637	0.55171
	3	0	0	1.631507	0.115564	-1.63151
	4	2	1.195229	0.931219	0.089856	1.068781
	5	1	0.92582	0.857825	0.087918	0.142175
29	1	1	0.935414	2.405198	0.159309	-1.4052
	2	1	0.935414	1.64923	0.132079	-0.64923

The CATMOD Procedure

Maximum Likelihood Predicted Values for Frequencies

		Observed		Predicted		
NOspecies	Infestation L	Frequency	Standard Error	Frequency	Standard Error	Residual
	3	1	0.935414	1.867526	0.134145	-0.86753
	4	2	1.224745	1.07627	0.104115	0.92373
	5	3	1.369306	1.001777	0.102126	1.998223
30	1	3	1.477098	3.26544	0.223383	-0.26544
	2	1	0.953463	2.259242	0.18484	-1.25924
	3	2	1.279204	2.571582	0.187506	-0.57158
	4	3	1.477098	1.496395	0.145263	1.503605
	5	2	1.279204	1.407341	0.142805	0.592659
31	1	2	1.195229	2.05154	0.145001	-0.05154
	2	1	0.92582	1.432161	0.119793	-0.43216
	3	1	0.92582	1.638633	0.121391	-0.63863
	4	1	0.92582	0.962762	0.093874	0.037238
	5	2	1.195229	0.914903	0.092469	1.085097
32	1	0	0	0.867922	0.063395	-0.86792
	2	0	0	0.611341	0.052313	-0.61134
	3	2	0.816497	0.703113	0.052961	1.296887
	4	1	0.816497	0.417113	0.040887	0.582887
	5	0	0	0.40051	0.040347	-0.40051
33	1	1	0.894427	1.427747	0.107784	-0.42775
	2	0	0	1.014717	0.088874	-1.01472
	3	3	1.095445	1.17311	0.089907	1.82689
	4	0	0	0.702682	0.069303	-0.70268
	5	1	0.894427	0.681744	0.068498	0.318256
34	1	2	1.095445	1.409019	0.109939	0.590981
	2	0	0	1.010418	0.090618	-1.01042
	3	1	0.894427	1.174214	0.091617	-0.17421
	4	1	0.894427	0.710164	0.070527	0.289836
	5	1	0.894427	0.696185	0.069813	0.303815
35	1	0	0	0.556142	0.044846	-0.55614
	2	1	0.707107	0.402403	0.036965	0.597597
	3	0	0	0.470067	0.037357	-0.47007
	4	1	0.707107	0.287053	0.028726	0.712947
	5	0	0	0.284336	0.028476	-0.28434
36	1	1	0.866025	1.097408	0.091439	-0.09741
	2	1	0.866025	0.801187	0.075402	0.198813
	3	0	0	0.940772	0.076184	-0.94077
	4	0	0	0.580067	0.058535	-0.58007

The SAS System

The CATMOD Procedure

Maximum Likelihood Predicted Values for Frequencies

		-------Observed------		------Predicted------		
	Infestation		Standard		Standard	
NOspecies	L	Frequency	Error	Frequency	Error	Residual
	5	2	1	0.580565	0.058104	1.419435
37	1	1	0.948683	2.70647	0.232972	-1.70647
	2	1	0.948683	1.9937	0.192262	-0.9937
	3	1	0.948683	2.35322	0.194247	-1.35322
	4	4	1.549193	1.465034	0.14917	2.534966
	5	3	1.449138	1.481575	0.148271	1.518425
38	1	0	0	1.868697	0.166134	-1.8687
	2	1	0.92582	1.388949	0.137259	-0.38895
	3	2	1.195229	1.647939	0.138696	0.352061
	4	1	0.92582	1.0359	0.106487	-0.0359
	5	3	1.309307	1.058515	0.105991	1.941485
39	1	2	0.816497	0.789845	0.072501	1.210155
	2	1	0.816497	0.592353	0.059989	0.407647
	3	0	0	0.706459	0.060636	-0.70646
	4	0	0	0.448389	0.046559	-0.44839
	5	0	0	0.462954	0.046408	-0.46295
40	1	1	0.894427	1.298113	0.122985	-0.29811
	2	0	0	0.982294	0.101943	-0.98229
	3	3	1.095445	1.177608	0.103095	1.822392
	4	1	0.894427	0.754675	0.079194	0.245325
	5	0	0	0.78731	0.079056	-0.78731
41	1	0	0	0.25598	0.025022	-0.25598
	2	0	0	0.195446	0.020785	-0.19545
	3	0	0	0.235525	0.021034	-0.23553
	4	0	0	0.152401	0.016169	-0.1524
	5	1	0	0.160648	0.016167	0.839352
42	1	0	0	0.757063	0.076322	-0.75706
	2	0	0	0.583234	0.06355	-0.58323
	3	2	0.816497	0.70649	0.064366	1.29351
	4	1	0.816497	0.461581	0.049532	0.538419
	5	0	0	0.491632	0.049609	-0.49163
43	1	0	0	0.248746	0.025852	-0.24875
	2	0	0	0.193356	0.021584	-0.19336
	3	0	0	0.235436	0.021883	-0.23544
	4	1	0	0.155313	0.016863	0.844687
	5	0	0	0.167148	0.01692	-0.16715