MULTINOMIAL REGRESSION IN INSECT CHOICE STUDIES: A CASE OF LEAF MINER PARASITOIDS' CHOICES

KEVIN KANYUIRA GIKONYO

MASTER OF SCIENCE

(Research Methods)

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

2013

Multinomial Regression in Insect Choice Studies: A Case of Leaf Miner Parasitoids' Choices

Kevin Kanyuira Gikonyo

A dissertation submitted to the Faculty of Agriculture in partial fulfillment of the requirements for the degree of Master of Science in Research Methods of Jomo Kenyatta University of Agriculture and Technology

2013

DECLARATION

This dissertation is my original work and has not been presented for the award of a degree in any other University.

Signature: …… **Date:**……..

Kevin Kanyuira Gikonyo (AG-332-1828/ 2010)

This dissertation has been submitted for examination with our approval as Supervisors.

Signature: …… **Date:** ……..

Dr. Elijah Ateka

Department of Horticulture, JKUAT

Signature: …… **Date:** ……..

Dr. Anthony Wanjoya

Department of Statistics and Actuarial Sciences, JKUAT

Signature: …… **Date:** ……..

Dr. Daisy Salifu

International Centre for Insect Physiology and Ecology (*icipe*),

ACKNOWLEDGEMENTS

My sincere gratitude goes to my supervisor Dr. Daisy Salifu (*icipe)* for introducing me to multinomial models and guidance during my study. Special thanks to my supervisors Dr. Anthony Wanjoya and Dr. Elijah Ateka (JKUAT) for their guidance, advice and encouragement during the entire study period.

I would like to appreciate Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) for their financial support during my studies. Many thanks to International Centre for Insect Physiology and Ecology (*icipe*) for awarding me an internship position at Duduville campus and providing data used in this study. I gratefully acknowledge Ministry of Fisheries Development for granting me a study leave.

To my family, colleagues and friends, I appreciate the support you gave me during my studies.

ABSTRACT

Researchers are often confronted with multinomial data in insect choice studies. Common choice models available to researchers for analysis of multinomial data include multinomial logit (MNL) and multinomial probit model (MNP). MNL relies on the Independence from Irrelevant Alternatives (IIA) assumption which is violated when choices are correlated resulting in overestimating the probability of selecting correlated alternatives. The more flexible MNP model relaxes IIA assumption and allows modelling correlated errors. Little evidence exists on the performance of multinomial logit and multinomial probit models on insect choice data. This study investigated the performance of the two models in terms of predictive accuracy and goodness of fit on choice data collected in a laboratory experiment involving leaf miner parasitoids. Sum of squared deviations of predicted probabilities from observed probabilities was used to evaluate predictive accuracy. Akaike Information Criterion and Bayesian Information Criterion were used to evaluate goodness of fit. The findings indicated that MNP resulted in a higher predictive accuracy than MNL. The observed predictive accuracy for MNP came with a cost on the goodness of fit since MNL had a better fit to the data than MNP model from the Bayesian Information Criterion statistics despite violation of IIA assumption. There was little evidence that imposing homoskedastic restriction on the covariance matrix of the MNP model improved predictive accuracy and goodness of fit. MNL and MNP models resulted in qualitatively similar predicted probabilities. These findings suggest recommending use of the more analytically-tractable MNL in modelling insect choice data when IIA assumption is violated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
BFGS	Broyden-Fletcher-Goldfarb-Shanno algorithm
BIC	Bayesian Information Criterion
GHK	Geweke-Hajivassiliou-Keane simulator
IIA	Independence from Irrelevant Alternatives
MACML	Maximum Approximate Composite Marginal Likelihood
MLE	Maximum Likelihood Estimation
MNL	MultiNomial Logit
MNP	MultiNomial Probit
MSL	Maximum Simulated Likelihood
MSM	Method of Simulated Moments

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Choice studies have been widely used to understand insect, host plants and parasitoid interactions (Hern and Dorn, 2001; Turlings *et al*., 2004; Richard and Davison, 2007). Given several alternatives, parasitoids or insects would choose one host and not the other. Choice studies where insects are evaluated on how they respond to more than two stimuli lead to multinomial response variables. The multinomial distribution is a generalization of binomial distribution to cases with more than two possible ordered or unordered outcomes. Given a response with more than two possible outcomes and independent trials with similar category probabilities for each trial, the distribution of counts in the various categories follows a multinomial distribution (Agresti, 2007).

There are several methods that are available for analysis of multinomial data. The most common form of categorical data analysis in biological sciences which give rise to frequency counts have been handled by constructing cross-tabulations or contingency tables and then using chi-square tests to examine associations between two or more categorical variables (Quinn and Keough, 2002). However, such an approach is not adequate for a study aimed at estimating the response given a change in explanatory variable since contingency tables are analyzed for association where neither variable is considered as a predictor or a response variable. The results are valid provided that fewer than 20% of cells have expected count below 5, and none are below 1 (Logan, 2010). Fisher's exact test extends the chi-square test in studies involving small samples sizes. Further the chi-square test for association between

variables is a test and not a model and therefore one is not able to obtain predicted values and their measure of precision for instance, standard errors.

Log-linear models for contingency tables have also been used in modeling the association between categorical response variables (Agresti, 2007). They are able to estimate log-odds and interaction effects and the residuals have to follow a poisson distribution (Logan, 2010). There is no distinction between response and explanatory variables in log-linear models (Quinn and Keogh, 2002). However, in studies with a large number of variables, log-linear models are limited because of the increase in possible associations and interactions of variables that restricts the range of good fitting models (Agresti, 2007).

Multinomial regression has been explored extensively in social science studies involving transport choice studies (McFadden, 1974; Bhat, 1998; Munziaga *et al*., 2000) and voter preferences among presidential candidates from more than two political parties (Alvarez and Nagler, 1998; Dow and Endersby, 2004; Kropko, 2008). Multinomial regression has also been used in animal behavioral studies involving alligator food preferences among alternative food choices (Agresti, 2007). Multinomial regression has potential application in insect behavioral studies where choice experiments are involved since there are several advantages over other methods. Some of the advantages are that one is able to model the relationship of multinomial responses with their explanatory variables, analyze association and estimate log-odds.

Commonly used multinomial regression models are multinomial logit (MNL) and multinomial probit (MNP) models Kropko (2008). The multinomial logit model relies on the independence from irrelevant alternatives (IIA) assumption which assumes modelled choices are not correlated Train (2003). Choices that do not share attributes are untenable in many behavioural studies leading to use of the more flexible multinomial probit model that relaxes

the limiting IIA assumption. This study seeks to evaluate the performance of MNL and MNP models for analysis of counts data arising from choice tests in insect behavioral studies.

1.2 Problem Statement and Justification

The use of multinomial regression is a challenge because of difficulties in implementing such models and interpretation of results due to the non-linear nature of the model (Long and Freese, 2001; Hoetker, 2007). The two commonly used multinomial regression models are multinomial logit and multinomial probit models. Multinomial logit model imposes the restrictive Independence from Irrelevant Alternatives (IIA) assumption which results in a very high joint probability of selecting similar or correlated alternatives. Multinomial probit model has a flexible error structure which relaxes IIA assumption and allows modelling correlated choices. However, MNP model has been found to be computationally intensive due to evaluation of multi-dimensional integrals. Little evidence exists on the performance of MNL and MNP models in insect choice data given the large data requirements of MNP model compared to MNL.

The study addressed the statistical analysis challenges in multinomial count data by generating a step by step approach that addressed the problem of data requirements and implementing multinomial models, estimating odds ratios or probabilities, and interpreting results. The study also provided empirical evidence on the performance of multinomial logit and multinomial probit models on insect count data.

1.3 Overall objective

To evaluate the performance of multinomial logit and multinomial probit models in the analysis of counts from insect choice studies.

1.3.1 Specific objectives

1. To evaluate the predictive accuracy of multinomial logit and multinomial probit models

2. To evaluate goodness of fit of multinomial logit and multinomial probit models

1.4 Study limitations

The MNL and MNP model performance findings are limited to one insect choice dataset and generalizing the findings to different datasets may require simulation studies.

CHAPTER TWO

LITERATURE REVIEW

Introduction

In this chapter, design of statistically efficient choice experiments and random utility maximization framework are reviewed. Multinomial logit and multinomial probit models are reviewed in detail. The chapter ends with a review of studies that have compared MNL and MNP.

2.1 Choice experiments

Insect behavior has been studied by setting up choice experiments that evaluate insect response to different stimuli. Where insects are evaluated on how they respond to more than two stimuli, the data generated is multinomial in nature. Choice analysis involves explaining variability in a behavioural response (Hensher *et al*., 2005). Choice variability is a result of observed influences and unobserved influences.

2.1.1 Efficient design of choice experiments

Generating statistically efficient choice experimental designs is the least understood process in choice modelling as observed by Hensher *et al*. (2005). Design of choice experiments shares design principles with other experimental studies and generally involves: identifying alternatives; identifying choice behavior influences (attributes); determining attribute levels; ensuring orthogonality of attributes (statistical independence); determining main and interaction effects; degrees of freedom required to estimate model; treatment combinations required (design degrees of freedom); blocking the design; and randomizing the choice sets.

Dow and Endersby (2004) underscore the importance of choice models capturing the process that generates choice data which necessitates a clear understanding by a choice analyst of experimental design used.

2.1.2 Random utility maximization framework

Individuals choose an alternative that maximizes their utility from a choice set. Choice models are usually derived from the utility-maximization framework (Train, 2003; Kropko, 2010) and resulting models are known as random utility models. This random utility maximization equation is of the form,

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{2.1}$$

where U_{ij} represents overall utility for an alternative, V_{ij} is the observed influences of utility and ε_{ij} is the unobserved influences (error).

The probability of an insect choosing alternative *i* over alternative *j* is equal to the probability that the utility of *i* being greater than (or equal to) the utility of *j* after evaluating all alternatives in a given choice set of *j=1,…,i,…J* alternatives (Hensher *et al*., 2005). This is given by,

$$\text{Prob}_{\text{i}} = \text{Prob}\left(U_i \geq U_j\right) \ \forall \ \ j \in \ j = 1,....., J; i \neq j) \tag{2.2}$$

The analyst's equation is of the form,

$$\text{Prob}_i = \text{Prob}\ [\left(V_i + \varepsilon_i\right) \geq \ \left(V_j + \varepsilon_j\right) \forall\ j \in j = 1,....., \ J; i \neq j] \tag{2.3}$$

Rearranging to reflect random utility maximization results in,

$$\text{Prob}_i = \text{Prob}\,[\,(\varepsilon_j - \varepsilon_i) \leq (V_i - V_j)\,\forall\, j \in\ j = 1,.....,\ J; i \neq j\,] \qquad (2.4)$$

Different choice models arise in relation to the assumed error structure of above ε_{ij} (Dow and Endersby, 2004).

2.2 Modelling multinomial data

Several methods exist for modelling multinomial data, 'traditional' methods of analyzing multinomial data include: analysis of frequency counts using chi-square test for contingency tables and log-linear models for contingency tables. This review focuses on describing multinomial logit and multinomial probit models in detail.

2.3 Multinomial regression

Multinomial regression models are applied in analyzing data where the categorical response variable has more than two possible outcomes while the independent variables could be continuous, categorical variables, or both (Hosmer and Lemeshow, 2000). The categorical response variable may be ordered or unordered. Ordered or ordinal response variables are unique values that represent rank order on some dimension, but there are not enough values to treat the variable as continuous. Unordered or nominal response variables are those whose values provide classification but provide no indication of order. This study reviewed nominal muiltinomial regression models.

2.3.1 Multinomial logit model

McFadden (1974) first introduced the multinomial logit model to explain the choice of transportation modes of urban commuters with the random utility model. MNL continues to be a popular choice model because choice probabilities formula has a closed form and is

readily interpretable and taste variation that relates to observed attributes can be represented by MNL (Train, 2003).

2.3.1.1 Model assumptions

Multinomial logit model assumes independence of irrelevant alternatives (IIA) which implies that the odds of choosing an alternative *i* relative to an alternative *j* are independent of the characteristics of or the availability of alternatives other than *i* and *j* (McFadden, 1973).

The IIA assumption requires that if a new alternative is available, then prior probabilities adjust precisely to retain original odds among all pairs of outcomes. In a hypothetical case where insects choose from 2 host plants A and B, the probability of choosing either plant under IIA $P_A=P_B=1/2$. Introducing another host plant C with similar characteristics to plant B, the probability of an insect choosing host plant C or B is the same $P_C/ P_B=1$. Under the MNL the probabilities would be $P_A=P_B=P_C=1/3$ while we would expect the probabilities to be $P_A=1/2$ and $P_B=P_C=1/4$. Maddala (1983) also observed that the MNL predicts a very high joint probability of selecting similar alternatives as observed in the above example which may not be appropriate in some applications. However, IIA assumption has an advantage when a large number of choices are considered. IIA allows a small subset of the choices to be used analyzed since relative probabilities in the choice subset are not affected by choices not included in the subset which significantly reduces computational time (Train, 2003).

This assumption can be tested using the Hausman-McFadden test (Hausman and McFadden, 1984). If a subset of choices is truly irrelevant, removing them from the model does not change the parameter estimates systematically though leading to inefficiency, the exclusion does not result in inconsistency. However, if remaining odds ratios are not truly independent from these alternatives; the parameter estimates obtained when these choices are included are

inconsistent (Greene, 2003). The Hausman-McFadden test has been criticized for giving inconsistent results when the base category is altered (Long and Freese, 2001).

The model also assumes choice error terms are independent and identically distributed (Train, 2003).

2.3.1.2 Limitations of multinomial logit model

Imposition of the independence of irrelevant alternatives (IIA) is restrictive for behavioral choice models since IIA limits the application of multinomial logit regression to choices that are correlated or share important qualities. MNL has a restricted substitution pattern due to IIA assumption which limits its application in studies interested in investigating the effect of dropping or adding some choices. Lastly, MNL is not able to represent random taste variation, and is also not applicable in analysis of panel choice data since error terms exhibit temporal correlation (Train, 2003).

2.3.2 Multinomial probit model

The model was proposed by Aitchison and Bennet (1970) and has a significant advantage over the multinomial logit model since MNP allows the modeling of correlated choices through the relaxing the IIA restriction. The multinomial probit model introduces additional parameters to the covariance matrix of the errors which increases flexibility of the error structure which allows any pattern of substitution, handles random taste variation, and can be applied in analysis of panel choice data (Train, 2003).

2.3.2.1 Model assumptions

The model assumes that the choice error terms have a multivariate normal distribution (Alvarez and Nagler, 1994; Long, 1997).

2.3.2.2 Limitations of multinomial probit model

MNP model's increased flexibility involves the evaluation of high dimensional multivariate normal integrals for solving probabilities which increases time before reaching convergence and becomes challenging especially if probability is close to zero or one (Cameron and Trivedi, 2005). This computational challenge has been slightly reduced with the development of new algorithms, advances in computing power and Bayesian estimation methods (Train, 2003). Greene (2003) observed the need of imposing additional restrictions on the error covariance matrix of MNP models estimated using maximum simulated likelihood to enhance convergence.

An alternative estimation procedure is method of simulated moments (MSM) though Cameron and Trivedi (2005) note an efficiency loss for MSM where low and large simulator draws are used which reduces computation. Bhat (2011) proposed the simpler maximum approximate composite marginal likelihood (MACML) estimation approach which has a computational time efficiency advantage relative to MSL approach. Bhat (1998) proposes imposing restrictions on the covariance matrix to reduce the number of parameters estimated. MNP model has been found to require a very large sample sizes to obtain reliable and precise estimates (Alvarez and Nagler, 1994; Dow and Endersby, 2004). The distribution of error terms has also been observed not to follow a normal distribution in some cases (Train, 2003).

2.4 Comparative studies on MNL and MNP models

Studies that compare the two models have mostly been in the field of political science (Alvarez and Nagler, 1994; Alvarez and Nagler, 1998; Quinn *et al*., 1999; Dow and Endersby, 2004; Kropko, 2008; Kropko, 2010) and transport studies (Munziaga *et al*., 2000). MNL and MNP models have been evaluated on the basis of precision of estimates, goodness of fit, time

taken before convergence, accuracy of predicted probabilities, rate of correct signs for coefficients, and implication of violating IIA assumption.

Their findings seem to contradict each other with regards to the performance of MNL and MNP models. Alvarez and Nagler (1994) contended with the finite sample behavior of both MNL and MNP models and found that in samples of less than 1000, both MNL and MNP accurately estimated parameters in the systematic component. However, the random component was weakly identified for MNP due to large sample size required to estimated covariance matrix parameters. Alvarez and Nagler (1998) found that MNP performed better than MNL by predicting more accurate probabilities after dropping or adding an alternative. Quinn *et al*. (1999) give a balanced assessment where MNP performs better on Dutch voter data while there is no difference between MNL and MNP on the British voter data. Their study compares the two models goodness of fit using the Bayes factor methodological approach.

MNL has been found to be more robust than MNP even in cases where IIA assumption has been violated (Dow and Endersby, 2004; Kropko, 2008; Kropko, 2010). However, Dow and Endersby (2004) underscore the importance of choice models capturing the process which generates the observed data while resulting in accurate estimates. MNL and MNP were compared on simulated transport data based on heteroskedasticity between options and between observations by Munziaga *et al*. (2000). They found that the MNL was fairly robust to homoskedasticity violations but justified the use of MNP in cases where heteroskedasticity between options was present. Munziaga *et al*. (2000) also note that the MNP requires very large sample sizes. Jones *et al*. (2010) compare MNL and MNP using student success in higher education and find no significant differences in conclusions arrived at by both models.

They however note that the MNP is susceptible to convergence problems unlike the MNL model.

Insect choice studies involving multinomial data have applied log-linear models in the analysis (Turlings *et al.*, 2004; Richard and Davison, 2007). Hern and Dorn (2001) mention multinomial logit models in analyzing insect response to different apple volatiles but instead apply log-linear and binomial models to their data. The binomial model was used after dropping one of the 3 choices in the study. They recommend further research on choice models in insect behavioural studies. Turlings *et al.* (2004) studied the application of log-linear models to wasp choice of volatiles in a six-arm olfactometer study. Richard and Davison (2007) extended the log-linear model proposed by Turlings *et al.* (2004) to insect choice studies with high overdispersion by proposing an inhomogeneous Markov chain model which explains overdispersion in analysis of count insect data.

This study seeks to provide more empirical evidence on the performance of MNL and MNP models by applying them on insect choice data while extending choice models research in insect behavioural studies.

CHAPTER THREE

METHODOLOGY

Introduction

This chapter describes the data used in the study and how the data was managed. Summary statistics, exploratory plots, MNL and MNP models are also described. The chapter ends by detailing predictive accuracy and goodness of fit methods used in the analysis.

3.1 Data Description

Secondary data used in this study came from a laboratory experiment conducted by Musundire *et al.* (2012) where parasitoids *Diglyphus isaea* (Walker) (Hymenoptera: Eulophidae) were allowed to either parasitise or host feed on larva of leaf miner flies reared on different leaf miner host plants. The study involved 3 leaf miner fly species: *Liriomyza huidobrensis* (Blanchard), *Liriomyza sativae* (Blanchard), and *Liriomyza trifolii* (Burgess) (Diptera: Agromyzidae) and 4 host plants considered to be of economic importance: *Phaseolus vulgaris, Pisum sativum, Solanum lycospersicum* and *Vicia faba*.

The aim of the study was to investigate whether leaf miner species influenced parasitoids choice of either host feeding or parasitizing leaf miner larva.

3.1.1 Experimental design

Four potted leaf miner host plants *P. vulgaris, P. sativum, S. lycospersicum* and *V. faba* were each infested with live late second to third instar larvae of *Liriomyza* species and placed in

ventilated Perspex cages (50 × 50 × 45 cm). *P. vulgaris*, *P. sativum*, *S. lycospersicum* plants used in the experiment were each two weeks old while was *S. lycospersicum* 5 weeks old.

3 generations before conducting the experiment, the 3 *Liriomyza* species were reared on each of the 4 host plants,to avoid bias resulting from rearing leaf miner on only one host plant. The *Liriomyza* were reared at a temperature of 27 ± 0.6^0C, relative humidity ranging between 27-35% and a 12L: 12D photoperiod.

50 male and female (sex ratio 1:1) adult *Liriomyza* aged 4 days were released to infest 16potted plants of each of the 4 host plant species placed in ventilated cages. The adult Liriomyza were given a 4 hour oviposition period after which infested host plants were transferred to a similar cage without adult Liriomyza where leaf miner larvae were allowed to develop until late second instar and third instar larval stages.

45 pre-mated *D. isaea* were then released per cage for 48 hours on leaf miner larvae infested host plants where they were allowed to mate and given a preoviposition period of 12 hours. Larvae were recorded as host fed once they became flaccid with black spots on their body as a result of stings of parasitoid females and parasitized when they were found with immatures of leaf miner flies. The experiment was replicated four times for each host plant and *Liriomyza* species with each of 4 cages constituting a replicate.

The response variable was number of parasitoids (counts) that parasitized or host fed leaf miner larva on a given host plant and explanatory variable was leaf miner species.

3.1.2 Data management

Before fitting the MNL model, data was organized in wide format with one row providing data for each choice situation for an individual parasitoid (Appendix 1).

For the MNP the data was organized in long format with one row for each alternative made by an individual parasitoid and since there were 3 host plant alternatives, the dataset had 3 rows for each choice made by parasitoids (Appendix 2).

Due to the small numbers of leaf miner larvae in *P. sativum* attributed to difficulties in rearing *L.sativae* and *L. trifolii* larva in *P. sativum*, *P. sativum* was excluded in the analysis both for host feeding and parasitism to ensure convergence (Agresti, 2007; Long and Freese, 2001). Agresti (2007) observed that empty cells as a result of zero counts result in infinite estimates and flat regions in the log likelihood leading to convergence difficulties.

3.2 Data Analysis

3.2.1 Summary statistics

The data was summarized using a contingency table and association tested between host plants and leaf miner species using chi square test for contingency tables. Fishers' exact test was used to analyze the association where fewer than 20% of cells in the contingency table had expected counts below 5 and none were below 1.

Two exploratory box plots were also used to show visualize patterns and check for outliers.

i. Box plot of total parasitoids that chose different host plants

ii. Box plot of parasitoids that chose different host plants including leaf miner species

3.2.2 Models

Nominal multinomial logit and multinomial probit models were fitted on the data.

3.2.2.1 Multinomial logit model

Two MNL models were fitted on the data one for host feeding and the other for parasitism cases.

The MNL model log odds equation was of the form,

$$\text{logit}\left(P_j\right) = \beta_0 + \beta_1 X + \varepsilon_j$$

where,

P_j = probability of choosing the jth host plant

β_0= constant term

β_1= leaf miner species parameter estimate

X= leaf miner species

ε_j = error terms

j = 1,…,3 host plant alternatives

Counts of parasitoids that host fed and parasitized leaf miner larvae were used as frequency weights for each host plant choice.

MNL models were fitted using default Stata settings and parameters estimated via maximum likelihood (MLE) implemented by the Newton-Raphson algorithm.

Testing IIA assumption

Hausman-McFadden test was used in testing IIA assumption for MNL model (Hausman and McFadden, 1984).

3.2.2.2 Multinomial probit model

Two MNP models estimated via maximum simulated likelihood (MSL) were fitted on the data for host feeding and parasitism cases. The simulation was implemented by the Geweke-Hajivassilou-Keane (GHK) simulator and optimization was via the Stata default Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. Hajivassiliou *et al*. (1996) compared 11 different simulation methods and came to the conclusion that the Geweke-Hajivassiliou-Keane (GHK) simulator performed better than the other simulation methods for MNP models.

The response variable was also host plant while explanatory case-specific (does not vary with choices) variable was leaf miner fly species.

$$\Phi^{-1}(P_j) = \beta_0 + \beta_1 X + \varepsilon_j$$

where,

P_j = probability of choosing the *j*th host plant

β_0= constant term

β_1 = leaf miner species parameter estimate

X = leaf miner species

ε_j = error terms

j = 1,…,3 host plant alternatives

Counts of parasitoids that host fed and parasitized leaf miner larvae were used as frequency weights for each host plant choice.

Two restrictions were imposed on the variance error structure of the MNP models:

i. Heteroskedastic variance error structure (default Stata setting) which accommodated correlated error terms which had different variance for each choice error.

ii. Homoskedastic variance error structure which forced the diagonal elements in the variance-covariance matrix to be 1. This restriction accommodated correlated errors only.

Both models allowed an unstructured correlation error structure which relaxed the IIA assumption.

Base Categories

P. vulgaris was used as the Stata default base category for host feeding case since it had the highest frequency counts while *V. faba* was used for parasitism case respectively.

3.3 Predictive accuracy evaluation

The predictive accuracy of predicted probabilities from MNL, homoskedastic MNP and unrestricted MNP models was evaluated using the sum of squared deviations (Maddala, 1983). The observed choice probabilities were generated in 0 and 1 binary format for the three host plants. When a parasitoid chose a host plant the observed probability was 1 and 0 otherwise. The three models were fitted and their predicted probabilities generated. The

predicted probabilities were then subtracted from observed probabilities, the deviations squared and the squared deviations summed to obtain the sum of squared deviations from predicted probabilities statistic (Appendix 4).

This approach has an advantage of being more robust by considering all probabilities compared to percent correctly predicted method which does not distinguish between predicted probabilities of 0.51 and 0.99.

Smaller sum of squared deviations indicate a higher predictive accuracy for the model.

3.4 Goodness of fit evaluation

Goodness of fit for the models was evaluated using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Both AIC and BIC select models that minimize the distance between fitted values and expected true values (Agresti, 2007). Lower AIC and BIC statistics indicate better fit.

3.4.1 Akaike Information Criterion (AIC)

The test was introduced by Akaike (1973) for the purpose of selecting an optimal model from within a set of proposed models. AIC measures the relative goodness of fit of competing statistical models taking into account the number of fitted parameters using the Kullback-Leibler information or distance (Burnham and Anderson, 2002).

The AIC selects the model that minimizes the distance between fitted values and expected true values (Agresti, 2007) and is of the form,

AIC has the form,

$$\mathrm{AIC} = -2\log L + 2k \tag{3.1}$$

However, the AIC has a tendency of selecting models with too many parameters in cases where the sample size is large.

3.4.2 Bayesian Information Criterion (BIC)

Bayesian Information Criterion was proposed by Schwarz (1978) who extended the AIC, arguing from a bayesian viewpoint. BIC has an advantage over AIC since BIC selects the correct model with a probability of 1 as the sample size increases or decreases as was demonstrated by Raftery (1986) by adding a constant to the likelihood function.

BIC has the form,

$$\mathrm{BIC} = -2\log L + k\log(n) \tag{3.2}$$

Where n is the sample size, L is the maximized likelihood and k is the number of regressors including the intercept.

BIC was used since it penalizes model complexity more than AIC (Raftery, 1986; Logan, 2010).

3.4.3 Ease of convergence

The ease of convergence was evaluated by counting the number of iterations before convergence of log-likelihood and simulated log-likelihood functions for MNL and MNP models respectively. Hensher *et al*. (2005) noted that number of iterations can be used to identify convergence difficulties with models whose iterations exceed 150 having high chances of not converging and when the models converge, the estimates are usually poor and of little use.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

In this chapter, summary statistics and exploratory plots are presented first. Predictive accuracy and goodness of fit results for MNL and MNP models are also presented. The chapter ends with a detailed discussion of the results.

4.0 RESULTS

4.1 Summary statistics

Table 1: Contingency table showing *D. isaea* counts for host feeding case

	Leaf miner species			
	L. huidobrensis	*L. sativae*	*L. trifolii*	Total
Host plant				
P. vulgaris	28	106	7	141
P. sativum	6	0	4	10
S. lycospersicum	3	16	88	107
V. faba	24	59	14	97
Total	61	181	113	355

Parasitoids prefer host feeding leaf miner larvae on *P. vulgaris* as a host plant and least prefer those on *P. sativum* (Table 1). Presence of sparse data is evident from counts of parasitoids that chose to host feed *P. sativum.* where only 0 and 4 parasitoids chose *P. sativum* infested with *L. sativae* and *L. trifolii* leaf miner larvae respectively. Using the chi-square for testing association between host plant and leaf miner species would not be appropriate since the test assumptions are not met. Fisher's exact test was used instead and revealed a significant

association between parasitoid choice of host plants and leaf miner species (χ^2= 204.56, *df* = 6, *P* <0.0001). These results explain overall association but do not explain which host plant is significantly associated to specific leaf miner species thus the chi-square test is limited in analysis of multinomial data.

The variable *P. sativum* which had sparse data was excluded in the analysis to ensure convergence was attained (Table 2).

Table 2: Contingency table of parasitoid counts and percentages for host feeding case (without *P. sativum*)

	Leaf miner species			
	L. huidobrensis	*L. sativae*	*L. trifolii*	Total
Host plant				
P. vulgaris	28	106	7	141
	50.91%	58.56%	6.42%	40.87%
S. lycospersicum	3	16	88	107
	5.45%	8.84%	80.73%	31.01%
V. faba	24	59	14	97
	43.64%	32.60%	12.84%	28.12%
Total	55	181	109	345
	100.00%	100.00%	100.00%	100.00%

The total percentage of parasitoids that chose to host feed leaf miner larvae were highest in *P. vulgaris* (40.87%) and lowest in *V. faba* (28.12%). *L. sativae* leaf miner larva had the highest counts (181) of host fed larvae while *L. huidobrensis* had the lowest count (55).

There was a highly significant association between host plant and leaf miner larvae species for host feeding by *D. isaea* parasitoids from the chi-square test (χ^2= 189.08, *df* = 4, *P*<0.0001).

Table 3: Contingency table of parasitoid counts for parasitism case

	Leaf miner species			
	L. huidobrensis	*L. sativae*	*L. trifolii*	Total
Host plant				
P. vulgaris	100	269	33	402
P. sativum	57	0	9	66
S. lycospersicum	53	64	139	256
V. faba	250	132	89	471
Total	460	465	270	1,195

The variable *P. sativum* was excluded from the analysis to ensure convergence since under *L. sativae* 0 counts were recorded as observed in Table 3. From Fisher's exact test, there was a significant association between host plant and leaf miner species under parasitism (χ^2= 395.40, *df* = 6, *P* <0.0001).

Table 4: Contingency table of parasitoid counts and percentages for parasitism case (without *P. sativum*)

	Leaf miner species			
	L. huidobrensis	*L. sativae*	*L. trifolii*	Total
Host plant				
P. vulgaris	100 24.81%	269 57.85%	33 12.64%	402 35.61%
S. lycospersicum	53 13.15%	64 13.76%	139 53.26%	256 22.67%
V. faba	250 62.03%	132 28.39%	89 34.10%	471 41.72%
Total	403 100.00%	465 100.00%	261 100.00%	1,129 100.00%

The total percentage of parasitoids that chose to parasitize larva was the highest in *V. faba* (41.72%) and lowest in *S. lycospersicum*. *L. sativae* leaf miner larva had the highest counts (465) of parasitized larvae while *L. trifolii* had the lowest counts (261).

There was a highly significant association between host plant and leaf miner species for parasitism (χ^2= 319.82, *df* = 4, *P* <0.0001). The significance results reveal overall association but do not reveal which host plant is significantly associated to specific leaf miner species thus the chi-square test is limited in analysis of multinomial data.

Overall, the total counts of parasitoids that chose to parasitize leaf miner larvae was higher (1,129) than the total number that chose to host feed (345) as observed in Tables 1 and 3 respectively. This indicated that *D. isaea* parasitoids parasitized 3 times as many leaf miner larvae as they host fed.

Box plots for exploratory data analysis

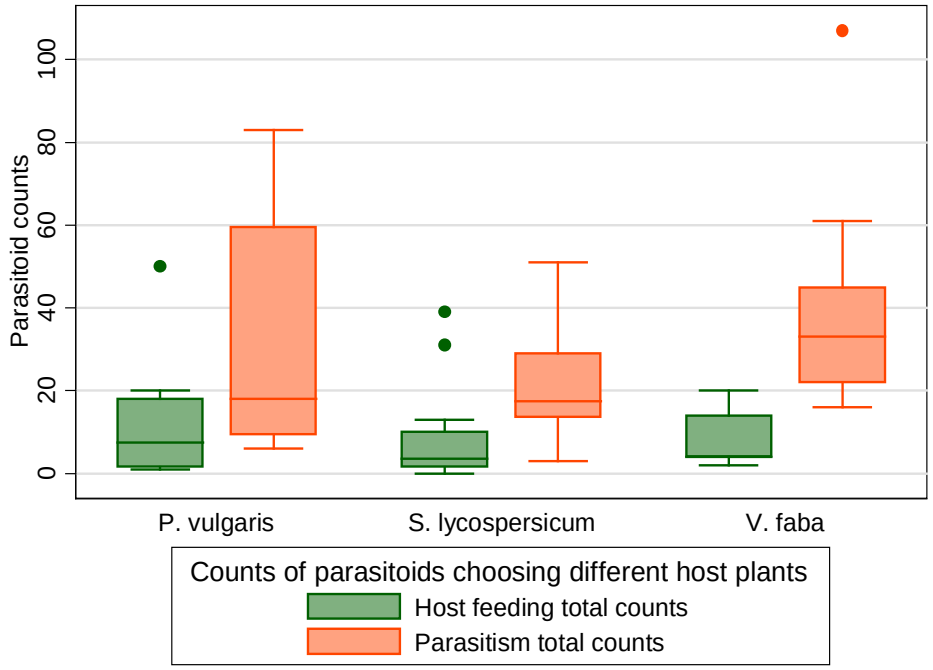

Figure 1: Box plot of total parasitoids that chose different host plants under host feeding and parasitism cases

Parasitoid choice of host plants displayed higher variability under parasitism case than for host feeding. Choice of *P. sativum* displayed very low variability under host feeding and highly skewed distribution under parasitism. An extreme outlier was observed in *V. faba* under parasitism prompting further examination (Figure 2).

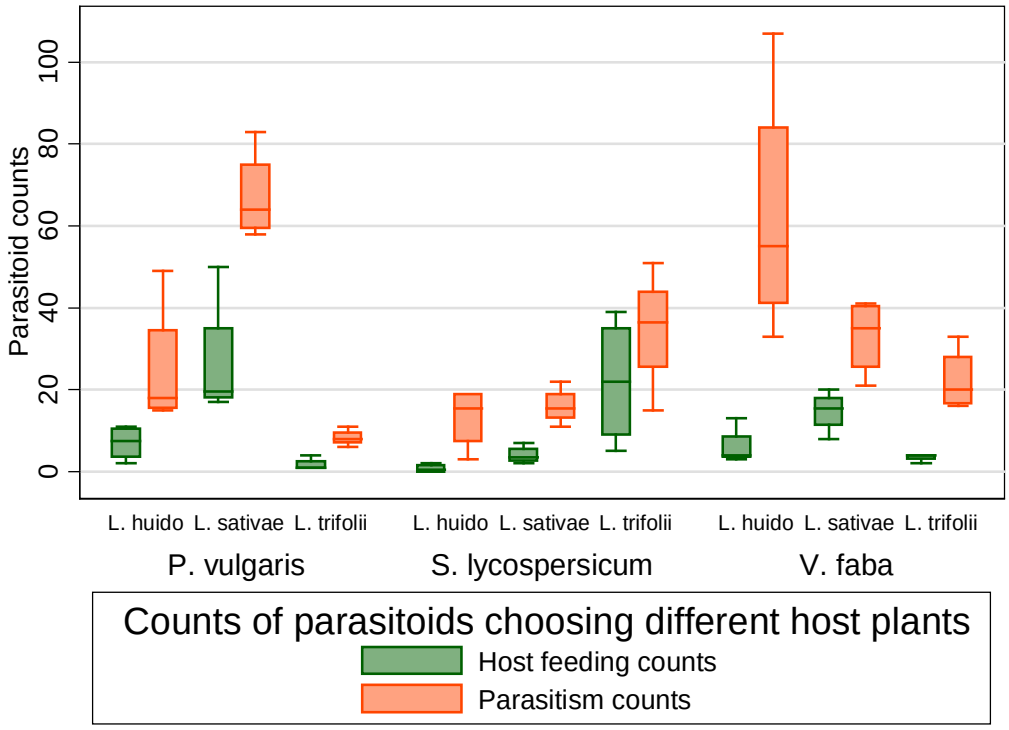

Figure 2: Box plot of parasitoids that chose different host plants including leaf miner species under host feeding and parasitism cases

The parasitoids choice of *V. faba* with *L. huidobrensis* leaf miner larvae was observed to have the highest variability for parasitism case which explained the extreme outlier observed in Figure 1.

4.2 Predictive accuracy

Table 5: Predictive accuracy for host feeding and parasitism case - sum of squared deviations from predicted probabilities

	Host feeding			Parasitism		
	MNL	MNP Unrestricted	MNP Homoskedastic	MNL	MNP Unrestricted	MNP Homoskedastic
$\sum$ (deviation) ^{2}	34.84	30.28	30.37	30.21	26.04	25.85

Unrestricted MNP model had the highest predictive accuracy for host feeding case while homoskedastic MNP had the highest predictive accuracy for parasitism case (Table 5). However, marginal differences in predictive accuracy were observed for MNP models with unrestricted and homoskedastic error structures for both parasitism and host feeding cases. MNL had the lowest predictive accuracy for both parasitism and host feeding cases respectively. The lower the sum of squared deviation, the higher the predictive accuracy for the respective choice model.

4.2.1 Predicted probabilities

Table 6: Host feeding parasitoid choice predicted probabilities for MNL, unrestricted MNP and homoskedastic MNP

	MNL			MNP Unrestricted			MNP Homoskedastic		
	L. huido	*L. sativae*	*L. trifolii*	*L. huido*	*L. sativae*	*L. trifolii*	*L. huido*	*L. sativae*	*L. trifolii*
P. vulgaris	0.620	0.518	0.120	0.526	0.556	0.080	0.589	0.520	0.113
S. lycospersicum	0.005	0.119	0.782	0.002	0.143	0.755	0.002	0.142	0.764
V. faba	0.375	0.363	0.098	0.472	0.301	0.165	0.409	0.337	0.123

Table 7: Parasitism parasitoid choice predicted probabilities for MNL, unrestricted MNP and homoskedastic MNP

	MNL			MNP Unrestricted			MNP Homoskedastic		
	L. huido	*L. sativae*	*L. trifolii*	*L. huido*	*L. sativae*	*L. trifolii*	*L. huido*	*L. sativae*	*L. trifolii*
P. vulgaris	0.365	0.376	0.307	0.374	0.378	0.263	0.366	0.374	0.302
S. lycospersicum	0.084	0.220	0.459	0.072	0.226	0.484	0.082	0.227	0.456
V. faba	0.551	0.404	0.234	0.554	0.395	0.252	0.552	0.400	0.242

The MNL, unrestricted MNP and homoskedastic MNP models reported qualitatively similar predicted probabilities for parasitoid choice of different host plants for host feeding and parasitism cases as observed in Tables 6 and 7.

Selected interpretation of estimated coefficients (Appendix 3).

Hausman-McFadden test of IIA assumption

H_0: Odds (Outcome-J versus Outcome-K) are independent of other alternatives.

Phaseolus vulgaris violated the IIA assumption (χ^2= 28.11, *df* = 2, *P* < 0.0001) under MNL for host feeding. Parasitoid choice of *P. vulgaris* host plants was correlated to *V. faba* and *S. lycospersicum.*

Vicia faba (χ^2= 8.922, *df* = 2, *P* = 0.012) violated the IIA assumption under MNL for parasitism. Parasitoid choice of *V. faba* host plants was correlated to *P. vulgaris* and *S. lycospersicum*.

4.3 Goodness of fit

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for the MNL and MNP models were used to compare goodness of fit in the table below. The number of iterations was also presented to give an indication of convergence ease.

Table 8: Multinomial logit and multinomial probit model goodness of fit statistics

	Host feeding			Parasitism		
	MNL	MNP Unrestricted	MNP Homoskedastic	MNL	MNP Unrestricted	MNP Homoskedastic
AIC	578.797	581.970	586.155	2282.433	2275.843	2284.915
BIC	594.171	606.680	610.866	2302.55	2312.61	2315.553
Iterations	5	25	20	4	25	20

From the low AIC and BIC values, it is evident that the MNL model had a better fit than unrestricted MNP and homoskedastic MNP respectively for host feeding case. Unrestricted MNP had a marginally better than homoskedastic MNP for host feeding case an indication that parasitoid choice of host plants was heteroskedastic. For the parasitism case, unrestricted MNP has a marginally better fit than MNL and homoskedastic MNP models respectively from the AIC statistic. However, BIC penalized both unrestricted MNP and homoskedastic MNP because of additional estimated parameters in the error covariance matrix resulting in a better fit for MNL model.

Imposing a homoskedastic error structure increased the ease of convergence for MNP model. MNL converged fastest while unrestricted MNP took longest to converge as observed in the number of log likelihood iterations.

4.4 DISCUSSION

4.4.1 Summary statistics

The chi-square test for contingency tables explains that there is a highly significant association (P <0.0001) between host plants and leaf miner flies larvae for both parasitism and host feeding cases respectively (Tables 1 and 3). However, because 44% of the cell sizes had counts less than 5, the Fisher's exact test was used to test for association between host plants and leaf miner species (Table 1 and 3) and also gave highly significant association results. From the results of the chi-square and Fisher's exact tests, we are not able to determine which species of leaf miner flies contribute most to the strong association. This breaking down of the chi-square test was also observed by Logan (2010).

The chi square test has potential use in insect choice studies during data exploration since the test gives an indication of patterns in the data as observed in Stout *et al*. (2010) who used the test in testing association between female cricket behavior and different sound frequencies made by male crickets. Contingency tables play a crucial role when fitting multinomial models on sparse data since they identify empty or small cells that would result in convergence difficulties due to model instability as was observed by Long and Freese (2001).

4.4.2 Predictive accuracy

MNP model has a higher predictive accuracy than MNL (Table 5). This finding can be explained by the presence of slight correlation in choices that is observed in host feeding and parasitism from Hausman--McFadden test. Correlation presence enables the MNP to consistently predict accurate estimates than MNL. This finding is supported by Kropko (2010)

who also observed that MNP has higher predictive accuracy especially when error terms follow a multivariate normal distribution.

Marginal differences in predictive accuracy are observed in MNP models with unrestricted and homoskedastic error structures for both host feeding and parasitism cases (Table 5). The heteroskedastic restriction on the covariance matrix seems to capture choice behavior marginally accurately than a homoskedastic restriction for host feeding case while the reverse is true for parasitism case. This finding agrees with Train (2003) who notes that covariance structure in MNP models depends on the specific situation being modeled. There is however little evidence from this study that imposing homoskedastic restrictions on the covariance matrix of the MNP model improves predictive accuracy. The consequence of forcing the diagonal elements to be 1 as is the case of homoskedastic restriction on the estimates may require further research by examining hessian condition to determine whether estimates converge at their global optimum. Researchers using MNP model should consider exploring different covariance matrix structures in an attempt to determine the most appropriate substitution pattern for their data.

The qualitatively similar predicted probabilities observed for both MNL and MNP models (Table 6 and 7) agree with Dow and Endersby (2004) who found that MNL and MNP models resulted in almost similar probabilities even in cases where choice correlation was present. Kropko (2010) also found little differences in predicted probabilities from MNL and MNP models compared under different correlation error structures.

Comparing parameter estimates from MNL and MNP without accounting for differences in scaling between the two models may lead to overestimating and underestimating effects.

MNL reports coefficients that are about 1.6 times larger than for MNP coefficients (Maddala, 1983). However, the significance of estimates is similar despite differences in scaling by the two models as was noted by Train (2003) who also found that the choice with the highest utility does not change no matter the scaling used. Parameter estimates from MNL and MNP should be re-scaled to avoid underestimating or overestimating effects when comparing the models.

4.4.3 Goodness of Fit

Multinomial logit fits the data better than MNP from the lower BIC statistic values for MNL model for both host feeding and parasitism cases (Table 9). MNL is observed to have a better fit to the data than both variations of MNP despite the presence of slight correlation as evidenced by IIA assumption violation. This finding agrees with Dow and Endersby (2004) who observes that IIA assumption being more of a logical decision making property and less of a statistical property is not particularly restrictive for most applications.

The AIC statistic for the parasitism case indicates that the unrestricted MNP has a marginally better fit than MNL and homoskedastic MNP models respectively. BIC penalizes the unrestricted MNP because of additional parameters estimated in the covariance matrix while the more parsimonious MNL is observed to fit the data better. In addition to being more parsimonious, MNL model's better fit can also be attributed to the tractable nature of MNL likelihood estimation compared to MNP that relies on simulation which may lose efficiency leading to convergence that is not at the global optimum. MNL model's simpler optimization leads to convergence at higher log likelihood and consequently results in a smaller AIC and BIC statistics which agrees with the observations of Trivedi (2009).

Imposing homoskedastic restriction on MNP model does not seem to improve fit for homoskedastic MNP as observed from both AIC and BIC statistics (Table 9) for parasitism and host feeding case. Though the MNL seems to fit the data better, the unrestricted MNP has a higher predictive accuracy. This observation underscores the need for researchers to scrutinize their estimates reliability in addition to measures of fit and agrees with Train (2003) who warns against choosing a competing model based only on fit statistics.

Restrictions on MNP covariance matrix reduce computational burden. The number of iterations before convergence are lower for homoskedastic MNP than unrestricted MNP for both host feeding and parasitism cases (Table 9). The reduced computational burden for homoskedastic MNP can be attributed to a reduction in the number of parameters estimated in the covariance matrix. This finding agrees with Greene (2003) who notes that imposing restrictions in the covariance matrix of MNP models enhances convergence for more than 3 choices. The finding also adds on the need of imposing restrictions even for 3 choices. Researchers choosing choice models for slightly correlated choices may benefit from MNL model's better fit over MNP. However, examining estimate accuracy and reliability should also guide the choice of model.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

Introduction

This chapter presents conclusions based on the study objectives followed by recommendations and future directions.

5.1 Conclusion

Multinomial probit had a higher predictive accuracy than multinomial logit model. Trading off simplicity and goodness of fit seems to result in a higher predictive accuracy for MNP model. There is evidence that both MNL and MNP models result in qualitatively similar predicted probabilities.

MNL had a better fit to the data than MNP despite violation of the IIA assumption. There was little evidence that imposing restrictions on the MNP covariance matrix improved predictive accuracy and goodness of fit though the homoskedastic restriction reduced computational burden.

5.2 Recommendations

The study findings suggest recommending use of the simpler and better fitting MNL in modelling insect choice data when IIA assumption is violated. However, more simulation studies should be conducted since MNL and MNP performance could vary under different scenarios. Future simulation studies should also evaluate the hessian condition of MNP estimates after imposing restrictions on the covariance matrix.

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis*, 2nd edn. New York: Wiley.

Aitchison, J., and Bennett, J. A. (1970). Polychotomous quantal response by maximum indicant, *Biometrika* **57(2)**: 253-262.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR* (pp. 267-281).

Alvarez, R. M. and Nagler, J. (1994). Correlated Disturbances in Discrete Choice Models: A Comparison of Multinomial Probit Models and Logit Models. Working Papers 914, California Institute of Technology, Division of the Humanities and Social Sciences.

Alvarez, R. M. and Nagler, J. (1998). When Politics and Models Collide: Estimating Models of Multiparty Elections. *American Journal of Political Science*, **42**:55-96.

Bhat, C. R. (1998). Accommodating flexible substitution patterns in multi-dimensional choice modeling: formulation and application to travel mode and departure time choice. *Transportation Research Part B: Methodological*, **32(7)**: 455-466.

Bhat, C. R. (2011). The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological*, **45:** 923-939.

Burnham, K. P., and Anderson, D. (2002). *Model selection and Multi Model Inference*, 2nd edn. Fort Collins: Springer Inc.

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

Dow, J. K. and Endersby, J. W. (2004). Multinomial Probit and Multinomial Logit: A Comparison of Choice Models for Voting Research. *Electoral Studies*, **23(1)**: 107-122.

Greene, W. (2003). *Econometric Analysis*, 5th ed. New Jersey: Prentice Hall.

Hajivassiliou, V., McFadden, D. and P. Ruud (1996). Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results, *Journal of Econometrics*, **72**: 85–134.

Hausman, J. and McFadden, D. (1984). Specification Tests for the Multinomial Logit Model, *Econometrica*, **52(5)**: 1219-1240.

Hensher, D. A., Rose, J. and Greene, W. (2005). *Applied Choice Analysis: A Primer*, Cambridge: Cambridge University Press.

Hern, A. and Dorn, S. (2001) Statistical modeling of insect behavioral responses in relation to the chemical composition of test extracts. *Physiol. Entomol*, **26**: 381–390.

Hoetker, G. (2007). The use of probit and logit models in strategic management research: Critical issues. *Strategic Management Journal*, **28**: 331-341.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edn. New York: Wiley.

Jones, W. D., Radcliffe, P. M., Huesman, R. L., and Kellogg, J. P. (2010). Redefining student success: Applying different multinomial regression techniques for the study of student graduation across institutions of higher education. *Research in Higher Education,* **51(2)**: 154-174.

Kropko, J. (2008). *Choosing between Multinomial Logit and Multinomial Probit Models for Analysis of Unordered Choice Data*. Paper Presented at the Annual Meeting of the MPSA Annual National Conference, Palmer House Hotel, Hilton, Chicago, IL, USA.

Kropko, J. (2010). *A Comparison of Three Discrete Choice Estimators*. Unpublished paper University of North Carolina.

Logan, M. (2010). *Biostatistical Design and Analysis Using R: A Practical Guide*.Chichester: John Wiley & Sons.

Long, S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Long, S. and Freese, J. (2001). *Regression Models for Categorical and Limited Dependent Variables Using Stata.* College Station, Texas: Stata Press.

Maddala, G. S. (1983). *Limited-dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*., New York: Academic Press.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, **3**: 303–328.

Munziaga, M. A., Heydecker, B. G., Ortuzar, J. D. D. (2000). Representation of heteroskedasticity in discrete choice models. *Transport Research Part B*: *Methodological*, **34(3)**: 219-240.

Musundire, R., Chabi-Olaye, A., Salifu, D., and Krüger, K. (2012) Host plant-related parasitism and host feeding activities of *Diglyphus isaea* (Hymenoptera: Eulophidae) on *Liriomyza huidobrensis* , *Liriomyza sativae* , and *Liriomyza trifolii* (Diptera: Agromyzidae). *Journal of Economic Entomology*, **105(1)**:161-168.

Quinn, G. P. and Keough, M. J. (2002). *Experimental design and data analysis for biologists*. New York: Cambridge University Press.

Quinn, K.M., Martin, A.D., and Whitford, A.B. (1999). Voter choice in multi-party democracies: a test of competing theories and models. *American Journal of Political Science*, **43**: 1231– 1247.

Raftery, M. A. E. (1986b). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, Series B,* **48**: 249-250.

Richard, I. and Davison, A. C. (2007). Statistical inference for olfactometer data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **56(4)**: 479-492.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6 (2)**: 461–464.

Stout, J., Navia, B., Jeffery, J., Samuel, L., Hartwig, L., Bultin, A., Chung, M., Wilson, J., Dashner, E., and Atkins, G., (2010). Plasticity of the phonotactic selectiveness of four

species of chirping crickets (*Gryllidae*): Implications for call recognition. *Physiol. Entomol.*, **35**: 99-116.

Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press.

Trivedi, P. K. (2009). *Microeconometrics using Stata*. Texas: Stata Press.

Turlings, T. C. J., Davison, A. C. and Tamo, C. (2004). A six-arm olfactometer permitting simultaneous observation of insect attraction and odour trapping. *Physiol. Entomol.*, **29**: 45–55.

APPENDICES

Appendix 1: Wide format choice data for MNL

list hplant species exptype rep hfed_t parasitized_t

	hplant	species	exptype	rep	hfed_t	parasi~t
1.	faba	huido	2	1	4	49
2.	faba	huido	2	2	3	61
3.	faba	huido	2	3	4	33
4.	faba	huido	2	4	13	107
5.	faba	sativae	2	1	8	30

Appendix 2: Long format choice data for MNP

list id possiblechoices choice hplant2 SPECIES rep hfed_t parasitized_t in 1/6, sepby(id)

	id	possib~s	choice	hplant2	SPECIES	rep	hfed_t	parasi~t
1.	1	faba	0	faba	huido	1	4	49
2.	1	French	0	faba	huido	1	4	49
3.	1	Tomato	1	faba	huido	1	4	49
4.	2	faba	0	faba	huido	2	3	61
5.	2	French	0	faba	huido	2	3	61
6.	2	Tomato	1	faba	huido	2	3	61

Appendix 3: Selected interpretation of estimated coefficients

Table 9: Multinomial logit model host feeding and parasitism estimates in odds ratios

	MNL	
	Host fed Host Plant	Parasitized Host Plant
P. vulgaris		
L. huidobrensis	0	0
	(.)	(.)
L. sativae	0	5.095***
	(.)	(0.810)
L. trifolii	0	0.927
	(.)	(0.218)
S. lycospersicum		
L. huidobrensis	0	0
	(.)	(.)
L. sativae	1.409	2.287***
	(0.936)	(0.491)
L. trifolii	117.33***	7.367***
	(84.876)	(1.497)
V. faba		
L. huidobrensis	0	0
	(.)	(.)
L. sativae	0.649	0
	(0.209)	(.)
L. trifolii	2.333	0
	(1.260)	(.)
N	345	1129

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(.) Standard errors missing since variable is used as a reference variable

Odds ratios interpretation for host feeding case

The odds of parasitoid choice of *L. sativae* to *L. huidobrensis* in *S. lycospersicum* relative to *P. vulgaris* would be expected to increase by a factor of 1.409 (Table 8). There is however no

statistically significant evidence that parasitoids choosing *S. lycospersicum* relative to *P. vulgaris* were more likely to prefer host feeding on *L. sativae* to *L. huidobrensis* leaf miner larvae ($z = 0.52$, $P = 0.606$).

The odds of parasitoid choice of *L. sativae* to *L. huidobrensis* in *V. faba* relative to *P. vulgaris* would be expected to increase by a factor of 0.649. There is however no statistically significant evidence that parasitoids choosing *V. faba* relative to *P. vulgaris* were less likely to prefer host feeding on *L. sativae* to *L. huidobrensis* leaf miner larvae ($z = -1.34$, $P = 0.180$).

Odds ratios interpretation for parasitism case

The odds of parasitoid choosing to parasitize *L. sativae* to *L. huidobrensis* in *P. vulgaris* relative to *V. faba* would be expected to increase by a factor of 5.095 (Table 8). There is statistically significant evidence that parasitoids choosing *P. vulgaris* relative to *V. faba* were more likely to prefer parasitizing *L. sativae* to *L. huidobrensis* leaf miner larvae ($z = 10.24$, $P < 0.0001$).

The odds of parasitoid choosing to parasitize *L. trifolii* to *L. huidobrensis* in *S. lycospersicum* relative to *V. faba* would be expected to increase by a factor of 7.367. There is statistically significant evidence that parasitoids choosing *S. lycospersicum* relative to *V. faba* were more likely to prefer parasitizing *L. trifolii* to *L. huidobrensis* leaf miner larvae ($z = 9.83$, $P < 0.0001$).

Appendix 4: Selected Stata commands

#Contingency table with host feeding counts, percentages, chi square test and Fisher's exact test

tabulate HPLANT SPECIES [fweight = hfed_t], chi2 col

tabulate HPLANT SPECIES [fweight = hfed _t], exact col

#MNL model for host feeding

mlogit HPLANT i.SPECIES [fweight = hfed_t]

#Fitting unrestricted MNP using for parasitism case

asmprobit choice [fweight = parasitized_t], case(id) alternatives(possiblechoices) casevars(SPECIES)

#Fitting homoskedastic MNP using for parasitism case

asmprobit choice [fweight = parasitized_t], case(id) alternatives(possiblechoices) casevars(SPECIES) stddev(homoskedastic)

#Goodness of Fit statistics for MNL and MNP models

estat ic

#Generating predicted probabilities

predict prob

#Calculating predictive accuracy for MNP homoskedastic model

predict homoskedastic

gen squares = (homoskedastic - choice)^2

gen sumsquares = squares

summarize sumsquares

display r(sum)

#Calculating predictive accuracy for MNL model

gen faba = (SPECIES == 1)

gen French = (SPECIES == 2)

```
gen Tomato = (SPECIES == 3)
predict p1 p2 p3, pr
gen squares1 = (p1 - faba)^2
gen squares2 = (p2 - French)^2
gen squares3 = (p3 - Tomato)^2
gen sumsquares = squares1 + squares2 + squares3
summarize sumsquares
display r(sum)
```