

# QUALITATIVE RESPONSE MODELS THEORY AND ITS APPLICATION TO FORESTRY

by

Alexandros A. Arabatzis

Dissertation submitted to the Faculty of the  
Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy  
in  
Forestry

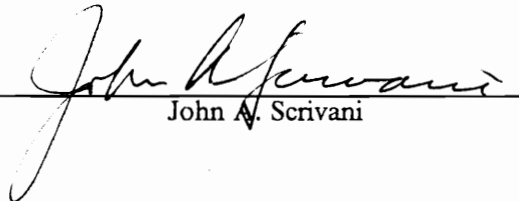
APPROVED:

  
Timothy G. Gregoire, Chairman

  
Harold E. Burkhardt

  
Richard G. Oderwald

  
Marion R. Reynolds, Jr.

  
John A. Scrivani

January, 1990

Blacksburg, Virginia

# QUALITATIVE RESPONSE MODELS THEORY AND ITS APPLICATION TO FORESTRY

by

Alexandros A. Arabatzis

Timothy G. Gregoire, Chairman

(ABSTRACT)

The focus of this dissertation is the theory of qualitative response models and its application to forestry related problems. Qualitative response models constitute a class of regression models used for predicting the result in one of a discrete number of mutually exclusive outcomes. These models, also known as discrete regression models, differ from the usual continuous regression models in that the response variable takes only discrete values. In forestry applications the use of such models has been largely confined to mortality studies where only the simplest kind of qualitative response models - a dichotomous (binary) dependent variable model - is applied. However, it is common in forestry to deal with many variables which are either discrete or recorded discretely and need to be formulated by more complex models involving polychotomous dependent variables. The estimation of such complex qualitative response models only recently has been made possible by the development of advanced computer technology.

The first objective of this study was to specify dichotomous and polychotomous response models that appear to be suitable for forestry applications and present methods of statistical analysis for these models. The models considered in this study were: the linear probability model, binary logit and probit, ordered and unordered multinomial logit and probit and McFadden's conditional logit. Special attention was paid to the following problems: i) how to motivate a qualitative response model which is theoretically correct and statistically manageable, ii) how to estimate and draw inferences about the model parameters, iii) what criteria to use when choosing among competing models and iv) how to detect outlying, high leverage and highly influential observations.

The second objective was to exemplify the utility of the above models by considering two, forestry related, case studies. Assessing the merchantability of loblolly pine trees growing on plantations in southern United States and modelling the incidence and spread of fusiform rust on loblolly and slash pine plantations in east Texas. The results demonstrated the potential of qualitative response models for meaningful implementation in a variety of forestry applications and also, suggested topics for future research work.

# Acknowledgements

I wish to express my deep and lasting gratitude to my major professor, Dr. Timothy G. Gregoire for his guidance, constructive criticism and continual encouragement throughout each stage of this study. I would also like to acknowledge the contribution of the other members of my graduate committee, Drs. Harold E. Burkhart, Richard G. Oderwald, Marion R. Reynolds, Jr. and John A. Scrivani. I am especially indebted to Dr. H. E. Burkhart for carefully reviewing parts of this dissertation and to Dr. J. David Lenhart at Stephen F. Austin State University for his comments and suggestions. Special thanks go to Dr. Robert E. Adams without whose efforts in securing funding this dissertation would not have been completed.

Data were provided by the Loblolly Pine Growth and Yield Cooperative at Virginia Tech and by the East Texas Pine Plantation Research Project at Stephen F. Austin State University. My deep appreciation is extended to the supporters of the above cooperatives.

During the past three years there have been some difficult moments for me and my wife, Alkmini. Without her, life would have been so much harder. In addition, two persons have never failed to support, encourage, listen patiently and understand. So it is to my parents, Athanasios and Irene, and to my wife, Alkmini Katsada, that I dedicate this dissertation.

# Table of Contents

**Introduction** ..... 1

**Binary Choice Models** ..... 6

2.1 Linear Probability Models ..... 10

2.2 Probit and Logit Models ..... 13

    2.2.1 Maximum Likelihood Estimation of the Logit Model ..... 14

    2.2.2 Maximum Likelihood Estimation of the Probit Model ..... 17

    2.2.3 Newton-Raphson and Method of Scoring Optimization Algorithms ..... 19

    2.2.4 Minimum Chi-Square Estimation Method ..... 20

**Multinomial Choice Models** ..... 22

3.1 Ordered Multinomial Choice Models ..... 22

    3.1.1 Maximum Likelihood Estimation for the OMNP Model ..... 24

    3.1.2 Maximum Likelihood Estimation for the OMNL Model ..... 27

3.2 Unordered Multinomial Choice Models ..... 29

    3.2.1 Unordered Multinomial Logit (UMNL) Model ..... 30

        3.2.1.1 Maximum Likelihood Estimation of the UMNL Model ..... 33

3.2.2	The Conditional Logit Model	36
3.2.3	The Independence of Irrelevant Alternatives (IIA) Property	37
3.2.4	Unordered Multinomial Probit Models	39
	<b>Inference and Model Selection</b>	<b>42</b>
4.1	Interpretation of Parameter Estimates	42
4.2	Tests for the General Linear Hypothesis	44
4.3	Criteria for Model Selection	48
4.3.1	Number of Wrong Predictions (WP)	49
4.3.2	Sum of Squared Residuals (SSR)	50
4.3.3	Weighted Sum of Squared Residuals (WSSR)	52
4.3.4	Prediction Success Index (PSI)	53
4.3.5	Likelihood Ratio Test (LR)	55
4.3.6	Akaike Information Criterion (AIC)	56
4.3.7	Theil's Information Inaccuracy of the Prediction	56
4.3.8	Discussion	59
4.4	Data Splitting and Model Validation	60
	<b>Outlier and Influence Diagnostics</b>	<b>63</b>
5.1	Introduction	63
5.2	Diagnostics for the detection of outliers	65
5.3	Diagnostics for High Leverage and Influence Observations	66
5.4	Diagnostics for Coefficient Sensitivity	68
	<b>Merchantability Models for Loblolly Pine</b>	<b>71</b>
6.1	Introduction	71
6.2	Data	75
6.3	Two-Product Logit Model	77

6.4 Three-Product Logit Model ..... 86

6.5 Concluding Remarks ..... 94

**Modeling Fusiform Rust Incidence in Loblolly and Slash Pine Plantations ..... 99**

7.1 Introduction ..... 99

7.2 Data ..... 102

7.3 Models that Predict Fusiform Rust Infection Levels ..... 105

    7.3.1 Dichotomous Models ..... 105

    7.3.2 Polychotomous Models ..... 120

    7.3.3 Discussion ..... 133

7.4 Models that Predict Fusiform Rust Transition Proportions ..... 136

    7.4.1 Discussion ..... 138

7.5 Concluding Remarks ..... 165

**Summary - Conclusions ..... 200**

**Bibliography ..... 203**

**Vita ..... 212**

# List of Illustrations

Figure 7.1. Index plot of $se_i$ vs $i$ (a) and of $d_i$ vs $i$ (b) for the binary logit model fitted to loblolly pine data. ....	112
Figure 7.2. Index plot of $h_{ii}$ vs $i$ (a) and $c^{-1}$ vs $i$ (b) for the binary logit model fitted to loblolly pine data. ....	113
Figure 7.3. Index plot of $se_i$ vs $i$ (a) and of $d_i$ vs $i$ (b) for the binary logit model fitted to slash pine data. ....	114
Figure 7.4. Index plot of $h_{ii}$ vs $i$ (a) and $c^{-1}$ vs $i$ (b) for the binary logit model fitted to slash pine data. ....	115
Figure 7.5. Standardized residuals plotted against age and landform for the binary logit model fitted to loblolly pine data. ....	116
Figure 7.6. Standardized residuals plotted against site index and landform for the binary logit model fitted to loblolly pine data. ....	117
Figure 7.7. Standardized residuals plotted against age and landform for the binary logit model fitted to slash pine data. ....	118
Figure 7.8. Standardized residuals plotted against average height and landform for the binary logit model fitted to slash pine data. ....	119
Figure 7.9. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of healthy loblolly pine trees. ....	168
Figure 7.10. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of branch infected loblolly pine trees. ....	169
Figure 7.11. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of stem infected loblolly pine trees. ....	170
Figure 7.12. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of dead loblolly pine trees. ....	171
Figure 7.13. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of healthy loblolly pine trees. ....	172
Figure 7.14. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of branch infected loblolly pine trees. ....	173



Figure 7.15. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of stem infected loblolly pine trees. . . . . 174

Figure 7.16. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of dead loblolly pine trees. . . . . 175

Figure 7.17. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of healthy loblolly pine trees. . . . . 176

Figure 7.18. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of branch infected loblolly pine trees. . . . . 177

Figure 7.19. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of stem infected loblolly pine trees. . . . . 178

Figure 7.20. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of dead loblolly pine trees. . . . . 179

Figure 7.21. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of healthy loblolly pine trees. . . . . 180

Figure 7.22. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of branch infected loblolly pine trees. . . . . 181

Figure 7.23. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of stem infected loblolly pine trees. . . . . 182

Figure 7.24. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of dead loblolly pine trees. . . . . 183

Figure 7.25. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of healthy slash pine trees. . . . . 184

Figure 7.26. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of branch infected slash pine trees. . . . . 185

Figure 7.27. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of stem infected slash pine trees. . . . . 186

Figure 7.28. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of dead slash pine trees. . . . . 187

Figure 7.29. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of healthy slash pine trees. . . . . 188

Figure 7.30. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of branch infected slash pine trees. . . . . 189

Figure 7.31. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of stem infected slash pine trees. . . . . 190

Figure 7.32. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of dead slash pine trees. . . . . 191

Figure 7.33. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of healthy slash pine trees. . . . . 192

Figure 7.34. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of branch infected slash pine trees. .... 193

Figure 7.35. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of stem infected slash pine trees. .... 194

Figure 7.36. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of dead slash pine trees. .... 195

Figure 7.37. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of healthy slash pine trees. .... 196

Figure 7.38. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of branch infected slash pine trees. .. 197

Figure 7.39. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of stem infected slash pine trees. .... 198

Figure 7.40. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of dead slash pine trees. .... 199

List of Tables

Table 4.1. Prediction success table for qualitative response models . . . . . 54

Table 6.1. Summary statistics for the 173 unthinned, lightly thinned and heavily thinned plots. 78

Table 6.2. Maximum likelihood coefficient estimates of the two product logit models fitted to unthinned, lightly and heavily thinned stands. . . . . 80

Table 6.3. Maximum likelihood coefficient estimates of the two-product logit models fitted to unthinned and thinned data. . . . . 81

Table 6.3a. Covariance matrix of the coefficient estimates of the binary logit model fitted to unthinned data. . . . . 82

Table 6.3b. Covariance matrix of the coefficient estimates of the binary logit model fitted to thinned data. . . . . 83

Table 6.4. Goodness of fit and prediction statistics of the two-product logit models fitted to unthinned and thinned data. . . . . 84

Table 6.5. Maximum likelihood coefficient estimates of the three-product ordered logit model fitted to pooled data. . . . . 88

Table 6.5a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 6.5. . . . . 89

Table 6.6. Maximum likelihood coefficient estimates of the three-product unordered logit model fitted to pooled data. . . . . 90

Table 6.6a. Covariance matrix of the coefficient estimates (sawtimber) of the UMNL model displayed on table 6.6. . . . . 91

Table 6.6b. Covariance matrix of the coefficient estimates (peelers) of the UMNL model displayed on table 6.6. . . . . 92

Table 6.7. Goodness of fit and prediction statistics of the unordered and ordered three product logit models fitted to pooled data. . . . . 93

Table 6.8. Actual and predicted proportions of trees by dbh class and product class for unthinned stands. . . . . 96

Table 6.9. Actual and predicted proportions of trees by dbh class and product class for thinned stands. . . . .	97
Table 7.1. Distribution of sample plots by species and age. . . . .	102
Table 7.2. Classification of survey plots by species and site preparation classes. . . . .	103
Table 7.3. Loblolly and slash pine plot summary statistics during the first measurement and remeasurement. . . . .	104
Table 7.4. Maximum likelihood coefficient estimates of the binary logit and probit models which predict rust infection on loblolly pine. . . . .	106
Table 7.4a. Covariance matrix of the coefficient estimates of the binary logit model displayed on table 7.4. . . . .	107
Table 7.5. Maximum likelihood coefficient estimates of the binary logit and probit models which predict rust infection on slash pine. . . . .	108
Table 7.5a. Covariance matrix of the coefficient estimates of the binary logit model displayed on table 7.5. . . . .	109
Table 7.6. Maximum likelihood coefficient estimates of the OMNL and OMNP models fitted to loblolly pine data. . . . .	121
Table 7.6a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.6. . . . .	122
Table 7.7. Maximum likelihood coefficient estimates of the UMNL model fitted to loblolly pine data. . . . .	123
Table 7.7a. Covariance matrix of the coefficient estimates (BRANCH) of the UMNL model displayed on table 7.7. . . . .	124
Table 7.7b. Covariance matrix of the coefficient estimates (STEM) of the UMNL model displayed on table 7.7. . . . .	125
Table 7.7c. Covariance matrix of the coefficient estimates (DEAD) of the UMNL model displayed on table 7.7. . . . .	126
Table 7.8. Maximum likelihood coefficient estimates of the OMNL and OMNP models fitted to slash pine data. . . . .	127
Table 7.8a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.8. . . . .	128
Table 7.9. Maximum likelihood coefficient estimates of the UMNL models fitted to slash pine data. . . . .	129
Table 7.9a. Covariance matrix of the coefficient estimates (BRANCH) of the UMNL model displayed on table 7.9. . . . .	130
Table 7.9b. Covariance matrix of the coefficient estimates (STEM) of the UMNL model displayed on table 7.9. . . . .	131

Table 7.9c. Covariance matrix of the coefficient estimates (DEAD) of the UMNL model displayed on table 7.9. ....	132
Table 7.10. General form of the rust infection plot transition matrix. ....	137
Table 7.11. Maximum likelihood coefficient estimates of the quatri-nomial ordered logit model that predicts the transitional proportions of healthy loblolly pine trees. . .	139
Table 7.11a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.11. ....	140
Table 7.12. Maximum likelihood coefficient estimates of the quatri-nomial unordered logit model that predicts transitional proportions of clear loblolly pine trees. ....	141
Table 7.12a. Covariance matrix of the coefficient estimates (CLR-BRA) of the UMNL model displayed on table 7.12. ....	142
Table 7.12b. Covariance matrix of the coefficient estimates (CLR-STEM) of the UMNL model displayed on table 7.12. ....	143
Table 7.12c. Covariance matrix of the coefficient estimates (CLR-DEAD) of the UMNL model displayed on table 7.12. ....	144
Table 7.13. Maximum likelihood coefficient estimates of the quatri-nomial ordered logit model that predicts the transitional proportions of branch infected loblolly pine trees. ....	145
Table 7.13a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.13. ....	146
Table 7.14. Maximum likelihood coefficient estimates of the tri-nomial unordered logit model which predicts transitional proportions of branch infected loblolly pine trees. . .	147
Table 7.14a. Covariance matrix of the coefficient estimates (BRA-STEM) of the UMNL model displayed on table 7.14. ....	148
Table 7.14b. Covariance matrix of the coefficient estimates (BRA-DEAD) of the UMNL model displayed on table 7.14. ....	149
Table 7.15. Maximum likelihood coefficient estimates of the binomial logit model that predicts the transitional proportions of stem infected loblolly pine trees. ....	150
Table 7.15a. Covariance matrix of the coefficient estimates of the binary logit model displayed on table 7.15. ....	151
Table 7.16. Maximum likelihood coefficient estimates of the quatri-nomial ordered logit model that predicts the transitional proportions of healthy slash pine trees. ....	152
Table 7.16a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.16. ....	153
Table 7.17. Maximum likelihood coefficient estimates of the quatri-nomial unordered logit model that predicts transitional proportions of clear slash pine trees. ....	154
Table 7.17a. Covariance matrix of the coefficient estimates (CLR-BRA) of the UMNL model displayed on table 7.17. ....	155

Table 7.17b. Covariance matrix of the coefficient estimates (CLR-STEM) of the UMNL model displayed on table 7.17. ....	156
Table 7.17c. Covariance matrix of the coefficient estimates (CLR-DEAD) of the UMNL model displayed on table 7.17. ....	157
Table 7.18. Maximum likelihood coefficient estimates of the tri-nomial ordered logit model that predicts the transitional proportions of branch infected slash pine trees. ...	158
Table 7.18a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.18. ....	159
Table 7.19. Maximum likelihood coefficient estimates of the tri-nomial unordered logit model which predicts transitional proportions of branch infected slash pine trees. ....	160
Table 7.19a. Covariance matrix of the coefficient estimates (BRA-STEM) of the UMNL model displayed on table 7.19. ....	161
Table 7.19b. Covariance matrix of the coefficient estimates (BRA-DEAD) of the UMNL model displayed on table 7.19. ....	162
Table 7.20. Maximum likelihood coefficient estimates of the binomial logit model that predicts the transitional proportions of stem infected slash pine trees. ....	163
Table 7.20a. Covariance matrix of the coefficient estimates of the binary logit model displayed on table 7.20. ....	164

# Chapter I

## Introduction

Qualitative response models, also known as quantal, categorical, logistic, qualitative choice or discrete regression models, are regression models in which the dependent variable is categorical, i.e., it can assume only a limited number of discrete values. As in classical regression analysis where the response variable is continuous, these models may be used to i) derive information regarding the role of each regressor in terms of its influence on the response variable, ii) explain the structure of the system (process) that generated the data observed by the researcher and iii) estimate (or predict) the response as a function of current (or future) observations. In qualitative response models however, the estimated (or predicted) response is actually the estimated probability that the response will assume a particular value. This is a fundamental difference between the qualitative and continuous regression models. As Nerlove and Press (1973) put it, it is not the value of the response variable that is important, but the probability that the response takes on a particular value.

Qualitative response models with a single regressor variable seem to have originated in mathematical psychology where they were being used for more than fifty years before they were either rediscovered or adapted by biometricians for use in biological assay (see Finney, 1971, sect. 3.1). The biological assay or bioassay is an experiment for estimating the potency of a substance by means

of the reaction that follows its application to living matter (Finney, 1978, sect. 1.1). Some of the early work in bioassay was due to Thomson (1919), Gaddum (1933), Bliss (1934a, b; 1935a, b; 1937), Berkson (1944) and Finney (1947).

Nevertheless, it was not until econometricians realized the variety of potential applications in behavioral sciences that the underlying probability and inference theory began to be fully developed and generalized to more complex problems (see Goldberger 1964, Zellner and Lee 1965, Theil 1967). Situations where the need for discrete regression models naturally arises are when studying the factors affecting, for example, a senator's decision on whether to vote yes or no on a particular piece of legislation, a consumer's choice on which among several shopping areas to visit or a couple's choice on the number of children. Over the past twenty years there has occurred considerable advancement in methods for modeling discrete choices. These new methods and models have been applied in economics (see Amemiya 1975, 1981, Manski 1981, McFadden 1981), travel demand (see Ben-Akiva and Lerman, 1985), transportation (see Domencich and McFadden, 1975), housing (Li, 1977) criminology (Witte, 1980), job location (Duncan, 1980), geography (Wrigley, 1982), and other fields.

In forestry research the use of qualitative response models has been confined to a relatively limited range of applications among which mortality studies are, by far, the most common (Monserud, 1976, Hamilton and Edwards, 1976, Hamilton, 1984 and 1986). Other applications include estimation of the probability of insect outbreak (Daniels et al., 1979), of insect spot becoming inactive (Reed et al., 1981), of a tree occurring into specified merchantability classes (Strub et al., 1986), modeling recreation choices (Stynes and Peterson, 1984), etc. Even though a considerable amount of data in forestry is naturally or artificially grouped into distinct classes (e.g., d.b.h. class, volume class, forest type, crown class, site class) the potential for broader application of such models is just now becoming apparent. Discrete regression models could serve, for example, in estimating the probability of fire occurrence at a certain forested area given certain climatic, vegetation, topographic, site and aspect characteristics, modeling the spread of a disease or infestation in a forest population, determining the factors affecting certain recreational activities in the forest, projecting



forest stand characteristics into the future as a part of an integrated forest yield and growth system, etc.

The inappropriateness of the classical regression analysis when the response is categorical is summarized in the following section. Situations when qualitative response models are appropriate arise when the response of an individual (whatever this might be, a person, a tree, an insect) can be classified into one of several categories. These categories, generally known as "alternatives", must be i) finite in number, ii) mutually exclusive and iii) collectively exhaustive. When a situation cannot be described as "qualitative" either because alternatives are not mutually exclusive or exhaustive, it is usually possible to redefine the set of alternatives in such a way that the redefined set meets the two criteria. The only truly restrictive criterion is the first one, namely, that the number of alternatives be finite. Typically, the set of alternatives available to the individual is denoted by a variable; for example, tree dbh is a non-negative number  $x$  which can be thought as representing the tree's choice. The variable  $x$  is obviously continuous in that, within any range, it can take an infinite number of values. Clearly the first criterion prohibits the use of qualitative modeling for continuous variables. However, many continuous variables can be represented without much loss of information by discrete variables. In many forest inventory situations, for example, the dbh values for the trees are not individually recorded but the tree frequencies are simply tallied by dbh classes (Avery and Burkhart 1983, Clutter et al. 1983). Since there is always a conceivable maximum value for these variables, qualitative response models are applicable. Whether or not to utilize this method, is of course a decision to be made by the researcher which will depend on the objective of research. Generally, when there is a small number of alternatives, questions such as "how many trees (will) fall in the 10-inch dbh class" can be fruitfully answered by applying qualitative response models.

The term "qualitative response models" designates a class of models, members of which are specific qualitative models such as probit and logit. All models in this class estimate the probability that an individual will choose a particular alternative from a set of alternatives, given the data observed by the researcher. The models in the class differ in two ways; first in the functional form that relates

the observed data to the probability (e.g., probit, logit) and second in the class of values assumed by the response variable. Fienberg(1980) distinguishes among discrete variables whose values are:

1. Dichotomous (e.g., yes or no, dead or alive)
2. Unordered polychotomous (e.g., mode of transport-car, bus or train)
3. Ordered polychotomous (e.g., old, middle aged, young)

Qualitative response models with dichotomous response are known as binary choice models whereas those with polychotomous response are known as multinomial choice models.

The focus of this dissertation is the theory of qualitative response models and its application to forestry related problems. In this first chapter the philosophy of qualitative response models has been introduced along with few brief historical remarks on the evolution of these models over the past sixty years. Chapter II presents the three major types of dichotomous response models namely, the linear probability, binary logit and probit formulations. The motivation and the estimation procedures given in this chapter provide the necessary theoretical background to understand the more complex polychotomous response models presented in chapter III. These models are the ordered multinomial logit and probit formulations, unordered multinomial logit and McFadden's conditional logit. The advantages and disadvantages of each of these models are also discussed in this chapter. Chapter IV refers to both dichotomous and polychotomous response models. It illustrates how to interpret and draw inferences about model parameters and also, it presents a variety of models selection and validation criteria. Chapter V deals with outlier and influence diagnostics specifically developed for binary logit models. The applicability of qualitative response models in forestry is illustrated in the following two chapters. In chapter VI, qualitative response models are employed to estimate the merchantability of loblolly pine trees growing on thinned and unthinned, site prepared, cut-over plantations throughout the southern United States. In chapter VII, these models are used to study the incidence and spread of fusiform rust in loblolly and slash pine plan-

tations in east Texas. Finally, chapter VIII summarizes the most important theoretical and empirical aspects of this dissertation and discusses the potential for further implementation of qualitative response models in forestry.

## Chapter II

### Binary Choice Models

The simplest of the qualitative response models are those in which the dependent variable is dichotomous or binary. For instance,  $y$  can be defined as 1 if a tree is alive and 0 if a tree is dead. One way to motivate binary choice models is to assume that there is an underlying response variable  $y_i^*$  defined by the linear regression relationship

$$y_i^* = \underline{\beta}' \underline{x}_i + \varepsilon_i^*$$

where,

$\underline{x}_i$  is the  $p \times 1$  vector of explanatory variables (continuous or discrete) observed at the  $i$ -th individual

$\underline{\beta}$  is the  $p \times 1$  vector of unknown regression coefficients

$\varepsilon_i^*$  is the error term for the  $i$ -th individual with  $E(\varepsilon_i^*) = 0$ .

In practice  $y_i^*$  is unobservable. It can be thought of as a random index for the  $i$ -th individual that defines its propensity to choose an alternative (Judge et al., 1988). What we observe is a Bernoulli random variable  $y$  defined by

$$y_i = 1 \quad \text{if } y_i^* > 0$$

$$y_i = 0 \quad \text{otherwise}$$

We will assume throughout this study that the random variables  $y_i$  are independently distributed except where noted. As mentioned in the introduction, interest lies with the estimation of the probability that the observed response ( $y$ ) takes on a particular value. The expectation of  $y_i$  is

$$\begin{aligned} E(y_i) &= 1 P(y_i = 1) + 0 P(y_i = 0) \\ &= P(y_i = 1) \\ &= P(y_i^* > 0) \\ &= P(\underline{\beta}' \underline{x}_i + \varepsilon_i^* > 0) \\ &= P(\varepsilon_i^* > -\underline{\beta}' \underline{x}_i) \\ &= 1 - F(-\underline{\beta}' \underline{x}_i) \end{aligned}$$

where  $F$  is the cumulative distribution function (cdf) of  $\varepsilon_i^*$ .

If the  $\varepsilon_i^*$ 's are independent and identically distributed (iid) uniform random variables defined in the open interval  $(-L, L)$  with  $L > 0$ , then  $F(-\underline{\beta}' \underline{x}_i) = 0$  if  $-\underline{\beta}' \underline{x}_i \leq -L$ ,

$$F(-\underline{\beta}' \underline{x}_i) = \int_{-L}^{-\underline{\beta}' \underline{x}_i} f(\varepsilon_i) d\varepsilon_i = \frac{-\underline{\beta}' \underline{x}_i + L}{2L} \quad \text{if } -L < -\underline{\beta}' \underline{x}_i < L$$

and  $F(-\underline{\beta}' \underline{x}_i) = 1$  if  $-\underline{\beta}' \underline{x}_i \geq L$ . It now follows that

$$\begin{aligned} E(y_i) &= P(y_i = 1) = p_i \\ &= 1 - F(-\underline{\beta}' \underline{x}_i) \\ &= 1 && \text{if } -\underline{\beta}' \underline{x}_i \leq -L && [2.1] \\ &= (\underline{\beta}' \underline{x}_i + L)/2L = \underline{\alpha}' \underline{z}_i && \text{if } -L < -\underline{\beta}' \underline{x}_i < L \\ &= 0 && \text{if } -\underline{\beta}' \underline{x}_i \geq L \end{aligned}$$

The above formulation is known as the linear probability model, so called because the probability  $p_i$  is expressed as a linear combination of the regressors  $\underline{x}_i$ .

If the cdf of  $\varepsilon_i$  is the logistic function (Johnson and Kotz, 1970) then we have the logit model. For this model,

$$F(-\underline{\beta}'\underline{x}_i) = \frac{\exp(-\underline{\beta}'\underline{x}_i)}{1 + \exp(-\underline{\beta}'\underline{x}_i)} = \frac{1}{1 + \exp(\underline{\beta}'\underline{x}_i)}$$

where  $-\infty < -\underline{\beta}'\underline{x}_i < \infty$ , and

$$\begin{aligned} E(y_i) &= P(y_i = 1) = p_i \\ &= 1 - F(-\underline{\beta}'\underline{x}_i) \\ &= \frac{\exp(\underline{\beta}'\underline{x}_i)}{1 + \exp(\underline{\beta}'\underline{x}_i)} \end{aligned} \quad [2.2]$$

In this case there is a closed-form expression for  $F$ , i.e., one that does not involve integrals explicitly. Not all distributions permit such a closed-form expression. For instance, in the probit model (also known as normit) we assume that the  $\varepsilon_i$  are iid  $N(0, \sigma^2)$ . In this case,

$$F(-\underline{\beta}'\underline{x}_i) = \Phi(-\underline{\beta}'\underline{x}_i) = \int_{-\infty}^{-\underline{\beta}'\underline{x}_i} (2\pi)^{-1/2} \exp(-\varepsilon_i^2/2\sigma^2) d(\varepsilon_i/\sigma)$$

A second motivation for binary choice models which is used extensively in the development of predictive models for human behavior, is based on the hypothesis of utility maximization (Domencich and McFadden 1975, Amemiya 1981). The theory of utility was developed during the 1930's and 1940's to describe an individual's preference among several alternative courses of action. The utility function is a decision rule which basically accounts for the internal mechanisms used by the decision maker to process the information available and to arrive at a unique choice

(DeGroot, 1970). It assumes that the "attractiveness" or utility derived from a particular choice can be expressed as a linear combination of attribute values which are specific to the individual (Judge et al., 1985). It is this index of attractiveness which is known as utility, a measure that the decision maker attempts to maximize through his choice. Depending on the area of application, this decision rule can be defined more specifically to minimize cost, to maximize profit or yield etc. We define  $U_{i0}$  and  $U_{i1}$  as the utilities associated with the  $y=0$  or  $1$  alternatives faced by the  $i$ -th individual. Then, assuming that each utility is a linear function of the individual's characteristics and a random disturbance we have,

$$U_{i0} = \underline{\gamma}_0' \underline{x}_i + \varepsilon_{i0}$$

$$U_{i1} = \underline{\gamma}_1' \underline{x}_i + \varepsilon_{i1}$$

The basic assumption is that the  $i$ -th individual chooses  $y_i = 1$  if  $U_{i1} > U_{i0}$  and  $y_i = 0$  otherwise. Then,

$$\begin{aligned} P(y_i = 1) &= P(U_{i1} > U_{i0}) \\ &= P(\underline{\gamma}_1' \underline{x}_i + \varepsilon_{i1} > \underline{\gamma}_0' \underline{x}_i + \varepsilon_{i0}) \\ &= P(\varepsilon_{i0} - \varepsilon_{i1} < (\underline{\gamma}_1 - \underline{\gamma}_0)' \underline{x}_i) \\ &= F(-\underline{\beta}' \underline{x}_i) \end{aligned}$$

where  $\underline{\beta} = (\underline{\gamma}_0 - \underline{\gamma}_1)$ . Thus, from the standpoint of utility, the linear probability, logit or probit models arise from assuming the uniform, normal or logistic cdf for  $\varepsilon_{i0} - \varepsilon_{i1}$  respectively.

A third motivation for binary choice models can be found in Finney (1947). It expresses the choice probabilities directly as a function of a set of explanatory variables, i.e.,

$$P(y_i = 1) = p_i = F(\underline{\beta}' \underline{x}_i) \tag{2.3}$$

The analyst's task is to select a proper form of  $F$  such that the implied constraint

$$0 \leq p_i \leq 1$$

is satisfied and the response curve concords with theoretical and empirical observations. Bliss (1934a) defined the probit model by selecting the normal cdf, and Berkson (1944, 1951), subsequently, noted the similarity between the normal and logistic curve, which enabled him to derive the logit model. The main difference of this motivation from the two, previously discussed, is that it makes no explicit assumption about the model's error structure. Indeed, the selection of  $F$  does not directly characterize the distribution of the error term, therefore no valid statistical inference can be made under this motivation. Although somewhat simplistic and not theoretically sound for subsequent statistical analysis, this kind of motivation is intuitively appealing and as such, it is frequently used by many authors for introducing the theory of qualitative response models to the reader.

## 2.1 *Linear Probability Models*

Recall from [2.1] that the linear probability model is defined

$$E(y_i) = P(y_i = 1) = \alpha' \mathbf{z}_i$$

It is reasonable to write the model in the usual regression framework as

$$y_i = \alpha' \mathbf{z}_i + \varepsilon_i$$

with  $E(\varepsilon_i) = 0$ . The conditional expectation  $E(y_i | \mathbf{z}_i) = \alpha' \mathbf{z}_i$ , defines the probability that the event will occur given  $\mathbf{z}_i$ . The calculated value of  $y_i$  from the regression equation,  $\hat{y}_i = \hat{\alpha}' \mathbf{z}_i$ , will then give the estimated probability that the event will occur given the particular value of the vector  $\mathbf{z}_i$ . An obvious defect of this formulation is that unless  $\hat{\alpha}' \mathbf{z}_i$  takes values within the interval  $[0, 1]$ , the esti-



mated probabilities can have values greater than one or be negative. A way out of this problem is to use an estimation procedure which automatically ensures these restrictions on  $\hat{y}_i$  for observations used in the fitting data. This can be accomplished by means of quadratic programming (Judge et al., 1985) or by applying restricted least squares estimation (Nerlove and Press, 1973). However, under either estimation procedure there is still no guarantee that predictions for values of  $z_i$  outside the sample range will fall within the  $[0, 1]$  interval (Judge et al., 1985). In addition, restricted least squares estimators are optimal only asymptotically, i.e., for large sample sizes, and require particularly complex calculations especially when heteroscedasticity is to be accounted for (Nerlove and Press, 1973).

Since  $y_i$  takes only two values, 1 and 0, the residuals  $\varepsilon_i$  also take two values only:

$$\begin{aligned}\varepsilon_i &= 1 - \alpha' z_i & \text{if } y_i = 1 \\ \varepsilon_i &= \alpha' z_i & \text{if } y_i = 0\end{aligned}$$

As already stated,  $P(y_i = 1) = E(y_i | z_i) = \alpha' z_i$  and  $P(y_i = 0) = 1 - \alpha' z_i$ . Since  $E(\varepsilon_i | z_i) = 0$ ,

$$\begin{aligned}Var(\varepsilon_i | z_i) &= (1 - \alpha' z_i)^2 (P(y_i = 1)) + (-\alpha' z_i)^2 (P(y_i = 0)) \\ &= (1 - \alpha' z_i)^2 (\alpha' z_i) + (\alpha' z_i)^2 (1 - \alpha' z_i) \\ &= \alpha' z_i (1 - \alpha' z_i) \\ &= E(y_i | z_i) (1 - E(y_i | z_i))\end{aligned}$$

Since  $E(y_i | z_i)$  varies with the levels of explanatory variables, the error variance is not homogeneous. In this case the use of OLS will result in estimated coefficients that will still be unbiased but will no longer have the minimum variance property among the class of linear unbiased estimators. Typically, the problem of unequal error variances can be solved by using weighted least squares (WLS) method. Goldberger (1964) suggested the following two stage estimation procedure. First estimate

$$y_i = \alpha' z_i + \varepsilon_i$$

by OLS. Next, compute  $\hat{y}_i(1 - \hat{y}_i)$  and use WLS; that is, define

$$\hat{w}_i = [\hat{y}_i(1 - \hat{y}_i)]^{1/2}$$

and then regress  $y_i/\hat{w}_i$  on  $\underline{z}_i/\hat{w}_i$ . This estimation procedure is generally known as Feasible Generalized Least Squares (FGLS) method (Fomby et al., 1984). Given that  $\hat{w}_i = \hat{y}_i(1 - \hat{y}_i)$  consistently estimates  $Var(\varepsilon_i | \underline{z}_i) = E(y_i | \underline{z}_i)[1 - E(y_i | \underline{z}_i)]$  (McGillivray, 1970), it can be shown that FGLS estimators are unbiased (Kakwani, 1967), consistent, asymptotically normal and asymptotically more efficient than OLS estimators (Zellner 1962, Schmidt 1976). Furthermore, in the case of normal errors, FGLS method yields estimators which are asymptotically equivalent to ML estimators provided that the ML estimator is used to estimate the  $Var(\varepsilon_i | \underline{z}_i)$  (Magnus, 1978) as is the case here. A problem with this procedure is that since OLS does not guarantee that  $\hat{y}_i$  will lie between 0 and 1, some of the estimates  $\hat{w}_i$  may be negative (Nerlove and Press, 1973). It must also be noted that this method only corrects for the heteroscedasticity noted above and does not avoid the inherent weakness of the linear probability models namely, that  $\underline{\alpha}'\underline{z}_i$  is not constrained to lie between 0 and 1. In view of the fact that the main use of linear probability models is in preliminary studies, Amemiya (1981) suggests to use OLS rather than FGLS estimators keeping in mind that the standard errors of the OLS estimators are biased because of the heteroscedasticity.

Another problem associated with the OLS estimation of linear probability models occurs when the majority of values of  $\underline{\alpha}'\underline{z}_i$  are either small or large, so that a preponderance of observed  $y_i$ 's are 0 or 1, respectively. This may result in a significant distortion of the estimated relationship from the true one (Nerlove and Press, 1973).

Finally, because the residuals are not normally distributed, the OLS and, in general, no method of estimation that is linear in the  $y_i$ 's, is fully efficient. That is, there exist non-linear estimation methods that are more efficient than the OLS and WLS methods (Cox, 1970). These methods are discussed next.

## 2.2 *Probit and Logit Models*

As mentioned before, probit and logit models are motivated by assuming that the cdf of  $\varepsilon_i^*$  is the normal and logistic function, respectively. The normality assumption associated with the probit model is neither stronger nor less strong than in any other branch of statistical inference (Finney, 1978). The true distribution may not be normal but in the absence of evidence favoring a specific alternative, the fact that the normal distribution accords fairly well with biological observations and is mathematically tractable (though not expressible in closed-form) makes the hypothesis of normality appealing. The central limit theorem also provides a major justification for probit analysis when means of several observations at each setting of the predictor variables are involved. Bliss (1934a, 1934b) was the first to present the probit model and the maximum likelihood estimation within the context of bioassay.

Berkson (1944) advocated the logistic function as an alternative to the normal cdf for modeling quantal response rates on different dose levels. He argued that “the logistic function is very near to the integrated normal curve, it applies to a wide range of physicochemical phenomena and therefore may have a better theoretic basis than the integrated normal curve”. The main advantage of the logistic distribution has been considered, for long, to be the fact that unlike the normal cdf, a closed form expression is readily available. Today however, given the fast computer hardware and efficient computing algorithms for evaluating the normal pdf, it is questionable how great this advantage is.

Because the normal and logistic cdf are very similar to each other except at the tails (Cox, 1970), one is not likely to obtain very different results using probit or logit analysis unless the sample is large (so that we have enough observations at the tails). However, as Amemiya (1981) pointed out, the estimates of  $\beta$  from the two methods are not directly comparable. The reason is that the variance of the standard normal variable is 1 whereas the variance of the standardized logistic distrib-

ution is  $\pi^2/3$ . Thus, the estimates of  $\underline{\beta}$  obtained by the logit model have to be multiplied by  $\sqrt{3}/\pi$  in order to be comparable to the estimates obtained by the probit model. Amemiya (1981) suggested that the logit estimates be multiplied by  $1/1.6 = 0.625$  instead of  $\sqrt{3}/\pi$  saying that this transformation produces a closer approximation of the logistic distribution to the distribution of the standard normal. He also showed that the coefficients of the linear probability model,  $\hat{\beta}_{LP}$  and the coefficients of the logit model,  $\hat{\beta}_L$  are related by the relationship:

$$\begin{aligned}\hat{\beta}_{LP} &\simeq 0.25\hat{\beta}_L && \text{except for the constant term} \\ \hat{\beta}_{LP} &\simeq 0.25\hat{\beta}_L + 0.5 && \text{for the constant term}\end{aligned}$$

Maximum likelihood is the method most commonly used to estimate the vector of regression coefficients,  $\underline{\beta}$ . The implementation of this method will be described first for the logit and second for the probit model.

### 2.2.1 Maximum Likelihood Estimation of the Logit Model

Recall the definition of the logit model,

$$\begin{aligned}E(y_i) &= P(y_i = 1) = p_i \\ &= 1 - F(-\underline{\beta}'\underline{x}_i) \\ &= \frac{\exp(\underline{\beta}'\underline{x}_i)}{1 + \exp(\underline{\beta}'\underline{x}_i)}\end{aligned}$$

where  $y_i$  was defined as

$$\begin{aligned}y_i &= 1 && \text{if } y_i^* > 0 \\ y_i &= 0 && \text{otherwise}\end{aligned} \tag{2.4}$$

In this case the observed values  $y_i$  are realizations of a Bernoulli process with probabilities given by [2.4] and varying from trial to trial (depending on  $\underline{x}_i$ ). Hence, the likelihood function is

$$L = \prod_{i=1}^n [F(-\underline{\beta}'\underline{x}_i)]^{1-y_i} [1 - F(-\underline{\beta}'\underline{x}_i)]^{y_i} \quad [2.5]$$

which, when  $F$  is assumed to be the logistic function,

$$\begin{aligned} L &= \prod_{i=1}^n \left[ \frac{1}{1 + \exp(\underline{\beta}'\underline{x}_i)} \right]^{1-y_i} \left[ \frac{\exp(\underline{\beta}'\underline{x}_i)}{1 + \exp(\underline{\beta}'\underline{x}_i)} \right]^{y_i} \\ &= \frac{\exp(\underline{\beta}' \sum_{i=1}^n \underline{x}_i y_i)}{\prod_{i=1}^n [1 + \exp(\underline{\beta}'\underline{x}_i)]} \end{aligned}$$

Define  $t' = \sum_{i=1}^n \underline{x}_i y_i$  and note that the likelihood function involves  $y_i$  only through  $t'$ , that is,  $t'$  is a sufficient statistic for  $\underline{\beta}$  given the observed vectors  $\underline{x}_i$ . To find the maximum likelihood (ML) estimator of  $\underline{\beta}$  we take the logarithm of  $L$ ,

$$\log L = \underline{\beta}' t' - \sum_{i=1}^n \log[1 + \exp(\underline{\beta}'\underline{x}_i)]$$

Hence,  $S(\underline{\beta}) = \partial \log L / \partial \underline{\beta} = 0$  gives,

$$S(\underline{\beta}) = - \sum_{i=1}^n \frac{\exp(\underline{\beta}' \underline{x}_i)}{1 + \exp(\underline{\beta}' \underline{x}_i)} \underline{x}_i + \underline{t}^*$$

These equations are non-linear in  $\underline{\beta}$ . Hence, in order to find  $\hat{\underline{\beta}}$  such that  $S(\underline{\beta}) = 0$  is satisfied we have to rely on some method of non-linear optimization such as the Newton-Raphson method. A necessary and sufficient condition for any iterative optimization procedure to converge to an absolute maximum (or minimum) regardless of the choice of initial values of the parameters is the global concavity (convexity) of the maximum likelihood function. Amemiya (1985, p.270) proves the global concavity of the maximum likelihood function for both logit and probit models. He also proves the consistency and asymptotic normality of the ML estimator of  $\underline{\beta}$  for both models.

Using the definition of the information matrix  $I(\underline{\beta})$ , we can write

$$\begin{aligned} I(\underline{\beta}) &= E \left( - \frac{\partial^2 \log L}{\partial \underline{\beta} \partial \underline{\beta}'} \right) \\ &= \sum_{i=1}^n \frac{\exp(\underline{\beta}' \underline{x}_i)}{[1 + \exp(\underline{\beta}' \underline{x}_i)]} \underline{x}_i \underline{x}_i' \end{aligned}$$

If the final converged ML estimates are denoted by  $\hat{\underline{\beta}}$ , then the asymptotic covariance matrix is consistently estimated by  $[I(\hat{\underline{\beta}})]^{-1}$ . Under the asymptotic normality of the ML estimator of  $\underline{\beta}$ , we can test hypotheses and construct confidence intervals about any one or all of the elements of  $\hat{\underline{\beta}}$ .

Using  $\hat{\underline{\beta}}$  we can estimate the probability  $p_i$  that the  $i$ -th observation is equal to 1. Denoting these estimated values by  $\hat{p}_i$  we have,

$$\hat{p}_i = \frac{\exp(\hat{\beta}' \underline{x}_i)}{1 + \exp(\hat{\beta}' \underline{x}_i)}$$

From  $S(\beta) = 0$  we obtain

$$\sum_{i=1}^n \hat{p}_i \underline{x}_i = \sum_{i=1}^n \underline{x}_i y_i$$

This means that if  $\underline{x}_i$  includes a constant term, the sum of the estimated probabilities is equal to  $\sum_{i=1}^n y_i$  i.e., the number of observations in the sample for which  $y = 1$ . Formally stated, the predicted frequency equals to the actual frequency.

### 2.2.2 Maximum Likelihood Estimation of the Probit Model

For the probit model the likelihood function is

$$\begin{aligned} L &= \prod_{i=1}^n [\Phi(-\underline{\beta}' \underline{x}_i)]^{1-y_i} [1 - \Phi(-\underline{\beta}' \underline{x}_i)]^{y_i} \\ &= \prod_{i=1}^n [1 - \Phi(\underline{\beta}' \underline{x}_i)]^{1-y_i} [\Phi(\underline{\beta}' \underline{x}_i)]^{y_i} \end{aligned}$$

and the log-likelihood is,

$$\log L = \sum_{i=1}^n (1 - y_i) \log[1 - \Phi(\underline{\beta}' \underline{x}_i)] + \sum_{i=1}^n y_i \log \Phi(\underline{\beta}' \underline{x}_i)$$

Differentiating  $\log L$  with respect to  $\underline{\beta}$  yields

$$S(\underline{\beta}) = \frac{\partial \log L}{\partial \underline{\beta}} = \sum_{i=1}^n \frac{[y_i - \Phi(\underline{\beta}' \underline{x}_i)]}{\Phi(\underline{\beta}' \underline{x}_i)[1 - \Phi(\underline{\beta}' \underline{x}_i)]} \phi(\underline{\beta}' \underline{x}_i) \underline{x}_i$$

where  $\phi(\cdot)$  is the probability density function of the standard normal distribution defined as

$$\frac{\partial \Phi(x)}{\partial x} = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

for  $-\infty < x < \infty$ . The ML estimator  $\hat{\underline{\beta}}$  can be obtained as a solution that satisfies the equations  $S(\underline{\beta}) = 0$ . Again, these equations are non-linear in  $\underline{\beta}$ , thus we have to rely on an iterative optimization procedure such as the Newton-Raphson method in order to solve them. The information matrix is,

$$\begin{aligned} I(\underline{\beta}) &= E\left(-\frac{\partial^2 \log L}{\partial \underline{\beta} \partial \underline{\beta}'}\right) \\ &= \sum_{i=1}^n \frac{[\Phi(\underline{\beta}' \underline{x}_i)]^2}{\Phi(\underline{\beta}' \underline{x}_i)[1 - \Phi(\underline{\beta}' \underline{x}_i)]} \underline{x}_i \underline{x}_i' \end{aligned}$$

As with the logit analysis, if  $\hat{\underline{\beta}}$  is the final converged ML estimate of  $\underline{\beta}$ , then  $[I(\hat{\underline{\beta}})]^{-1}$  is the corresponding covariance matrix of  $\hat{\underline{\beta}}$  which can be used to conduct any test of significance or construct confidence intervals about any element of  $\hat{\underline{\beta}}$ .



### 2.2.3 Newton-Raphson and Method of Scoring Optimization Algorithms

Because of the global concavity of the likelihood function of both logit and probit models, any non-linear iterative optimization procedure can be used to calculate ML estimators. The two most commonly used iterative methods are the Newton-Raphson method and the method of scoring. Given an initial estimate  $\hat{\beta}_1$ , the second round estimate  $\hat{\beta}_2$  in each method is defined as follows:

#### Newton - Raphson Method

$$\hat{\beta}_2 = \hat{\beta}_1 - \left[ \frac{\partial^2 \log L}{\partial \beta \partial \beta'} \bigg|_{\hat{\beta}_1} \right]^{-1} \frac{\partial \log L}{\partial \beta} \bigg|_{\hat{\beta}_1}$$

#### Method of Scoring

$$\begin{aligned} \hat{\beta}_2 &= \hat{\beta}_1 - \left[ E \frac{\partial^2 \log L}{\partial \beta \partial \beta'} \bigg|_{\hat{\beta}_1} \right]^{-1} \frac{\partial \log L}{\partial \beta} \bigg|_{\hat{\beta}_1} \\ &= \left[ \sum_{i=1}^n \frac{\hat{f}_i^2}{\hat{F}_i(1 - \hat{F}_i)} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \sum_{i=1}^n \frac{\hat{f}_i}{\hat{F}_i(1 - \hat{F}_i)} \mathbf{x}_i [y_i - \hat{F}_i + \hat{f}_i(\beta' \mathbf{x}_i)] \end{aligned}$$

where  $\hat{F}_i = F(\hat{\beta}_1' \mathbf{x}_i)$ ,  $\hat{f}_i = f(\hat{\beta}_1' \mathbf{x}_i)$  and  $\log L$  as in [2.5]. The third round estimator  $\hat{\beta}_3$  is obtained by substituting  $\hat{\beta}_2$  for  $\hat{\beta}_1$  in the right hand side of the above expressions. This procedure is to be repeated until the sequence of estimated thus obtained converges. Note that the two methods differ only in that in the method of scoring the expectation is taken of the matrix of second derivatives.

An interesting interpretation of the method of scoring estimate as a non-linear weighted least squares estimate is due to Walker and Duncan (1967). Writing a binary choice model directly as a function of the explanatory variables, we obtain

$$y_i = F(\beta' \mathbf{x}_i) + \varepsilon_i$$

which is a heteroscedastic non-linear regression model with  $E(\varepsilon_i|\underline{x}_i) = 0$  and  $Var(\varepsilon_i|\underline{x}_i) = F(\underline{\beta}'\underline{x}_i)(1 - F(\underline{\beta}'\underline{x}_i))$ . Expanding  $F(\underline{\beta}'\underline{x}_i)$  in a Taylor series about  $\underline{\beta} = \hat{\underline{\beta}}_1$  and rearranging terms, we have

$$y_i - \hat{F}_i + \hat{f}_i(\hat{\underline{\beta}}_1'\underline{x}_i) \simeq \hat{f}_i(\underline{\beta}'\underline{x}_i) + \varepsilon_i$$

Thus the second expression of  $\hat{\underline{\beta}}_2$  in the method of scoring iteration can be interpreted as the weighted least squares estimator of  $\underline{\beta}$  applied to the above model with the  $Var(\varepsilon_i|\underline{x}_i)$  estimated by  $\hat{F}_i(1 - \hat{F}_i)$ . For this reason, Walker and Duncan (1967) called the method of scoring iteration in the qualitative response models as the non-linear weighted least squares (NLWLS) iteration.

#### 2.2.4 Minimum Chi-Square Estimation Method

In the preceding discussion the maximum likelihood estimation method was presented for both logit and probit models. This method of estimation is applicable whether repeated observations are available for each cell or not. A cell is defined to be a particular setting of values of the explanatory (independent) variables. When several (repeated) observations are available for each cell, one can use an alternative estimation procedure known as the minimum chi-square method (Berkson, 1944). Its main advantage is that any statistical computer package which includes multiple linear regression fitting can be used to estimate the regression coefficients for the probit or logit models. The minimum chi-square method will produce estimates asymptotically equivalent to maximum likelihood estimates if the number of observations per cell is sufficiently large (at least 30, as a rule of thumb) for the asymptotic theory to be in effect. However, even in the case where the analyst designs the experiment and therefore controls the values of independent variables, it is not always possible to generate many observations per cell due to cost and time limitations. For example, in the case of four independent variables each of which takes four distinct values, the required total number of observations is at least  $(4^4)(30) = 7680$ . Moreover, when at least one of the independent variables is continuous, a case frequently occurring in forestry research, the large

number of empty cells prohibits the implementation of minimum chi-square method. Grouping the data into artificially constructed intervals does not guarantee that the number of observations per cell will be sufficiently large and always results in loss of information. For these reasons the minimum chi-square method of estimation will not be discussed in this study.

## Chapter III

# Multinomial Choice Models

Multinomial choice models constitute a class of qualitative response models for which the response is a polychotomous variable, i.e., it can be classified into many categories. As mentioned previously, we distinguish between ordered and unordered polychotomous variables. Because model specification and analysis differ according to the type of the response variable involved, the cases of ordered and unordered responses will be discussed separately.

### *3.1 Ordered Multinomial Choice Models*

Models of this type consider individuals who are grouped into  $m > 2$  ordered categories such as, for example, dead, severely affected, unaffected; old middle age, young; less than high school, high school and college education. Motivation for such models is provided by considering an ordered categorical variable as a coarsely measured version of an underlying continuous and unobservable

random variable  $y^*$  (McCullagh, 1980). It is then reasonable to assume that the ordered categories correspond to non-overlapping and exhaustive intervals of the real line.

Suppose that  $y^*$  for the  $i$ -th individual ( $i = 1, 2, \dots, n$ ) is expressed as

$$y_i^* = \underline{\beta}' \underline{x}_i + \varepsilon_i^*$$

where  $\underline{x}_i$  is the vector of explanatory variables for the  $i$ -th individual,  $\underline{\beta}$  is the vector of the corresponding parameters and  $\varepsilon_i^*$  is the residual. Then, assume that the  $i$ -th individual belongs to the  $j$ -th ( $j = 1, 2, \dots, m$ ) category if

$$\alpha_{j-1} < y_i^* < \alpha_j$$

where  $\alpha_1 = -\infty$ ,  $\alpha_m = \infty$  and  $\alpha_1 < \alpha_2 < \dots < \alpha_m$  partition the real line into successive intervals.

In practice,  $y_i^*$  is not observed. Instead, a Bernoulli random variable,  $y_{ij}$ , is observed, where

$$\begin{aligned} y_{ij} &= 1 && \text{if the } i\text{-th individual falls in the } j\text{-th category} \\ y_{ij} &= 0 && \text{otherwise.} \end{aligned}$$

Hence, if  $p_{ij}$  denotes the probability for the  $i$ -th individual to belong to the  $j$ -th category,

$$\begin{aligned} p_{ij} &= P(y_{ij} = 1) \\ &= P(\alpha_{j-1} < y_i^* < \alpha_j) \\ &= P(\alpha_{j-1} - \underline{\beta}' \underline{x}_i < \varepsilon_i^* < \alpha_j - \underline{\beta}' \underline{x}_i) \\ &= F(\alpha_j - \underline{\beta}' \underline{x}_i) - F(\alpha_{j-1} - \underline{\beta}' \underline{x}_i) \end{aligned} \tag{3.1}$$

where  $F$  is the cdf of  $\varepsilon_i^*$ . This is the general form of an ordered multinomial choice model.

If  $F$  is the standard normal cdf, [3.1] defines the ordered multinomial probit (OMNP) model, and if  $F$  is the logistic cdf, it defines the ordered multinomial logit (OMNL) model also known as the proportional odds model (McCullagh, 1980). The OMNP model was first considered by Aitchison

and Silvey (1957) and Ashford (1959) in the biological assay context. It has been applied, among others, by Gurland, Lee and Dahm (1960) to study the response of an insect (dead, moribund, alive) subject to various dosage levels of an insecticide, David and Legg (1975) to explain the price of a home by several socioeconomic household characteristics, and McKelvey and Zavoina (1975) to analyze the determinants of congressional voting (yes, no, abstain) on the 1965 medicare bill. The OMNL model was introduced by Cox (1970) and has been applied only to a limited number of studies. Among these studies is that by Deacon and Shapiro (1975) to analyze the voting behavior of Californians to two different referenda and that by Sheffi (1979) to study trip generation by elderly individuals. McCullagh (1980), Anderson and Philips (1981), Anderson (1984) and Greenland (1985) provide in-depth theoretical discussions about the properties of unordered multinomial choice models.

### 3.1.1 Maximum Likelihood Estimation for the OMNP Model

The likelihood and log-likelihood functions for the OMNP model are given respectively by

$$L = \prod_{i=1}^n \prod_{j=1}^m [\Phi_{i,j} - \Phi_{i,j-1}]^{y_{ij}}$$

and

$$\log L = \sum_{i=1}^n \sum_{j=1}^m y_{ij} \log[\Phi_{i,j} - \Phi_{i,j-1}]$$

where

$$\Phi_{i,j} = P[\varepsilon_i^* < \alpha_j - \beta' \underline{x}_i] = \Phi(\alpha_j - \beta' \underline{x}_i) - \Phi(\alpha_{j-1} - \beta' \underline{x}_i)$$

and  $\Phi$  is the cdf of a standard normal random variable. As usual, the objective is to find the values of  $\alpha_j$  and  $\underline{\beta}$  such that  $L$  or  $\log L$  is maximized. For this, we differentiate with respect to  $\alpha_k$  ( $k = 1, 2, \dots, m$ ) and  $\underline{\beta}$  and then set the derivatives equal to zero. Noting that

$$\frac{\partial \Phi_{i,j}}{\partial \underline{\beta}} = \phi_{i,j} \underline{x}_i$$

and

$$\frac{\partial \Phi_{i,j}}{\partial \alpha_k} = \xi_{j,k} \phi_{i,j}$$

where  $\phi_{i,j}$  is the standard normal pdf evaluated at  $\tau_{i,j} = \alpha_j - \underline{\beta}' \underline{x}_i$  and  $\xi_{j,k} = 1$  if  $j = k$  and 0 otherwise, we obtain

$$\frac{\partial \log L}{\partial \underline{\beta}} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}(\phi_{i,j-1} - \phi_{i,j})}{\Phi_{i,j} - \Phi_{i,j-1}} \underline{x}_i = 0 \quad [3.2]$$

and

$$\frac{\partial \log L}{\partial \alpha_k} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{\Phi_{i,j} - \Phi_{i,j-1}} [\xi_{j,k} \phi_{i,j} - \xi_{j-1,k} \phi_{i,j-1}] = 0 \quad [3.3]$$

Equations [3.2] and [3.3] are non-linear in the parameters since  $\Phi_{i,j}$  and  $\phi_{i,j}$  are non-linear functions of  $\alpha_j$  and  $\underline{\beta}$ . Thus, we must rely on an iterative optimization algorithm such as the Newton-Raphson procedure or the method of scoring in order to derive ML estimates of the parameters. Pratt (1981) showed that the log-likelihood function of the model in [3.1] is globally concave if  $f$ , the derivative of  $F$ , is positive and  $\log F$  is concave, a condition that McCullagh (1980) has shown

that it is satisfied for both OMNP and OMNL models. Hence, it is certain that the iterative algorithm will converge to the global maximum of the likelihood function.

Given that

$$\frac{\partial \phi_{ij}}{\partial \underline{\beta}} = \tau_{ij} \phi_{ij} \underline{x}_i$$

and

$$\frac{\partial \phi_{ij}}{\partial \alpha_k} = \xi_{j,k} \tau_{ij} \phi_{ij}$$

the second partial derivatives can be written as

$$\begin{aligned} \frac{\partial^2 \log L}{\partial \underline{\beta} \partial \underline{\beta}'} &= \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{(\Phi_{ij} - \Phi_{ij-1})^2} \\ &\times [(\Phi_{ij} - \Phi_{ij-1})(\tau_{ij-1} \phi_{ij-1} - \tau_{ij} \phi_{ij}) - (\phi_{ij-1} - \phi_{ij})^2] \underline{x}_i \underline{x}_i' \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 \log L}{\partial \underline{\beta} \partial \alpha_k} &= \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{(\Phi_{ij} - \Phi_{ij-1})^2} \\ &\times [(\Phi_{ij} - \Phi_{ij-1})(\tau_{ij} \phi_{ij} \xi_{j,k} - \tau_{ij-1} \phi_{ij-1} \xi_{j-1,k}) - (\phi_{ij-1} - \phi_{ij})(\phi_{ij} \xi_{j,k} - \phi_{ij-1} \xi_{j-1,k})] \underline{x}_i \end{aligned}$$

and



$$\begin{aligned}
\frac{\partial^2 \log L}{\partial \alpha_k \partial \alpha_l} &= \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{(\Phi_{ij} - \Phi_{i,j-1})^2} \\
&\times [(\Phi_{ij} - \Phi_{i,j-1})(\tau_{i,j-1} \phi_{i,j-1} \xi_{j-1,k} \xi_{j-1,l} - \tau_{ij} \phi_{ij} \xi_{j,k} \xi_{j,l}) \\
&- (\phi_{ij} \xi_{j,k} - \phi_{i,j-1} \xi_{j-1,k})(\phi_{ij} \xi_{j,l} - \phi_{i,j-1} \xi_{j-1,l})]
\end{aligned}$$

Denoting by  $\hat{\alpha}_j$  and  $\hat{\underline{\beta}}$  the ML estimates of  $\alpha_j$  and  $\underline{\beta}$  respectively, we evaluate the matrix of second partial derivatives of  $\log L$  at  $\hat{\alpha}_j$  and  $\hat{\underline{\beta}}$ . This matrix, with the sign reversed, is the information matrix, and the inverse of the information matrix gives the estimates of the asymptotic variance-covariance matrix of the parameter estimates. One can use this matrix to perform any desired tests of significance or construct confidence intervals about the parameters of interest.

### 3.1.2 Maximum Likelihood Estimation for the OMNL Model

The procedure for obtaining ML estimates of the parameters of the OMNL model is similar to that of the OMNP model. Specifically, let  $L_{ij}$  denote the logistic cdf of  $\varepsilon_i^*$  written as

$$L_{ij} = P(\varepsilon_i^* < \alpha_j - \underline{\beta}' \underline{x}_i) = \frac{\exp(\alpha_j - \underline{\beta}' \underline{x}_i)}{1 + \exp(\alpha_j - \underline{\beta}' \underline{x}_i)}$$

Notice that

$$\frac{\partial L_{ij}}{\partial \underline{\beta}} = L_{ij}(1 - L_{ij})\underline{x}_i \quad [3.4]$$

and

$$\frac{\partial L_{ij}}{\partial \alpha_k} = \xi_{j,k} L_{ij} (1 - L_{ij}) \quad [3.5]$$

The likelihood function of the OMNL model is

$$L = \prod_{i=1}^n \prod_{j=1}^m [L_{ij} - L_{ij-1}]^{y_{ij}}$$

and the log-likelihood function is thus

$$\log L = \sum_{i=1}^n \sum_{j=1}^m y_{ij} \log(L_{ij} - L_{ij-1})$$

Using [3.4] and [3.5] we obtain

$$\frac{\partial \log L}{\partial \underline{\beta}} = \sum_{i=1}^n \sum_{j=1}^m y_{ij} [1 - L_{ij} - L_{ij-1}] \underline{x}_i = 0 \quad [3.6]$$

and

$$\frac{\partial \log L}{\partial \alpha_k} = \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{L_{ij} - L_{ij-1}} [\xi_{j,k} L_{ij} (1 - L_{ij}) - \xi_{j-1,k} L_{ij-1} (1 - L_{ij-1})] = 0 \quad [3.7]$$

Equations [3.6] and [3.7] must be solved iteratively because they are non-linear functions of the parameters. Pratt's (1981) result guarantees the convergence of the optimization algorithm to the global maximum of L or logL. The second partial derivatives of logL can be written as follows:

$$\frac{\partial \log L}{\partial \underline{\beta} \partial \underline{\beta}'} = - \sum_{i=1}^n \sum_{j=1}^m y_{ij} [L_{i,j}(1 - L_{i,j}) - L_{i,j-1}(1 - L_{i,j-1})] \underline{x}_i \underline{x}_i'$$

$$\frac{\partial \log L}{\partial \underline{\beta} \partial \underline{\beta}'} = - \sum_{i=1}^n \sum_{j=1}^m y_{ij} [\xi_{j,k} L_{i,j}(1 - L_{i,j}) - \xi_{j-1,k} L_{i,j-1}(1 - L_{i,j-1})] \underline{x}_i \underline{x}_i'$$

and

$$\begin{aligned} \frac{\partial^2 \log L}{\partial \alpha_k \partial \alpha_l} &= \sum_{i=1}^n \sum_{j=1}^m \frac{y_{ij}}{(L_{i,j} - L_{i,j-1})^2} \\ &\times [\xi_{j,k} \xi_{j,l} L_{i,j}^3 (1 - L_{i,j}) - \xi_{j,k} \xi_{j,l} L_{i,j} L_{i,j-1} (1 - L_{i,j})^2 - \xi_{j-1,k} \xi_{j-1,l} L_{i,j}^2 L_{i,j-1} (1 - L_{i,j}) \\ &+ \xi_{j-1,k} \xi_{j-1,l} L_{i,j} L_{i,j-1} (1 - L_{i,j})(1 - L_{i,j-1}) + \xi_{j,k} \xi_{j-1,l} L_{i,j} L_{i,j-1} (1 - L_{i,j})(1 - L_{i,j-1})] \end{aligned}$$

The asymptotic variance-covariance matrix of the ML estimates can now be derived and inferences about the significance of subsets of the parameters are possible.

### 3.2 *Unordered Multinomial Choice Models*

A typical model of this kind considers individuals who must fall into one of  $m \geq 2$  distinct and mutually exclusive categories. When  $m = 2$ , the problem reduces to the specification of a binary choice model of the type presented in the preceding section. Suppose now that there are  $m > 2$  categories. Unlike binary choice models for which three alternative ways of motivation have been discussed, utility maximization provides the only meaningful basis for the motivation of

multinomial choice models, suitable for the definition of both logit and probit formulations. Let the utility that the  $i$ -th ( $i = 1, 2, \dots, n$ ) individual derives by choosing the  $j$ -th ( $j = 1, 2, \dots, m$ ) alternative be expressed as

$$U_{ij} = \beta_j' \mathbf{x}_i + \varepsilon_{ij} \quad [3.8]$$

where  $\beta_j' \mathbf{x}_i$  is a nonstochastic linear function of explanatory variables and unknown parameters and  $\varepsilon_{ij}$  is an unobservable random variable, with  $E(\varepsilon_{ij}) = 0$ . Given that each individual is a utility maximizer, i.e., it chooses the alternative for which the associated utility is highest, the probability that the  $j$ -th alternative is chosen can be expressed as follows:

$$\begin{aligned} p_{ij} &= P[U_{ij} = \max(U_{i1}, U_{i2}, \dots, U_{im})] \\ &= P\left[\bigcap_{j \neq k} (U_{ij} > U_{ik})\right] \\ &= P\left[\bigcap_{j \neq k} (\varepsilon_{ik} - \varepsilon_{ij} < (\beta_j' - \beta_k') \mathbf{x}_i)\right] \end{aligned} \quad [3.9]$$

for  $j, k = 1, 2, \dots, m$ . This is the general expression of the unordered multinomial choice model, so called because the observations are assumed to be generated by a multinomial process with probabilities  $p_{ij}$ . As is evident from [3.9], the kind of multinomial choice model, logit or probit, depends on the distribution assumed for the error terms  $\varepsilon_{ij}$ . Clearly, distributions that produce convenient distributions under subtraction, are popular candidates for the distribution of the error term (Judge et al., 1985).

### 3.2.1 Unordered Multinomial Logit (UMNL) Model

McFadden (1974) was the first to show that a necessary and sufficient<sup>1</sup> condition for the UMNL model to arise from [3.8] is that the  $\varepsilon_{ij}$  's are independently and identically distributed (iid) with the standardized type I extreme value or log-Weibull distribution (Johnson and Kotz, 1972). His proof was based on a result by Gumbel (1962) that the difference of two independent random variables, each having the same type I extreme value distribution, is distributed as a logistic random variable. The cdf of a standardized type I extreme value random variable  $X$  is defined as

$$P[X \leq x] = \exp(-\exp(-x))$$

for  $-\infty < x < \infty$ . Under McFadden's proof, the individual choice probabilities in [3.9] can then be rewritten as

$$p_{ij} = \frac{\exp(\beta_j' x_i)}{\sum_{j=1}^m \exp(\beta_j' x_i)}$$

To satisfy  $\sum_{j=1}^m p_{ij} = 1$ , we set  $\beta_1 = 0$  obtaining

$$p_{i1} = \frac{1}{1 + \sum_{j=2}^m \exp(\beta_j' x_i)}$$

and

$$p_{ij} = \frac{\exp(\beta_j' x_i)}{1 + \sum_{j=2}^m \exp(\beta_j' x_i)}$$

for  $j = 2, 3, \dots, m$ . This formulation is known as the unordered multinomial logit model. Mantel (1966), Bock (1969) and Press (1972) developed in detail the theoretical background of this model.

---

<sup>1</sup> The first proof of the necessary part is due to A. Marley as reported in Luce and Suppes (1965). McFadden rediscovered this and proved the sufficient part.

It has been used, among others, by Theil (1969) to study choices of transportation modes, by Uhler and Cragg (1971) to study the structure of asset portfolios of households and by Schmidt and Strauss (1975) to study the determinants of occupational choice.

A second motivation for the UMNL model which was introduced by Theil (1969) and used extensively before McFadden's utility motivation was published, is based on the idea of expressing individual choice probabilities in binary form, as a direct generalization of the binary logit model. To illustrate this, let

$$\frac{p_{ij}}{p_{i1} + p_{ij}} = F(\beta_j' x_i)$$

for  $j = 2, 3, \dots, m$ . It is then implied that

$$\frac{p_{ij}}{p_{i1}} = \frac{F(\beta_j' x_i)}{1 - F(\beta_j' x_i)} = G(\beta_j' x_i)$$

for  $j = 2, 3, \dots, m$ . Since

$$\begin{aligned} \sum_{j=2}^m G(\beta_j' x_i) &= \sum_{j=2}^m \frac{p_{ij}}{p_{i1}} \\ &= \frac{1 - p_{i1}}{p_{i1}} \end{aligned}$$

we can write

$$p_{i1} = \frac{1}{1 + \sum_{j=2}^m G(\beta_j' x_i)} \quad [3.10]$$

and

$$p_{ij} = \frac{G(\beta_j' x_i)}{1 + \sum_{j=2}^m G(\beta_j' x_i)} \quad [3.11]$$

for  $j = 2, 3, \dots, m$ . If  $F$  is chosen to be the logistic cdf, then

$$F(\beta_j' x_i) = \frac{\exp(\beta_j' x_i)}{1 + \exp(\beta_j' x_i)}$$

therefore,

$$G(\beta_j' x_i) = \exp(\beta_j' x_i) \quad [3.12]$$

By substituting [3.12] into [3.10] and [3.11], the expression of UMNL model is obtained.

### 3.2.1.1 Maximum Likelihood Estimation of the UMNL Model

Recall the selection probabilities as defined by the UMNL model:

$$p_{i1} = \frac{1}{1 + \sum_{j=2}^m \exp(\beta_j' x_i)}$$

and

$$p_{ij} = \frac{\exp(\beta_j' x_i)}{1 + \sum_{j=2}^m \exp(\beta_j' x_i)}$$

for  $j = 2, 3, \dots, m$ . Given this formulation, we shall now consider ML estimation of the model parameters,  $\beta_j$ , based on a sample of size  $n$ . Each of the  $n$  individuals is assumed to fall into one of the  $m$  categories with probabilities  $p_{ij}$  ( $j = 1, 2, \dots, m$ ). By defining a set of dummy variables  $y_{ij}$  as

$$y_{ij} = 1 \quad \text{if the } i\text{-th individual falls in the } j\text{-th category}$$

$$y_{ij} = 0 \quad \text{otherwise.}$$

we are able to express the likelihood function of the UMNL model in the form

$$L = \prod_{i=1}^n \prod_{j=1}^m p_{ij}^{y_{ij}}$$

and the logarithm of the likelihood function,

$$L = \sum_{i=1}^n \sum_{j=1}^m y_{ij} \log p_{ij}$$

To maximize  $\log L$  we differentiate with respect to  $\beta_k$  ( $k = 1, 2, \dots, m$ ) and set the resulting matrix of derivatives equal to zero. Notice first that  $\partial p_{ij} / \partial \beta_k = -p_{ij} p_{ik} \mathbf{x}_i$ ,  $\partial p_{ik} / \partial \beta_k = p_{ik}(1 - p_{ik}) \mathbf{x}_i$  and  $\sum_{j=1}^m y_{ij} = 1$ . Then,

$$\begin{aligned} \frac{\partial \log L}{\partial \beta_k} &= \sum_{i=1}^n \left[ \frac{y_{ik}}{p_{ik}} p_{ik}(1 - p_{ik}) - \sum_{\substack{j=1 \\ j \neq k}}^m \frac{y_{ij} p_{ij} p_{ik}}{p_{ij}} \right] \mathbf{x}_i \\ &= \sum_{i=1}^n \left[ y_{ik} - y_{ik} p_{ik} - p_{ik} \left( \sum_{\substack{j=1 \\ j \neq k}}^m y_{ij} \right) \right] \mathbf{x}_i \\ &= \sum_{i=1}^n [y_{ik} - y_{ik} p_{ik} - p_{ik}(1 - y_{ik})] \mathbf{x}_i \\ &= \sum_{i=1}^n [y_{ik} - p_{ik}] \mathbf{x}_i \end{aligned}$$



Thus, the equations to solve for obtaining the ML estimates for  $\underline{\beta}_k$  are

$$\sum_{i=1}^n (y_{ik} - p_{ik}) \underline{x}_i = 0 \quad [3.13]$$

Notice the similarity of the above equations with those obtained in the case of binary logit model. The interpretation is similar. For instance, if  $\underline{x}_i$  contains a constant term, then the predicted and actual frequencies are identical for each one of the  $m$  categories.

Equations [3.13] are nonlinear in the parameters  $\underline{\beta}_k$ . Consequently, in order to obtain solutions we will have to employ a non-linear optimization procedure such as the Newton-Raphson or the method of scoring. Notice that by differentiating [3.13] we obtain

$$\frac{\partial^2 \log L}{\partial \underline{\beta}_k \partial \underline{\beta}_k'} = - \sum_{i=1}^n p_{ik} (1 - p_{ik}) \underline{x}_i \underline{x}_i' \quad [3.14]$$

and

$$\frac{\partial^2 \log L}{\partial \underline{\beta}_k \partial \underline{\beta}_l'} = \sum_{i=1}^n p_{ik} p_{il} \underline{x}_i \underline{x}_i' \quad [3.15]$$

which, interestingly, do not depend on  $y_{ik}$ . Since  $p_{ik} > 0$  by the definition of the UMNL model, the matrix of second derivatives is negative definite hence, we can conclude that the log-likelihood function is globally concave and any non-linear optimization algorithm will eventually converge to the maximum.

The asymptotic variance-covariance matrix of the ML estimates can be obtained by the elements of the inverse of the information matrix, whose diagonal blocks are given by equation [3.14] and non-diagonal blocks by equation [3.15], both expressions with signs reversed.

### 3.2.2 The Conditional Logit Model

McFadden (1974) suggested a different type of an unordered multinomial logit model which he called "the conditional logit model". The individual choice probabilities arising from this formulation are defined as

$$p_{i1} = \frac{1}{1 + \sum_{j=2}^m \exp(\underline{\beta}' \underline{x}_{ij})}$$

and

$$p_{ij} = \frac{\exp(\underline{\beta}' \underline{x}_{ij})}{1 + \sum_{j=2}^m \exp(\underline{\beta}' \underline{x}_{ij})}$$

for  $j = 2, 3, \dots, m$ .

Recall that under the UMNL formulation the probabilities of different choices for each individual are expressed as functions of  $\underline{\beta}_j' \underline{x}_i$ , the product of a coefficient vector specific to each alternative and the vector of explanatory variables featuring individual characteristics. Under McFadden's conditional logit model, the individual choice probabilities are expressed as functions of  $\underline{\beta}' \underline{x}_{ij}$ , the product of a coefficient vector common to all alternatives and the vector of explanatory variables which is specific for each individual and alternative. As McFadden (1974) described it, the main difference between the conditional logit model and the UMNL model is that the conditional logit model considers the effects of choice characteristics on the determinants of choice probabilities as well, whereas the UMNL model considered here makes the choice probabilities dependent on individual characteristics only.

The fact that the coefficient vector remains constant across alternatives makes the conditional logit model useful in predicting the probability of choice for an alternative not considered in the estimation procedure, but for which we are given the vector of characteristics  $\mathbf{x}_{ij}$ . This is considered as the main advantage of the conditional logit model over the UMNL model.

The specification of the conditional logit model is particularly convenient for applications in human behavior. Its popularity among sociologists and socio-economists is such that many authors refer to it as the UMNL model, thus contributing into a great deal of confusion in the literature. On the other hand, applications of conditional logit model in forest biometry are not expected to be as widespread as those of the UMNL model. The reason lies with the definition of the conditional logit model which requires the explanatory variables to vary across alternatives for each individual. It is easy to specify and acquire such data, when studying the behavior of individuals for which concepts such as cost, profit, loss can be meaningfully defined, but it is very difficult, if not impossible, to obtain when studying the behavior of non-sentient organisms such as trees. For instance, in studying a consumer's choice of transportation mode, it is sensible to define independent variables specific to the mode such as time and cost involved for riding a bus, train or driving own car but it is not clear what independent variables would characterize a tree according to the category it belongs to, when studying, say, individual tree mortality. The data foresters usually have access to, with the possible exception of forest economists, varies across individuals (trees) but not across alternatives, so that for the  $i$ -th individual,  $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \dots = \mathbf{x}_{im} = \mathbf{x}_i$ , indicating that the UMNL specification is more suitable for forestry applications than the conditional logit model.

### 3.2.3 The Independence of Irrelevant Alternatives (IIA) Property

Despite the fundamental difference in their specification, the UMNL and conditional logit model have been preferred over other theoretical possibilities because of computational simplicity. The primary disadvantage of UMNL and conditional logit models is a property termed the "independ-

ence of irrelevant alternatives” which was first pointed out by Arrow (1951).<sup>2</sup> The IIA property holds that for a specific individual the ratio of choice probabilities of any two alternatives is unaffected by the presence of any other alternative. Consider for example the ratio of choice probabilities for two alternatives,  $k$  and  $l$  :

$$\frac{p_{ik}}{p_{il}} = \frac{\exp(\beta_k' x_i)}{\exp(\beta_l' x_i)}$$

It is clear that the ratio of probabilities for the two alternatives does not depend on any alternative other than  $k$  and  $l$ . This seemingly simple property has some important ramifications. When the alternatives are not distinct enough, the IIA property may cause the model to produce odd and erroneous predictions.

As an illustration consider McFadden’s famous example of alternative modes of transportation. Suppose a population faces the alternatives of traveling by car and by bus, and two-thirds choose to use car. Suppose now that a distinction is made between blue buses and red buses. As it affects choice between modes of transportation, the color of the bus is irrelevant. One would expect two out of three persons to choose car when offered the alternative of a blue bus, and the same when offered a choice between car and the red bus. Given that color is irrelevant to those who choose bus (1/2 blue, 1/2 red), then 1/4 of the entire population must prefer blue bus over car in order to maintain a two-to-one preference for car. In turn, this implies that 1/2 of the entire population must prefer car over bus regardless of color, which is paradoxical. Consequently, two-thirds of the population will still choose car and the remainder will split between the bus alternatives. Under the logit formulation however, only half the population will use car in order for the overall population proportion to remain two-thirds in favor of the car.

To understand fully the IIA property and its implications one must go back to the original assumptions from which both the UMNL and the conditional logit models are derived. The core

---

<sup>2</sup> As Hausman and Wise (1978) state, this terminology is somewhat misleading. In fact, “independence of relevant alternatives” or “independence among alternatives” might be more descriptive of the property.

of the problem is in the assumption that the error terms  $\varepsilon_{ij}$  are mutually independent (Lerman and Manski, 1981). This assumption requires that the sources of errors contributing to individual utilities must do so in a way such that the overall error terms are independent across individuals and alternatives. There are many circumstances where the alternatives are defined to be distinct enough so that this assumption is satisfied reasonably well. There are cases however, such as the red bus/blue bus paradox, for which this assumption is wholly implausible since the error terms of the red bus and blue bus are more reasonably assumed to be perfectly correlated.

The IIA property is not uniquely associated with the definition of the UMNL and the conditional logit models (Ben-Akiva and Lerman, 1985). In fact, although models other than the ones considered in this study might produce different results, any model based on the assumption of independent error terms would necessarily yield counterintuitive results for the red bus/ blue bus problem.

As mentioned earlier, the IIA property is not a negative factor for the logit formulation when the alternatives are distinct enough. One sensible way to avoid the unpleasant implications of IIA property, is to merge alternatives for which there exists profound similarities (Kmenta, 1986). Hausman and McFadden (1984) proposed a test for the IIA hypothesis in the conditional logit model which can be used whenever the alternatives are not similar enough so that can be merged but not so different that the assumptions of independent error terms is questionable. No such test exists for the UMNL model however.

### 3.2.4 Unordered Multinomial Probit Models

One form of unordered multinomial probit model can be derived in a manner similar to that of UMNL model by assuming that the error terms  $\varepsilon_{ij}$  in [3.1] are iid normal random variables with mean 0 and constant variance  $\sigma^2$ . This assumption leads to a probit model known as "the inde-

pendent probit model" which was first proposed by Aitchison and Bennett (1970). As explained previously, this model, like any other model which is based on the assumption of iid error terms, features the IIA property. Computationally, the estimation of parameters is more problematic under the independent probit than under the UMNL model since the normal cdf cannot be expressed in closed form and therefore,  $m-1$ -tuple integrals<sup>3</sup> arising from [3.2] must be evaluated at each step of the iterative procedure employed for the maximization of the likelihood function (Amemiya, 1985). In addition, because of the well recognized close relationship between the normal and logistic distributions, parameter estimates derived by the independent probit model are usually very similar to those derived by the UMNL model. As a result, only a few, very limited number of applications of the independent probit model appear in the literature.

Lerman and Manski (1981) proposed a probit analogue of McFadden's conditional logit model that does not exhibit the IIA property. This model is known as the "dependent probit model" because it assumes that the error terms in [3.1] are independent across individuals (i) but not across alternatives (j) following a multivariate normal distribution with  $m \times 1$  vector of means,  $\underline{0}$  and a  $m \times m$  variance-covariance matrix  $\Sigma$ . Lerman and Manski applied this model to explain modal choice among driving a car, sharing rides and riding a bus, for 557 workers in the area of Washington, D.C. A conditional logit model was also fitted to the data for comparison. The conclusions were that i) probit and logit estimates did not differ by much, ii) the asymptotic variance-covariance matrix of the estimated coefficients could not be accurately derived, iii) based on the Akaike information criterion, an increase in the log-likelihood function in the probit model as compared to the logit was not large enough to compensate for the loss in degrees of freedom due to the estimation of  $\Sigma$  and iv) the probit estimation took 1400 percent more CPU time per iteration than the logit estimation.

Hausman and Wise (1978) considered a slightly less general model than that of Lerman and Manski by imposing certain zero specification to the error variance-covariance matrix  $\Sigma$ . They then applied

---

<sup>3</sup> Taking advantage of the independent error terms, the evaluation of  $m-1$ -tuple integrals is substantially simplified to the evaluation of one integral at a time, a simple task by today's computing standards.

their model to the same data that Lerman and Manski used and compared the results with those of a conditional logit and independent conditional<sup>4</sup> logit models. The conclusions from their study were that i) conditional logit and independent conditional probit gave similar results, ii) the dependent conditional probit differed significantly from the other two models and iii) the dependent conditional probit model fitted best.

Amemiya (1985) interpreted the apparent discrepancy between the conclusions of Hausman and Wise and those of Lerman and Manski as an indication of the crucial role that the specification of the error variance-covariance matrix may play in multinomial probit formulations. Despite the different conclusions however, both studies recognized that the dependent conditional probit model can be applied only for small number of alternatives (at most three or four) because the computations involve the evaluation of multiple integrals at each step of the iterative method used for estimation. Other works by Dutt (1976), Daganzo, Bouthelier and Sheffi (1977) and Daganzo (1979) have contributed in resolving some of the computational problems. However, only a very small number of applications appeared in the literature and there is still no evidence to suggest in which situations the greater generality of the dependent conditional probit model is worth the additional computational problems resulting from its use (Ben-Akiva and Lerman, 1985).

In forestry, all versions of the multinomial probit model presented here, are of very limited use. The independent probit because it yields estimates similar to the UMNL model with added computational difficulty and the conditional probit, independent or dependent, because the availability of explanatory variables specific to individuals and alternatives is very limited in forestry. Given this circumstance, unordered multinomial probit models will be excluded from further consideration in the remainder of this study.

---

<sup>4</sup> The independent conditional logit model is the direct probit analogue of the conditional logit model, exhibiting the IIA property.

## Chapter IV

### Inference and Model Selection

#### *4.1 Interpretation of Parameter Estimates*

An important difference between classical regression models and qualitative response models lies with the interpretation of the influence which the explanatory variables exert upon the response variable. Unlike linear regression analysis, the coefficients in qualitative response models do not indicate the change in the response given a unit increase in the corresponding independent variable. Rather, these coefficients reflect the impact of a change in an independent variable on  $F^{-1}(p_{ij})$ , where  $F$  indicates the kind of model considered, i.e., linear probability, probit or logit model. Consider first a dichotomous response. Recall that the probability that an event will occur is generally defined by

$$\begin{aligned} p_i &= P(y_i = 1) \\ &= F(\underline{\beta}' \underline{x}_i) \end{aligned}$$



for the  $i$ -th ( $i = 1, 2, \dots, n$ ) individual. Taking the partial derivative of  $p_i$  with respect to a particular independent variable  $x_{ij}$ , we obtain

$$\frac{\partial p_i}{\partial x_{ij}} = f(\beta' \underline{x}_i) \beta_j$$

where  $f(\cdot)$  is the pdf corresponding to  $F(\cdot)$ .

It is now clear that although the sign of the corresponding coefficient indicates the direction of change in the probability, the magnitude of this change is a function not only of the magnitude of the coefficient  $\beta_j$ , but also, of the values of all the independent variables and associated coefficients. Since  $f(\beta' \underline{x}_i)$  is defined to be the derivative of  $F(\beta' \underline{x}_i)$ , it is reasonable to think of this measure as representing the steepness of the cdf  $F$ . Naturally, the steeper the cdf, the greater the effect of a change in the value of an explanatory variable (Fomby et al., 1984).

Steinberg (1987) considered the interpretation of coefficient estimates in UMNL models. He noted that each set of coefficient estimates may be viewed as representing the outcome of a binary logit model in which a two choice problem has been analyzed; in each case the pair of choices consists of the "reference" choice and one of the remaining choices. Under this viewpoint, it becomes clear that results are dependent upon the coding of the dependent variable. Thus, like in linear regression models with dummy variables, the coefficient estimates represent shifts in the probability outcome relative to the "reference" choice. Given that the analyst defines the "reference" choice to be the one for which inference is most important, the dependency of the UMNL coefficient estimates on the coding of the dependent variable is usually beneficial. There are situations however, where information concerning a contrast between any two non-reference choices is needed. To avoid re-coding and re-fitting the UMNL model, Steinberg suggested the construction of a table of derivatives which provides information about the change in probability of each choice with respect to each regressor variable, independently of the code assigned to the dependent variable. More specifically, he expressed the partial derivative of the probability that the  $i$ -th individual will select the  $j$ -th alternative with respect to the  $k$ -th regressor variable as follows:

$$\frac{\partial p_{ij}}{\partial x_{ik}} = p_{ij} \left( \beta_{jk} - \sum_{l=1}^{J-1} p_{il} \beta_{lk} \right).$$

From the above expression, the reader can verify that the influence of the  $k$ -th regressor can be decomposed into two parts. The first is a direct effect which is given by the coefficient  $\beta_{jk}$  and is proportional to the probability of the  $j$ -th choice. The second part is the negative of a weighted average of the coefficients  $\beta_{lk}$  of  $x_{ik}$  corresponding to the other choices. Therefore, even though the direct effect  $\beta_{jk}$  might be positive, the overall impact might be positive or negative depending on the magnitude of the impact this regressor exerts on the remaining choices. A convenient way to present this information is a table having as rows the regressor variables and as columns the alternatives considered by the problem at hand. The entries in this table are the derivative values averaged over all individuals.

## 4.2 Tests for the General Linear Hypothesis

Individual or joint hypothesis tests about coefficients can be conducted by relying on the asymptotic normality of ML estimators. A general linear hypothesis is defined by

$$H_0: Q' \underline{\beta} = \underline{c}$$

where  $Q$  is a  $p \times q$  matrix of known constants and  $\underline{c}$  is a  $q$ -vector of known constants, each appropriately determined by the research objectives. It is implicitly assumed that  $p < q$  and  $Q$  is a full row rank matrix.

Two asymptotic tests, not specifically designed for qualitative response models, are most commonly used to test the general linear hypothesis, Wald's test and the likelihood ratio test. Both tests assume that the alternative hypothesis specifies

$$H_1: Q'\underline{\beta} \neq \underline{c}.$$

Let  $\hat{\underline{\beta}}$  be a ML estimator and let  $\hat{V}(\hat{\underline{\beta}})$  be a consistent estimator of its asymptotic variance-covariance matrix. Then, Wald's test of the hypothesis

$$H_0: Q'\underline{\beta} = \underline{c} \quad \text{vs.}$$

$$H_1: Q'\underline{\beta} \neq \underline{c}$$

is based on the test statistic  $W$ , known as Wald statistic (Wald, 1943) and defined by

$$W = (Q'\hat{\underline{\beta}} - \underline{c})'[Q'(\hat{V}(\hat{\underline{\beta}}))Q]^{-1}(Q'\hat{\underline{\beta}} - \underline{c})$$

Under the null hypothesis,  $W$  is asymptotically distributed as a chi-square random variable with  $q$  degrees of freedom ( $\chi_q^2$ ). The decision rule at the  $\alpha$  level of significance is

$$\text{Reject } H_0 \quad \text{if } W > \chi_{q(1-\alpha)}^2$$

$$\text{Accept } H_1 \quad \text{otherwise}$$

For the special case  $q = 1$ , Wald statistic is reduced to the form

$$\sqrt{W} = \frac{Q'\hat{\underline{\beta}} - \underline{c}}{\sqrt{Q'(\hat{V}(\hat{\underline{\beta}}))Q}}$$

which, under the null hypothesis, is asymptotically distributed as a standard normal random variable. This form is to be preferred since it allows for one sided alternative hypotheses such as,  $Q'\underline{\beta} > \underline{c}$  or  $Q'\underline{\beta} < \underline{c}$ . Amemiya (1981) reports that some authors prefer to assume that  $\sqrt{W}$  is asymptotically distributed as a central  $t$  random variable with  $n - p$  degrees of freedom ( $t_{n-p}$ ), instead

of standard normal. He argues however, that when  $n - p$  is reasonably large, it does not make much practical difference.

The likelihood ratio test statistic is defined by

$$LR = 2[\log L(\hat{\underline{\beta}}_F) - \log L(\hat{\underline{\beta}}_R)]$$

where  $\hat{\underline{\beta}}_R$  denotes the constrained (reduced model) ML estimator obtained by maximizing the log-likelihood function with respect to  $\underline{\beta}$  subject to the constraint  $Q'\underline{\beta} = \underline{c}$  and  $\hat{\underline{\beta}}_F$  is the unconstrained (full model) ML estimator of  $\underline{\beta}$ . Like the Wald statistic,  $LR$  is asymptotically distributed as  $\chi^2_q$  under  $H_0$ . The decision rule for the  $\alpha$  level of significance specifies

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } LR > \chi^2_{q(1-\alpha)} \\ \text{Accept } H_1 & \text{otherwise.} \end{array}$$

A definite advantage of Wald's test over the likelihood ratio test, especially due to the iterative nature of maximum likelihood estimation, is that it requires fitting the model only under the alternative hypothesis (in contrast to the likelihood ratio test, which requires fitting the model twice). However, as Rao (1965) points out, little is known about the comparative power functions of the two tests. Hauck and Donner (1977) found Wald's test to behave in an aberrant manner for testing hypotheses regarding a single parameter in a binary logit model. In particular, the authors found that i) for any sample size, Wald's test statistic decreases to zero as the distance between the parameter estimate and null value increases and ii) the power of Wald's test based on its asymptotic distribution decreases to the significance level for alternatives far from the null value. Consequently, the analyst has no way of knowing whether a small value of  $W$  indicates that the actual value of the parameter is near or very far from the null value. Based on their findings, the authors recommended the use of likelihood ratio test instead. Jennings (1986) showed that the poor performance of Wald's test may be due to inadequately defined observed information matrices. More specifically, he argued that since inference based on the observed information matrix (e.g. Wald's test) can be viewed as a quadratic approximation of the likelihood surface, this approximation may not

be satisfactory in some instances. By examining the cubic term of the Taylor expansion of the log-likelihood, the author developed a measure to judge the adequacy of the inference obtained from the observed information matrix. He also suggested one approach for transforming the parameters when poor inference is detected.

Large sample approximations for confidence intervals for predicted probabilities can be computed using the fact that the linear estimate  $\hat{\beta}_j' \underline{x}_i$  follows asymptotically the normal distribution with mean  $\beta_j' \underline{x}_i$  and variance  $\underline{x}_i' \text{Var}(\hat{\beta}_j) \underline{x}_i$ . Then, one can first compute upper and lower confidence limits for the parameter  $\beta_j' \underline{x}_i$  and then transform them using the inverse logistic function. More specifically, if  $\hat{y}_j = \hat{\beta}_j' \underline{x}_0$  is the estimated response for alternative  $j$  at a given vector  $\underline{x}_0$ , the predicted probability for the  $j$ th alternative is given by

$$\hat{p}_j = \frac{\exp(\hat{y}_j)}{1 + \exp(\hat{y}_j)}$$

Also, because  $\hat{p}_j$  is a one-to-one monotonically increasing function of  $\hat{y}_j$  and

$$\hat{y}_{jL} = \hat{y}_j - z_{1-\alpha/2} \sqrt{\underline{x}_0' (\hat{V}(\hat{\beta}_j)) \underline{x}_0}$$

and

$$\hat{y}_{jU} = \hat{y}_j + z_{1-\alpha/2} \sqrt{\underline{x}_0' (\hat{V}(\hat{\beta}_j)) \underline{x}_0}$$

are respectively the  $(1 - \alpha/2)100$  percent lower and upper confidence limits for  $y_j$ , the corresponding limits for the probability of the  $j$ th alternative  $p_j$  are:

$$\hat{p}_{jL} = \frac{\exp(\hat{y}_{jL})}{1 + \exp(\hat{y}_{jL})} \quad \text{and} \quad \hat{p}_{jU} = \frac{\exp(\hat{y}_{jU})}{1 + \exp(\hat{y}_{jU})}.$$

### 4.3 *Criteria for Model Selection*

The problem of choosing a model among several alternatives is often solved by means of goodness of fit criteria. These criteria are summary statistics, like  $R^2$  in the familiar regression analysis, which indicate the accuracy by which a model approximates the observed data (Myers, 1986). In the case of qualitative response models, accuracy can be judged either in terms of the fit between estimated and observed frequencies (if such information is available) or in terms of the model's ability to forecast observed responses. It should be emphasized at this point that expressing a model's goodness of fit by a single norm is not an easy problem to solve. In fact, it is more complicated in the case of qualitative response models than it is in regression analysis. This is because comparing the values between a continuous and a discrete variable (e.g. the estimated choice probabilities vs. the actual choices in QR models) is generally more difficult than comparing the values between two continuous variables (e.g. the estimated and observed responses in regression models). It is the difference in the scale of measurement of the two variables which led Hensher and Johnson (1981) to argue that it does not make sense to use "residuals" (i.e., the difference between the actual choice and the estimated choice probability for all individuals) in the calculation of a measure of fit analogous to  $R^2$ .

Amemiya (1981) lists several goodness of fit criteria which have been proposed over the past fifteen years. The most frequently encountered criteria in econometric and biometric research are presented below.

### 4.3.1 Number of Wrong Predictions (WP)

This criterion is defined by

$$WP = \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \hat{y}_{ij})^2$$

where  $\hat{y}_{ij}$  is the estimated analogue of  $y_{ij}$  namely,

$$\begin{aligned} \hat{y}_{ik} &= 1 & \text{if } \hat{p}_{ik} = \max(\hat{p}_{ij}) \ (j = 1, 2 \dots m) \\ \hat{y}_{ik} &= 0 & \text{otherwise} \end{aligned}$$

for  $k = 1, 2, \dots, m$  and  $\hat{p}_{ik}$  is the estimate of  $p_{ik}$ , i.e., the probability that the  $i$ -th individual belongs to the  $j$ -th category. WP gives the number of wrong predictions since  $(y_{ij} - \hat{y}_{ij})^2 = 1$  if and only if  $y_{ij} \neq \hat{y}_{ij}$ .

For the dichotomous case where  $j = 0$  or  $1$ , WP has the form

$$WP = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where

$$\begin{aligned} \hat{y}_i &= 1 & \text{if } \hat{p}_i \geq 0.5 \\ \hat{y}_i &= 0 & \text{otherwise.} \end{aligned}$$

This criterion is typically used in discriminant analysis where an individual must be classified as belonging to one of several groups according to its observed characteristics. A problem with WP is that events with zero or near zero probability of occurrence are treated the same as events with

much higher but not maximum probability of occurrence. Thus, when an event  $y_i = 1$  takes place, a person who estimated its probability to be 0.49 and a person who estimated it to be 0 are equally penalized. A major disadvantage is that if we are dealing with an event which happens with high probability or a low probability, most models will do well by this criterion.

### 4.3.2 Sum of Squared Residuals (SSR)

SSR is defined for multinomial choice models as

$$SSR = \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \hat{p}_{ij})^2$$

and for binary choice models as

$$SSR = \sum_{i=1}^n (y_i - \hat{p}_i)^2$$

This is a criterion corresponding to the residual sum of squares in linear regression analysis from which  $R^2$  is derived. In fact, Lave (1970) used  $SSR$  to define an analogue of  $R^2$  for the case of a dichotomous response variable as

$$R_L^2 = 1 - \frac{\sum_{i=1}^n (y_i - F(\hat{\beta}' \underline{x}_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



where  $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ . He also suggested the use of a related statistic,  $R_{LA}^2$ , defined as

$$R_{LA}^2 = 1 - \frac{(n-p)}{n} (1 - R_L^2)$$

where  $R_L^2$  has been defined previously and  $p$  is the number of parameters estimated. This criterion is directly analogous to the "adjusted  $R^2$ " used in linear regression analysis (see Draper and Smith, 1981). It measures the percent variability explained by the model, corrected for the degrees of freedom (Goldberger, 1973).

Morrison (1972) argued that the low  $R_L^2$  values which have been frequently obtained by many authors need not imply that the model is not good. He derived an upper bound on  $R_L^2$  (which was less than 1) based on the assumption that the response variable ( $p$  in our case) followed a beta distribution. Goldberger (1973) disagreed with Morrison by pointing out that  $R_L^2$  as a measure of the proportion of the variance explained by the binary choice model should only be bounded by 0 and 1. Apart from this controversy however, both authors failed to realize a major deficiency associated with the use of SSR as a goodness of fit criterion. This deficiency stems from the fact that the use of of SSR ignores the heteroscedastic nature of qualitative response models.

Even though the binary model is estimated by methods that account for its error heterogeneity by assigning weights to individuals inversely proportional to their estimated variance (see the interpretation of the method of scoring iteration above), SSR weighs equally all individuals entering the model. The following criterion is specifically defined to overcome this deficiency of SSR.

### 4.3.3 Weighted Sum of Squared Residuals (WSSR)

WSSR is defined as

$$WSSR = \sum_{i=1}^n \sum_{j=1}^m \frac{(y_{ij} - \hat{p}_{ij})^2}{\hat{p}_{ij}}$$

for multinomial choice models and as

$$WSSR = \sum_{i=1}^n \frac{(y_i - F(\hat{\beta}'x_i))^2}{F(\hat{\beta}'x_i)(1 - F(\hat{\beta}'x_i))}$$

for binary response models.

WSSR is a natural and intuitively appealing criterion, given the error heterogeneity of qualitative response models. Its objective is to attach a higher cost to the error made in predicting a random variable with a smaller variance since such a random variable should be easier to predict than the one with a larger variance (Amemiya, 1981). Because the variance of the residuals is unknown, an estimate is used instead. Thus, WSSR assigns each residual a weight which is inversely proportional to its estimated variance.

Although it is obvious how the error variance is incorporated in the expression of WSSR for binary models, the expression of WSSR for multinomial models may not be as clear. For this reason, the derivation of WSSR, as provided by Amemiya (1981), is presented below.

Define first the  $m-1$  vectors  $Y_i = (y_{i2}, y_{i3}, \dots, y_{im})'$  and  $P_i = (p_{i2}, p_{i3}, \dots, p_{im})'$ . Also define  $D_i$  as an  $(m-1) \times (m-1)$  diagonal matrix whose  $j$ -th diagonal element is  $p_{ij}$ . Then, the error variance-covariance matrix is by definition

$$V_i = E(Y_i - P_i)(Y_i - P_i)' = D_i - P_i P_i'$$

with inverse (see Maddala 1977, p. 446, eq. A-16)

$$V_i^{-1} = D_i^{-1} + \frac{1}{p_{i1}} \mathbf{1}\mathbf{1}'$$

where  $\mathbf{1}$  is an  $m$ -vector of ones. The weighted sum of squared residuals is now defined as

$$\sum_{i=1}^n (Y_i - P_i)' [V_i^{-1}] (Y_i - P_i) = \sum_{i=1}^n \sum_{j=1}^m \frac{(y_{ij} - p_{ij})^2}{p_{ij}}$$

and WSSR is obtained by substituting  $\hat{p}_{ij}$  for  $p_{ij}$  in the above expression.

#### 4.3.4 Prediction Success Index (PSI)

McFadden et al. (1977) looked at the proportion of successful predictions of the choices made by all individuals and derived a measure for assessing a model's predictive ability which they called "prediction success index" and expressed as

$$PSI = \sum_{i=1}^m \left[ \frac{N_{ii}}{N_{..}} - \left( \frac{N_{.i}}{N_{..}} \right)^2 \right]$$

where  $N_{ij}$ ,  $N_{.i}$  and  $N_{..}$  are given by the prediction success table (Table 4.1) below. In this table,  $N_{ij}$  refers to the number of individuals who chose alternative  $i$  but had been predicted to choose alternative  $j$ .  $N_{ii}$  refers to the number of correct predictions for alternative  $i$ .

**Table 4.1. Prediction success table for qualitative response models**

		Predicted Choice				Observed Count
		1	2	...	m	
Observed Choice	1	$N_{11}$	$N_{12}$	...	$N_{1m}$	$N_{1.}$
	2	$N_{21}$	$N_{22}$	...	$N_{2m}$	$N_{2.}$
	3	$N_{31}$	$N_{32}$	...	$N_{3m}$	$N_{3.}$
	$\vdots$					
	m	$N_{m1}$	$N_{m2}$	...	$N_{mm}$	$N_{m.}$
Predicted Count		$N_{.1}$	$N_{.2}$	...	$N_{.m}$	$N_{..}$

This index is non-negative with a maximum value of

$$1 - \sum_{i=1}^m \left( \frac{N_{.i}}{N_{..}} \right)^2$$

and can be normalized so as to have a maximum value of one.

The rationale behind the expression of PSI is as follows:  $N_{.i}/N_{..}$  is the proportion of sample individuals predicted to choose alternative  $i$ .  $N_{ii}/N_{.i}$  is the proportion of predictions for alternative  $i$  that were correct. For the  $i$ -th category, McFadden et al. expressed the success index  $PSI_i$  as

$$PSI_i = \frac{N_{ii}}{N_{.i}} - \frac{N_{.i}}{N_{..}}.$$

The expression of PSI is then the weighted average of  $PSI_i$  with weights  $N_{.i}/N_{..}$ . Clearly, the higher the value of PSI, the greater the predictive capability of the model.

### 4.3.5 Likelihood Ratio Test (LR)

This criterion, defined earlier as a procedure for testing the general linear hypothesis in binary choice models, is especially suitable for comparison among nested models, i.e., models for which its variables is a subset of the variables of the another. Based on this test, McFadden (1974) suggested a  $R^2$ -like norm called Pseudo -  $R^2$  and defined as

$$Pseudo - R^2 = 1 - \frac{\log L(\hat{\beta}_{ML})}{\log L(\hat{\beta}_o)}$$

where  $\log L(\hat{\beta}_{ML})$  is the value of the log-likelihood function evaluated at  $\hat{\beta}_{ML}$ , the ML estimator of  $\beta$  and  $\log L(\hat{\beta}_o)$  is the value of the log-likelihood function evaluated under the constraint that all coefficients except the constant term are zero. This measure is 1 when the model is a perfect predictor, i.e.,  $\hat{p}_i = 1$  when  $y_i = 1$  and  $\hat{p}_i = 0$  when  $y_i = 0$  and is 0 when  $\log L(\hat{\beta}_{ML}) = \log L(\hat{\beta}_o)$ . Between these limits the value of Pseudo-  $R^2$  has no obvious intuitive meaning. Using concepts from information theory, Hauser (1978) interpreted Pseudo-  $R^2$  as the percent of "uncertainty" in the data explained by the empirical results. His derivation however is not valid, since  $\underline{x}_i$  was treated as a random vector following a probability distribution  $p(\underline{x}_i)$ . Clearly, this is a violation of the basic assumption of qualitative response models theory which calls for  $\underline{x}_i$  to be a fixed vector of observations taken on the i-th individual.

Hensher and Johnson (1981) noted that values of Pseudo-  $R^2$  between 0.2 and 0.4 should be considered as extremely good fits so that the analyst should not be looking for values in excess of 0.9 as is often the case when using  $R^2$  in ordinary regression. They also proposed a measure similar to Pseudo-  $R^2$  adjusted for the model's degrees of freedom which, they claim, improves Pseudo-  $R^2$  as a model selection criterion. The adjusted Pseudo-  $R^2$  is given by

$$Adj. \text{ Pseudo- } R^2 = 1 - \frac{\log L(\hat{\beta}_{ML})/(m-1) - k}{\log L(\hat{\beta}_o)/m - 1}$$

where  $m$  refers to the number of alternatives faced by the individuals and  $k$  is the total number of parameters in the model.

#### 4.3.6 Akaike Information Criterion (AIC)

$$AIC = -\log L(\hat{\beta}_{ML}) + p$$

Akaike (1973) proposed this criterion to aid model selection. It is simple to calculate and most important, it accounts for the degrees of freedom available to the model. One is to select the model for which AIC is smallest. Amemiya (1980) suggests the use of AIC whenever the competing models are not nested and therefore, the likelihood ratio test does not apply.

#### 4.3.7 Theil's Information Inaccuracy of the Prediction

Theil (1967, 1970 and 1971) derived a measure of lack of fit which he called "the information inaccuracy of the forecast", using concepts from information theory. The basic ideas of this theory that are necessary to understand the reasoning of this criterion are summarized as follows.<sup>5</sup>

Let  $p$  be the probability that some event  $E$  will take place. Suppose that at some point of time a message comes which states that the event actually took place. The information content of this message as defined in information theory is equal to

$$I(p) = \log\left(\frac{1}{p}\right) = -\log(p).$$

---

<sup>5</sup> This discussion is based on Theil, 1967 and Gallagher, 1968.

This is a monotonically decreasing function of  $p$ , declining from  $\infty$  ( $p=0$ ) to zero ( $p=1$ ). The decreasing character is motivated by the consideration that the message is more informative when the event was less probable before the message comes in. The logarithmic form is useful because of the additive property of stochastically independent events: The information content of the message which states that the two independent events both occurred is equal to the sum of the information contents of the two separate messages, one dealing with the first event and the other with the second.

Consider a set of  $m$  events,  $E_j$  ( $j=1,2, \dots m$ ) with associated probabilities of occurrence  $p_j$ . It is said that these events form a complete system if it is certain that exactly one of them will occur, i.e.,  $\sum_{j=1}^m p_j = 1$ . When we receive a definite message stating that  $E_j$  has occurred, the information content of the message is, as before,  $I(p_j) = -\log(p_j)$ . Before the message is received we do not know, of course, how large this information content will be, since it may be any one of the numbers  $I(p_j)$  for  $j=1,2, \dots m$ . However, an expected information content, also known as entropy of the information system, can be calculated before the message comes in, as

$$H = \sum_{j=1}^m p_j I(p_j) = - \sum_{j=1}^m p_j \log(p_j)$$

It is obvious that  $H$  cannot be negative since all individual terms of the summation are non-negative. Using Lagrangian optimization, it can be shown (Theil, 1967) that  $H$  is maximized when all events are equally likely to occur, i.e.,  $p_1, p_2, \dots, p_m = 1/m$ . This is a natural result because when all events are equiprobable, the message which states what actually happened is expected to contain more information than in any situation where it is known that some events are more probable than others. Or equivalently, there is a maximum of uncertainty when all  $m$  possibilities are equal to  $1/m$ ; and the more uncertainty there is prior to the message the more information is expected to be delivered by the message. It is under this logic that uncertainty and expected information are

regarded as dual concepts by some authors (Shannon 1948, Shannon and Weaver 1949, Goldman 1953, Attneave 1959 and Theil 1967). Thus, it has been established that

$$0 \leq H \leq \log m.$$

Suppose now that instead of receiving a definite message stating what event actually occurred, a message is received which implies that the event probabilities have changed so that some events become more probable and some less probable. That is, the message transforms the prior probabilities  $p_j$  into the posterior probabilities  $q_j$  where, again,  $\sum_{j=1}^m q_j = 1$ . Considering one particular event,  $E_j$  with prior and posterior probabilities  $p_j$  and  $q_j$  respectively, the information content of the message is defined as the difference

$$I(q_j; p_j) = I(p_j) - I(q_j) = \log \left( \frac{q_j}{p_j} \right).$$

However, such an indirect message is not restricted to one event  $E_j$  but it states that each  $E_j$  ( $j = 1, 2, \dots, m$ ) has its own posterior probability  $q_j$ . Taking the expectation over the separate information contents we find that

$$EI(q; p) = \sum_{j=1}^m q_j \log \left( \frac{q_j}{p_j} \right)$$

is the expected information content of the indirect message. Theil (1967) shows that  $EI(q; p)$  is always positive except when  $q_j = p_j$  for all  $j$ , in which case the value of  $EI(q; p)$  is zero. This is not surprising considering that the expected information content of an indirect message vanishes when it leaves all prior probabilities unchanged. Notice also, that  $EI(q; p)$  may be infinitely large when  $q_j > p_j = 0$  for some  $j$ . The prior probability then specifies that  $E_j$  has zero probability of occurrence, but this is raised to a positive value by the message. Theil (1967) interprets this as an increase in



the prior probability by a factor of “infinity” and that the analyst would be also “infinitely surprised” by the message.

Theil (1967, 1971) used the concept of expected information of an indirect message to derive a “natural” index for comparing the observed sample proportion of each alternative with the corresponding predicted share. More specifically, he assumed that the model predictions  $p_1, p_2, \dots, p_m$  are available before the realizations  $q_1, q_2, \dots, q_m$  are available, so that he could reasonably argue that  $p_j$  are the prior probabilities which the message transforms into the observed (posterior) probabilities  $q_j$ . When the message has zero expected information, then  $q_j = p_j$  and hence predicts perfectly. When its expected information is very small, the predictions are accurate although not perfect. When the expected information is large enough so that at least some of the  $p_j$ 's differ substantially from the corresponding  $q_j$ 's, the predictions as a whole are very inaccurate. Therefore, he called

$$EI(q; p) = \sum_{j=1}^m q_j \log \left( \frac{q_j}{p_j} \right)$$

the information inaccuracy of the forecasts  $p_j$  with respect to the observed  $q_j$  and regarded it as a measure of the model's lack of fit.

#### 4.3.8 Discussion

A number of scalar model selection criteria have been presented above with brief comments on each. It should be emphasized that a blind reliance on any one of these criteria is, of course, ill-advised. During the process of model selection, there are numerous theoretic and statistical factors that must be taken under consideration and certainly, no scalar criterion can accomodate all of

them. Every criterion is optimal in its own merits, i.e., it evaluates only a limited aspect of a model's performance. Thus, there are criteria such as SSR, WSSR and Theil's information inaccuracy of the prediction, which focus on the model's quality of fit, others like WP and PSI which evaluate the predictive ability of the model and finally, there are also criteria like LR and AIC which are related to a test of the hypothesis that one or more of the regression coefficients are zero. Directly related to the latter criteria is Pseudo- $R^2$  which attempts to assess the proportion of the variance of the dependent variable explained by the independent variables.

One should not expect a single criterion to be optimal for every occasion. It is our suggestion that the analyst should select three or four criteria and then compare the results. Our preference lies with criteria such as WSSR, AIC, PSI and Theil's information inaccuracy of the prediction because they are based on a sound theoretical justification and, together, they consider all aspects of a model's performance.

## ***4.4 Data Splitting and Model Validation***

The model selection criteria discussed in the previous section, although they are meaningful measures of a model's goodness of fit, they do not assess the model's actual predictive ability. An established practice in regression analysis is to combine the notion of "selection of best model" with model validation. The latter suggests a search for a type of model checking against independent data, i.e., evaluation of each candidate model by predicting response values that are independent of the data which built the model. The concept of model validation is discussed in detail by Snee (1977) and Montgomery and Peck (1982). Stone (1974) gives a brief history of the development of model validation ideas.

Data splitting is a model validation method. It refers to the partitioning of the original data set into two subsets, a fitting or estimation data set and a validation or prediction data set. The former is used for estimating the model and the latter to apply the fitted model in order to derive response values which can then be meaningfully compared to the true values. Then, norms such as WP, PSI or Theil's information inaccuracy of the prediction can be used to determine the best predicting model. Once a model is selected, it must be fitted again, using the original, full, data set making use of all available information.

Even though data splitting is motivated by a great deal of common sense, the actual methodology for subdividing the data deserves special attention. In general, any procedure that determines how the data set is to be partitioned, should be designed according to the particular application (Myers, 1986). In some cases, the criterion for data splitting is obvious. For example, if data are collected sequentially in time, it seems reasonable to pick a point in time to divide the data into two subsets. Normally, the most recent observations are assigned to the validation data set so that valuable information regarding the model's forecasting ability is obtained. There are many circumstances however, where no obvious variables, such as time, exist to serve as a basis to split the data. For situations of this type, Snee (1977) recommends the use of the DUPLEX algorithm, developed by R. W. Kennard. The objective of this algorithm is to divide the data into two sets which cover approximately the same region and exhibit similar statistical properties. Snee developed a rule to determine whether the two subsets are similar, which is based on the determinant of the  $P'P$  matrix,  $|P'P|$ , where  $P$  is the design or data matrix of the explanatory variables, with its columns standardized and orthonormalized. According to Kennard and Stone (1969) and Weisberg (1985),  $|P'P|$  is directly related to the volume of the smallest ellipsoid that contains all data points and also, it is a useful scalar measure of the statistical properties of the corresponding data set.

Using  $|P'P|$ , Snee considered the quantity

$$\left[ \frac{|P'P| \text{ for the estimation data set}}{|P'P| \text{ for the validation data set}} \right]^{1/p}$$

where  $p$  is the number of variables in the data matrix, as a measure of the relative statistical properties and the region covered by the estimation and validation data sets.

The DUPLEX algorithm, as applied by Snee, consists of the following steps:

- Step 1** All data points are standardized and orthonormalized. A  $n \times n$  triangular matrix having as elements the Euclidean distances between the row and column data points is constructed.
- Step 2** The two data points which are farthest apart are assigned to the estimation data set.
- Step 3** The two of the remaining data points which are farthest apart are assigned to the validation data set.
- Step 4** The point with the largest distance from the points of step 2 is assigned to the estimation data set.
- Step 5** The point with largest distance from the points of step 3 is assigned to the validation data set.

Steps 4 and 5 are repeated until all data points are assigned to one of the two data sets.

A half and half partition appears to be the most popular data splitting strategy. However, Snee (1977) recommends that one not to try to split a data set in half unless  $n$  (total sample size) is greater than  $2p + 25$  in order to provide an adequate number of degrees of freedom for model estimation.

# Chapter V

## Outlier and Influence Diagnostics

### *5.1 Introduction*

The maximum likelihood method of fitting qualitative response models exhibits certain optimality properties if all pertinent model assumptions hold true. When however some of the assumptions are violated, the maximum likelihood fit may be badly affected (Hoaglin, Mosteller and Tukey, 1983). The lack of resistance of the maximum likelihood fit to violations of assumptions is most effectively mirrored by its sensitivity to outlying responses and, also, to data points that are extreme in the design space (high leverage observations) (Pregibon, 1981). Consequently, it is of vital importance to the analyst to be able to detect and assess the degree of discrepancy between the model assumed and the data observed.

In classical linear regression analysis this task is usually accomplished by computing measures, known as diagnostics, which aid in highlighting data points that may reflect violation of assumptions (outliers), exert undue influence on regression statistics (high leverage observations) or do both

of the above (high influence observations). The related literature is extensive (see, for example, Belsley, Kuh and Welsch, 1980, Draper and Smith, 1981, Cook and Weisberg, 1982 and Myers, 1986) and most of the standard regression computer packages output diagnostic information about both the identification and assessment of the extent of influence for each data point.

On the contrary, there is very little in the literature with respect to diagnostic measures for qualitative response models mainly due to the following two reasons. First, the area of qualitative response models has not yet been fully covered and most of the research effort is still directed toward improved model specifications and error structures rather than diagnostic development. Second, qualitative response models only recently have been applied to data obtained in observational studies. In contrast to controlled experimentation, data from observational studies are much more likely to include outlying responses and/or extreme data points, thus, the need for diagnostic development emerged much later than the initial applications of qualitative response models.

As a result of the above, only a limited number of studies on qualitative response model diagnostics appear in the literature and most of them consider the case of dichotomous logistic regression analysis. The two most important contributions to this matter are due to Pregibon (1981) and Cook and Weisberg (1982). Both studies are based on Walker and Duncan (1960) iterative weighted least squares (NLWLS, see section 2.2.3) interpretation of the method of scoring algorithm for maximum likelihood estimation. The proposed diagnostics are simple extensions of existing linear regression diagnostics with minor modifications in order to accommodate the heteroscedastic nature of binary choice models and the iterative fashion of the NLWLS procedure. Perhaps, the absence of diagnostic tools for multinomial choice models (ordered or unordered) may be in part due to the fact that no straightforward least squares analogy similar to NLWLS exists for such models.

In the following sections a number of diagnostics for the detection of outliers, high leverage points and high influence points are described as they have been proposed by Pregibon (1981) and Cook and Weisberg (1982).

## 5.2 *Diagnostics for the detection of outliers*

In most applications of regression analysis (linear or non-linear), residual analysis is customary in order to shed light on possible violations of assumptions. In particular, the outlier analysis is a methodology specifically designed to highlight data points that appear not to follow the proposed model fit to the balance of the data (Myers, 1986). The conditions detected by outlier diagnostics are errors in the y-direction, i.e., model shifts that produce anomalies in the measured response. The symptom is a residual which is larger than would realistically be produced by chance.

The influence of outlying observations in the least squares fit of linear regression models is meaningfully described by Myers (1986) as "the "fallout" that results from the regression being pulled toward the errant measured response". This same type of influence, maybe stronger, occurs also to models fitted by the maximum likelihood method. Pregibon (1981) and Cook and Weisberg (1982) suggest that the two most useful expressions of residuals are:

- the standardized residuals

$$se_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1 - \hat{y}_i)}}$$

Note the  $se_i^2$  are the individual components of WSSR (see section 4.3.3), i.e.,  $WSSR = \sum_{i=1}^n se_i^2$ .

- the individual components ( square root ) of  $-2\log L$ ,

$$d_i = -\sqrt{2} \sqrt{\hat{\beta}' x_i y_i - \log[1 + \exp(\hat{\beta}' x_i)]}$$

Note that  $d_i$  is defined for all values of  $y_i$  and also,  $-2\log L = \sum_{i=1}^n d_i^2$ .

Since both WSSR and  $-2\log L$  are measures of the goodness of fit of the model, large  $se_i$  and  $d_i$  indicate observations poorly accounted for by the model. Suspect points are ordinarily set aside for further checking in order to determine whether these observations were mistakenly recorded during the data taking process. The concern is that an erroneous observation will exert an undue amount of influence on the regression results, influence which is counterproductive. On the other hand, if the data point does reveal a model deficiency, then it is quite likely that it has arisen from an unusual combination of circumstances which may be of great interest.

### ***5.3 Diagnostics for High Leverage and Influence***

#### ***Observations***

The fact that an observation provides an outlier does not necessarily mean that the observation is influential in fitting the chosen model. Indeed, not all high influence observations are due to errors in the y-direction. Influence can also occur when a single observation is extreme in the x-direction, i.e., it lies at a disproportionate distance from the data center, even though it is a proper observation and does not necessarily represent evidence of model fallacy. Such data points, which are extreme in the x-direction or design space are frequently known as high leverage observations. High leverage observations do not always have a negative impact on the fitted model. For example, if a data point lies far apart from the body of the data but it follows the trend suggested by the majority of observations, then this remote data point actually reinforces the fitted model and enhances its performance.

In linear regression, the diagnostic that provides information regarding what data points exert high leverage is the diagonal element of the HAT matrix which is defined as

$$H = X(X'X)^{-1}X'$$



where  $X$  is the data or design matrix. The HAT diagonals are the quadratic forms

$$h_{ii} = \underline{x}_i'(X'X)^{-1}\underline{x}_i$$

where  $\underline{x}_i$  is the vector of measurements in the independent variables at the  $i$ th data point. Large values of  $h_{ii}$  indicate observations that are extreme in the design space ( $x$ -direction) and thus, they have the potential for exerting undue influence on at least one regression coefficient (Hoaglin and Welsch, 1978).

In order to account for the heteroscedastic error structure of binary logistic regression models, Pregibon (1981) employed the generalized least squares version of the HAT matrix,  $H^*$  (Pregibon, 1980) which can be expressed as

$$H^* = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$$

where  $W$  is the variance-covariance matrix of the error terms in the model. In binary logistic regression models,  $W$  is a diagonal matrix with elements  $w_{ii} = \hat{y}_i(1 - \hat{y}_i)$ , corresponding to the  $i$ th observation (see also the discussion on NLWLS procedure, section 2.2.3). Large values of  $h_{ii}^*$ , the diagonal elements of  $H^*$ , are useful in detecting high leverage observations.

Draper and Smith (1982) view the influence of a data point as the offspring of a collaboration between the leverage and the nature of the fit of the model to the point in question (residual). In this context, the combination that produces the greatest influence is a data point that contains a large HAT diagonal accompanied by a relatively large residual. Consequently, the joint examination of  $se_i$ ,  $d_i$ , and  $h_{ii}^*$  will call attention to observations that i) are not well explained by the model and/or ii) are dominating some aspect of the fit. Pregibon (1981) suggests that index plots of these quantities, i.e., plots of  $se_i$  vs  $i$ ,  $d_i$  vs  $i$  and  $h_{ii}^*$  vs  $i$ , are most useful for identifying these observations.

## 5.4 Diagnostics for Coefficient Sensitivity

The  $se_i$ ,  $d_i$  and  $h_{ii}^*$  reveal which individual observations have potential for exerting excessive influence. The quantities however, do not allow easy diagnosis of which model components are influenced and to what degree. In linear regression analysis, this type of information is provided by diagnostics that measure the changes that would occur in the coefficients if each observation were deleted from the data set but without actually removing it from the data. The development of these diagnostics is described in detail by Belsley, Kuh and Welsch (1980). Pregibon (1981) extended this methodology to the case of binary logistic regression models. In particular, he re-expressed the log-likelihood function of the binary logit model by including an indicator variable  $v_i$  which takes the values 1 or 0 according to the presence or absence of the  $i$ -th observation in the estimation data set. The augmented log-likelihood function,  $\log L^*$  is written as

$$\log L^* = \hat{\underline{\beta}}' \sum_{i=1}^n v_i \underline{x}_i y_i - \sum_{i=1}^n v_i \log[1 + \exp(\hat{\underline{\beta}}' \underline{x}_i)]$$

The maximum likelihood estimator of  $\underline{\beta}$  will now be a function of  $v_i$  and can be obtained by maximizing  $\log L^*$ . Although this approach for model perturbation is intuitively reasonable, it is also computationally timely and costly to carry out since for each observation considered, a new maximum likelihood estimate must be derived. Pregibon (1981) explains that the analyst's main concern lies with the detection of strong individual effects with a minimal effort. Small changes in the coefficients are not important whereas for large changes it is usually sufficient to know their direction and relative magnitude with respect to other observations' changes. Therefore, the derivation of highly precise estimators may be unnecessary. Under this logic, Pregibon (1981) suggested the use of Chamber's (1973) one-step estimator for  $\underline{\beta}$  which is, simply, the estimator of  $\underline{\beta}$  obtained by ter-

minating the maximum likelihood estimation procedure after the first iteration. Cook and Weisberg (1982) express the one-step estimator as

$$\hat{\underline{\beta}}_{(i)}^1 = \hat{\underline{\beta}} - \frac{(X'WX)^{-1} \underline{x}_i(y_i - \hat{y}_i)}{1 - h_{ii}^*}$$

where,

$\hat{\underline{\beta}}_{(i)}^1$  is the one-step estimator obtained by excluding the  $i$ -th observation from the estimation data set and

$\hat{\underline{\beta}}$  is the full data set maximum likelihood (fully iterated) estimator of  $\underline{\beta}$ .

The diagnostic for individual coefficient sensitivity is then,

$$\Delta_i \hat{\underline{\beta}}' = \frac{(X'WX)^{-1} \underline{x}_i(y_i - \hat{y}_i)}{1 - h_{ii}^*}.$$

Pregibon (1981) discussed the accuracy of this one-step approximation and concluded that although it tends to underestimate the fully iterated value, it still clearly highlights influential observations. He also found that index plots of  $\Delta_i \hat{\underline{\beta}}' / s.e.(\hat{\underline{\beta}})$  are generally useful in detecting observations that are causing instability.

The role of  $\Delta_i \hat{\underline{\beta}}^1$  is to ascertain which observations influence specific regression coefficient estimates. However, it is possible for changes in some coefficients to be offset by changes in other coefficients in such a way that the fitted values change very little. In addition, when a large number of explanatory variables are included in the model, it becomes difficult to determine if an observation is unduly influencing the fit by examining sensitivity plots for each element of the parameter vector. Thus, it is useful to consider a single, overall measure of the influence that each data point exerts on the set of coefficients. Pregibon (1981) adapted Cook's D, an overall influence measure for linear regression models due to Cook (1977). More specifically, he expressed the boundary of an asymptotic confidence region for the parameter vector  $\underline{\beta}$  as

$$-2[\log L(\underline{\beta}' \underline{x}_i, y) - \log L(\hat{\underline{\beta}}' \underline{x}_i, y)] = c$$

and replaced  $\underline{\beta}$  by  $\hat{\underline{\beta}}_{(i)}$ , the maximum likelihood estimator of  $\underline{\beta}$  obtained by deleting the  $i$ -th observation from the estimation data set. The resulting scalar,  $c_i$ , is then a measure of the influence of the  $i$ -th data point on the estimated coefficient vector  $\hat{\underline{\beta}}$ . Using the one-step approximation to  $\hat{\underline{\beta}}_{(i)}$ ,  $\hat{\underline{\beta}}_{(i)}^1 c_i$  becomes

$$c_i^1 = \frac{se_i^2 h_{ii}^*}{1 - h_{ii}^{*2}}$$

a quantity which is very easy to calculate. Again, index plots of  $c_i^1$  are strongly recommended.

# **Chapter VI**

## **Merchantability Models for Loblolly Pine**

### ***6.1 Introduction***

In this section we illustrate the applicability of qualitative response models to forestry by focusing on the problem of modeling the merchantability of loblolly pine trees growing in thinned and unthinned, cutover, site-prepared plantations. Models of this type, generally known as merchantability models, are used to i) determine stand and tree characteristics that greatly influence tree merchantability, i.e., tree quality as judged by the class of wood product the tree produces and ii) predict the probability that an individual tree will produce a certain product (e.g. pulpwood, sawtimber or peelers) given certain stand and tree characteristics.

The information obtained by merchantability models, especially when they are incorporated into individual tree growth and yield simulation systems or into diameter distribution based yield prediction systems, is considered to be extremely valuable in management decision making. To better appreciate the utility of such models, recall, for example, that a diameter distribution system pre-

dicts the number of surviving trees per unit area by dbh class and these data are then used in conjunction with an individual tree volume equation to estimate the yield per unit area by dbh class. Because the tree volume equation is normally constructed by measurements taken from sampled trees of merchantable size throughout the study area, it is implicitly assumed that the predicted tree frequencies by dbh class refer to merchantable trees and so does the predicted yield. Even though this assumption is considered to be sufficient for many purposes, the yield per unit area is clearly overestimated since not all surviving trees are merchantable, a fact which is especially true for smaller dbh classes. Evidently, such optimistic yield predictions may seriously impact management decisions. However, the yield over-prediction can be corrected, to a certain degree, by the use of a merchantability model which will predict the probability of a tree being merchantable. In this way, the number of merchantable trees per acre will be a fraction of the initially predicted number of surviving trees per acre and managers will be furnished with more realistic information.

Merchantability models can prove very useful in assessing the effect of thinning operations in terms of the quality of end-products produced by the stand. For instance, decisions about thinning operations are currently based on growth and yield estimates that account for the effect of thinning on the stand diameter distribution only, and not on the quality of end products produced by the stand. However, it is precisely this type of information which, combined with information on stand diameter distribution, will greatly contribute to more realistic economic analysis about thinning necessity and intensity.

Strub et al. (1986) considered a non-linear discrete regression model (other than logit or probit) to classify loblolly pine trees according to suitability for sawtimber. Using graphical techniques the authors identified tree dbh and average height of dominant and codominant trees as important predictor variables in the presence of which, stand age and number of surviving trees per acre were found to be non-significant. Their data base, consisting of measurements taken on old-field plantation loblolly pine trees from the states of Virginia, Delaware, Maryland and North Carolina contained only observations from unthinned stands; therefore, no thinning effect could be assessed. In addition, only two product classes were considered, namely pulpwood and sawtimber. Even

though this type of classification is sufficient for many purposes, for growth and yield prediction systems a more detailed classification may be in order if meaningful comparisons among alternative management regimes are to be made.

Burkhart and Bredenkamp (1989) pooled the data from two successive enumerations on unthinned, lightly thinned and heavily thinned loblolly pine plantations, to develop equations to estimate the proportions of trees by dbh class that qualify as peelers, sawtimber or pulpwood. After grouping the data into half-inch dbh classes, the authors fit a modified Chapman-Richards equation, with the midpoint of each dbh class as the independent variable, to each thinning treatment separately in order to estimate i) the proportion of solid wood products (peelers and sawtimber) by dbh class and ii) the proportion of peelers by dbh class. The proportion of pulpwood was then obtained as the complement of the proportion of solid wood products (all proportions sum to one) and the proportion of sawtimber by subtracting the proportion of peelers from that of solid wood products. They concluded that i) total tree height, stand density and site quality had negligible influence on the probability of a tree being allocated to a particular product class after the effect of dbh was removed, ii) the proportion of peelers by dbh class increased with thinning but no significant difference was detected between light and heavy thinning and iii) the proportion of solid wood products by dbh class remained relatively constant regardless of thinning intensity. Because the estimated equation for peelers exhibited illogical crossing for thinned and unthinned stands the authors constrained the shape parameter of the Chapman-Richards equation to be equal for both thinned and unthinned stands.

The results derived by this study are somewhat unexpected. If significant thinning effect is present for the estimation of peelers one would expect the same to be true for the estimation of solid wood products (peelers and sawtimber). Also, it is implied that the proportion of sawtimber is affected by thinning in a counter active manner with respect to the effect on the proportion of peelers, so that the overall solid wood proportion by dbh class remains unaffected by thinning. In addition, the absence of significant thinning effect for the estimation of the proportion of solid wood products implies that the estimation of the proportion of pulpwood is also free of thinning effect.

Grouping the observations into dbh classes always results in a loss of information. Such a loss may be responsible for small power in testing for significant thinning effects. Another reason may be that the Chapman-Richards equations are not quite appropriate for modeling the merchantability of loblolly pine trees. The illogical crossing of peelers equations for thinned and unthinned stands seems to support such a possibility.

It is believed that qualitative response models theory exhibits certain characteristics that make it very attractive for modelling tree merchantability. First, it does not require grouping of the observations into dbh classes thus avoiding any loss of information. Second, the probabilities of all product classes are estimated simultaneously under the same model and not separately. Third, the levels of thinning intensities enter the model as dummy variables thus facilitating all pertinent significance tests. Finally, variable screening can be performed in a routine manner in order to identify important stand and tree characteristics that significantly contribute to the estimation of tree merchantability.

The present study investigates the applicability of qualitative response models as alternative formulations for modelling tree merchantability and assessing the effect of thinning operations to the quality of end-products of forest stands. More specifically, this study utilized the data Burkhart and Bredenkamp (1989) used, to i) calculate probit and logit estimates for the probability that a loblolly pine tree will produce solid wood products, ii) fit a multinomial logit model with an unordered trichotomous response variable to estimate the probabilities of a tree being classified as peelers, sawtimber or pulpwood and iii) fit a multinomial logit model with an ordered trichotomous response variable by making use of the natural ordering of the three product classes as defined in terms of minimum dbh requirements. Finally, the results derived by this study were compared with those derived by Burkhart and Bredenkamp (1989) to see if any discrepancies arise and appropriate conclusions were drawn.



## 6.2 *Data*

The data used in this study were tree measurements taken on plots from 173 cutover, site-prepared plantations throughout the Piedmont and Coastal Plain physiographic regions in the southeastern United States. At each plantation three plots, similar in density and site quality were established. Then, no thin, light thin (approximately 1/3 of the basal area removed) and heavy thin (approximately 1/2 of the basal area removed) treatments were randomly assigned to the plots.

Thinnings, both light and heavy, were from below. Occasionally, entire rows were removed in order to provide access to remaining trees. The maximum permissible row removal was 1 row in 5. In selecting trees for removal, a trade-off between quality and spacing was required. Trees that were forked, leaning, diseased and that would not leave an excessive opening in the canopy if removed, were to be cut regardless of size. On the other hand, in order to avoid large openings in the canopy, smaller trees of inferior quality had to remain in the stand. Despite the detailed thinning instructions, the definition of thinning is by nature subjective; therefore, it was interpreted differently by different operators. In some cases, the light thinning by one operator was in fact heavier than the heavy thinning conducted by another operator.

Four site preparation classes were defined to summarize the wide variety of site preparation treatments:

- Site preparation class 1: Tilled; debris moved
- Site preparation class 2: Tilled; debris not moved
- Site preparation class 3: Not tilled; debris moved
- Site preparation class 4: Not tilled; debris not moved

where tilled is either disked or bedded or both.

Plot characteristics recorded were: physiographic region (PR), level of thinning intensity (TI), site preparation class (SP), age defined as number of years since planting, average height of dominant and codominant trees (HD) in feet, basal area per acre (BA) in square feet/acre and number of surviving trees per acre (N). In addition, the quadratic mean diameter (QMD) was computed from the observations on BA and N and also the site index (SI) according to a model developed by Amateis and Burkhart (1985),

$$\ln(HD) = \ln(SI) (25/A)^{-0.02205} \exp(-2.83285 (A^{-1} - 25^{-1}))$$

with base age 25 years.

Each tree in a plot was measured for dbh to the nearest tenth of an inch, total height (TH) to the nearest foot and height to the base of live crown (HTCR) in feet. The crown ratio (CR) was also computed as

$$CR = 1 - \frac{HTCR}{TH}.$$

Each tree was visually inspected for assessing its stem quality. More specifically, each tree was first classified as "Single-Stemmed" or "Forked", then as "Normal Top" or "Broken Top", and the bole was inspected and put into one of four categories, "Straight", "Bole Sweep", "Butt Sweep" and "Short Crook". Each tree was also checked as having "No Disease or Insect Damage" or "Disease or Insect Damage". Finally, each tree was classified as peelers, sawtimber or pulpwood according to the following criteria (Burkhart and Bredenkamp, 1989).

**Peelers**        minimum dbh 10.6 inches with a minimum of two peeler bolts, each 8 feet 7 inches long with a minimum 8.6 inside bark diameter at small end.

**Sawtimber** chip-n-saw material; not qualified for peelers; minimum dbh 7.6 inches with a minimum of one 16-foot sawlog (including a 4-inch trim allowance) to a minimum 6 inch inside bark diameter at the small end.

**Pulpwood** material not qualified for peelers or sawtimber; minimum dbh 4.6 inches.

In the present study, only trees with dbh greater than or equal to 7.6 inches were used. Additional details concerning the data can be found in Burkhart et al (1985) and Burkhart (1987). Table 6.1 displays summary statistics for the stand and tree variables considered in this study.

Due to the plethora of available observations, data-splitting is ideal for model validation. The DUPLEX algorithm (Snee, 1977, see section 4.4) has been applied to partition each of the three data sets, unthinned, lightly thinned and heavily thinned, into two, equal size data sets, one for model fitting and one for model validation.

According to the merchantability criteria defined above, trees with dbh between 7.6 and 10.5 inches produce either pulpwood or sawtimber whereas trees with dbh larger than or equal to 10.6 inches can produce any one of the three product classes. Consequently, two different types of logistic regression models were considered for modelling the merchantability of loblolly pine trees. One with a dichotomous response, pulpwood or solid wood (sawtimber or peelers) and one with a trichotomous response, pulpwood, sawtimber and peelers.

### ***6.3 Two-Product Logit Model***

To estimate the probability that a loblolly pine tree produces solid wood products such as sawtimber or peelers, a binary logit model was fitted separately to the data from the unthinned, lightly thinned and heavily thinned plots. Starting with all available plot and tree variables, in-

Table 6.1. Summary statistics for the 173 thinned, lightly thinned and heavily thinned plots.

stand variables	unthinned			lightly thinned			heavily thinned		
	min.	mean	max.	min.	mean	max.	min.	mean	max.
AGE	14.00	22.55	31.00	14.00	22.23	31.00	14.00	21.96	31.00
BA	24.62	145.87	205.20	30.71	113.82	193.56	18.97	95.79	145.22
N	90.00	475.79	935.00	89.00	314.40	633.00	44.00	244.93	479.00
QMD	4.83	7.62	10.63	5.48	8.22	11.19	5.35	8.55	11.92
SINDEX	43.38	61.07	80.11	37.93	61.14	81.57	37.40	60.59	82.85
tree variables									
DBH	7.60	8.92	15.30	7.60	9.02	18.40	7.60	9.11	17.20
TH	27.00	58.03	89.00	32.00	57.18	90.00	28.00	56.22	93.00
HTCR	14.00	35.84	71.00	11.00	33.82	67.00	10.00	32.39	65.00
CR	0.09	0.39	0.70	0.15	0.42	0.77	0.08	0.43	0.78
Number of trees	2656			5604			5223		

cluding some transformations and interactions, variable screening was based on the model selection and validation criteria discussed in Chapter 4. The final models, which were found to perform best in terms of goodness-of-fit and prediction are presented on Table 6.2.

To test whether thinning treatments have a significant effect on the merchantability of loblolly pine trees, one thinning indicator variable was added to the model which was then fitted to the pooled data of all three thinning treatments. Both likelihood ratio and Wald's tests indicated the presence of significant thinning effect (p-values were 0.0004 for the LR test and 0.0002 for the Wald's test). A similar procedure, testing for significant differences between light and heavy thinning, failed to reject the null hypothesis (p-values were 0.9333 for LR test and 0.9564 for Wald's test).

Given the above conclusions, the data from lightly thinned and heavily thinned stands were pooled into one "thinned" data set. Again, variable screening was conducted and the finally selected model is shown on Table 6.3 together with the model for unthinned stands. Tables 6.3a and 6.3b present the variance-covariance matrices for the two models and Table 6.4 displays the corresponding goodness-of-fit and validation statistics. Influence diagnostics computed for both models did not indicate any particularly strong influential data points.

As seen from Table 6.3 the stand characteristics that appear to play an important role in estimating the merchantability of loblolly pine trees are: stand age, basal area per acre, number of trees per acre, type of site preparation and the physiographic region where the stand is located. The absence of site index as a measure of site quality, although at first surprising, is probably because similar information, and perhaps more valuable, is provided by individual tree characteristics such as height to the base of live crown and crown ratio. Note that Burkhart and Bredenkamp (1989) in their study, also found the site index to be non-significant.

An interesting feature of the models is the presence of three measures of stand density namely, basal area per acre, number of trees per acre and quadratic mean diameter. Since the latter is a function of both basal area and number of trees per acre, it would seem that the inclusion to the model of

Table 6.2 Maximum likelihood coefficient estimates of the two-product logit models fitted to unthinned, lightly thinned and heavily thinned stands

predictor variables	unthinned		lightly thinned		heavily thinned	
	estimate	std. error	estimate	std. error	estimate	std. error
INTERCEPT	-19.007584	1.997119	-16.686243	1.587293	-18.892145	1.759692
AGE	0.057412	0.019465	0.096253	0.014026	0.072865	0.014990
BA	-0.033262	0.006403	-0.018923	0.006654	-0.051487	0.008988
N	0.007550	0.001847	0.005124	0.002234	0.015237	0.003196
QMD	0.904006	0.235186	0.339124	0.182762	0.811011	0.202767
SP1	0.583125	0.163212	0.208235	0.106599	0.450387	0.116977
PR	0.299023	0.119062	0.283251	0.079330	0.172114	0.084395
DBH	0.810235	0.072350	0.806231	0.049453	0.778001	0.051820
HTCR	0.069124	0.015561	0.072014	0.011405	0.721359	0.012353
CR	6.165230	1.269427	6.965023	0.873878	6.465001	0.940389
STRAIGHT	0.429065	0.154404	0.814045	0.114761	0.737852	0.121501
BOLE	-0.980124	0.164169	-0.768125	0.121104	-0.989231	0.129282
BUTT	-1.928236	0.469714	-0.755036	0.308657	-0.973124	0.348255
DOM	1.466215	0.305753	0.648129	0.076376	0.601023	0.078187
FORK	-0.453843	0.219391	-0.492235	0.178411	-0.597365	0.220971
DISEASE	-1.122642	0.141283	-0.868325	0.099920	-1.215300	0.106228

STRAIGHT = 1 if stem form is "straight"; 0 otherwise.

BOLE = 1 if stem form is "bole sweep"; 0 otherwise.

BUTT = 1 if stem form is "butt sweep"; 0 otherwise.

DOM = 1 if tree is dominant or codominant; 0 otherwise.

FORK = 1 if stem is forked; 0 otherwise.

DISEASE = 1 if tree is diseased; 0 otherwise.

SP1 = 1 if site preparation class is 1; 0 otherwise;

PR = 1 physiographic region is Piedmont; 0 if Coastal Plain.

**Table 6.3. Maximum likelihood coefficient estimates of the two-product logit models fitted to unthinned and thinned data.**

predictor variables	Unthinned		Thinned	
	estimate	std. error	estimate	std. error
INTERCEPT	-19.007451	1.997119	-16.474120	1.095629
AGE	0.058014	0.019465	0.084129	0.010239
BA	-0.033521	0.006403	-0.031003	0.005519
N	0.007121	0.001847	0.009121	0.001956
QMD	0.904172	0.235186	0.358003	0.124572
SP1	0.583001	0.163212	0.327124	0.078461
PR	0.295111	0.119062	0.192358	0.057530
DBH	0.810298	0.072345	0.811214	0.035527
HTCR	0.069123	0.015561	0.079114	0.008157
CR	6.165042	1.269428	7.149128	0.635875
STRAIGHT	0.429111	0.154404	0.779541	0.083066
BOLE	-0.981224	0.164169	-0.877123	0.088134
BUTT	-1.928566	0.469714	-0.508158	0.224241
DOM	1.465879	0.305753	0.590224	0.053636
FORK	-0.453753	0.219391	-0.539354	0.138457
DISEASE	-1.122653	0.141283	-1.028543	0.072267

Table 6.3a. Covariance matrix of the coefficient estimates of the binary logit model fitted to unthinned data.

	INTERCEPT	AGE	BA	N	QMD	SPI	PR
INTERCEPT	3.98848443						
AGE	4.4949E-03	3.7889E-04					
BA	5.0537E-03	1.4014E-05	4.0998E-05				
N	-1.7027E-03	-2.3746E-06	-9.8402E-06	3.4114E-06			
QMD	-1.3195E-01	-1.8033E-04	-7.3368E-04	2.3582E-04	5.5312E-02		
SPI	-8.9303E-03	5.4861E-04	-4.6361E-05	1.1725E-05	6.5796E-04	2.6638E-02	
PR	-2.4224E-03	-7.9479E-04	6.9128E-05	-1.2239E-05	-9.1200E-04	-5.6503E-03	1.4176E-02
DBH	-2.7793E-02	-2.1185E-04	-8.7675E-06	3.3849E-06	-5.8007E-04	4.6560E-04	5.4431E-04
HTCR	-5.1286E-03	-8.1573E-05	-3.4656E-06	-1.8491E-06	-5.5121E-04	-1.8148E-04	3.6316E-04
CR	-7.0109E-01	3.4630E-03	2.8662E-04	-1.9393E-04	-3.3421E-02	-1.2470E-02	3.6922E-03
STRAIGHT	-3.0072E-02	6.8828E-05	-5.0947E-06	4.4206E-06	-3.2965E-04	-2.2287E-04	1.0096E-03
BOLE	-6.3567E-03	-4.7544E-05	3.3576E-05	-4.6197E-06	-1.2204E-04	5.6056E-04	1.9525E-04
BUTT	-5.5665E-02	1.9543E-04	-1.2889E-05	6.5267E-06	1.1668E-03	1.7436E-03	1.3261E-03
DOM	-1.5893E-02	2.7657E-04	-4.5288E-05	1.3952E-05	3.3137E-03	-5.1100E-04	1.0905E-04
FORK	-1.4870E-02	6.2480E-05	-1.0164E-05	1.9952E-06	-1.8578E-05	-2.2693E-03	1.1601E-03
DISEASE	2.6507E-04	7.2873E-05	-4.7557E-05	1.6747E-05	1.7813E-03	-5.7136E-03	3.8821E-04
	DBH	HTCR	CR	STRAIGHT	BOLE	BUTT	DOM
DBH	5.2338E-03						
HTCR	-2.5030E-04	2.4214E-04					
CR	-3.0616E-02	1.5938E-02	1.26942804				
STRAIGHT	-9.8446E-04	-4.9243E-05	-3.4384E-03	2.3841E-02			
BOLE	-5.9598E-04	-1.5224E-04	-7.9517E-03	2.0006E-02	2.6951E-02		
BUTT	-6.1394E-04	-1.3908E-04	-6.4075E-03	2.0302E-02	2.0427E-02	2.8803E-02	
DOM	-1.9969E-03	-1.0554E-04	-5.8419E-03	-1.4695E-03	-1.3775E-03	-5.6073E-04	9.3485E-02
FORK	-1.2184E-03	2.7926E-04	2.8705E-02	7.4669E-04	1.1848E-03	1.4547E-03	1.8509E-03
DISEASE	-1.6086E-03	-1.2814E-04	-3.3164E-03	-2.9497E-04	1.0806E-03	1.6087E-03	-5.5456E-05
FORK							
DISEASE		DISEASE					
	4.8132E-02						
	5.1203E-04	1.9961E-02					



Table 6.3b. Covariance matrix of the coefficient estimates of the binary logit model fitted to thinned data.

	INTERCEPT	AGE	BA	N	QMD	SP1	PR
INTERCEPT	1.20040291						
AGE	-1.4917E-03	1.0484E-04					
BA	2.8505E-03	6.1420E-06	3.0459E-05				
N	-9.9244E-04	-7.5324E-07	-5.6027E-06	3.8259E-06			
QMD	-6.6190E-02	-8.9476E-05	-3.8448E-04	1.2765E-04	1.5518E-02		
SP1	-3.2995E-03	9.2253E-05	-2.0808E-05	5.7218E-06	3.4759E-04	6.1561E-03	
PR	3.6074E-04	-2.3168E-04	1.2267E-05	-2.6668E-06	-1.8531E-04	-1.1787E-03	3.3097E-03
DBH	-6.8718E-03	-1.6835E-05	-4.1945E-06	1.4390E-07	-2.9834E-04	2.2964E-04	2.8781E-05
HTCR	-5.1286E-03	-8.1573E-05	-3.4656E-06	-1.8491E-06	-5.5121E-04	-1.8148E-04	3.6316E-04
CR	-7.0109E-01	3.4630E-03	2.8662E-04	-1.9393E-04	-3.3421E-02	-1.2470E-02	3.6922E-03
STRAIGHT	-1.6200E-03	-2.4125E-05	-2.6751E-06	-2.5554E-07	-1.3832E-04	-4.3070E-05	1.0817E-04
BOLE	-4.2973E-03	-2.2804E-05	-6.7341E-06	3.5971E-06	1.3477E-04	5.0629E-05	6.5174E-05
BUTT	-5.5235E-03	-1.0585E-05	7.6480E-06	-2.3100E-06	5.0652E-05	-9.3743E-05	7.4423E-04
DOM	-1.6487E-03	1.6454E-05	-3.0249E-06	1.5382E-06	7.2209E-04	-2.4298E-04	1.7485E-04
FORK	-6.1776E-04	1.4842E-05	-7.9246E-06	1.2810E-06	2.0855E-04	-1.3083E-04	4.4907E-05
DISEASE	3.2108E-03	-1.7404E-06	3.1610E-06	-4.3431E-07	4.1670E-05	-1.2818E-03	1.0921E-04
	DBH	HTCR	CR	STRAIGHT	BOLE	BUTT	DOM
DBH	1.2622E-03						
HTCR	-7.2549E-05	6.6537E-05					
CR	-7.7387E-03	4.5313E-03	4.0431E-01				
STRAIGHT	2.7799E-04	1.0053E-05	7.3926E-05	6.9001E-03			
BOLE	-2.5994E-04	-4.8327E-06	-1.0984E-03	6.2358E-03	7.7676E-03		
BUTT	-1.4189E-04	-6.8344E-06	-8.5078E-05	6.2621E-03	6.2744E-03	5.5423E-02	
DOM	-5.2791E-04	-2.5618E-05	-1.6707E-03	-1.6052E-04	-3.1383E-04	1.8048E-04	2.8768E-03
FORK	-4.0345E-04	5.0778E-06	3.2837E-03	-2.0670E-05	2.1123E-04	5.8444E-04	2.9246E-04
DISEASE	-5.8365E-03	-3.0474E-05	-7.5234E-04	-2.2447E-04	9.8771E-05	1.8877E-05	-1.2812E-04
FORK		DISEASE					
DISEASE	1.9170E-02						
	1.7145E-04	5.2225E-03					

**Table 6.4.** Goodness of fit and prediction statistics of the two-product logit models fitted to unthinned and thinned data.

	Unthinned	Thinned
logL	-1309.29	-5181.06
AIC	1324.29	5196.06
pseudo- $R^2$	0.24	0.24
WP <sup>1</sup>	26.38%	28.32%
WP <sup>2</sup>	29.97%	31.53%
PSI <sup>1</sup>	75.91%	76.85%
PSI <sup>2</sup>	79.35%	80.21%
EI(q;p) <sup>1</sup>	17.21	198.46
EI(q;p) <sup>2</sup>	9.20	154.39
SSR <sup>1</sup>	243.78	1002.94
SSR <sup>2</sup>	433.58	1711.90
WSSR <sup>1</sup>	1810.92	5972.49
WSSR <sup>2</sup>	3119.37	11496.87

<sup>1</sup> Based on the validation data set

<sup>2</sup> Based on the full data set

percentages refer to the total number of tress in the data set.

all three measures may lead to overspecification of the model. Instead, we prefer to think of the quadratic mean diameter as a meaningful expression of the interaction taking place between the basal area and the number of trees per acre; the interaction which happens to be an important predictor variable in the model formulations considered here.

An attractive characteristic of the models are the positive and negative signs of the number of trees and basal area per acre coefficient estimates respectively. Their joint contribution results into a "balanced" effect of the stand density so that the merchantability of loblolly pine trees is not a monotonically decreasing or increasing function of the stand density.

Site preparation class 1 (tilled; debris moved) appears to significantly improve the merchantability of loblolly pine. The interpretation, however, of the role of the site preparation treatment must be treated with caution and in light of the type of data analyzed. More specifically, as Burkhart et al. (1985) pointed out, site preparation treatments were subjectively chosen, based on methods in use and conditions on the ground at the time of plantation establishment. As a result, the positive effect demonstrated by site preparation class 1 may not be exclusively due to this particular method of site preparation but rather, due to the combined effect of several other factors, known or unknown, which are confounded with the effect of this method. For example, a closer look at the data base revealed that almost three times as many plots prepared with method 1 exist in Piedmont than in Coastal Plain. Considering now that both models show a significant positive effect of the Piedmont physiographic region, we can conclude that the positive effect of site-preparation class 1 may be, at least in part, due to the larger frequency with which it is encountered in Piedmont. The above discussion aims in warning the reader about the type of incorrect interpretations that can occur when the data are obtained from observational studies and not from designed experiments. Our decision to include the site preparation class 1 effect in the models rests entirely upon the significant improvement in the prediction performance of these models when this effect is taken into account. In particular, the inclusion of this effect resulted into an increase of the number of correct predictions in the validation set from 68.1 percent to 75.9 percent for unthinned stands and from 68.4 percent to 76.7 percent for thinned stands (Table 6.4).

The significance of the physiographic region mentioned earlier, agrees, in part, with the conclusions reached by Burkhardt et al. (1985). In their study, the authors considered only unthinned stands and found that height/age relationships were significantly different for the Coastal Plain and Piedmont.

Important tree characteristics are: tree dbh, height to the base of the live crown, crown ratio, stem form, insect damage or disease and whether the tree is dominant or codominant. The signs of all coefficient estimates are logical. Total tree height does not enter the model directly; it is a functional part of the computation of crown ratio.

The joint significance of all variables in each model was tested by the likelihood ratio and the Wald's tests. For both models, the tests strongly rejected the null hypotheses at all conventional significance levels. Finally, as can be judged by the values of WP(validation) and PSI(validation) statistics in Table 6.4, the two models performed similarly with respect to prediction.

## **6.4 *Three-Product Logit Model***

To estimate the probability that a loblolly pine tree will fall into any one of the three product categories, two different models, an ordered and an unordered trinomial logit model, were considered. As with the two-product case, the two models were first fitted to the data from each thinning treatment separately and then two tests were conducted, one for significant thinning effect and one for significant difference between light and heavy thinning. Because both tests rejected the null hypotheses, the data from all thinning treatments were pooled together and the two models were again fitted to the new, pooled data set. Model selection and validation procedures have been applied at all stages of the analysis. The maximum likelihood coefficient estimates and the corresponding variance-covariance matrices of the finally selected models are shown on Tables 6.5, 6.5a,

(ordered logit) and 6.6, 6.6a and 6.6b (unordered logit). Table 6.7 displays the corresponding goodness-of-fit and validation statistics for these models.

The absence of significant thinning effect is an interesting result especially in contrast to the significant thinning effect which was detected by the two-product logit models of the previous section. It is probably due to the fact that the effect of thinning is generally more noticeable in smaller dbh classes or younger plots for even aged stands, than in larger dbh classes or older stands. The role of thinning, is basically to provide extra growing space to the promising trees in a stand. However, the effect of thinning diminishes if it is applied to relatively older stands, where the trees have already achieved their "social" status, i.e., they are dominant, codominant, intermediate or suppressed. It is therefore not surprising that a significant thinning effect was present for trees with dbh larger than or equal to 7.6 inches and absent for trees with dbh larger than or equal to 10.6. The above discussion also explains why the status of a tree in a stand, dominant or codominant, is no longer considered to be an important factor in estimating the merchantability of loblolly pine trees.

With the exception of the dummy variables indicating if a tree is dominant or codominant and whether site preparation method 1 was applied, the same explanatory variables as in the two product models were considered. The discussion of the previous section concerning the interpretation of the role of predictor variables, apply here as well.

In choosing between the ordered and unordered models observe that i) the standard errors are smaller for the coefficient estimates in the ordered model than those in the unordered model and ii) all goodness of fit and prediction performance criteria in Table 6.7 are in favor of the unordered model. The situation is analogous to that of comparing restricted and unrestricted models in the usual linear regression analysis. In particular, while unrestricted OLS fit minimizes the sum of squared residuals, its variance-covariance matrix is larger than that of restricted OLS fit of the same model (Kmenta, 1986). The increase in efficiency of the restricted coefficient estimates is attributed to the prior information provided by the restrictions. In our case, the restrictions imposed in the ordered model is that of common slope coefficients for all merchantability product classes. How-

**Table 6.5.** Maximum likelihood coefficient estimates of the three-product ordered logit model fitted to pooled data.

predictor variables	estimate	std. error
INT.1 (SAW)	-17.742879	1.847988
INT.2 (PEEL)	-20.162521	1.863876
AGE	0.159421	0.023564
BA	-0.037521	0.007091
N	0.014803	0.002487
QMD	0.787113	0.160731
PR	0.221674	0.158672
DBH	0.729856	0.076826
HTCR	0.012536	0.017389
CR	1.725884	0.647323
STRAIGHT	1.167995	0.172550
BOLE	-0.792632	0.181215
BUTT	-0.887452	0.495175
FORK	-0.827899	0.280083
DISEASE	-1.443220	0.149593

Table 6.5a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 6.5.

	INT.1	INT.2	AGE	BA	N	QMD	PR
INT.1	3.41505965						
INT.2	3.44492154	3.47403374					
AGE	-8.5344E-03	-8.7765E-03	5.5526E-04				
BA	7.9730E-03	8.0287E-03	3.0336E-05	35.0367E-05			
N	-3.2347E-03	-3.2566E-03	-4.0766E-06	-1.6247E-05	6.1852E-06		
QMD	-1.8500E-01	-1.8628E-01	-4.3155E-04	-9.8139E-04	3.6914E-04	2.5834E-02	
PR	-2.5275E-02	-2.5972E-02	-1.3233E-03	-9.1635E-07	6.3479E-06	2.0948E-03	2.5178E-02
DBH	-3.6816E-02	-3.7820E-02	-1.4128E-05	-1.8636E-05	3.2841E-06	-5.6092E-04	8.0401E-05
HTCR	-1.2475E-02	-1.2501E-02	-9.9545E-05	-1.8614E-05	2.5847E-06	-3.2892E-04	5.8074E-04
CR	-1.53844025	-1.5487E-03	4.0769E-03	-5.6247E-04	8.8506E-05	-3.1401E-02	6.9412E-03
STRAIGHT	-7.1092E-02	-7.3137E-02	1.2369E-05	-1.4621E-04	4.9823E-05	3.8294E-03	5.9545E-04
BOLE	-4.5583E-02	-4.5828E-02	-1.4461E-04	-8.4161E-05	3.0219E-05	2.5745E-03	9.4616E-04
BUTT	-1.5254E-02	-1.2743E-02	3.3238E-04	1.8976E-04	-6.0994E-05	-4.0958E-03	6.3133E-03
FORK	-2.3071E-02	-2.1140E-02	1.5008E-04	-4.9915E-05	1.6028E-05	8.3647E-04	1.3346E-03
DISEASE	1.3809E-02	1.6799E-02	3.0867E-04	-1.1582E-05	-5.9742E-08	-2.0806E-04	-1.8756E-03

	DBH	HTCR	CR	STRAIGHT	BOLE	BUTT	FORK
DBH	5.9022E-03						
HTCR	-2.6001E-04	3.0238E-04					
CR	-2.8881E-02	2.4976E-02	4.1903E-01				
STRAIGHT	1.5014E-03	-4.9877E-07	-2.0349E-03	2.9774E-02			
BOLE	4.0398E-04	-5.9634E-05	1.3352E-03	2.3805E-02	3.2839E-02		
BUTT	3.9154E-04	2.5120E-04	-4.8593E-03	2.3040E-02	2.2123E-02	2.4520E-01	
FORK	8.0401E-05	-1.5624E-05	1.2010E-02	2.8479E-03	3.2950E-03	1.4534E-03	7.8446E-02
DISEASE	-1.8182E-03	-7.1375E-05	1.5545E-03	-1.8333E-03	2.2580E-04	-1.0168E-03	2.0811E-03

DISEASE	
DISEASE	2.2378E-02

Table 6.6. Maximum likelihood coefficient estimates of the three-product unordered logit model fitted to pooled data.

predictor variables	Sawtimber		Peelers	
	estimate	std. error	estimate	std. error
INTERCEPT	-11.584254	3.765320	-25.181478	3.777590
AGE	0.108841	0.045352	0.242111	0.045646
BA	-0.025763	0.015564	-0.052143	0.015481
N	0.010201	0.005864	0.203541	0.005845
QMD	0.423551	0.354878	1.018112	0.352625
PR	0.278002	0.113596	0.478152	0.134530
DBH	0.031501	0.014965	0.757111	0.145869
HTCR	0.079845	0.031784	0.054213	0.031793
CR	10.369235	2.879575	7.396224	2.917120
STRAIGHT	1.253224	0.459600	1.983475	0.516989
BOLE	-1.478002	0.211826	-1.829445	0.214202
BUTT	-1.197554	0.604968	-3.856112	1.151613
FORK	-0.485323	0.413670	-1.374541	0.455466
DISEASE	-1.486896	0.223265	-2.278520	0.237446



Table 6.6a. Covariance matrix of the coefficient estimates (sawtimber) of the UMNL model displayed on table 6.6.

	INTERCEPT	AGE	BA	N	QMD	PR	DBH
INTERCEPT	14.17763471						
AGE	-3.6116E-02	2.06568-03					
BA	3.1157E-02	8.9669E-05	2.4224E-04				
N	-1.1966E-02	-2.0550E-05	-6.0415E-05	3.4386E-05			
QMD	-7.4887E-01	-1.8248E-03	-3.9094E-03	1.4589E-03	1.2594E-01		
PR	-2.2542E-02	-7.1617E-03	3.2055E-05	3.0310E-05	7.0946E-03	1.2904E-02	
DBH	-1.1901E-01	-1.3120E-04	-5.4466E-05	6.6128E-06	-1.8549E-03	4.5136E-05	2.2395E-04
HTCR	-3.8503E-02	-3.1677E-04	-5.9063E-05	4.4025E-06	-1.9520E-03	2.1222E-03	-1.0662E-03
CR	-4.62378541	2.2342E-02	-8.2364E-04	-4.1578E-04	-1.9878E-01	4.8010E-04	-1.0936E-01
STRAIGHT	-1.9415E-01	9.6239E-04	-1.3235E-04	6.8382E-05	9.6005E-03	4.6086E-04	3.3810E-03
BOLE	-1.4257E-01	-6.2880E-04	-1.8111E-04	7.5597E-05	6.5672E-03	1.3726E-03	2.4833E-03
BUTT	-7.4886E-02	2.1628E-03	6.6407E-04	-2.0147E-04	-1.2304E-02	1.5054E-02	2.7015E-03
FORK	-3.3182E-02	1.2769E-03	-9.7699E-06	1.7882E-05	1.0648E-03	3.2449E-03	-2.2564E-03
DISEASE	7.0704E-02	1.4998E-04	3.3189E-05	-3.1320E-05	-5.3583E-03	1.7674E-03	-4.2755E-03

	HTCR	CR	STRAIGHT	BOLE	BUTT	FORK	DISEASE
HTCR	1.0102E-03						
CR	9.0664E-02	8.29195218					
STRAIGHT	-5.5996E-04	-3.4070E-02	2.1123E-01				
BOLE	-2.7888E-04	-5.1556E-04	7.4856E-02	4.4870E-02			
BUTT	1.3591E-04	-1.4773E-02	7.8469E-02	7.3923E-02	4.4870E-02		
FORK	-4.8602E-04	2.9658E-02	2.5124E-03	5.4720E-03	-2.4067E-03	3.6599E-01	
DISEASE	1.1546E-05	2.9183E-02	-1.2595E-02	8.6563E-03	9.7879E-04	1.1340E-02	5.0008E-02

Table 6.6b. Covariance matrix of the coefficient estimates (peelers) of the UMNL model displayed on table 6.6.

	INTERCEPT	AGE	BA	N	QMD	PR	DBH
INTERCEPT	14.27018621						
AGE	-4.7092E-02	2.08368-03					
BA	4.8431E-02	8.4379E-05	2.3966E-04				
N	-1.9721E-02	-7.2175E-05	-9.0407E-05	3.4164E-05			
QMD	-1.11808125	-1.2091E-03	-5.2863E-03	2.0665E-03	1.2434E-01		
PR	-5.0279E-02	-5.6483E-03	-2.2644E-04	8.7675E-05	1.3663E-02	1.8098E-02	
DBH	-2.0020E-01	2.7404E-04	-1.5894E-04	4.4722E-05	1.5324E-04	1.3143E-02	2.1278E-02
HTCR	-4.3681E-02	-3.3447E-04	-8.9421E-05	1.7897E-05	-7.4061E-04	1.3966E-03	-1.5032E-03
CR	-5.68369100	2.0931E-02	-5.4030E-03	1.4778E-03	-4.1441E-02	-1.0539E-01	-1.3007E-01
STRAIGHT	-3.4335E-01	1.8580E-03	-6.8094E-04	2.6615E-04	1.8355E-02	5.6730E-03	6.7833E-03
BOLE	-1.0500E-01	-3.4117E-04	-4.2031E-04	1.1002E-04	9.0371E-03	-1.3396E-03	1.0916E-03
BUTT	-1.0602E-01	1.4584E-03	9.3376E-05	5.1713E-05	1.7433E-03	8.9178E-03	8.6066E-04
FORK	-8.7143E-02	2.1742E-04	-1.7544E-04	4.2435E-05	-1.4669E-03	-3.1372E-03	7.4179E-03
DISEASE	1.5195E-01	-2.5937E-04	1.9070E-04	-9.3919E-05	-7.3492E-03	-2.0208E-03	-1.0013E-02

	HTCR	CR	STRAIGHT	BOLE	BUTT	FORK	DISEASE
HTCR	1.0108E-03						
CR	8.8170E-02	8.50958909					
STRAIGHT	-5.0275E-04	-2.5497E-02	2.6728E-01				
BOLE	-7.4163E-04	-7.1527E-02	9.9289E-02	4.5882E-02			
BUTT	-1.1208E-03	-6.0126E-02	1.0142E-01	9.7430E-02	1.32621250		
FORK	-2.7635E-04	1.5669E-02	5.4128E-03	1.1546E-02	2.6482E-03	2.0745E-01	
DISEASE	3.0974E-04	3.3086E-02	-2.2139E-02	8.8147E-04	-8.8147E-04	-1.0060E-02	5.6381E-02

Table 6.7. Goodness of fit and prediction statistics of the unordered and ordered three product logit models fitted to pooled data.

	Ordered	Unordered
logL	-1254.20	-956.99
AIC	1267.20	969.99
pseudo- $R^2$	0.22	0.28
WP <sup>1</sup>	26.11%	28.95%
WP <sup>2</sup>	27.88%	30.11%
PSI <sup>1</sup>	68.91%	73.42%
PSI <sup>2</sup>	71.95%	75.88%
EI(q;p) <sup>1</sup>	170.28	114.38
EI(q;p) <sup>2</sup>	110.53	87.59
SSR <sup>1</sup>	2404.57	1711.90
SSR <sup>2</sup>	4806.39	3852.83
WSSR <sup>1</sup>	5118.92	4025.36
WSSR <sup>2</sup>	9100.61	7438.59

<sup>1</sup> Based on the validation data set

<sup>2</sup> Based on the full data set

percentages refer to the total number of trees in the data set.

ever, if the prior information is not correct, i.e., the true model is unordered having different slope coefficients for each merchantability class, then the use of an ordered model can result into serious biases in the estimation of the corresponding probabilities (Amemiya, 1985). Even though in this case the ordering of the merchantability product classes does indeed seem to be "natural", it does not necessarily imply common slope coefficients for all product classes. As Anderson (1984) put it, even when the response variable is putatively ordered, there is no guarantee that an ordered model is appropriate. In his article, Anderson recommends that the analyst should always begin with the general (unrestricted) unordered model and use an ordered model only if it shows better fit than the unordered model. Note that both LR and Wald tests rejected the hypothesis that the two models are the same at all conventional levels of significance. Amemiya (1985) cautions in the use of ordered models by pointing out that the cost of using an unordered model when the true model is ordered is loss of efficiency rather than consistency. Thus, from a conservative point of view, it is much safer to use the unordered model than the ordered. In light of the above discussion and the fact that all criteria indicate better prediction performance of the unordered model, we endorse the use of unordered models in modelling the merchantability of loblolly pine trees.

## ***6.5 Concluding Remarks***

Two logistic regression models, one dichotomous and one unordered trichotomous were developed to estimate the merchantability of loblolly pine trees growing on thinned and unthinned stands. Model selection and validation procedures were employed to highlight important stand and tree characteristics that influence loblolly pine merchantability. To illustrate the utility of these models, Tables 6.8 and 6.9 have been constructed showing the actual and predicted proportions of the three product categories by dbh class for unthinned and thinned plots respectively. The data used for the derivation of these tables are the validation data sets used for model selection during the analysis. As the reader can verify by looking at these tables, the predicted proportions follow, in general, the

trend of the actual proportions even though an over-prediction of sawtimber and peelers is apparent. In any case, the incorporation of these models into individual tree yield simulation systems or into diameter distribution based growth and yield prediction systems will result into more realistic yield per acre estimates which in turn will facilitate management decision making.

One final point must be emphasized. It concerns the basic assumption in logistic regression analysis that the individual tree observations are independent of each other. Although this is a reasonable assumption for trees growing on different clusters, its validity is questionable for trees in the same cluster which, are most likely correlated. The beta-binomial distribution has been used as a model for correlated binary observations in biomedical applications such as ophthalmology and cardiology. Initially, these models assumed positive intra-cluster correlation and clusters of equal size consisting of a small number of observations, two or three (Griffiths 1973, Williams 1975 and Paul 1979). Rosner (1984) proposed a more general model also based on the beta-binomial distribution but allowing for unequal cluster sizes and including the binary logit model as a special case when the observations are uncorrelated. However, this model still assumed positive intra-cluster correlation and the maximum likelihood estimation procedure was becoming computationally intensive for clusters with size five or more. Qu et al. (1987) generalized Rosner's model to allow for negative, zero or positive intra-cluster correlation by using the Polya-Eggenberger distribution instead of the beta-binomial. They also proposed a derivative free estimation procedure which was shown to be feasible for clusters of size less than ten. For larger sizes, the required computer time was impractical. Connolly and Liang (1988) suggested an even more general model which, depending on the form of an appropriately selected distribution function could result in Qu et al. model as a special case. The authors also suggested a class of easily computed estimating functions which have high efficiency compared to the computationally intensive Newton-Raphson methodology. In illustrating this method however, the authors considered clusters with size no larger than six. With regard to probit formulations, Ochi and Prentice (1984) described a correlated binary probit model in which the necessary likelihood derivatives could be reduced to linear combinations of equally correlated normal probabilities for which some existing approximations could be applied. Even this

Table 6.8. Actual and predicted proportions of trees by dbh class and product class for unthinned stands.

dbh class	pulpwood		sawtimber		peelers	
	actual	predicted	actual	predicted	actual	predicted
7.6-8.0	0.65	0.61	0.35	0.39	0.00	0.00
8.1-8.5	0.56	0.50	0.44	0.50	0.00	0.00
8.6-9.0	0.48	0.43	0.52	0.57	0.00	0.00
9.1-9.5	0.39	0.34	0.61	0.66	0.00	0.00
9.6-10.0	0.26	0.19	0.74	0.81	0.00	0.00
10.1-10.5	0.23	0.15	0.77	0.85	0.00	0.00
10.6-11.0	0.20	0.11	0.63	0.55	0.17	0.34
11.1-11.5	0.14	0.05	0.51	0.45	0.35	0.50
11.6-12.0	0.10	0.02	0.36	0.30	0.54	0.68
12.1-12.5	0.04	0.00	0.32	0.28	0.64	0.72
12.6-13.0	0.01	0.00	0.21	0.15	0.79	0.85
13.1-13.5	0.00	0.00	0.28	0.10	0.72	0.90
13.6-14.0	0.00	0.00	0.16	0.07	0.84	0.93
14.1-14.5	0.00	0.00	0.00	0.00	1.00	1.00
14.6-15.0	0.00	0.00	0.01	0.00	0.99	1.00
15.1-	0.00	0.00	0.00	0.00	1.00	1.00

Table 6.9. Actual and predicted proportions of trees by dbh class and product class for thinned stands.

dbh class	pulpwood		sawtimber		peelers	
	actual	predicted	actual	predicted	actual	predicted
7.6-8.0	0.59	0.51	0.41	0.49	0.00	0.00
8.1-8.5	0.49	0.40	0.51	0.60	0.00	0.00
8.6-9.0	0.37	0.21	0.67	0.79	0.00	0.00
9.1-9.5	0.27	0.12	0.73	0.88	0.00	0.00
9.6-10.0	0.20	0.08	0.80	0.92	0.00	0.00
10.1-10.5	0.14	0.04	0.86	0.96	0.00	0.00
10.6-11.0	0.11	0.03	0.79	0.86	0.10	0.11
11.1-11.5	0.09	0.02	0.58	0.64	0.33	0.34
11.6-12.0	0.07	0.01	0.36	0.39	0.57	0.60
12.1-12.5	0.03	0.00	0.21	0.28	0.76	0.72
12.6-13.0	0.00	0.00	0.13	0.19	0.87	0.81
13.1-13.5	0.00	0.00	0.04	0.08	0.96	0.92
13.6-14.0	0.00	0.00	0.00	0.00	1.00	1.00
14.1-14.5	0.00	0.00	0.00	0.00	1.00	1.00
14.6-15.0	0.00	0.00	0.01	0.00	0.99	1.00
15.1-	0.00	0.00	0.00	0.00	1.00	1.00

method, however, was found to be computationally intensive and the approximations were becoming poorer as cluster sizes were increasing.

Thus, although a considerable body of work exists on correlated binary responses, efficient computational procedures are needed in order for this theory to be applied in forestry applications where large cluster sizes are frequently encountered. In addition, there is practically nothing in the literature with regard to the consequences from this violation of the independence assumption. It is only the analogy to the general linear models theory that allows us to speculate that the coefficient estimates are expected to remain asymptotically unbiased but with their estimate of error inflated. However, it is unknown to what extent this error inflation occurs or what exactly are the consequences to the model's overall prediction ability. To answer these questions further analytical or simulation studies are required. Finally, the problem of analyzing correlated polychotomous responses has not yet been investigated and is a topic of future research efforts.



## **Chapter VII**

# **Modeling Fusiform Rust Incidence in Loblolly and Slash Pine Plantations**

### ***7.1 Introduction***

Loblolly pine and slash pine are the major pine species in the southern United States with as much as two million acres planted annually. Fusiform rust, caused by the fungus *Cronartium quercuum* (Berk.) Miyabe ex Shirai f. sp. fusiforme, is responsible for more damage in these species than any other pathogen. Annual yield losses in excess of 560 million board feet of saw timber and 200 million cubic feet of total volume have been reported which represent approximately 10 percent of the annual harvest of these species (Powers et al. 1974). Assessing the incidence and the spread of this disease is therefore of vital interest to forest managers so that detection, control and prevention practices can be efficiently deployed.

Due to the great economic impact of fusiform rust, several attempts have been made to determine site factors that contribute to rust infection and to predict the occurrence of the disease in loblolly and slash pine plantations. The modeling approaches taken by these studies, however, suffer to one degree or another by violations of assumptions most as we discuss below.

Wells and Dinus (1978) developed linear probability models of the type discussed in section 2.1, using the proportion of stem infected trees at age 5 as predictor, to predict rust associated mortality at the age of 10 for slash pine and rust resistant and rust susceptible loblolly seed sources. Schmidt et al. (1979) also employed linear probability models to predict the proportion of rust infected loblolly and slash pine trees at ages 6 through 10 as a function of the proportion of rust infected trees at an earlier age. In both studies, the authors overlooked some of the problems associated with the application of linear probability models to predict proportions of trees falling into two mutually exclusive classes. As discussed in section 2.1, the most serious weakness of such models is that the predicted proportions are not restricted to fall within the  $[0,1]$  interval. This deficiency usually becomes profound when the models are applied to data other than those used to fit the models. In addition, because the error variance is not homogeneous, the OLS coefficient estimates are unbiased but not of minimum variance among the class of linear unbiased estimators.

Borders and Bailey (1986) applied OLS methodology to regress the logit of the proportion of trees per acre with one or more fusiform rust galls on stand age, site index and geographic location. The logit transformation, while ensuring that the predicted values will lie between 0 and 1, does not however eliminate the unequal variances of the error terms. Thus, as mentioned previously, the OLS coefficient estimates are unbiased but not fully efficient. It is worth mentioning that most computer software packages performing logit or probit analysis, use these OLS estimators as starting values for the iterative search for maximum likelihood estimators.

Nance et al. (1981) employed a Markov model, suggested by Arvanitis and Amateis (1978), to assess the effect of fusiform rust on slash pine mortality. Using this model, the authors predicted the number of surviving trees at any age (both infected and uninfected) based on the number of trees

present at age 3 and rust associated mortality also at age 3. Subsequently, the same model was applied to loblolly pine plantations (Shoulders and Nance, 1987). It is generally known that Markov process based models produce valid estimates only if the transition probabilities between two specific states remain stationary over time. In regard to mortality associated with fusiform rust, this assumption implies that the probability of a healthy tree dying during a fixed period of time must remain the same regardless of stand conditions. In view of the ever-changing stand and disease dynamics and the myriad of interacting factors present in a forest stand, it may be very difficult to adopt such an assumption.

Thus, even though some progress has been made in modeling fusiform rust incidence, the models employed have one or more of their basic assumptions violated and, with the exception of the Markov formulation, cannot handle multinomial situations where trees are classified into more than two categories according to the severity of rust infection. It is the purpose of this study to illustrate the applicability of qualitative response models as alternative formulations in modelling fusiform rust occurrence. In the first part of the study, dichotomous and polychotomous, logit and probit models were applied to permanent plot data from loblolly and slash pine plantations in East Texas to predict the proportion of trees per acre with various levels of rust infection at a given age. In the second part, the same data were used to fit qualitative response models that predict transitional proportions, i.e., proportions of trees moving from one level of rust infection to another.

It is believed that the models suggested in this study are useful from several standpoints. They identify sites with a high probability of fusiform rust incidence, making it possible for forest managers to allocate usage and/or manage appropriately on a short term basis. Perhaps more importantly, the development of such models will help determine factors which contribute to rust occurrence and spread, and provide specific long range direction for alleviating or managing these factors.

## 7.2 Data

Data were provided by the East Texas Pine Plantation Research Project consisting of 256 site-prepared loblolly and slash pine plantations. The survey design has been documented by Lenhart et al. (1985). During the dormant seasons of 1982-1984, 178 plots were permanently established in loblolly pine plantations and 78 in slash pine plantations. Because the general practice in evaluating the effect of fusiform rust is to base its impact on the percentage of infected stems by age 5 (Wells and Dinus 1978, Schmidt et al. 1979, Schmidt and Klapproth 1982) plantations of age less than 5 years were excluded from the study leaving 70 plots for loblolly pine and 38 plots for slash pine. Table 7.1 displays the age distribution of sample plots at the time of establishment.

Table 7.1. Distribution of sample plots by species and age.

species	plantation age (years)													total
	5	6	7	8	9	10	11	12	13	14	15	16	17	
loblolly	11	4	11	9	3	9	8	7	2	1	2	2	1	70
slash	3	9	8	4	2	1	4	1	1	4	0	1	0	38

The various site preparation treatments applied to the plantations were summarized in the following classes:

**KGSNW**      KG bladed/sheared not windrowed

**KGSW**        KG bladed/sheared not windrowed

**KGSBD**      KG bladed/sheared bedded

**KGSBN**      KG bladed/sheared burned

**MISC**        various other treatments

Table 7.2 shows the distribution of site preparation classes in the sample plots of both species.

Table 7.2. Classification of survey plots by species and site preparation classes.

	loblolly	slash
KGSWBD	4	2
KGSBN	8	8
KGSNW	0	2
KGSW	53	25
MISC	5	1
TOTAL	70	38

Plot characteristics recorded were age, average height of dominant and codominant trees (HEIGHT) in feet, site index (SINDEX) at base age 25 in feet, number of trees per acre (N) and basal area per acre (BA) in squared feet per acre (Table 7.3). The quadratic mean diameter (QMD) was computed from observations of N and BA. An additional measure of stand density, known as spacing index or relative spacing (Clutter et al. 1983), was also computed as

$$RS = \frac{\sqrt{43560/\text{number of trees per acre}}}{\text{average height of dominant and codominant trees}}$$

Each pine tree in a plot was visually inspected for rust infection and classified into one of the following three categories:

**CLEAR**      the tree is healthy, free of rust infection

**BRANCH**    galls exist on a live or dead branch at a distance of more than 12 inches from the stem

**STEM**        galls exist on stem or on a live branch within 12 inches from the stem

Three years after the first measurement the plots were revisited and measurements were taken on the same plot variables (Table 7.3). Each tree in a plot was again checked for rust infection and classified as CLEAR, BRANCH, STEM or DEAD if the tree was found dead as a result of rust infection.

Table 7.3. Loblolly and slash pine plot summary statistics during the first measurement and remeasurement.

LOBLOLLY						
plot variables	first measurement			remeasurement		
	min.	mean	max.	min.	mean	max.
HEIGHT	7.00	29.41	53.00	20.00	39.39	67.00
SINDEX	29.00	68.91	100.00	32.00	72.07	101.00
N	139.00	444.04	749.00	139.00	436.19	740.00
BA	0.00	49.86	116.00	10.00	77.39	140.00
QMD	0.00	4.12	7.09	2.39	5.60	8.10
RS	0.09	0.30	0.48	0.08	0.25	0.31
SLASH						
HEIGHT	8.00	28.29	54.00	17.00	38.52	60.00
SINDEX	27.00	66.11	99.00	37.00	68.64	84.00
N	133.00	345.00	1002.00	112.00	151.50	989.00
BA	0.00	35.39	107.00	5.00	52.58	129.00
QMD	0.00	3.99	7.72	1.96	5.37	7.57
RS	0.08	0.36	0.49	0.07	0.29	0.35

## ***7.3 Models that Predict Fusiform Rust Infection Levels***

### **7.3.1 Dichotomous Models**

As a first step in this study, it was decided to develop models that predict the proportion of rust infected and dead trees at a given age for loblolly and slash pine plantations. The remeasurement data were classified into two categories; one containing healthy trees and one containing branch infected, stem infected and dead trees. Two binary choice models, a logit and a probit, were considered and the finally selected models along with the corresponding goodness of fit statistics are shown on Tables 7.4 and 7.5 for loblolly and slash pine respectively. Tables 7.4a and 7.5b display the variance-covariance matrices for the coefficient estimates of the two logit models.

The predictor variables used by these models are described below:

**NORTH:** 1 if plantation is located in North-East Texas, i.e., in Cass, Harrison, Marion, Panola, Red River or Rusk Counties; 0 otherwise

**KGSW:** 1 if site preparation class is KGSW; 0 otherwise

**KGSNW:** 1 if site preparation class is KGSNW; 0 otherwise

**KGSBN:** 1 if site preparation class is KGSBN; 0 otherwise

**KGSBD:** 1 if site preparation class is KGSBD; 0 otherwise

**SLOPE:** 1 if plantation lies on slope terrain; 0 otherwise

**FLAT:** 1 if plantation lies on flat terrain; 0 otherwise

Table 7.4. Maximum likelihood coefficient estimates of the binary logit and probit models which predict rust infection on loblolly pine.

predictor variables	LOGIT		PROBIT	
	estimate	std. error	estimate	std. error
INTERCEPT	3.988322	0.534601	2.067210	0.298115
NORTH	-0.381866	0.051186	-0.21192	0.027775
KGSW	0.191307	0.071649	0.099893	0.039108
KGSBN	0.299342	0.083436	0.154439	0.045889
KGSBD	0.615450	0.100303	0.331718	0.054959
SLOPE	0.199984	0.047273	0.466000	0.035055
FLAT	-0.098114	0.014730	-0.118291	0.026582
AGE	-0.014638	0.002133	-0.053595	0.008215
SINDEX	-0.014638	0.002133	-0.007810	0.001207
QMD	-0.097964	0.035891	-0.053664	-0.019982
RS	-0.778364	0.105695	-0.420165	0.058534
CLEAR0	-0.004954	0.000325	-0.002709	0.000181
INFECT0	0.004673	0.000356	0.002864	0.000205
logL = -14108.86		logL = -14111.02		
AIC = 14120.86		AIC = 14123.02		
pseudo- $R^2$ = 0.243		pseudo- $R^2$ = 0.240		
SSR = 0.352847		SSR = 0.353758		
WSSR = 2.393087		WSSR = 2.383570		



Table 7.4a. Covariance matrix of the coefficient estimates of the binary logit model displayed on table 7.4.

	INTERCEPT	NORTH	KGSW	KGSBN	KGSBD	SLOPE	FLAT
INTERCEPT	2.2580E-01						
NORTH	3.6894E-03	2.6200E-03					
KGSW	3.9020E-03	2.0429E-04	5.1336E-03				
KGSBN	3.2478E-04	-7.7447E-05	4.6256E-03	6.9616E-03	1.0061E-02		
KGSBD	1.0204E-02	3.9025E-04	4.8363E-03	4.4557E-03	7.0220E-04	2.2347E-03	
SLOPE	-2.9164E-03	-1.6875E-04	3.3520E-04	5.0597E-05	9.8930E-05	1.4348E-03	2.1697E-04
FLAT	-1.4643E-03	2.3168E-04	-5.4929E-05	-4.8490E-04	-5.6558E-06	1.0968E-05	-6.0168E-05
AGE	-3.6636E-03	5.7068E-05	1.0001E-04	1.9067E-04	-1.6971E-05	-2.0905E-06	-5.5126E-06
SINDEX	-6.7511E-04	1.2009E-05	-8.8995E-06	7.8591E-06	-8.4628E-04	4.6437E-05	7.6932E-05
QMD	-6.9522E-03	-4.9887E-04	-5.5900E-04	-6.0313E-04	-2.9429E-03	2.9202E-04	2.7594E-05
RS	-5.1898E-02	-1.3891E-03	-2.0481E-03	-1.5592E-03	-1.1003E-05	9.5985E-07	1.8829E-06
CLEAR0	-1.5099E-04	-1.6938E-06	-6.3096E-06	-4.1512E-06	-8.2165E-06	3.7100E-07	5.1756E-07
INFECT0	-1.1948E-04	-2.0000E-06	-8.8522E-06	-6.1833E-06			

	AGE	SINDEX	QMD	RS	CLEAR0	INFECT0
AGE	4.5497E-06					
SINDEX	2.4741E-05	4.5497E-06				
QMD	-3.2044E-04	-3.3143E-05	1.2882E-03			
RS	2.2647E-04	6.6372E-05	2.4685E-03	1.1171E-02		
CLEAR0	5.2188E-07	1.9381E-07	6.9059E-06	3.0363E-05	1.0563E-07	
INFECT0	-6.9538E-07	6.9557E-08	7.7904E-06	2.8739E-05	8.2556E-08	1.2674E-07

**Table 7.5.** Maximum likelihood coefficient estimates of the binary logit and probit models which predict rust infection on slash pine.

	LOGIT		PROBIT	
<u>predictor variables</u>	<u>estimate</u>	<u>std. error</u>	<u>estimate</u>	<u>std. error</u>
INTERCEPT	-0.356198	0.154162	-0.194749	0.090714
KGSW	-0.914583	0.150005	-0.562631	0.089390
KGSBN	-0.832987	0.167654	-0.519618	0.099227
KGSBD	-1.565833	0.154702	-0.964353	0.091953
KGSNW	-0.867938	0.177564	-0.538487	0.106308
SLOPE	0.543587	0.169851	0.306855	0.094091
FLAT	-0.485316	0.095910	-0.297962	0.056909
AGE	0.103497	0.016174	0.060431	0.009480
HEIGHT	-0.033378	0.011936	-0.019063	0.007139
BA	-0.021554	0.003110	-0.013244	0.001876
QMD	0.487000	0.046947	0.292044	0.002769
INFECT0	0.003679	0.000290	0.002265	0.000175
logL = -7770.73		logL = -7771.71		
AIC = 7781.74		AIC = 7788.71		
pseudo- $R^2$ = 0.248		pseudo- $R^2$ = 0.244		
SSR = 0.149110		SSR = 0.151193		
WSSR = 0.745477		WSSR = 0.751730		

Table 7.5a. Covariance matrix of the coefficient estimates of the binary logit model displayed on table 7.5.

	INTERCEPT	KGSW	KGSBN	KGSBD	KGSNW	SLOPE	FLAT
INTERCEPT	2.3766E-02						
KGSW	-1.4237E-02	2.2502E-02					
KGSBN	-1.2201E-02	2.3301E-02	2.8108E-02				
KGSBD	-1.5339E-02	1.9248E-02	1.9481E-02	2.3933E-02			
KGSNW	-1.5923E-02	2.2139E-02	2.3084E-02	1.8166E-02	3.1529E-02		
SLOPE	7.6257E-05	-6.2991E-04	-1.2643E-03	-1.0892E-03	-7.3187E-04	2.8849E-02	
FLAT	9.2986E-04	2.6019E-04	5.7861E-04	3.2060E-03	-4.4010E-03	6.0089E-04	9.1987E-03
AGE	-6.1439E-04	-2.4401E-05	-5.9405E-05	-7.3160E-04	3.1656E-04	1.6769E-04	-7.4997E-04
HEIGHT	-6.1246E-05	-6.3266E-04	-7.9733E-04	-9.1497E-05	-6.5190E-04	-1.7316E-04	5.0764E-06
BA	6.5700E-05	2.3188E-04	3.0244E-04	6.6117E-05	2.4990E-04	1.7077E-06	-3.0296E-05
QMD	-7.8778E-04	1.3618E-03	1.2571E-03	1.6091E-03	7.2998E-04	8.2867E-04	1.7830E-03
INFECT0	-3.0373E-06	-1.6688E-05	-1.9450E-05	-9.3699E-06	-1.4537E-05	-9.7738E-07	2.0083E-06

	AGE	HEIGHT	BA	QMD	INFECT0
AGE	2.6160E-04				
HEIGHT	-8.7994E-05	1.4247E-04			
BA	1.2962E-05	-2.9800E-06	9.6721E-06		
QMD	-2.1565E-05	3.5960E-04	3.0362E-05	2.2040E-03	
INFECT0	-6.2841E-07	7.9342E-07	-3.8172E-07	5.6491E-07	8.4100E-08

<b>AGE:</b>	age at remeasurement
<b>HEIGHT:</b>	average height of dominant and codominant trees during the initial measurement
<b>SINDEX:</b>	site index; base age 25 years
<b>BA:</b>	basal area per acre
<b>QMD:</b>	quadratic mean diameter
<b>RS:</b>	relative spacing
<b>CLEAR0:</b>	initial number of healthy trees per acre
<b>INFECT0:</b>	initial number of infected trees per acre (branch and stem infected)

Because the models predict proportions and not individual tree probabilities, model selection criteria that evaluate model performance based on individual correct predictions such as WP, PSI and Theil's information inaccuracy of the prediction cannot be meaningfully employed. Comparison between probit and logit models is meaningful only on the basis of SSR, WSSR and pseudo- $R^2$  statistics, since the log-likelihood functions of the two formulations are not comparable. The values of these statistics in Tables 7.4 and 7.5 indicate that logit models fit the data slightly better than probit models. This improvement, although it does not seem to be significant for practical purposes, influenced our decision to favor the logit model over the probit.

Influence diagnostics were then calculated for the logit models fitted to loblolly and slash pine data. In particular, for each plot observation the following diagnostics were computed as shown in Chapter 5:

- the standardized residual,  $se_i$ , and the corresponding component of  $-2\log L$ ,  $d_i$ , to identify outlying plot observations (see section 5.1),

- the diagonal element,  $h_{ii}^*$ , of the HAT matrix  $H^*$ , to identify high leverage plot observations (see section 5.2) and
- the scalar  $c_i^{-1}$  to measure the influence which the particular plot observation exerts on the estimated coefficient vector (see section 5.3).

Figures 7.1-7.2 and 7.3-7.4 display the index plots of these diagnostics, i.e., plots of  $se_i$ ,  $d_i$ ,  $h_{ii}^*$  and  $c_i^{-1}$  vs plot number  $i$ , for loblolly and slash pine respectively.

Focusing first on loblolly pine logistic diagnostics, plot number 49 is seen to be an obvious outlier (Figure 7.1 a and b). However, the leverage of this plot is small as the corresponding  $h_{ii}^*$  value indicates in Figure 7.2a. The same figure identifies plot observations 13, 47, 52 and 63 as high leverage data points. But again, these plots exhibit only small residual values (Figure 7.1 a and b). Figure 7.2b shows that plots 35, 52 and 64 are the most influential observations to the estimated coefficient vector.

Turning now to slash pine logistic diagnostics, plots 14, 19, 35 and 36 were identified as outliers (Figure 7.3 a and b); plots 2, 9 and 17 as high leverage observations (Figure 7.4a) and plots 3, 13, 30 and 35 as highly influential to the estimated vector of coefficient estimates (Figure 7.4b).

Careful checking on the values of the plot observations which have been reported as outliers, high leverage or high influence data points showed no evidence of illogical or mistakenly recorded measurements. Hence, these plots were considered to contain valid observations and no action was taken against their presence in the estimation data set.

To provide a better insight on how well the models perform, Figures 7.5, 7.6, 7.7 and 7.8 display plots of the standardized residuals versus age, landform and site index or average height of dominant and codominant trees for loblolly and slash pine respectively. The landform of each site is also indicated in these plots. The models appear to perform equally well for both species and no ap-

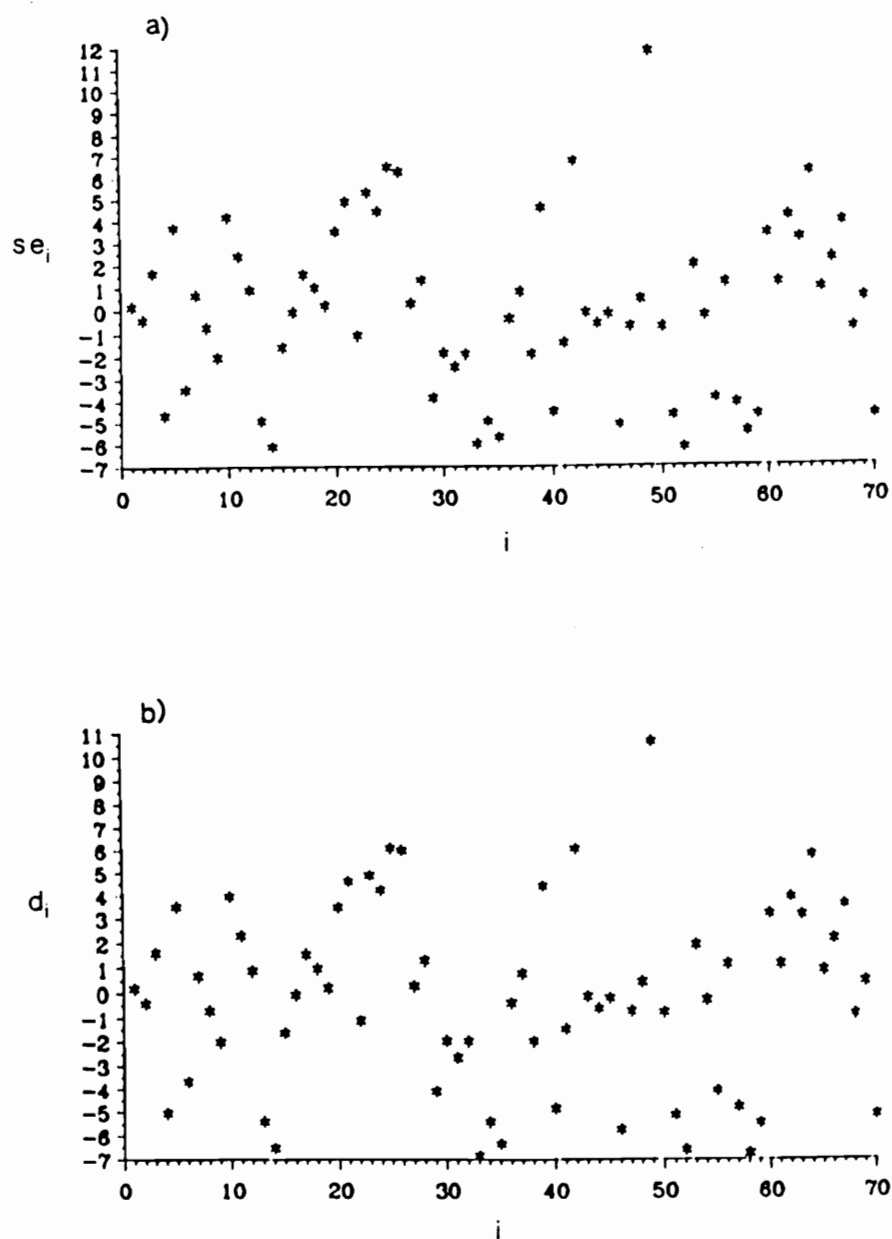


Figure 7.1. Index plot of  $se_i$  vs  $i$  (a) and of  $d_i$  vs  $i$  (b) for the binary logit model fitted to loblolly pine data.

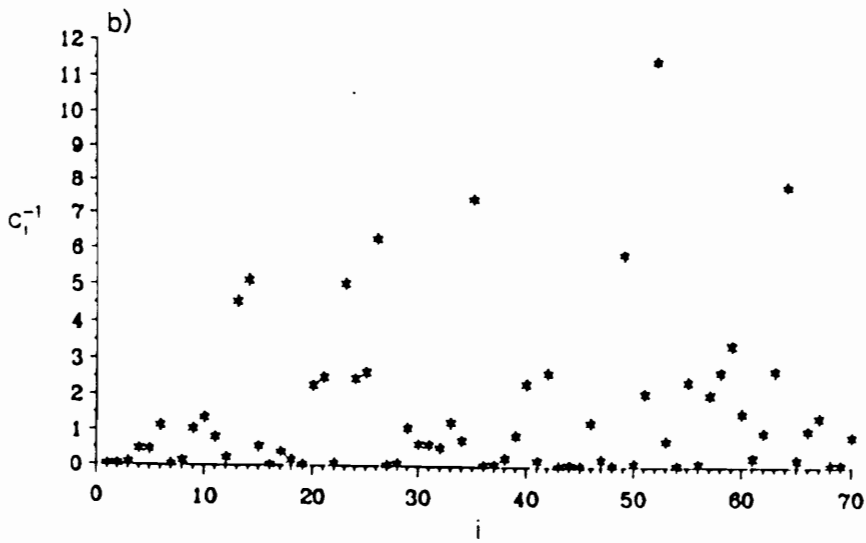
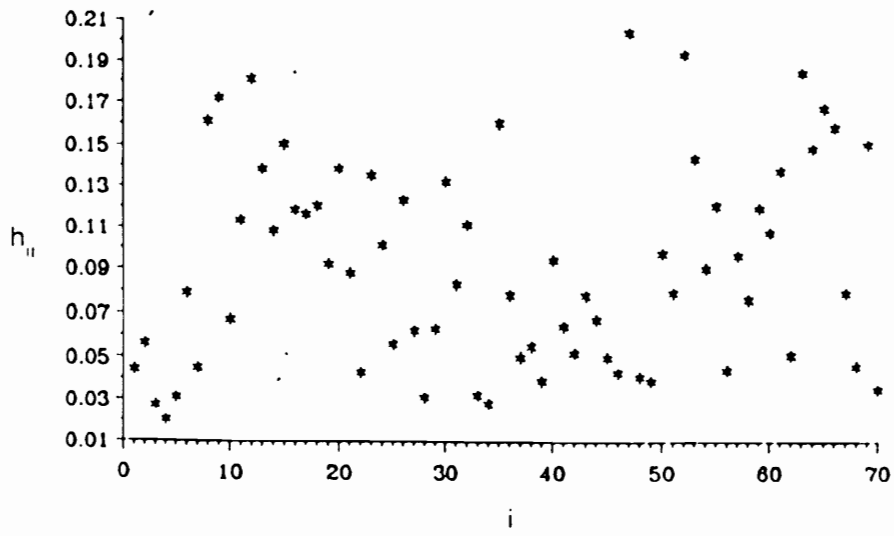


Figure 7.2. Index plot of  $h_{ii}$  vs  $i$  (a) and  $c_i^{-1}$  vs  $i$  (b) for the binary logit model fitted to loblolly pine data.

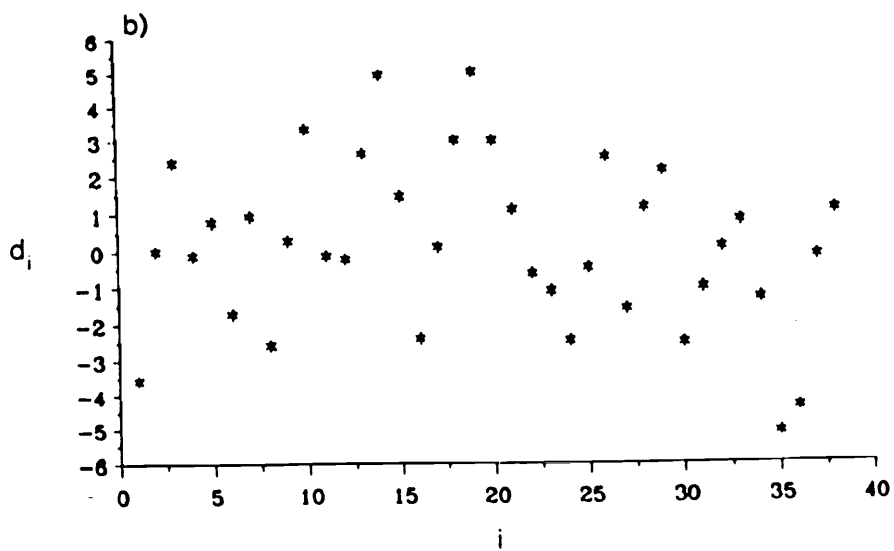
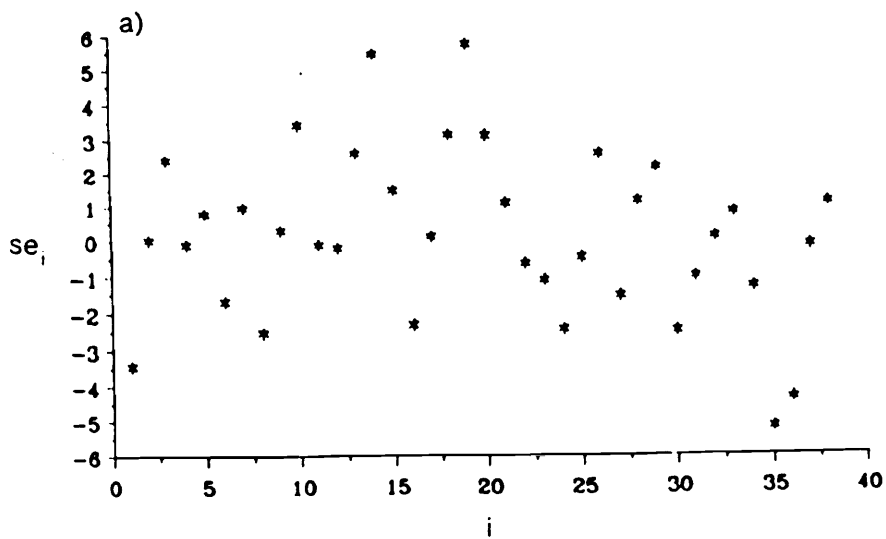


Figure 7.3. Index plot of  $se_i$  vs  $i$  (a) and of  $d_i$  vs  $i$  (b) for the binary logit model fitted to slash pine data.



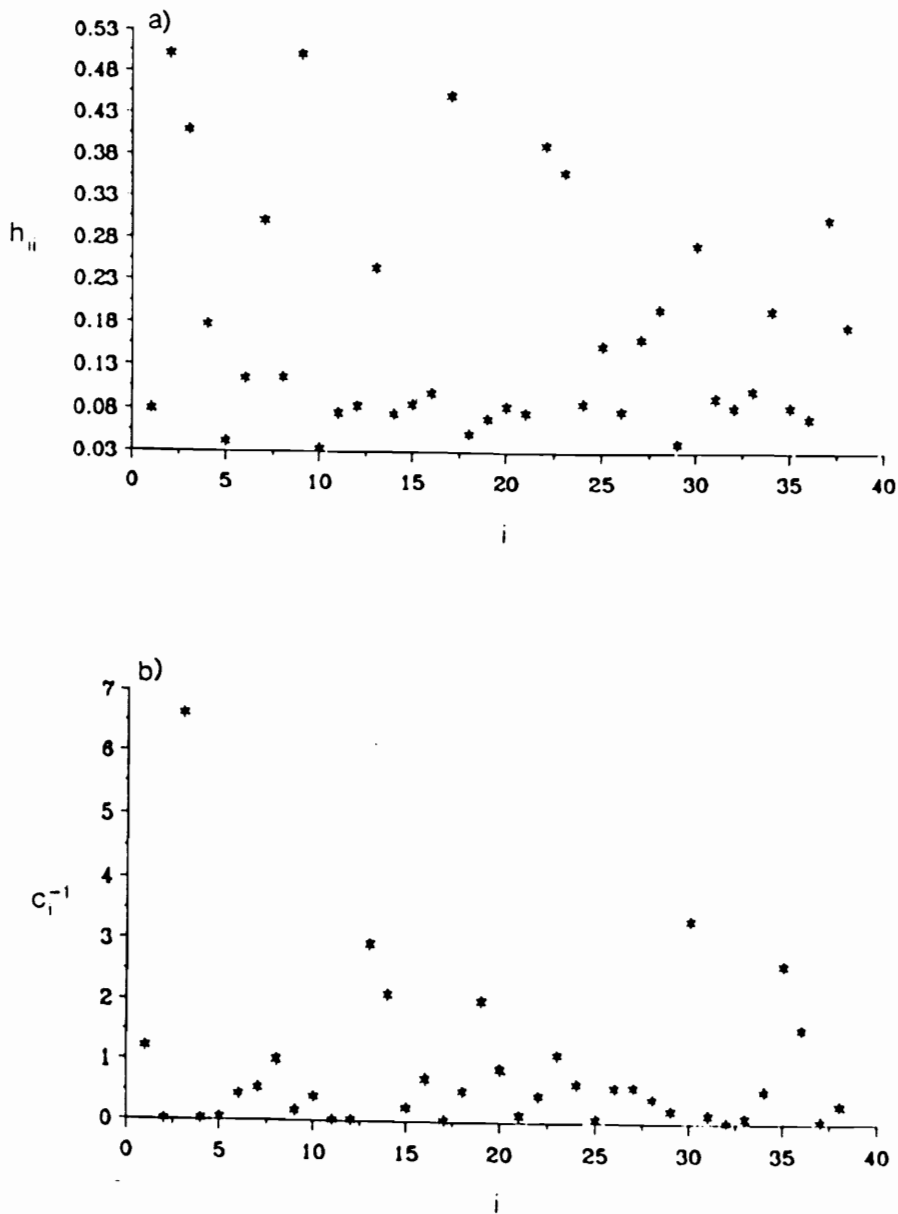
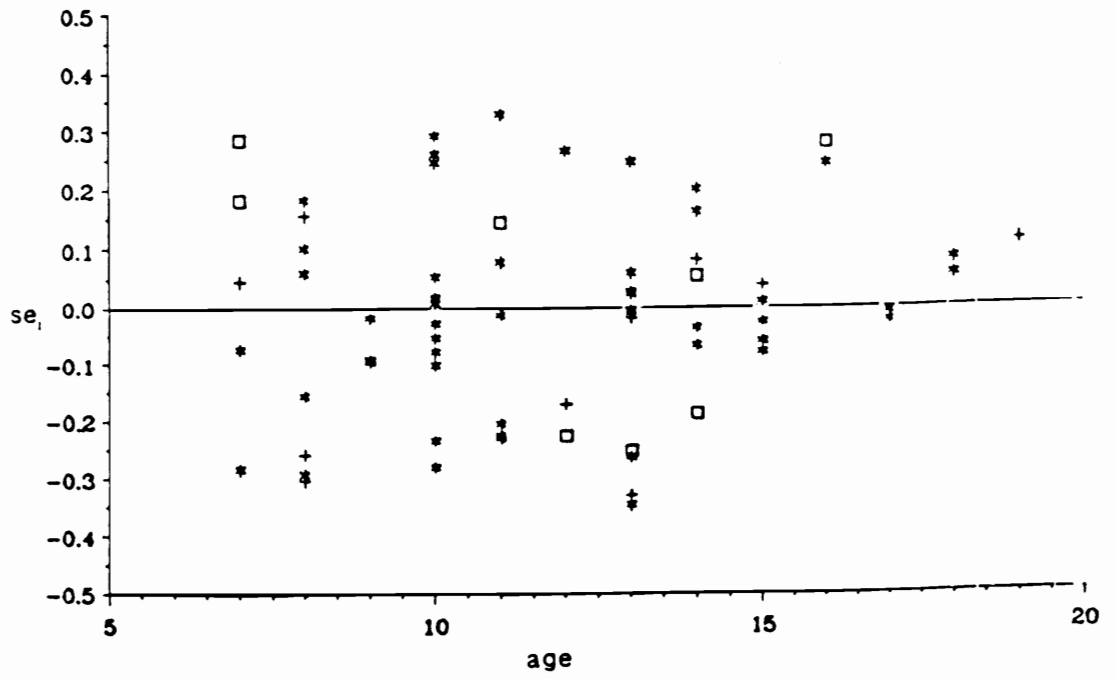
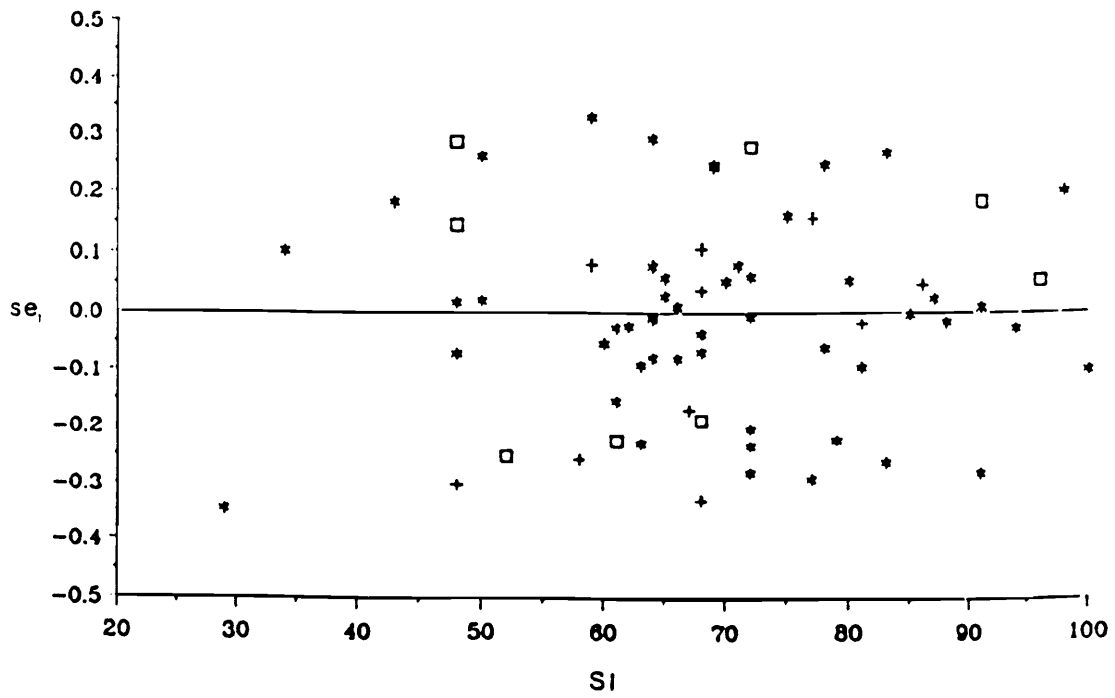


Figure 7.4. Index plot of  $h_{ii}$  vs  $i$  (a) and  $c_i^{-1}$  vs  $i$  (b) for the binary logit model fitted to slash pine data.



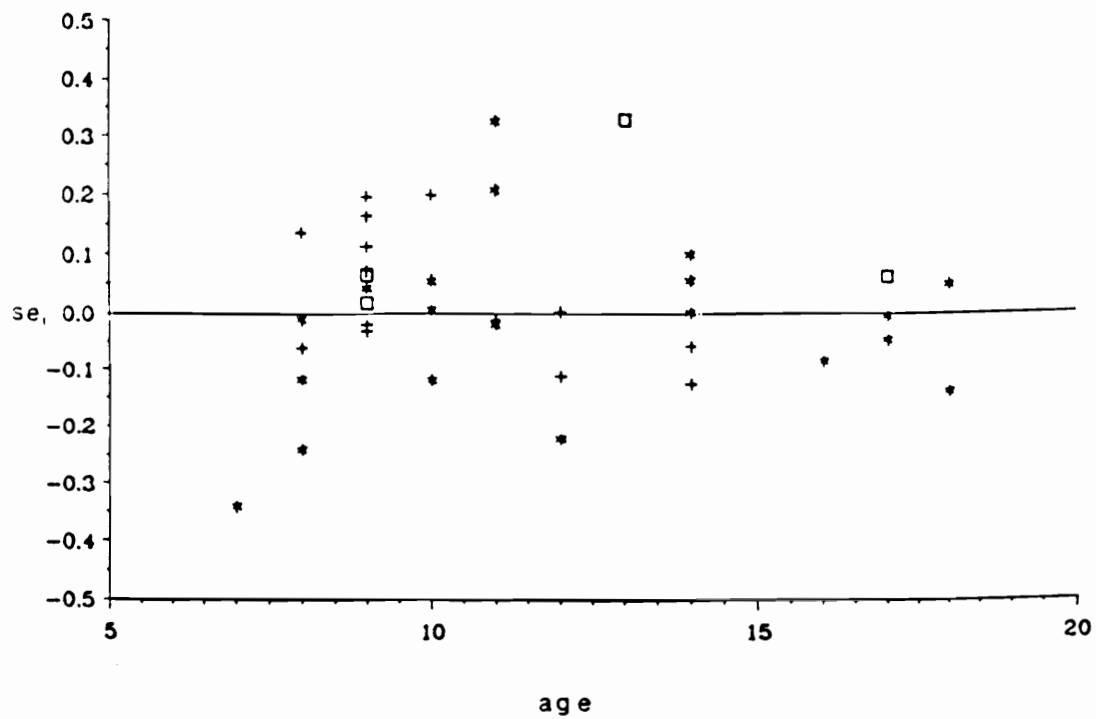
\* : slope site. + : flat site. □ : ridge.

Figure 7.5. Standardized residuals plotted against age and landform for the binary logit model fitted to loblolly pine data.



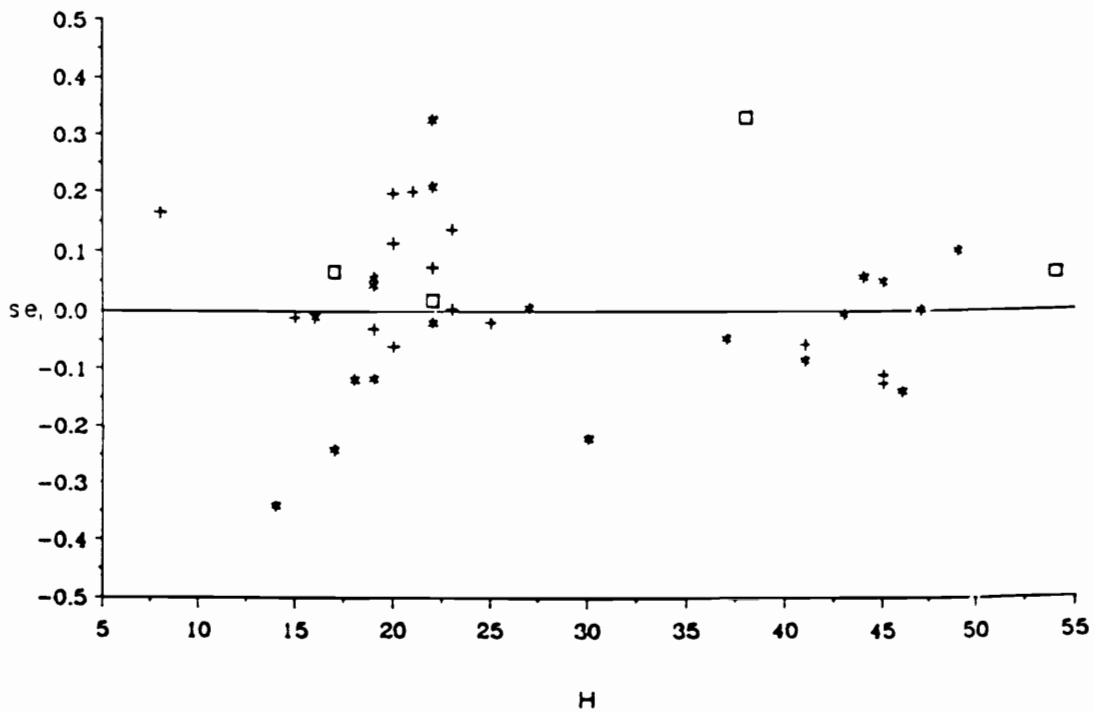
\* : slope site. + : flat site. □ : ridge.

Figure 7.6. Standardized residuals plotted against site index and landform for the binary logit model fitted to loblolly pine data.



\* : slope site. + : flat site. □ : ridge.

Figure 7.7. Standardized residuals plotted against age and landform for the binary logit model fitted to slash pine data.



\* : slope site. + : flat site. □ : ridge.

Figure 7.8. Standardized residuals plotted against average height and landform for the binary logit model fitted to slash pine data.

parent trend was detected. The observations which were found to be associated with the largest residuals were checked again but nothing peculiar about them was revealed.

### 7.3.2 Polychotomous Models

To obtain a better insight on the levels of fusiform rust infection, three types of multinomial choice models namely, ordered probit (OMNP), ordered logit (OMNL) and unordered logit (UMNL) were considered for predicting the proportions of trees that are branch infected, stem infected and dead at a given age. For each of the above models applied to the data of each species, variable screening was conducted on the basis of criteria such as SSR, WSSR, pseudo-  $R^2$ , AIC and also, the maximized value of log-likelihood functions. Tables 7.6-7.7 and 7.8-7.9 show the finally selected forms matrices of OMNP, OMNL and UMNL models, fitted to loblolly and slash pine data respectively. Tables 7.6a, 7.7(a,b,c), 7.8a and 7.9(a,b,c) display the corresponding estimated variance-covariance matrices of the logit formulations.

As seen from these tables, the three models fitted to each species contain explanatory variables which are the same as those used by the binary logit and probit models in Tables 7.4 and 7.5. The only notable difference is that the initial number of rust infected trees per acre (INFECT0) used by the binary models has now been decomposed into its two components, the initial number of branch infected trees per acre (BRANCH0) and the initial number of stem infected trees per acre (STEM0).

As the reader can verify, the logit formulations, OMNL and UMNL, outperform OMNP for both species. In choosing between OMNL and UMNL, all goodness of fit statistics, including AIC, strongly favor UMNL. The superiority of UMNL was also witnessed in testing the hypothesis that there is no significant difference between the two logit models. Both LR and Wald tests rejected this hypothesis at all conventional levels of significance for both species.

Table 7.6. Maximum likelihood coefficient estimates of the OMNL and OMNP models fitted to loblolly pine data.

predictor variables	OMNL		OMNP	
	estimate	std. error	estimate	std. error
INT-1 (BRANCH)	4.265233	0.565751	1.655300	0.031707
INT-2 (STEM)	3.870174	0.565622	0.217552	0.000825
INT-3 (DEAD)	1.614921	0.566553	1.274312	0.003011
NORTH	-0.383306	0.051175	-0.200879	0.003145
KGSW	0.223195	0.071429	0.104515	0.004298
KGSBN	0.299639	0.083088	0.134831	0.004298
KGSBD	0.401695	0.112184	0.396656	0.005427
SLOPE	0.766370	0.064001	0.359554	0.003197
FLAT	-0.153229	0.050131	-0.025384	0.002526
AGE	-0.071811	0.014661	-0.025324	0.000835
SINDEX	-0.014200	0.002185	-0.006002	0.000122
QMD	-0.157199	0.036573	-0.074800	0.001959
RS	-0.886945	0.110611	-0.395236	0.006283
CLEAR0	-0.005216	0.000343	-0.002495	0.000018
BRANCH0	0.007772	0.000985	-0.001604	0.000050
STEM0	0.005893	0.000423	0.003578	0.000022
logL = -19074.46		logL = -19196.72		
AIC = 19087.46		AIC = 19208.72		
pseudo- $R^2$ = 0.246		pseudo- $R^2$ = 0.241		
SSR = 0.768999		SSR = 0.810976		
WSSR = 7.861294		WSSR = 10.698772		

Table 7.6a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.6.

	INT.1	INT.2	INT.3	NORTH	KGSW	KGSBN	KGSBD
INT.1	3.2007E-01						
INT.2	3.1201E-01	3.1993E-01					
INT.3	3.1165E-01	3.1160E-01	3.2098E-01				
NORTH	2.9851E-03	2.9894E-03	2.9990E-03	2.6189E-03			
KGSW	4.2610E-03	4.2566E-03	4.2427E-03	1.8976E-04	5.1021E-03		
KGSBN	-5.2234E-04	-5.2734E-04	-5.4086E-04	-4.5623E-05	4.5704E-03	6.3036E-03	9.9058E-03
KGSBD	1.2550E-02	1.2542E-02	1.2506E-02	3.1940E-03	4.8514E-03	4.3525E-03	7.5570E-05
SLOPE	-1.6661E-03	-1.6562E-03	-1.6203E-03	-1.7805E-04	3.6036E-04	1.7712E-05	1.7783E-04
FLAT	2.1355E-03	2.1490E-03	2.1927E-03	1.6802E-04	3.1400E-04	-5.4655E-04	1.7783E-04
AGE	-3.4407E-03	-3.4396E-03	-3.4402E-03	5.0556E-05	1.1000E-04	1.8981E-04	1.9248E-05
SINDEX	-7.5649E-04	-7.5626E-04	-7.5590E-04	1.4298E-05	-9.3857E-06	1.1610E-05	-2.4350E-05
QMD	-8.1776E-03	-8.1735E-03	-8.1541E-03	-4.6262E-04	-5.9050E-04	-5.7810E-04	-9.6499E-04
RS	-5.6873E-02	-5.6855E-02	-5.6798E-02	-1.2561E-03	-2.1326E-03	-1.4051E-03	-3.4073E-03
CLEAR0	-1.6772E-04	-1.6763E-04	-1.6736E-04	-1.1837E-06	-6.6181E-06	-3.5906E-06	-1.2660E-05
BRANCH0	-2.7162E-04	-2.7162E-04	-2.7156E-04	1.6103E-06	-1.2322E-05	-1.8342E-06	-2.2570E-05
STEM0	-7.7451E-05	-7.7564E-05	-7.7972E-05	-2.9400E-06	-7.7121E-06	-7.2107E-06	-4.4257E-06
SLOPE	4.0961E-03						
FLAT	1.4903E-03	2.5131E-03					
AGE	1.6278E-05	-3.9363E-05	2.1494E-04				
SINDEX	-5.3039E-06	-1.5993E-05	2.3683E-05	4.7742E-06			
QMD	1.3738E-05	-9.8155E-05	3.2037E-04	-2.7881E-05	1.3376E-03		
RS	6.1024E-05	-6.6380E-04	1.9370E-04	1.9370E-04	8.3739E-05	1.2235E-02	
CLEAR0	1.4480E-07	-5.9004E-07	4.0773E-07	2.5311E-07	7.6651E-06	3.3696E-05	1.1765E-07
BRANCH0	-5.1517E-05	-1.6898E-05	-1.5916E-06	5.7487E-07	1.4893E-05	5.8421E-05	1.8439E-07
STEM0	1.7887E-07	4.7645E-06	-4.1697E-07	-6.5474E-08	5.7618E-06	2.0540E-05	5.4509E-09
BRANCH0		STEM0					
BRANCH0	7.7023E-07						
STEM0	-9.5653E-07	1.7893E-07					



Table 7.7. Maximum likelihood coefficient estimates of the UMNL model fitted to loblolly pine data.

predictor variables	BRANCH		STEM		DEAD	
	estimate	std. error	estimate	std. error	estimate	std. error
INTERCEPT	7.976794	1.003823	4.208575	0.692400	-10.045421	2.133754
NORTH	-0.544619	0.087542	-0.309321	0.064233	-0.456870	0.193252
KGSW	0.080439	0.109769	0.384377	0.097746	-0.074521	0.218088
KGSBN	0.101326	0.128898	0.494029	0.111336	-0.134300	0.287292
KGSBD	-0.058592	0.201172	0.717329	0.136825	0.576993	0.266684
SLOPE	0.716086	0.135649	0.821542	0.074520	-0.075421	0.183179
FLAT	-0.746697	0.102489	-0.007467	0.059120	-0.567580	0.127643
AGE	-0.309710	0.027162	-0.053382	0.018492	-0.221107	0.044505
SINDEX	-0.044598	0.004053	0.011803	0.002683	0.023733	0.007257
QMD	0.098220	0.066269	-0.246445	0.044738	0.029034	0.125988
RS	-1.388200	0.199021	-0.955524	0.132993	0.479509	0.437394
CLEAR0	-0.060079	0.000616	-0.006312	0.000125	-0.002101	0.001213
BRANCH0	0.004868	0.001618	-0.001260	0.001245	0.003668	0.003501
STEM0	-0.000550	0.000801	0.008150	0.000509	0.007864	0.001452

logL = -18606.45

AIC = 18645.45

pseudo- $R^2$  = 0.248

SSR = 0.641868

WSSR = 6.852381

Table 7.7a. Covariance matrix of the coefficient estimates (BRANCH) of the UMN/L model displayed on table 7.7.

	INTERCEPT	NORTH	KGSW	KGSBN	KGSBD	SLOPE	FLAT
INTERCEPT	1.00776606						
NORTH	1.1407E-02	7.6636E-03					
KGSW	2.3370E-02	5.1394E-04	1.2049E-02				
KGSBN	8.0575E-03	-5.7337E-04	1.0156E-02	1.6615E-02			
KGSBD	5.4415E-02	1.2463E-03	1.1611E-02	1.0132E-02	4.0470E-02	1.8401E-02	
SLOPE	-8.7182E-03	-2.5325E-04	9.7522E-04	-1.4973E-04	2.4076E-03	7.4087E-03	1.0504E-02
FLAT	-1.1866E-04	7.5580E-04	1.2711E-04	-1.5283E-03	8.8695E-04	4.5449E-05	-1.0926E-04
AGE	-1.0900E-02	1.0582E-04	9.4466E-05	2.5820E-04	-3.0532E-04	-2.7169E-05	-3.9402E-05
SINDEX	-2.4376E-03	1.7364E-05	-6.5667E-05	7.7321E-06	-1.1706E-04	7.3146E-05	-6.4486E-04
QMD	-2.8681E-02	-1.2289E-03	-1.5980E-03	-1.4963E-03	-3.0397E-03	5.2359E-04	-1.6451E-03
RS	-1.8999E-01	-3.7197E-03	-7.3476E-03	-4.6575E-03	-1.2282E-02	2.0121E-06	7.1927E-07
CLEAR0	-5.5011E-04	-5.6388E-06	-2.2156E-05	-1.2592E-05	-4.4396E-05	-1.3663E-05	-6.6670E-05
BRANCH0	-7.9646E-04	-9.7213E-07	-3.9364E-05	-2.6973E-06	-5.4198E-05	8.6124E-06	2.3679E-05
STEM0	-2.7395E-04	-7.9825E-06	-2.1993E-05	-2.0583E-05	-1.6940E-05		

	AGE	SINDEX	QMD	RS	CLEAR0	BRANCH0	STEM0
AGE	7.3777E-04						
SINDEX	8.0660E-05	1.6427E-05					
QMD	-1.0962E-03	-9.8444E-05	4.3916E-03				
RS	5.9409E-04	2.7948E-03	9.1406E-03	3.9609E-02			
CLEAR0	6.5891E-07	7.1076E-07	2.7488E-05	1.1299E-04	3.7946E-07		
BRANCH0	-4.1544E-06	1.6805E-06	4.2170E-05	1.7472E-04	5.2889E-07	2.6179E-06	
STEM0	-3.3923E-06	-3.1580E-07	2.3645E-05	7.3622E-05	2.0637E-07	-2.3726E-07	6.4160E-07

Table 7.7b. Covariance matrix of the coefficient estimates (STEM) of the UMINL model displayed on table 7.7.

	INTERCEPT	NORTH	KGSW	KGSBN	KGSBD	SLOPE	FLAT
INTERCEPT	4.8164E-01						
NORTH	3.8411E-03	4.1259E-03					
KGSW	4.0220E-03	2.0452E-04	9.5543E-03				
KGSBN	-3.3729E-03	-2.0495E-04	8.7988E-03	1.2396E-02			
KGSBD	1.6679E-02	3.8016E-04	9.1039E-03	8.3335E-03	1.8721E-02		
SLOPE	-1.6290E-03	-3.0942E-04	6.2075E-04	9.0392E-05	1.3168E-03	2.5438E-03	3.4952E-03
FLAT	5.3882E-03	2.8339E-04	1.2551E-04	-9.3724E-04	5.4839E-04	2.0874E-03	-1.0211E-04
AGE	-45.5297E-03	7.4242E-05	1.4470E-04	2.9267E-04	9.9669E-06	1.8956E-05	-3.1480E-05
SINDEX	-1.1572E-03	2.7901E-05	1.4425E-05	1.6379E-05	-4.0307E-05	-9.2096E-06	-1.4914E-04
QMD	-1.1117E-02	-7.1998E-04	-8.2570E-04	-8.0519E-04	-1.3632E-03	-5.1424E-05	-1.3316E-03
RS	-8.2400E-02	-1.8636E-03	-3.0763E-03	-1.9804E-03	-4.9770E-03	-6.3621E-05	-1.7592E-06
CLEAR0	-2.5015E-04	-1.0602E-06	-9.3464E-06	-5.0091E-06	-1.8351E-05	-6.9901E-08	-2.7047E-05
BRANCH0	-4.0808E-04	5.1673E-06	-1.8116E-05	-3.0648E-06	-3.6345E-05	-9.0695E-06	6.7977E-06
STEM0	-1.0523E-04	-4.5947E-06	-1.1570E-05	-1.1037E-05	-5.3343E-06	1.6977E-06	

	AGE	SINDEX	QMD	RS	CLEAR0	BRANCH0	STEM0
AGE	3.4195E-04						
SINDEX	3.7703E-05	7.1985E-06					
QMD	-5.0809E-04	-4.4856E-05	2.0015E-03				
RS	3.0292E-04	1.2258E-04	1.7613E-02	1.7687E-02			
CLEAR0	7.6155E-07	3.9850E-07	1.0946E-05	4.9327E-05	1.5625E-08		
BRANCH0	-2.8590E-06	8.9729E-07	2.2792E-05	8.7296E-05	2.8416E-07	1.5500E-06	
STEM0	-5.2009E-07	-1.2194E-07	8.0116E-06	2.8606E-05	7.4633E-08	-1.7272E-07	2.5908E-07

Table 7.7c. Covariance matrix of the coefficient estimates (DEAD) of the UMNL model displayed on table 7.7.

	INTERCEPT	NORTH	KGSW	KGSBN	KGSBD	SLOPE	FLAT
INTERCEPT	4.55290613						
NORTH	3.3406E-02	3.7346E-02					
KGSW	7.2195E-02	2.7911E-03	4.7562E-02				
KGSBN	8.8098E-04	5.7087E-03	4.1219E-02	8.2537E-02			
KGSBD	1.4976E-01	4.2829E-03	4.6558E-02	4.0747E-02	7.1120E-02		
SLOPE	-2.5875E-02	-2.8501E-03	3.4199E-03	3.3723E-04	5.0953E-03	1.2650E-03	
FLAT	-1.7116E-02	5.9162E-04	-2.4489E-03	-3.7615E-03	-1.8562E-03	8.7347E-03	2.3016E-02
AGE	-4.3584E-02	1.2955E-03	1.9360E-03	3.4261E-03	2.2298E-03	1.8049E-04	1.3209E-04
SINDEX	-9.9146E-03	2.4957E-04	-4.9097E-05	3.0813E-04	-9.9209E-05	-8.3444E-06	4.4576E-05
QMD	-1.3815E-01	-6.8201E-03	-9.3484E-03	-1.0292E-02	-1.5037E-02	8.1571E-04	1.1392E-05
RS	-8.7618E-01	-1.7599E-02	-2.8047E-02	-2.0393E-02	-4.2782E-02	3.2702E-03	3.6538E-04
CLEAR0	-2.3145E-03	-1.7792E-05	-8.8417E-05	-4.4205E-05	-1.4917E-04	4.5903E-06	1.0876E-05
BRANCH0	-4.2045E-03	-8.4824E-06	-1.5475E-04	-5.5290E-05	-3.4023E-04	-2.0148E-05	-1.1592E-04
STEM0	-1.3343E-03	-3.3296E-05	-1.0043E-04	-7.3500E-05	-5.8750E-05	1.4834E-05	4.6767E-05

	AGE	SINDEX	QMD	RS	CLEAR0	BRANCH0	STEM0
AGE	1.9807E-03						
SINDEX	2.4231E-04	5.2664E-05					
QMD	-2.7807E-03	-2.1813E-02	1.5873E-02				
RS	4.0666E-03	1.2532E-03	3.9710E-02	1.9131E-01			
CLEAR0	9.5170E-06	3.7536E-06	1.0109E-04	4.7418E-04	1.4714E-06		
BRANCH0	-1.9490E-05	6.7000E-06	2.5281E-04	9.0681E-04	2.7074E-06	1.2257E-05	
STEM0	1.1295E-05	1.3354E-06	5.2922E-05	3.0412E-04	7.0322E-07	-8.7966E-07	2.1083E-06

Table 7.8. Maximum likelihood coefficient estimates of the OMNL and OMNP models fitted to slash pine data.

predictor variables	OMNL		OMNP	
	estimate	std. error	estimate	std. error
INT-1 (BRANCHI)	-0.150355	0.131325	-0.033264	0.008399
INT-2 (STEM)	-0.440165	0.131365	0.174821	0.000791
INT-3 (DEAD)	-3.740128	0.137026	2.072110	0.002520
KGSW	-0.634831	0.130203	-0.378077	0.007949
KGSBN	-0.467229	0.144032	-0.287272	0.008982
KGSBD	-1.443341	0.132852	-0.907530	0.008225
KGSNW	-0.531860	0.158168	-0.347225	0.009508
SLOPE	0.592344	0.134243	0.328183	0.009813
FLAT	-0.464763	0.084898	-0.269049	0.005154
AGE	0.079445	0.014797	0.050330	0.001002
HEIGHT	-0.042496	0.012738	-0.025871	0.000719
BA	-0.013138	0.003495	-0.006847	0.000192
QMD	0.439954	0.042325	0.244994	0.002660
BRANCH0	-0.000747	0.000696	-0.005781	0.000042
STEM0	0.004624	0.000278	0.002619	0.000015
logL = -13086.92		logL = -13102.23		
AIC = 13098.92		AIC = 13114.23		
pseudo- $R^2$ = 0.216		pseudo- $R^2$ = 0.203		
SSR = 0.528301		SSR = 0.530849		
WSSR = 3.356088		WSSR = 3.9827535		

Table 7.8a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.8.

	INT.1	INT.2	INT.3	KGSW	KGSBN	KGSBD	KGSNW
INT.1	1.7246E-02						
INT.2	1.7172E-02	1.7527E-02					
INT.3	1.7220E-02	1.7257E-02	1.8776E-02				
KGSW	-9.3740E-03	-9.3570E-03	-9.2697E-03	1.6953E-02			
KGSBN	-7.9764E-03	-7.9614E-03	-7.9507E-03	1.7319E-02	2.-745E-02		
KGSBD	-1.1017E-02	-1.0980E-02	-1.0777E-02	1.3169E-02	1.3263E-02	1.7650E-02	
KGSNW	-1.0506E-02	-1.0491E-02	-1.0470E-02	1.7016E-02	1.7418E-02	1.2433E-02	2.5017E-02
SLOPE	7.9471E-04	7.8950E-04	7.0432E-04	1.0144E-05	-5.5413E-04	-7.8423E-04	1.1649E-04
FLAT	7.9926E-04	8.1094E-04	8.6403E-04	1.2045E-04	3.4701E-04	2.3305E-03	-3.1077E-03
AGE	-2.9296E-04	-2.9464E-04	-3.1476E-04	1.7047E-04	1.3326E-04	-5.7481E-04	4.5392E-04
HEIGHT	-1.6972E-04	-1.6864E-04	-1.5289E-04	-6.8093E-04	-7.6451E-04	4.5900E-06	-7.8027E-04
BA	8.0662E-05	8.1025E-05	8.2991E-05	2.4158E-04	2.8116E-04	2.7831E-05	2.8181E-04
QMD	-4.7810E-04	-4.8952E-04	-6.1991E-04	1.1813E-03	1.0323E-03	1.1102E-03	8.9588E-04
BRANCH0	-1.6648E-05	-1.6678E-05	-1.5409E-05	-2.8972E-05	-2.7031E-05	-4.6273E-07	-3.6068E-05
STEM0	-1.4879E-06	-1.5996E-06	-2.8193E-06	-1.2209E-05	-1.4705E-05	-8.2382E-06	-9.0701E-06

	SLOPE	FLAT	AGE	HEIGHT	BA	QMD	BRANCH0
SLOPE	1.8021E-02						
FLAT	7.4280E-04	7.2077E-03					
AGE	1.5160E-04	-5.3886E-04	2.1895E-04				
HEIGHT	-2.5163E-04	-1.4025E-05	-1.0810E-04	1.6226E-04			
BA	4.2137E-05	-1.9276E-05	2.3324E-05	-3.8710E-05	1.2215E-05		
QMD	8.0149E-04	1.3675E-03	3.1868E-05	-3.3809E-04	4.1783E-05	1.7914E-03	
BRANCH0	-1.2612E-05	-7.0194E-07	-3.8289E-06	5.0606E-06	-1.6626E-05	-1.9028E-06	4.8442E-07
STEM0	-2.3520E-07	1.2930E-06	2.6854E-06	-6.9955E-07	3.7637E-07	6.0348E-06	8.3796E-03

STEM0	STEM0	
STEM0	7.7284E-08	

Table 7.9. Maximum likelihood coefficient estimates of the UMNL model fitted to slash pine data.

predictor variables	BRANCH		STEM		DEAD	
	estimate	std. error	estimate	std. error	estimate	std. error
INTERCEPT	-3.742810	0.295232	-0.430302	0.162457	-2.741815	0.301218
KGSW	-0.257786	0.263235	-0.877076	0.158742	-0.754868	0.320173
KGSBN	-1.323084	0.330372	-0.818603	0.175124	-0.441504	0.348768
KGSBD	-0.401780	0.264059	-1.704881	0.163579	-3.104621	0.392600
KGSNW	0.133579	0.332495	-0.748652	0.189961	-0.791888	0.397751
SLOPE	0.996363	0.336726	0.316201	0.175469	1.181194	0.219732
FLAT	-0.574241	0.200206	-0.480496	0.099370	-0.583664	0.199810
AGE	0.238733	0.041121	0.115020	0.017703	0.196408	0.034376
HEIGHT	-0.068111	0.037263	-0.057416	0.014374	-0.108360	0.030980
BA	0.004404	0.009329	-0.016693	0.003911	0.001443	0.009151
QMD	0.338453	0.121754	0.539176	0.049372	0.552964	0.087344
BRANCH0	0.006755	0.001780	0.000822	0.000816	-0.001083	0.001576
STEM0	-0.001724	0.000675	0.004410	0.000319	0.005943	0.000615

logL = -12850.92

AIC = 12886.92

pseudo- $R^2$  = 0.225

SSR = 0.355634

WSSR = 2.633930

Table 7.9a. Covariance matrix of the coefficient estimates (BRANCH) of the UMNL model displayed on table 7.9.

	INTERCEPT	KGSW	KGSBN	KGSBD	KGSNW	SLOPE	FLAT
INTERCEPT	8.7221E-02						
KGSW	-2.6454E-02	6.9293E-02					
KGSBN	-8.5018E-03	7.3772E-02	1.0915E-01				
KGSBD	-4.2182E-02	4.1431E-02	3.7283E-02	6.9727E-02			
KGSNW	-4.0582E-02	7.1558E-02	7.3466E-02	3.7089E-02	1.1055E-01		
SLOPE	-9.9622E-03	-3.1753E-03	-5.6373E-03	5.0626E-03	2.2255E-03	1.1338E-01	
FLAT	1.0007E-02	4.7714E-02	8.8033E-03	1.7292E-02	-1.3174E-02	9.0267E-03	4.0082E-02
AGE	-1.7310E-03	2.6214E-03	2.7534E-03	-4.2910E-03	4.8791E-03	-1.9329E-03	-3.4376E-03
HEIGHT	-1.7582E-03	-5.5128E-03	-6.8963E-03	9.1819E-04	-6.1989E-03	3.2039E-03	-8.4316E-02
BA	6.6506E-04	1.6561E-03	2.1422E-03	-2.0312E-04	1.8707E-03	-5.5156E-03	-6.9458E-05
QMD	-2.1715E-03	1.0562E-02	1.0343E-02	7.8789E-03	9.5846E-03	-9.7867E-03	1.0811E-02
BRANCH0	-1.1583E-04	-2.1671E-04	-2.4994E-04	2.7767E-05	-2.3753E-04	1.3643E-04	-5.0201E-05
STEM0	-1.1368E-05	-5.8753E-05	-7.4461E-05	-2.2963E-05	-4.2894E-05	1.2045E-06	7.0195E-06

	AGE	HEIGHT	BA	QMD	BRANCH0	STEM0
AGE	1.6909E-03					
HEIGHT	-9.8790E-04	1.3885E-03				
BA	2.3390E-04	-3.3379E-04	8.7030E-05			
QMD	4.6591E-04	-2.9630E-03	4.5319E-04	1.4824E-02		
BRANCH0	-3.4973E-05	4.7092E-05	-1.2985E-05	-7.5237E-05	3.1684E-06	
STEM0	-6.6603E-06	2.8209E-06	-1.4684E-06	-1.2822E-05	1.4594E-07	4.5563E-07



Table 7.9b. Covariance matrix of the coefficient estimates (STEM) of the UMINI model displayed on table 7.9.

	INTERCEPT	KGSW	KGSBN	KGSBD	KGSNW	SLOPE	FLAT
INTERCEPT	2.6392E-02						
KGSW	-1.5217E-02	2.5199E-02					
KGSBN	-1.3137E-02	2.5839E-02	3.0668E-02				
KGSBD	-1.7278E-02	2.1042E-02	2.1416E-02	2.6758E-02			
KGSNW	-1.6557E-02	2.5228E-02	2.5823E-02	1.9766E-02	3.6085E-02		
SLOPE	6.0786E-04	-1.2228E-04	-1.0871E-03	-1.7004E-03	2.5359E-04	3.0789E-02	
FLAT	1.1142E-03	2.9447E-04	6.3858E-04	3.2520E-03	-4.4206E-03	6.5166E-04	9.8744E-03
AGE	-5.0516E-04	8.8146E-05	-4.3547E-06	-8.4995E-04	5.3085E-04	3.4607E-04	-7.8715E-04
HEIGHT	-2.4937E-04	-8.3102E-04	-9.3831E-04	-1.0479E-05	-9.7311E-04	-3.8038E-04	-1.0834E-05
BA	1.2327E-04	2.9998E-04	3.5588E-04	4.8303E-05	3.5793E-04	5.8983E-05	-2.7812E-05
QMD	-5.8902E-04	1.5966E-03	1.3674E-03	1.5433E-03	1.0856E-03	1.1432E-03	1.9086E-03
BRANCH0	-2.0725E-05	-3.4462E-05	-3.0833E-05	1.2596E-06	4.6447E-05	-2.0817E-05	1.4199E-06
STEM0	-1.0378E-06	-1.6345E-05	-1.9981E-05	-1.1519E-05	-1.2254E-05	1.3148E-06	2.2497E-06

	AGE	HEIGHT	BA	QMD	BRANCH0	STEM0
AGE	3.3140E-04					
HEIGHT	-1.3585E-04	2.0661E-04				
BA	2.6505E-05	-4.8194E-05	1.5296E-05			
QMD	2.8924E-05	-4.4840E-04	5.1649E-05	2.4376E-03		
BRANCH0	-5.2992E-06	6.4989E-06	-2.1097E-06	-6.3492E-05	6.6586E-07	
STEM0	-2.2134E-07	-1.0378E-06	-1.6345E-05	-1.9981E-05	-1.1519E-05	1.0176E-07

Table 7.9c. Covariance matrix of the coefficient estimates (DEAD) of the UMNL model displayed on table 7.9.

	INTERCEPT	KGSW	KGSBN	KGSBD	KGSNW	SLOPE	FLAT
INTERCEPT	9.0732E-02						
KGSW	5.3512E-02	1.0251E-01					
KGSBN	4.5841E-02	1.1003E-01	1.2164E-01				
KGSBD	5.8324E-02	6.9781E-02	6.7260E-02	1.5413E-01			
KGSNW	5.7749E-02	1.0911E-01	1.1345E-01	6.8389E-02	1.5821E-01		
SLOPE	-2.0405E-03	-1.2894E-03	-9.6870E-04	-4.3109E-04	-2.5364E-03	4.8282E-02	
FLAT	5.8080E-03	7.9598E-03	7.9365E-03	9.1964E-03	1.9097E-02	-3.6576E-03	3.9924E-02
AGE	-1.6353E-03	-1.7204E-03	-2.1781E-03	-3.1760E-02	-2.6088E-03	9.1960E-05	-2.8628E-03
HEIGHT	2.7558E-04	4.5538E-03	5.5815E-03	7.8194E-04	5.0982E-03	-6.0452E-04	1.0221E-03
BA	-2.7542E-04	-1.8564E-03	-2.2150E-03	-5.5649E-05	-2.1088E-03	1.5697E-04	-3.9788E-04
QMD	-3.8541E-03	-2.6400E-03	-2.4593E-03	-1.8900E-03	-1.7134E-03	1.8173E-03	-3.9994E-03
BRANCH0	9.7234E-05	1.3819E-04	1.5496E-04	2.1795E-05	1.7816E-04	-4.8083E-05	5.1987E-06
STEM0	-1.4246E-05	-7.2069E-05	-7.6295E-05	-4.8075E-05	-5.9977E-05	8.9036E-07	-4.5504E-06

	AGE	HEIGHT	BA	QMD	BRANCH0	STEM0
AGE	1.1817E-03					
HEIGHT	-7.8768E-04	9.5976E-04				
BA	1.7394E-04	-2.6153E-04	8.3741E-05			
QMD	7.5612E-04	-1.5395E-03	1.5815E-04	7.6290E-03		
BRANCH0	-2.4396E-05	2.8121E-05	-9.1082E-06	-2.0846E-05	2.4838E-06	
STEM0	1.0463E-06	-2.4006E-07	9.4351E-07	1.1054E-05	-1.6012E-07	3.7823E-07

Figures 7.9-7.24 for loblolly pine and 7.25-7.40 for slash pine at the end of this chapter show the standardized residuals of the OMNL and UMNL models for each level of rust infection plotted against age, landform and site index (loblolly) or average height of dominant and codominant trees (slash). The better goodness-of-fit performance of the UMNL over the OMNL model can be verified from these graphs. No particular trends with respect to age, site quality and landform of the plantation were found to be present.

### 7.3.3 Discussion

The binary models in Tables 7.4 and 7.5 are useful in highlighting important site factors and also, in describing their role on the occurrence of rust infection and mortality. In addition, the UMNL models given in Tables 7.7 and 7.9, containing the same predictors as the binary models, can provide further insight on the disease dynamics by focusing on the effects of predictor variables on each level of rust infection separately.

The geographic location of loblolly pine plantations was found to significantly affect the occurrence of fusiform rust. In particular, as shown in Tables 7.3 and 7.8, loblolly pine plantations in South-East Texas seem to be more susceptible to rust infection than plantations located in North-East Texas. No such geographic trend was found to be significant for slash pine. Note that Wells and Dinus (1978) also reported geographic variation as an important factor affecting the resistance of loblolly pine to fusiform rust.

The effect of site preparation treatments on the incidence of fusiform rust appears to be positive on loblolly pine (Table 7.4) and negative on slash pine (Table 7.5). More specifically, Table 7.7 for loblolly pine shows a significant positive effect of KGSW, KGSBN and KGSBD on the proportion of stem infected trees. On the other hand, Table 7.9 for slash pine shows a significant negative effect

of KGSW, KGSNW and KGSBN on the proportion of stem infected and dead trees and of KGSBD on the proportion of branch and stem infected trees.

The negative impact that site preparation treatments seem to exert on rust infection rate on slash pine contrasts what seems to be a general belief, that increased site preparation intensity usually results in increased rust incidence (Schmidt et al. 1988, Burton et al. 1985, May et al. 1973, Miller 1972). For loblolly pine, the relationship between rust incidence and site preparation has not been thoroughly examined and is not as well defined as for slash pine. However, the increase in rust infection rate with site preparation noted by this study agrees in general with the conclusions reached by Zutter et al. (1987) in studying the effects of various weed control treatments on rust incidence in loblolly pine plantations in Alabama, Georgia and North Carolina. Based on their findings, the authors concluded that site preparation methods that aim in increased loblolly pine growth are usually accompanied by increased rust incidence and severity. In any case, the results obtained in this study should not be faithfully trusted for meaningful interpretations of the role of site preparation methods on fusiform rust occurrence. The reason is that site preparation treatments were subjectively assigned to plantations on the basis of equipment availability, previous land use, vegetation and soil factors. Thus, what the models indicate as significant site preparation effect may in fact be the combined effects of many other unidentified factors which are confounded with site preparation treatments. This probably explains the unexpected negative effect of site preparation methods on rust incidence in slash pine.

Site landscape was also found to be an important factor with similar effect on the occurrence of fusiform rust in both species (Tables 7.4, 7.5, 7.7 and 7.9). In particular, SLOPE sites are seen to be positively related to all levels of rust infection, whereas FLAT sites are negatively related. Considering that SLOPE sites are better drained than FLAT sites, the above results are in agreement with previous studies reporting lower fusiform rust incidence on poorly drained sites and higher on well drained sites (Hollis et al. 1975, May et al. 1973, Schmidt et al. 1988).

Stand density enters the models of the two species in a different way. For loblolly pine (Tables 7.4 and 7.7) through quadratic mean diameter (QMD) and relative spacing (RS) and in slash pine through QMD and basal area per acre (BA). As suggested by the opposite signs of the corresponding coefficient estimates, it is very difficult to assign an overall positive or negative role to stand density, a suggestion which is in agreement with the conclusions of Wakeley (1969) and Miller (1972) who reported no consistent relationship between the incidence of fusiform rust and spacing in plantations of slash pine in Georgia.

Site quality, expressed as site index on loblolly pine (Tables 7.4 and 7.7) and average height of dominant and codominant trees on slash pine (Tables 7.5 and 7.9), was found to be negatively related to all levels of rust infection except for the positive effect that site index has on the proportion of dead loblolly pine trees (Table 7.9). This, overall negative relation is in disagreement with results from previous studies (Borders and Bailey 1986, Nance et al. 1981) and with what is generally accepted, i.e., that rust incidence is positively related to increased growth (Schmidt et al. 1988, Zutter et al. 1987). No meaningful explanation for this discrepancy can be offered at this point.

Fusiform rust infection rate at all levels was found to decrease with time in loblolly pine (Tables 7.4 and 7.7) and increase with time in slash pine (Tables 7.5 and 7.9). Similar conclusions were reached by Hunt and Lenhart (1986) after compiling data from four surveys on loblolly and slash pine plantations in East Texas between 1969 and 1984, the 1984 survey containing the initial measurements used in this study. They, too, found rust incidence increasing with time on slash pine and either decreasing or about constant on loblolly pine.

The initial number of healthy trees (CLEAR0) is seen to be a significant predictor variable, negatively related to all levels of rust infected loblolly pine trees (Tables 7.4 and 7.7). Given the declining rust infection rate over time on loblolly pine, this is an expected result. The larger the initial number of healthy trees, the smaller the proportions of branch infected, stem infected or dead trees are expected. For slash pine, where rust infection is increasing with time, the effect of CLEAR0 was found to be positive but not significant.

The two components of INFECT0 namely, the initial number of branch infected (BRANCH0) and stem infected (STEM0) trees are required by the UMNL models of both species (Tables 7.7 and 7.9). BRANCH0 is positively related to the proportion of branch infected loblolly and slash pine trees. STEM0 is negatively related to the proportion of branch infected trees and positively related to the proportions of stem infected and dead trees of both species. Clearly, the larger the initial number of stem infected trees the higher are expected to be the proportions of stem infected and dead trees. The positive effect of BRANCH0 on the proportion of branch infected trees is also meaningful given the negative effects of CLEAR0 (in loblolly only, Table 7.7) and STEM0 variables for this level of infection.

## ***7.4 Models that Predict Fusiform Rust Transition***

### ***Proportions***

To further understand the dynamics of fusiform rust incidence in loblolly and slash pine plantations, more detailed information than that provided by the models developed in the previous section is often needed. For instance, knowledge of the proportion of trees which were initially classified as healthy and subsequently became infected or died as a result of rust infection, and of the effects that site factors have on these proportions would provide a better insight on the potential damage this disease can cause and, also, could suggest alternative management practices for limiting the spread of the disease. Similarly, information about the proportion of branch infected trees moving to a higher level of infection is regarded to be extremely useful to direct management actions to prevent further devaluation of such, lightly infected trees. Throughout this study, the proportions of trees moving from one level of rust infection to another will be called transitional proportions.

Using the permanent plot data from loblolly and slash pine plantations in East Texas, the transitional proportions of each plot were summarized in a transition matrix of the form given in Table 7.10.

Table 7.10. General form of the rust infection plot transition matrix.

initial measurement	first remeasurement			
	CLEAR	BRANCH	STEM	DEAD
CLEAR	CLR-CLR	CLR-BRA	CLR-STEM	CLR-DEAD
BRANCH	-	BRA-BRA	BRA-STEM	BRA DEAD
STEM	-	-	STEM-STEM	STEM-DEAD

The rows of this matrix correspond to the initial levels of rust infection and the columns to the levels of rust infection at the time of the first remeasurement. The  $ij$ -th element in this matrix is the proportion of trees, initially classified as falling into the  $i$ -th level of rust infection (CLEAR, BRANCH or STEM), which three years later were classified as falling into the  $j$ -th level of rust infection (CLEAR, BRANCH, STEM or DEAD). For example, BRA-DEAD is the proportion of branch infected trees that died. Note that because a tree cannot move from a higher level of infection to a lower, the elements of the lower triangular part of the transition matrix are not defined. Also, the elements of each row of the transition matrix must sum to one since they are proportions of the number of trees having a common initial level of rust infection.

The objective of this part of the study was to develop qualitative response models that predict the transitional proportions at each row of the transition matrix. The models finally selected for their goodness of fit performance among a variety of logit and probit, unordered and ordered models were:

- quatri-nomial ordered and unordered logit for the first row of the transition matrix

- tri-nomial ordered and unordered logit for the second row of the transition matrix and
- a binomial logit for the third row of the transition matrix.

These models are shown fitted in Tables 7.11-7.15 for loblolly pine and in Tables 7.16-7.20 for slash pine.

### 7.4.1 Discussion

The transition of healthy and branch infected loblolly pine trees to higher levels of infection (i.e., CLR = BRA, CLR-STEM, BRA-STEM in Tables 7.11-7.14) is seen to increase as we move towards South-East Texas. No such geographic effect was found to be significant on slash pine.

As before, no meaningful conclusions concerning the effect of site preparation classes on transitional proportions could be drawn due to the subjective manner with which site preparation treatments were assigned to plantations of both species.

Transition to higher levels of rust infection was found to be positively related to SLOPE sites and negatively related to FLAT sites for all models fitted to the data of both species. Here again, soil drainage is considered to be the driving factor for these relations.

For loblolly pine, the proportion of clear trees becoming branch infected and stem infected (CLR-BRA and CLR-STEM, Table 7.12) and of stem infected trees that died (STEM-DEAD, Table 7.15) is decreasing over time. On the other hand, the proportion of branch infected trees moving to higher levels of rust infection (BRA-STEM and BRA-DEAD, Table 7.14) is increasing with time. For slash pine the effect of age is reversed. The proportions of clear and stem infected trees moving to higher infection levels (CLR-BRA, CLR-STEM, CLR-DEAD in Table 7.17 and



Table 7.11. Maximum likelihood coefficient estimates of the quatri-nomial ordered logit model that predicts the transitional proportions of healthy loblolly pine trees.

predictor variables	estimate	std. error
INT.1 (CLR-BRA)	6.102617	0.602737
INT.2 (CLR-STEM)	5.590692	0.602516
INT.3 (CLR-DEAD)	3.832865	0.603644
NORTH	-0.548252	0.069522
KGSW	0.288359	0.084850
KGSBN	0.418701	0.099696
KGSBD	0.666601	0.121588
SLOPE	0.942497	0.074243
FLAT	-0.163629	0.060837
AGE	-0.205122	0.028366
HEIGHT	0.043460	0.009567
SINDEX	-0.026131	0.003401
QMD	-0.315349	0.046082
RS	-1.061606	0.091540
CLEAR0	-0.005378	0.000302

---

logL = -13252.09

AIC = 13264.09

pseudo- $R^2$  = 0.209

SSR = 0.617542

WSSR = 8.071242



Table 7.12. Maximum likelihood coefficient estimates of the quadri-nomial unordered logit model fitted to loblolly pine data.

predictor variables	CLR-BRA		CLR-STEM		CLR-DEAD	
	estimate	std. error	estimate	std. error	estimate	std. error
INTERCEPT	9.024620	0.998488	6.138524	0.819534	-9.730852	2.050060
NORTH	-0.749205	0.105100	-0.572776	0.099093	0.118150	0.209983
KGSW	0.245446	0.123863	0.539658	0.131924	-0.005095	0.225333
KGSBN	0.381282	0.143819	0.663226	0.152069	0.241784	0.299963
KGSBD	-0.442839	0.263479	0.673764	0.196575	0.552375	0.276644
SLOPE	0.763786	0.147977	1.162402	0.095721	0.240751	0.202606
FLAT	-0.787930	0.118164	-0.091684	0.082285	-0.667319	0.156003
AGE	-0.467815	0.046222	-0.142510	0.038810	0.105711	0.091946
HEIGHT	0.018224	0.015961	0.038851	0.012950	0.078774	0.029633
SINDEX	-0.057631	0.005623	-0.017474	0.004482	0.009035	0.012614
QMD	0.221109	0.076372	-0.556322	0.063489	-0.225258	0.140155
RS	-1.355270	0.155463	-1.294513	0.120204	0.605964	0.318031
CLEAR0	-0.005415	0.000515	-0.007067	0.000407	0.002846	0.000908

logL = -12795.43

AIC = 12831.43

pseudo- $R^2$  = 0.258

SSR = 0.462578

WSSR = 6.979003

Table 7.12a. Covariance matrix of the coefficient estimates (CLR-BRA) of the UMN model displayed on table 7.12.

	INTERCEPT	NORTH	KGSW	KGSBN	KGSBD	SLOPE	FLAT
INTERCEPT	9.9698E-01						
NORTH	-2.1560E-02	1.1046E-02					
KGSW	1.0994E-02	-8.1088E-04	1.5342E-02				
KGSBN	4.0304E-03	-1.0074E-03	1.3592E-02	2.0684E-02			
KGSBD	-3.6480E-02	6.4877E-04	-1.4430E-02	-1.3785E-02	6.9421E-02		
SLOPE	1.3693E-02	-5.1660E-04	1.2219E-03	1.0482E-04	-2.4144E-03	2.1897E-02	
FLAT	-7.4907E-03	1.1012E-03	-1.1761E-04	-1.2339E-03	2.5384E-04	-8.3208E-03	1.3963E-02
AGE	-3.3025E-02	1.0226E-03	-8.1314E-04	-3.4901E-04	3.5973E-04	-3.4853E-04	4.7300E-05
HEIGHT	6.8882E-03	-5.5000E-04	2.5835E-04	1.4062E-04	-2.3550E-04	9.9789E-05	-7.7665E-05
SINDEX	-4.2001E-03	1.0152E-04	-1.4072E-04	-5.3473E-05	9.2374E-05	-1.0755E-05	1.9565E-05
QMD	2.3100E-02	-2.5800E-03	7.7968E-04	1.0636E-03	-1.0790E-03	1.1269E-04	-6.8814E-04
RS	-1.3497E-01	3.5231E-03	-3.4117E-03	-3.1483E-03	8.6023E-03	-4.8781E-04	4.0549E-04
CLEAR0	-4.0160E-04	6.5597E-06	-1.1494E-05	-8.0888E-06	3.5191E-05	-2.8233E-06	5.5297E-06

	AGE	HEIGHT	SINDEX	QMD	RS	CLEAR0
AGE	2.1365E-03					
HEIGHT	-5.2983E-04	2.5475E-04				
SINDEX	2.2386E-04	-5.4454E-05	3.1618E-05			
QMD	-3.0679E-04	7.4557E-04	-1.1839E-05	5.8327E-03		
RS	2.6554E-03	-4.4118E-04	3.6923E-04	-5.0182E-03	2.4169E-02	
CLEAR0	7.1250E-06	1.8006E-06	1.0227E-06	1.7856E-07	-6.4397E-05	2.6523E-07

Table 7.12b. Covariance matrix of the coefficient estimates (CLR-STEM) of the UMNL model displayed on table 7.12.

	INTERCEPT	NORTH	KGSW	KGSBN	KGSBD	SLOPE	FLAT
INTERCEPT	6.7164E-01						
NORTH	-1.9137E-02	9.8194E-03					
KGSW	6.6231E-03	-4.5487E-04	1.7404E-02				
KGSBN	1.2899E-02	-9.3516E-04	1.6474E-02	2.3125E-02			
KGSBD	6.9344E-03	-5.0455E-04	1.6033E-02	1.5495E-02	3.8642E-02		
SLOPE	3.1649E-03	-3.0617E-04	1.1112E-03	2.1142E-04	2.2168E-03	3.9473E-03	
FLAT	-7.7951E-03	1.0089E-03	-2.2032E-04	-1.7289E-03	-5.0234E-04	-2.9488E-03	6.7708E-03
AGE	-2.3614E-02	9.3273E-04	-4.4541E-04	-6.6877E-05	-2.6092E-04	-3.3119E-04	5.0461E-04
HEIGHT	5.1841E-03	-4.5005E-04	2.7802E-04	2.3940E-04	4.7167E-05	2.7743E-05	-5.0641E-05
SINDEX	-2.7865E-03	6.4094E-05	-7.5802E-05	-2.6948E-05	-4.5382E-05	-7.9548E-06	7.0121E-05
QMD	-1.5600E-02	2.3941E-03	-9.3607E-04	-1.2050E-03	-1.3068E-03	-9.0752E-05	2.1145E-04
RS	-8.2258E-02	3.6953E-03	-1.1616E-03	-6.1069E-04	-3.7163E-03	-2.8693E-05	1.4904E-03
CLEAR0	-2.5214E-04	3.1739E-06	-5.5094E-06	-3.7635E-06	-1.7723E-05	-1.5499E-07	1.0525E-06

	AGE	HEIGHT	SINDEX	QMD	RS	CLEAR0
AGE	1.5062E-03					
HEIGHT	-3.7528E-04	1.6770E-04				
SINDEX	1.5531E-04	-3.7167E-05	2.0088E-05			
QMD	2.4389E-04	-4.9412E-04	5.9615E-06	4.0309E-03		
RS	1.8847E-03	-4.2485E-04	2.2538E-04	3.7038E-03	1.4494E-02	
CLEAR0	5.5014E-06	-1.4013E-06	7.3000E-07	1.0533E-05	3.7580E-05	1.6565E-07

Table 7.12c. Covariance matrix of the coefficient estimates (CLR-DEAD) of the UMNL model displayed on table 7.12.

	INTERCEPT	NORTH	KGSW	KGSBN	KGSBD	SLOPE	FLAT
INTERCEPT	4.20274601						
NORTH	-8.9225E-02	4.4093E-02					
KGSW	3.1058E-03	5.6226E-03	5.0775E-02				
KGSBN	-4.0637E-02	6.6866E-03	4.5983E-02	8.9978E-02	7.6532E-02		
KGSBD	-1.4362E-02	5.5479E-03	-4.7748E-02	4.5192E-02	9.0802E-03	4.1049E-02	
SLOPE	-8.6763E-03	2.8486E-03	-6.9656E-03	2.5566E-03	-6.4832E-03	-9.4670E-03	2.4337E-02
FLAT	4.9466E-03	-2.5616E-03	1.5549E-03	-4.3995E-03	2.1355E-03	2.9373E-05	-3.6165E-03
AGE	-1.4749E-01	3.8508E-03	-6.1326E-04	2.3325E-03	6.9550E-03	5.2782E-05	-1.0117E-03
HEIGHT	-3.4995E-02	2.2768E-03	-1.6603E-03	1.3154E-03	1.4800E-04	2.9891E-05	-3.7006E-04
SINDEX	-2.0122E-02	4.3711E-04	-2.8458E-04	1.9874E-04	-1.3592E-02	-2.3496E-04	1.5639E-03
QMD	1.0651E-01	-1.2235E-02	1.0731E-02	-1.4421E-02	1.5553E-02	9.2780E-04	-2.4481E-03
RS	-5.3806E-01	1.8687E-02	-8.5186E-03	1.0990E-02	8.3278E-05	1.1745E-05	-2.7610E-05
CLEAR0	-1.2912E-03	1.3473E-05	-3.9574E-05	2.0486E-05			
	AGE	HEIGHT	SINDEX	QMD	RS	CLEAR0	
AGE	8.4541E-03						
HEIGHT	2.3173E-03	8.7811E-04					
SINDEX	1.0984E-03	2.0122E-02	1.5911E-04				
QMD	-1.8839E-03	-2.2718E-03	-1.3529E-04	1.9643E-02			
RS	1.1229E-02	2.7302E-03	1.4947E-03	-2.4891E-02	1.0114E-01		
CLEAR0	2.1802E-05	4.2045E-06	3.2430E-06	-5.1839E-05	2.2068E-04	8.2446E-07	

Table 7.13. Maximum likelihood coefficient estimates of the quatri-nomial ordered logit model that predicts the transitional proportions of branch infected loblolly pine trees.

predictor variables	estimate	std. error
INT.1 (BRA-STEM)	-7.610186	2.139148
INT.2 (BRA-DEAD)	-11.854576	2.180963
NORTH	-0.903714	0.352658
SLOPE	1.074169	0.207306
FLAT	-1.025874	0.185241
AGE	1.049363	0.161902
HEIGHT	-0.419189	0.061349
SINDEX	0.133388	0.021584
QMD	0.528343	0.216072
RS	-1.216635	0.356701
STEM0	-0.004693	0.001529

---

logL = -459.28

AIC = 468.28

pseudo- $R^2$  = 0.199

SSR = 10.031994

WSSR = 98.793379

Table 7.13a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.13.

	INT.1	INT.2	NORTH	SLOPE	FLAT	AGE	HEIGHT
INT.1	4.57595417						
INT.2	4.97013521	4.75659961					
NORTH	3.9424E-01	4.0353E-01	1.2437E-01				
SLOPE	-6.9553E-02	-8.2478E-02	-1.8490E-02	4.2976E-02			
FLAT	1.6082E-01	1.6939E-01	4.1520E-02	-1.0701E-01	3.4314E-02		
AGE	-3.2308E-01	-3.3105E-01	-2.7355E-02	1.4065E-02	-1.7953E-02	2.6212E-02	
HEIGHT	1.0469E-01	1.0783E-01	1.1865E-02	-6.2342E-03	7.5449E-02	-1.0200E-02	3.7637E-03
SINDEX	-4.3794E-02	-4.4495E-02	-3.5877E-03	2.9319E-03	-3.3005E-03	3.7449E-03	-1.3203E-03
QMD	-1.9408E-01	-1.9730E-01	-3.9270E-02	1.3142E-02	-2.3065E-02	1.4269E-02	-9.7557E-03
RS	2.5999E-01	2.5117E-01	2.9042E-02	-3.1517E-21	1.85436E-02	-8.72550E-03	6.1454E-03
STEM0	3.3373E-04	3.8431E-04	7.7970E-05	-6.4424E-05	1.3663E-04	-9.8679E-05	3.4286E-05

	SINDEX	QMD	RS	STEM0
SINDEX	4.6587E-04			
QMD	1.7250E-03	4.6687E-02		
RS	-1.6533E-03	-1.4329E-03	1.2724E-01	
STEM0	-1.4151E-05	-1.7440E-06	2.9172E-04	2.3378E-06



Table 7.14. Maximum likelihood coefficient estimates of the tri-nomial unordered logit model which predicts transitional proportions of branch infected loblolly pine trees.

predictor variables	BRA-STEM		BRA-DEAD	
	estimate	std. error	estimate	std. error
INTERCEPT	-8.533462	2.358365	-28.612847	12.215820
NORTH	-0.913029	0.366794	-0.965512	0.256234
SLOPE	1.074169	0.207306	0.594602	0.040722
FLAT	-1.241152	0.230190	-0.998191	0.620590
AGE	1.276880	0.179165	1.140251	0.697792
HEIGHT	-0.516187	0.068704	-0.401486	0.255759
SINDEX	0.162534	0.024594	0.234079	0.114268
QMD	0.545386	0.136716	1.699430	0.402358
RS	-1.644194	0.394803	0.540369	1.462532
STEM0	-0.005155	0.001649	-0.012172	0.007742

$\log L = -420.63$

AIC = 438.63

pseudo- $R^2 = 0.214$

SSR = 5.351839

WSSR = 34.948338

Table 7.14a. Covariance matrix of the coefficient estimates (BRA-STEM) of the UMNL model displayed on table 7.14.

	INTERCEPT	NORTH	SLOPE	FLAT	AGE	HEIGHT	SINDEX
INTERCEPT	5.56188547						
NORTH	4.4025E-01	1.3454E-01					
SLOPE	-5.1563E-02	-1.9888E-02	4.2976E-02				
FLAT	1.6143E-01	4.5788E-02	-1.1160E-01	5.2987E-02			
AGE	-3.8942E-01	-3.1006E-02	1.5098E-02	-1.8741E-02	3.2100E-02		
HEIGHT	1.2821E-01	1.3553E-02	-6.7462E-03	7.8747E-03	-1.2444E-02	4.7202E-03	
SINDEX	-5.3764E-02	-4.1279E-03	3.0741E-03	-3.3495E-03	4.6048E-03	-1.6381E-03	6.0486E-04
QMD	-2.3620E-01	-4.3887E-02	1.1311E-02	-2.3881E-02	1.7046E-02	-1.1459E-02	2.0933E-03
RS	2.9666E-01	3.0141E-02	-3.6409E-02	1.7901E-01	1.1788E-02	7.6337E-02	2.0557E-03
STEM0	5.3335E-04	9.5063E-05	-6.1097E-05	1.4508E-04	-1.2543E-04	4.3053E-05	-1.8072E-05

	QMD	RS	STEM0
QMD	1.8691E-02		
RS	-2.5709E-03	1.5587E-01	
STEM0	-1.2745E-06	3.4558E-04	2.7192E-06

Table 7.14b. Covariance matrix of the coefficient estimates (BRA-DEAD) of the UMNL model displayed on table 7.14.

	INTERCEPT	NORTH	SLOPE	FLAT	AGE	HEIGHT	SINDEX
INTERCEPT	149.22625822						
NORTH	9.3016E-01	6.5656E-02					
SLOPE	-2.36140100	-9.8370E-01	1.6583E-03				
FLAT	2.59944600	1.29378000	-6.7111E-01	3.8513E-01			
AGE	-9.67789000	-1.9988E-03	1.2185E-01	-1.7786E-01	4.8691E-01		
HEIGHT	3.15882000	3.5384E-02	-2.9499E-02	4.1770E-02	-2.0173E-01	6.5413E-02	
SINDEX	-1.59639000	-5.9216E-02	1.9516E-02	-1.2651E-02	9.0766E-02	-3.1300E-02	1.3057E-02
QMD	-8.97865000	-1.5374E-01	2.8347E-02	-6.2368E-02	4.9482E-01	-1.9391E-01	7.7369E-02
RS	-8.14632500	-2.18667000	5.2581E-01	-6.4933E-01	1.5116E-01	-1.1280E-02	3.8991E-03
STEM0	1.7260E-02	6.6616E-03	-1.0084E-03	2.2470E-03	-1.6918E-03	7.0767E-04	-3.7881E-04
	QMD	RS	STEM0				
QMD	1.6189E-01						
RS	1.6810E-01	2.13899985					
STEM0	-1.1438E-03	-6.3226E-03	5.9939E-05				

Table 7.15. Maximum likelihood coefficient estimates of the binomial logit model that predicts the transitional proportions of stem infected loblolly pine trees.

STEM-DEAD		
predictor variables	estimate	std. error
INTERCEPT	10.144089	2.367079
AGE	-0.329695	0.104875
SINDEX	-0.034081	0.014906
QMD	0.271324	0.104667
RS	-4.293221	0.667400
STEM0	-0.005733	0.001660

---

logL = -399.65

AIC = 404.65

pseudo- $R^2$  = 0.268

SSR = 0.860859

WSSR = 15.043567

Table 7.15a. Covariance matrix of the coefficient estimates of the binary logit model displayed on table 7.15.

	INTERCEPT	AGE	SINDEX	QMD	RS	STEM0
INTERCEPT	5.60306403					
AGE	-2.0231E-01	1.0999E-02				
SINDEX	-2.6606E-02	1.3344E-03	2.2219E-04			
QMD	2.1493E-01	-1.7524E-02	-2.4234E-03	1.0955E-02		
RS	-1.31286000	3.3242E-02	3.6387E-03	-2.0239E-02	4.4542E-01	
STEM0	-9.3848E-04	-1.8729E-05	-4.7544E-06	1.1850E-04	4.4288E-04	2.7556E-06

Table 7.16. Maximum likelihood coefficient estimates of the quatri-nomial ordered logit model that predicts the transitional proportions of healthy slash pine trees.

<u>predictor variables</u>	<u>estimate</u>	<u>std. error</u>
INT.1 (CLR-BRA)	0.808099	0.226399
INT.2 (CLR-STEM)	-1.232745	0.226694
INT.3 (CLR-DEAD)	-4.362624	0.246543
KGSW	0.024262	0.006479
KGSBD	-0.470318	0.156412
SLOPE	0.590624	0.113353
FLAT	-0.495984	0.142231
AGE	0.126856	0.025901
HEIGHT	-0.149834	0.019540
BA	0.016889	0.004865
QMD	0.698453	0.079777
CLEAR0	-0.007971	0.001090
BRANCH0	-0.004905	0.001090
STEM0	-0.001172	0.000412

---

logL = -5244.01

AIC = 5255.01

pseudo- $R^2$  = 0.198

SSR = 1.285955

WSSR = 7.055818

Table 7.16a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.16.

	INT.1	INT.2	INT.3	KGSW	KGSBD	SLOPE	FLAT
INT.1	5.1257E-02						
INT.2	-5.1288E-02	5.1390E-02					
INT.3	-5.0965E-02	5.1140E-02	6.0783E-02				
KGSW	5.0018E-03	-5.0024E-03	-4.9810E-03	4.1977E-05			
KGSBD	-10.4163E-02	1.4293E-02	1.4696E-02	-4.8048E-03	2.4465E-02		
SLOPE	3.5044E-03	-3.5419E-03	-3.6993E-03	1.7654E-03	-8.0349E-03	1.2849E-02	
FLAT	-2.3842E-03	2.3537E-03	2.2971E-03	-3.9539E-03	1.1236E-02	-2.7638E-03	2.0230E-02
AGE	1.1630E-03	-1.1710E-03	-1.1957E-03	2.3767E-04	-3.0064E-03	5.8565E-04	-1.2321E-03
HEIGHT	-9.4052E-04	9.2899E-04	8.8494E-04	-1.7235E-04	1.5427E-03	-9.6514E-04	2.7234E-04
BA	6.5441E-04	-6.5284E-04	-6.4480E-04	1.2655E-05	-1.0649E-04	9.2449E-05	-5.1313E-07
QMD	3.9748E-03	-4.0256E-03	-4.2021E-03	1.8159E-03	-3.1002E-03	4.3216E-03	-4.7566E-03
CLEAR0	-5.4942E-05	5.5114E-05	5.5466E-05	-2.1954E-06	7.2741E-05	-1.6702E-05	3.3348E-06
BRANCH0	-5.0660E-05	5.0086E-05	4.7363E-05	-8.4300E-06	5.3424E-05	-4.8784E-05	1.3806E-05
STEM0	-2.3925E-05	2.3785E-05	2.3247E-05	-2.1951E-07	1.5648E-05	-2.5089E-07	9.2666E-06

	AGE	HEIGHT	BA	QMD	CLEAR0	BRANCH0	STEM0
AGE	6.7086E-04						
HEIGHT	-2.7879E-04	3.8181E-04					
BA	4.5452E-05	-6.7843E-05	2.3668E-05				
QMD	1.2291E-04	-9.8361E-04	4.2818E-05	6.3644E-03			
CLEAR0	-1.3724E-06	6.0076E-07	-5.6919E-07	-1.1251E-05	1.1881E-06		
BRANCH0	-7.6723E-06	9.9111E-06	-2.6047E-06	-1.9831E-05	4.5627E-08	1.1881E-06	
STEM0	-5.3510E-07	7.7841E-07	-7.6473E-07	-7.2600E-06	3.2184E-08	1.0797E-08	1.6974E-07

Table 7.17. Maximum likelihood coefficient estimates of the quadri-nomial unordered logit model fitted to slash pine data.

predictor variables	CLR-BRA		CLR-STEM		CLR-DEAD	
	estimate	std. error	estimate	std. error	estimate	std. error
INTERCEPT	-5.312562	0.609893	-0.575828	0.261256	-0.002141	0.000452
KGSW	0.296858	0.147222	-0.098209	0.086245	0.556400	0.270967
KGSBD	-3.682073	0.896229	-1.884640	0.255152	-8.960232	4.954352
SLOPE	0.090808	0.238325	0.541846	0.122158	0.862989	0.377640
FLAT	-0.262754	0.268741	-0.663576	0.154594	-2.073054	0.491694
AGE	0.145132	0.062105	0.111445	0.028681	-0.074257	0.100174
HEIGHT	-0.144706	0.042662	-0.168784	0.021588	-0.081642	0.076351
BA	0.003096	0.009444	0.023717	0.005367	0.045575	0.014880
QMD	0.971361	0.175343	0.700472	0.087692	0.517542	0.292386
CLEAR0	0.007652	0.001382	0.002078	0.000431	-0.006533	0.002532
BRANCH0	0.002138	0.002103	-0.009747	0.001256	-0.013566	0.004002
STEM0	-0.000241	0.000094	-0.002141	0.000451	-0.006150	0.002165

logL = -5124.87

AIC = 5157.87

pseudo- $R^2$  = 0.232

SSR = 1.112989

WSSR = 6.538713



Table 7.17a. Covariance matrix of the coefficient estimates (CLR-BRA) of the UMNLL model displayed on table 7.17.

	INTERCEPT	KGSW	KGSBD	SLOPE	FLAT	AGE	HEIGHT
INTERCEPT	3.7193E-01						
KGSW	-7.2669E-03	2.1674E-02					
KGSBD	4.2230E-01	-2.4603E-02	8.0323E-01				
SLOPE	-2.3124E-02	6.9979E-03	-4.8040E-02	5.6799E-02	7.2222E-02		
FLAT	5.2193E-02	-1.7044E-02	1.3469E-01	-1.1881E-02	-8.6337E-03		
AGE	-1.7978E-02	9.8550E-04	-4.2469E-02	4.1772E-03	3.8570E-03		
HEIGHT	4.0446E-03	-1.1322E-03	2.1844E-02	-5.1666E-03	-1.8932E-03		1.8200E-03
BA	-2.8530E-03	2.2914E-04	-2.1076E-04	4.2902E-04	-2.8415E-04	2.1911E-04	-2.8568E-04
QMSD	-5.1984E-02	7.0453E-03	-7.4074E-02	2.2242E-02	-1.0457E-02	2.8884E-03	-5.1998E-03
CLEAR0	-7.7208E-04	5.1491E-06	-1.2847E-03	8.7626E-05	-1.5381E-04	4.9220E-05	-2.5661E-05
BRANCH0	2.3744E-04	-6.2112E-05	2.1432E-04	-2.1814E-04	9.3569E-05	-4.4982E-05	4.8779E-05
STEM0	-3.5297E-04	1.3878E-05	-5.1211E-04	8.6119E-06	-3.8853E-05	1.2461E-05	-1.0327E-05

	BA	QMSD	CLEAR0	BRANCH0	STEM0
BA	8.9189E-05				
QMSD	1.2240E-04	3.0745E-02			
CLEAR0	2.6960E-06	1.4098E-04	1.9099E-06		
BRANCH0	1.2221E-05	8.0632E-05	5.4074E-08	4.4226E-06	
STEM0	-1.4503E-06	-7.6639E-05	-8.4233E-07	-5.9230E-09	8.8360E-09

Table 7.17b. Covariance matrix of the coefficient estimates (CLR-STM) of the UMNL model displayed on table 7.17.

	INTERCEPT	KGSW	KGSBD	SLOPE	FLAT	AGE	HEIGHT
INTERCEPT	6.8255E-02						
KGSW	7.1594E-03	7.4382E-03					
KGSBD	1.3324E-02	5.9831E-03	6.5103E-02				
SLOPE	-4.0609E-03	-2.4610E-03	-8.9661E-03	1.4923E-02			
FLAT	5.1040E-03	4.8191E-03	1.2982E-02	-2.9779E-03	2.3885E-02		
AGE	-1.6041E-03	-2.5122E-04	-3.2229E-03	6.1134E-04	-1.5064E-03	8.2260E-04	
HEIGHT	1.4468E-03	1.7878E-04	1.7037E-03	-1.0559E-03	2.2257E-04	-3.3957E-04	4.6604E-04
BA	-9.1755E-04	-4.3730E-06	-1.3342E-04	1.1032E-04	-4.4506E-05	5.3311E-05	-8.7503E-05
QMSD	-4.4937E-03	-2.1003E-03	-3.3484E-03	4.6619E-03	-5.7112E-03	1.7549E-04	-1.1913E-03
CLEAR0	-6.5315E-05	-2.5878E-06	-8.0609E-05	1.9331E-05	-2.3713E-06	1.1694E-06	-2.9986E-07
BRANCH0	7.6370E-05	5.4245E-06	7.1284E-05	-5.5352E-05	3.9178E-06	-9.1976E-06	1.2083E-05
STEM0	3.1625E-05	1.6741E-06	1.2628E-05	-2.2092E-06	1.4043E-05	-7.4692E-07	6.4424E-07

	BA	QMSD	CLEAR0	BRANCH0	STEM0
BA	2.8805E-05				
QMSD	4.5076E-05	7.6899E-03			
CLEAR0	7.8384E-07	1.3001E-05	1.8576E-05		
BRANCH0	-3.3142E-06	-2.1780E-05	-6.1487E-08	1.5775E-06	
STEM0	-1.1986E-07	-7.1361E-06	-3.1776E-08	1.9034E-09	2.0340E-07

Table 7.17c. Covariance matrix of the coefficient estimates (CLR-DEAD) of the model displayed on table 7.17.

	INTERCEPT	KGSW	KGSBD	SLOPE	FLAT	AGE	HEIGHT
INTERCEPT	1.8063E-07						
KGSW	-5.4786E-02	7.3277E-02					
KGSBD	4.2609E-03	-7.4554E-02	2.4546E01				
SLOPE	-5.5857E-02	6.1154E-03	-9.4006E-02	1.4261E-01			
FLAT	1.4615E-02	-4.2782E-02	9.3348E-02	-4.5446E-02	2.4176E-01		
AGE	4.4229E-02	-3.7324E-03	1.2508E-02	-3.9502E-03	2.3600E-02	1.0035E-02	
HEIGHT	1.6452E-02	-7.7794E-04	7.3978E-03	-9.8248E-03	5.0504E-03	5.4177E-03	5.8295E-01
BA	-3.9456E-03	1.9841E-04	-2.0267E-04	1.0386E-03	-3.8647E-04	-6.0613E-04	-8.5902E-04
QMSD	-1.2710E-01	1.6876E-02	-1.3353E-02	4.7722E-02	-2.7838E-02	-1.1892E-02	-1.8166E-02
CLEAR0	1.7323E-03	-3.7356E-05	5.6836E-04	-2.5925E-04	2.8600E-04	1.1832E-04	8.7529E-05
BRANCH0	5.1201E-04	-1.5036E-04	3.8342E-04	-6.3146E-04	1.6389E-04	1.0606E-04	1.2135E-04
STEM0	9.2090E-04	-5.5287E-05	2.4179E-04	-1.0016E-04	1.6416E-04	6.5388E-05	7.6189E-05

	BA	QMSD	CLEAR0	BRANCH0	STEM0
BA	-8.5902E-04	2.2141E-04			
QMSD	-1.8166E-02	1.4240E-03	8.5490E-02		
CLEAR0	8.7529E-05	-2.1198E-06	-4.1749E-04	6.4110E-06	
BRANCH0	1.2135E-04	-2.8408E-05	-3.2613E-04	3.0353E-06	1.6016E-05
STEM0	7.6189E-05	-4.4649E-06	-2.9908E-04	3.5795E-06	2.0152E-06
					4.6872E-06

Table 7.18. Maximum likelihood coefficient estimates of the tri-nomial ordered logit model that predicts the transitional proportions of branch infected slash pine trees.

<u>predictor variables</u>	<u>estimate</u>	<u>std. error</u>
INT.1 (BRA-STEM)	4.366378	0.426427
INT.2 (BRA-DEAD)	-1.608412	0.404897
KGSW	-0.702556	0.186926
SLOPE	0.753012	0.449421
FLAT	-0.781718	0.237119
AGE	-0.536261	0.058478
HEIGHT	0.321517	0.050710
BA	-0.508301	0.010207
QMD	-0.641535	0.168070

---

logL = -672.49

AIC = 680.49

pseudo- $R^2$  = 0.204

SSR = 1.451386

WSSR = 8.815158

Table 7.18a. Covariance matrix of the coefficient estimates of the OMNL model displayed on table 7.18.

	INT.1	INT.2	KGSW	SLOPE	FLAT	AGE	HEIGHT
INT.1	1.8184E-01						
INT.2	-1.4197E-01	1.6394E-01					
KGSW	-3.2761E-02	2.4925E-02	3.4941E-02				
SLOPE	4.4310E-04	-8.5559E-03	-1.9501E-03	2.0198E-01			
FLAT	-4.1995E-02	3.9847E-02	1.0143E-02	-2.6874E-02	5.6225E-02		
AGE	-1.4666E-02	9.6313E-03	2.1644E-03	-2.7910E-03	9.9232E-03	3.4197E-03	
HEIGHT	9.8597E-04	-2.7996E-03	-1.0774E-03	5.9841E-03	-1.5322E-03	-1.7911E-03	2.5715E-03
BA	-4.8676E-04	1.0936E-03	2.0980E-04	-6.1494E-04	1.4395E-04	3.0567E-04	-3.7595E-04
QMSD	-2.8605E-03	7.2035E-03	1.7432E-03	-2.4754E-02	2.8834E-02	1.0106E-03	-6.7488E-03
	BA	QMSD					
BA	1.0418E-04						
QMSD	7.1469E-04	2.8248E-02					

Table 7.19. Maximum likelihood coefficient estimates of the tri-nomial unordered logit model which predicts transitional proportions of branch infected slash pine trees.

predictor variables	BRA-STEM		BRA-DEAD	
	estimate	std. error	estimate	std. error
INTERCEPT	5.413192	0.565512	3.266681	1.218662
NORTH	-0.913029	0.366794	-0.965512	0.256234
KGSW	-1.207113	0.246908	-1.011320	0.478076
SLOPE	0.758685	0.199004	0.292145	0.065847
FLAT	-0.385959	0.306286	-3.012251	0.836844
AGE	-0.545872	0.075677	-1.322820	0.194627
HEIGHT	0.379607	0.071287	0.752104	0.116575
BA	-0.066517	0.013515	-0.093241	0.023635
QMSD	-1.023071	0.232075	-1.599621	0.433272

logL = -644.22

AIC = 660.22

pseudo- $R^2$  = 0.225

SSR = 1.166012

WSSR = 4.407802

Table 7.19a. Covariance matrix of the coefficient estimates (BRA-STEM) of the UMNL model displayed on table 7.19.

	INTERCEPT	NORTH	KGSW	SLOPE	FLAT	AGE	HEIGHT
INTERCEPT	3.1980E-01						
NORTH	-2.2154E-03	1.3454E-01					
KGSW	-1.3367E-01	2.5298E-03	6.0964E-02				
SLOPE	4.5080E-02	-1.5483E-03	-3.0993E-02	3.9603E-02			
FLAT	-1.8630E-01	2.2541E-02	7.2484E-02	-7.3084E-02	1.0932E-01		
AGE	-4.7889E-02	1.1536E-04	1.3837E-02	-7.2371E-04	2.9588E-02	5.0818E-03	
HEIGHT	1.6824E-02	-2.5512E-04	-5.3533E-03	9.9608E-03	-4.5337E-03	-4.7901E-03	5.0818E-03
BA	-3.1674E-03	2.4215E-04	2.5114E-03	-4.8840E-02	3.5556E-02	6.1890E-03	-8.1295E-04
QMSD	-2.7309E-02	2.1145E-02	2.5141E-03	-4.8840E-02	3.5557E-02	6.1890E-03	-1.3097E-02
	BA	QMSD					
BA	1.8266E-04						
QMSD	1.6827E-03	5.3859E-01					

Table 7.19b. Covariance matrix of the coefficient estimates (BRA-DEAD) of the UMN model displayed on table 7.19.

	INTERCEPT	NORTH	KGSW	SLOPE	FLAT	AGE	HEIGHT
INTERCEPT	1.48513707						
NORTH	-5.9146E-03	6.5656E-02					
KGSW	-3.6607E-03	2.5481E-03	2.2856E-01				
SLOPE	1.1478E-05	-2.3454E-04	-1.9123E-05	4.3358E-03			
FLAT	-3.4530E-03	2.3221E-04	1.2287E-03	-3.4323E-04	7.0031E-01		
AGE	-1.2458E-03	1.2312E-04	2.2387E-04	-2.3319E-04	2.1876E-04	3.7880E-02	
HEIGHT	2.2141E-04	-2.5512E-05	-3.1111E-05	2.8521E-06	-8.5412E-06	-4.0258E-06	1.3590E-02
BA	-2.2584E-05	2.1450E-05	4.8145E-06	-5.2358E-06	2.2548E-07	-7.5842E-07	-2.3841E-06
QMSD	-3.5289E-02	1.7524E-02	3.6387E-03	-2.4234E-03	1.7524E-03	1.3344E-04	-2.5895E-04
	BA	QMSD					
BA	5.5861E-04						
QMSD	-2.5895E-04	2.5689E-03					



Table 7.20. Maximum likelihood coefficient estimates of the binomial logit model that predicts the transitional proportions of stem infected slash pine trees.

STEM-DEAD		
predictor variables	estimate	std. error
INTERCEPT	-1.122129	0.325029
SLOPE	0.854689	0.143496
FLAT	-0.443695	0.216910
AGE	0.090650	0.031460
HEIGHT	-0.068326	0.024471
BA	0.023910	0.009005
CLEAR0	-0.002129	0.000532
BRANCH0	-0.005316	0.001570

---

logL= -290.62

AIC= 297.62

pseudo- $R^2$ = 0.209

SSR= 0.223809

WSSR= 2.030853

Table 7.20a. Covariance matrix of the coefficient estimates of the binary model displayed on table 7.20.

	INTERCEPT	SLOPE	FLAT	AGE	HEIGHT	BA	CLEAR0
INTERCEPT	1.0564E-01						
SLOPE	8.0812E-03	2.0591E-02					
FLAT	4.4641E-03	5.2781E-03	4.7050E-02				
AGE	-1.1478E-05	-6.6094E-04	-2.9123E-03	9.8973E-04			
HEIGHT	-5.7730E-03	-1.7047E-04	1.2819E-03	-5.0593E-04	5.9883E-04		
BA	2.4149E-03	1.0960E-04	-2.7556E-04	1.2419E-04	2.1022E-04	8.1090E-05	
CLEAR0	-1.2031E-04	-3.5000E-06	3.1521E-06	-3.4014E-06	7.8742E-06	3.0536E-06	2.8302E-7
BRANCH0	-2.3211E-04	-3.1450E-05	-3.8163E-05	-1.6110E-05	2.0538E-05	-7.6229E-06	1.1841E-07

BRANCH0	
BRANCH0	2.4649E-06

STEM-DEAD in Table 7.20) are increasing over time whereas, the proportions of branch infected trees (BRA-STEM, BRA-DEAD in Table 7.19) are decreasing over time.

Site index had a significant effect on loblolly pine transitional proportions but not on those of slash pine. More specifically, site index was found to be negatively related to CLR-BRA, CLR-STEM (Table 7.12) and STEM-DEAD (Table 7.15) proportions and positively related to BRA-STEM and BRA-DEAD (Table 7.14) proportions. Once again, we remain skeptical about the negative relation of site quality with the infection rates of clear loblolly pine trees. There is enough evidence in the literature to convince us for the opposite. Further examination of the data failed to reveal any patterns that perhaps could explain this unexpected result.

The average height of dominant and codominant loblolly pine trees was found to be positively related to CLR-STEM and CLR-DEAD (Table 7.12) and negatively related to BRA-STEM and BRA-DEAD (Table 7.14) proportions. In slash pine, it had a negative effect on CLR-BRA, CLR-STEM (Table 7.17) and STEM-DEAD (Table 7.20) and a positive effect on BRA-STEM and BRA-DEAD (Table 7.19) proportions. Overall, it seems that healthy large loblolly pine trees are more susceptible to fusiform rust than similar slash pine trees.

Finally, the specific effect of stand density although significant it was very difficult to evaluate in a meaningful manner. As mentioned previously, there seems to be no consistent relationship between stand density and transitional proportions on both loblolly and slash pine.

## ***7.5 Concluding Remarks***

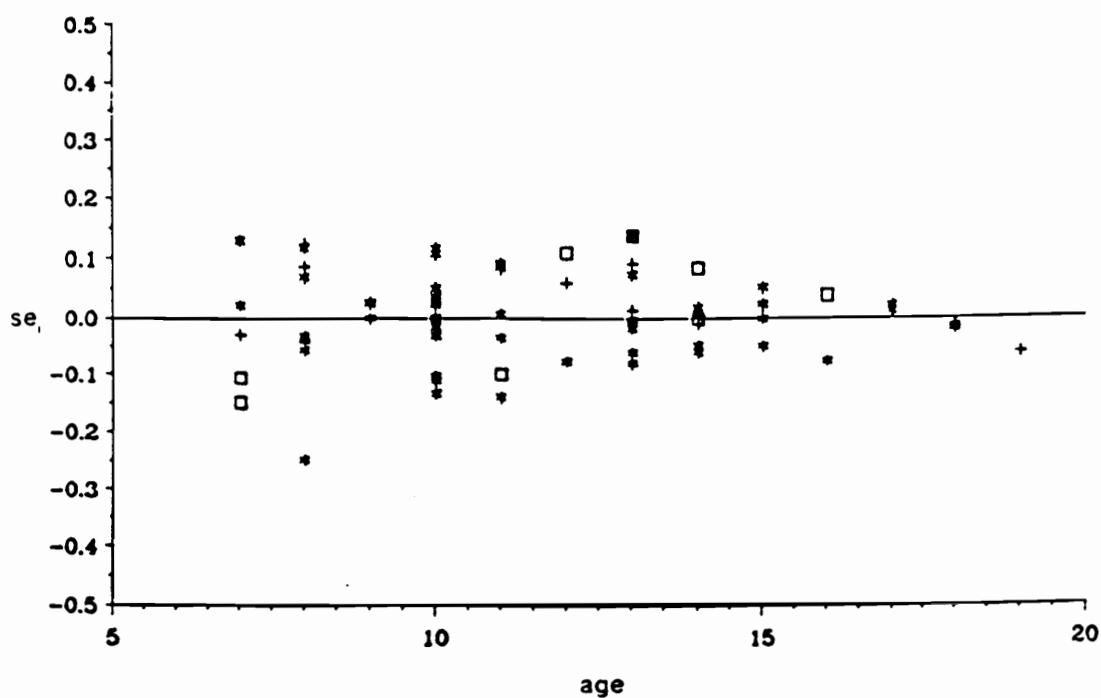
The primary goal in this chapter was to exemplify the applicability of qualitative response models in studying the incidence and spread of fusiform rust in loblolly and slash pine plantations in East

Texas. In the first part of the study, these models were used to predict the proportion of trees per acre at three levels of rust infection (branch infected, stem infected and dead) at a given age. In the second part, the transitional proportions at each row of the transition matrix were modeled. In all cases, logit formulations fit the data better than probit. Among multinomial logit models, the unordered performed better than the ordered with regard to goodness of fit criteria, but the estimates in the ordered models have smaller standard errors. Our feelings against the use of ordered models remain the same as those expressed in section 6.4 in modeling the merchantability of loblolly pine trees. Perhaps in the present study they are more justifiable because, intuitively, one would not expect site factors to affect in the same way the behavior of the disease at various stages of infection as the ordered models imply.

All models indicated that fusiform rust infection rates were decreasing with time on loblolly pine and increasing on slash pine. Loblolly pine plantations in South-East Texas appeared to be more susceptible to fusiform rust than plantations in North-East Texas. Rust incidence was found to increase on well-drained soils and decrease on moderately and on poorly drained soils. Site quality was, to our surprise, negatively related to rust incidence. No meaningful explanation can be given on this matter.

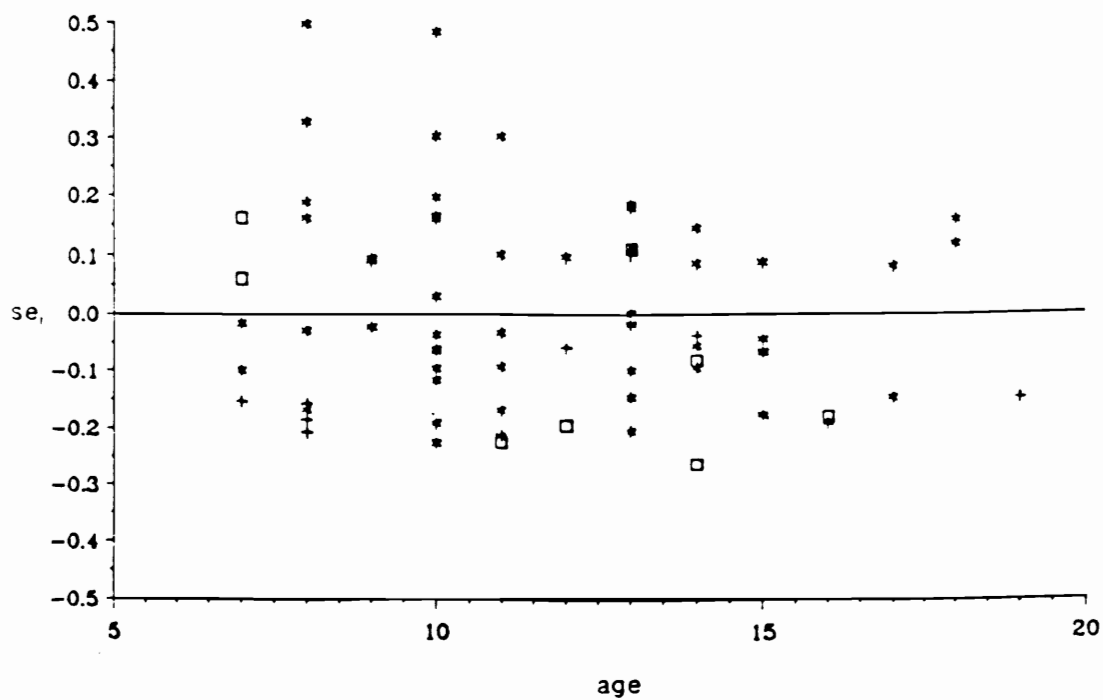
Obviously, rust infection levels are highly variable from stand to stand and the usual survey variables used in the models above, explain at best, only a portion of this variation. Apparently, other more directly associated factors are influencing rust incidence and spread. For instance, it has been reported (Froelich and Snow 1986, Hollis et al. 1975) that the presence of large numbers of oaks (*Quercus* spp.) scattered among the planted pines is usually associated with a high incidence of fusiform rust infection. Thus, it is possible that variables such as associated oak volume or oak leaf area would explain additional variation in modelling rust incidence. However, these variables are not, and probably will never be, incorporated into forest inventories because of time and money limitations (Borders and Bailey 1986). Consequently, the use of models that include such explanatory variables as the above mentioned, would be limited due to lack of appropriate, readily available data.

It is our belief that rust prediction models such as those developed in this study provide valuable insights into site-rust hazard relations and management. Perhaps more importantly, these models can easily be incorporated into existing growth and yield prediction systems or even to more sophisticated decision guideline models such as forest stand simulators to provide realistic estimates for expected value calculations of the volumetric or financial impact of fusiform rust in loblolly and slash pine plantations.



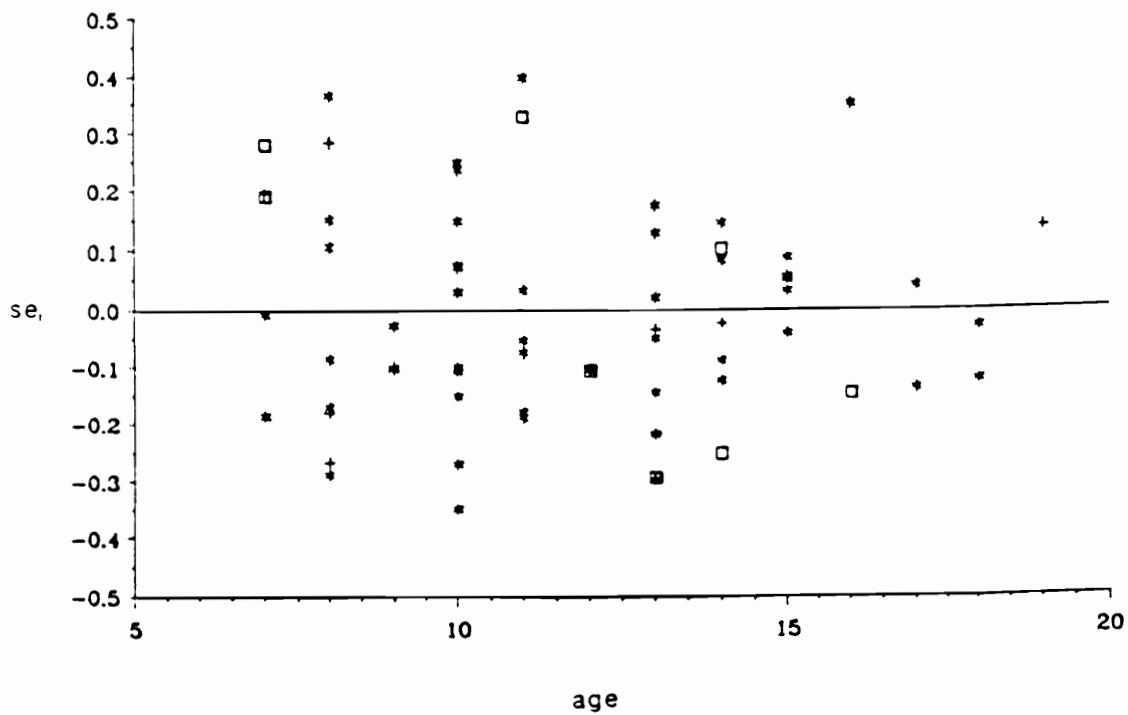
\* : slope site. + : flat site. □ : ridge.

Figure 7.9. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of healthy loblolly pine trees.



\* : slope site. + : flat site. □ : ridge.

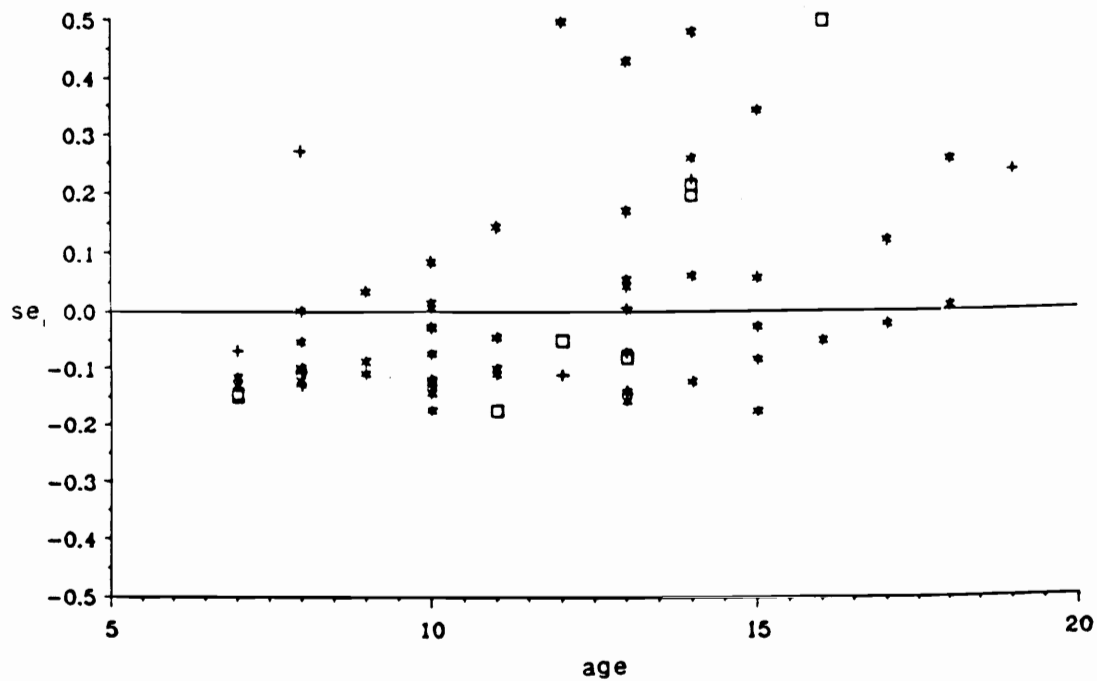
Figure 7.10. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of branch infected loblolly pine trees.



\* : slope site. + : flat site. □ : ridge.

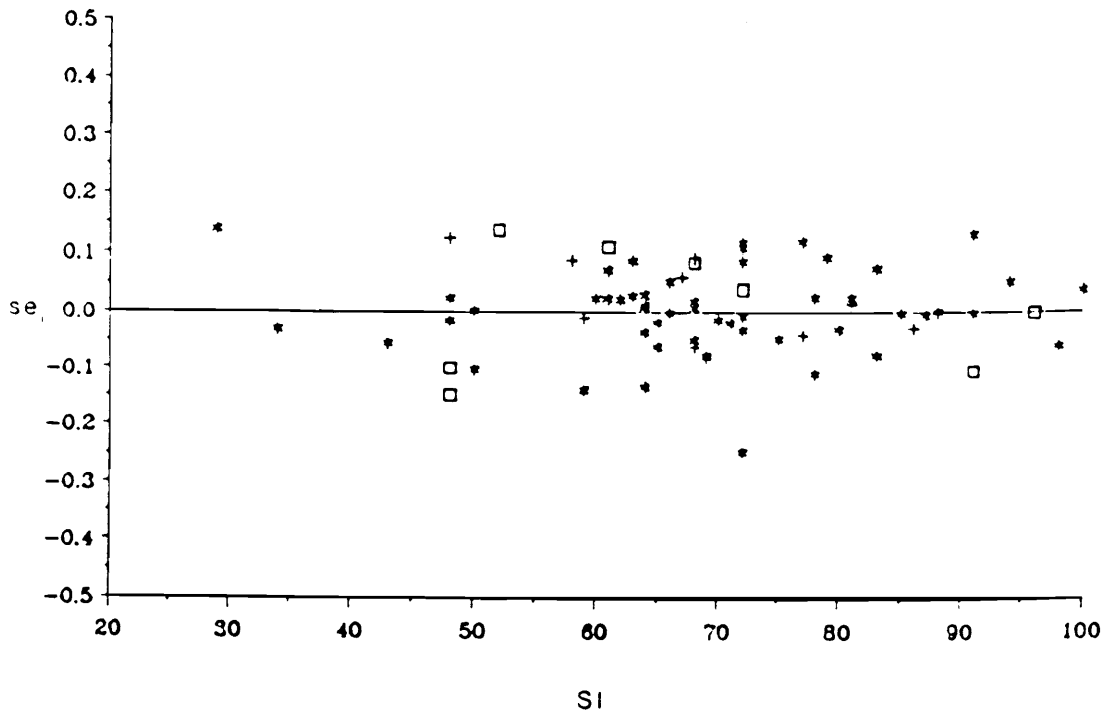
Figure 7.11. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of stem infected loblolly pine trees.





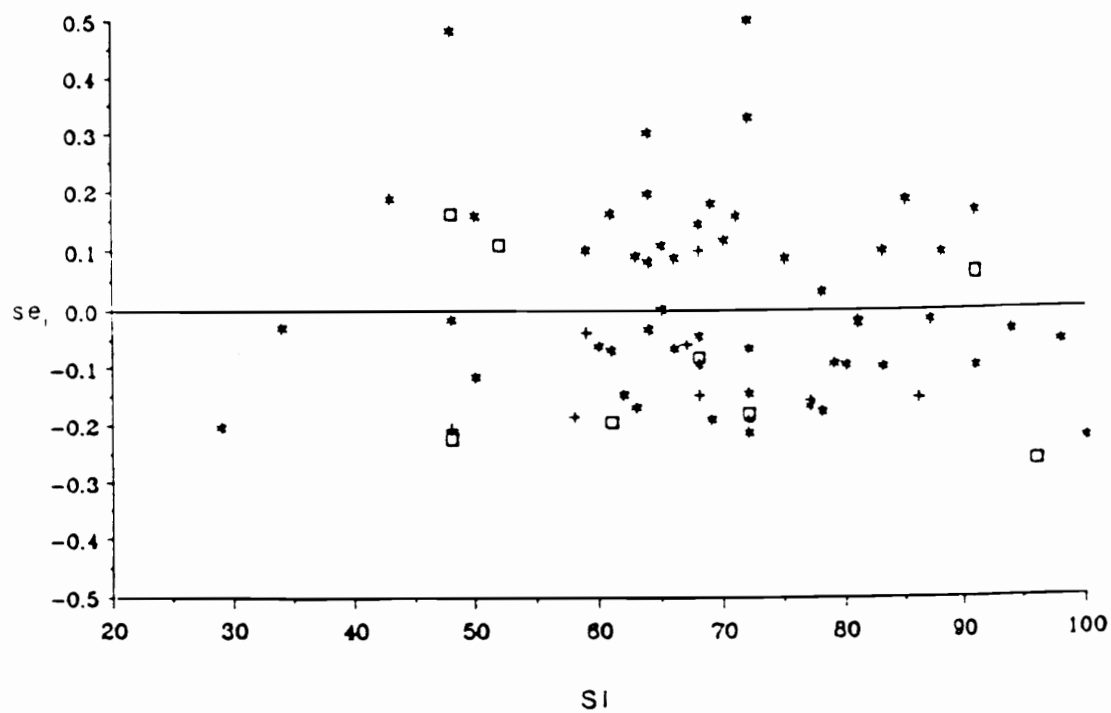
\* : slope site. + : flat site. □ : ridge.

Figure 7.12. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of dead loblolly pine trees.



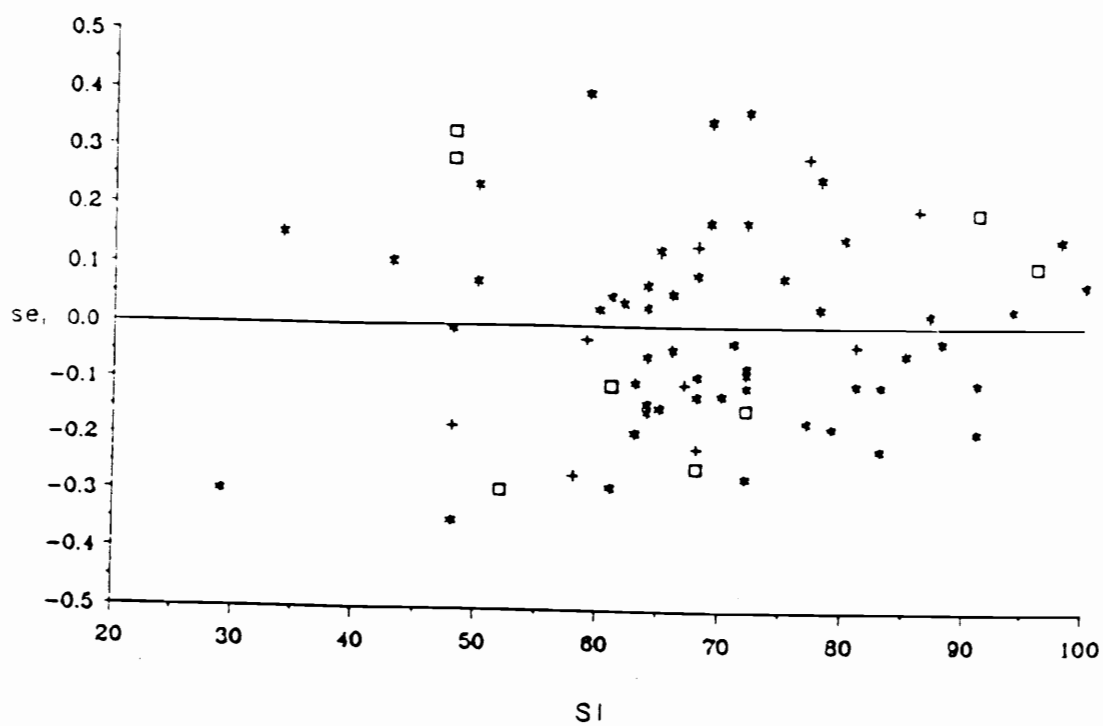
\* : slope site. + : flat site. □ : ridge.

Figure 7.13. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of healthy loblolly pine trees.



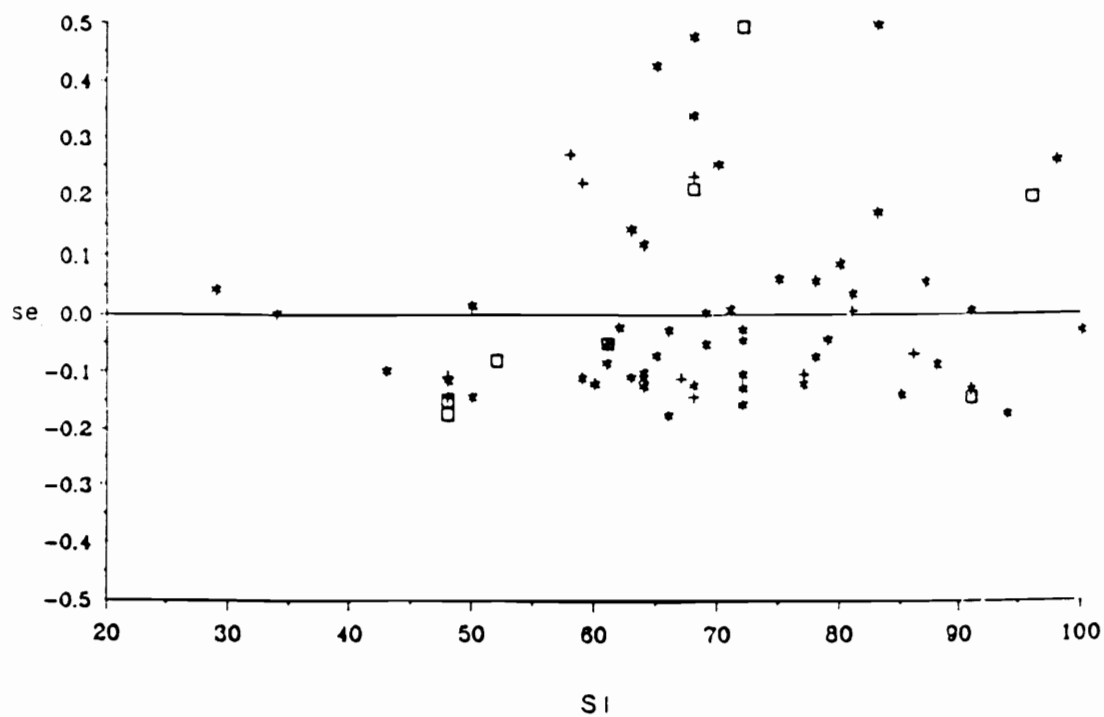
\* : slope site. + : flat site. □ : ridge.

Figure 7.14. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of branch infected loblolly pine trees.



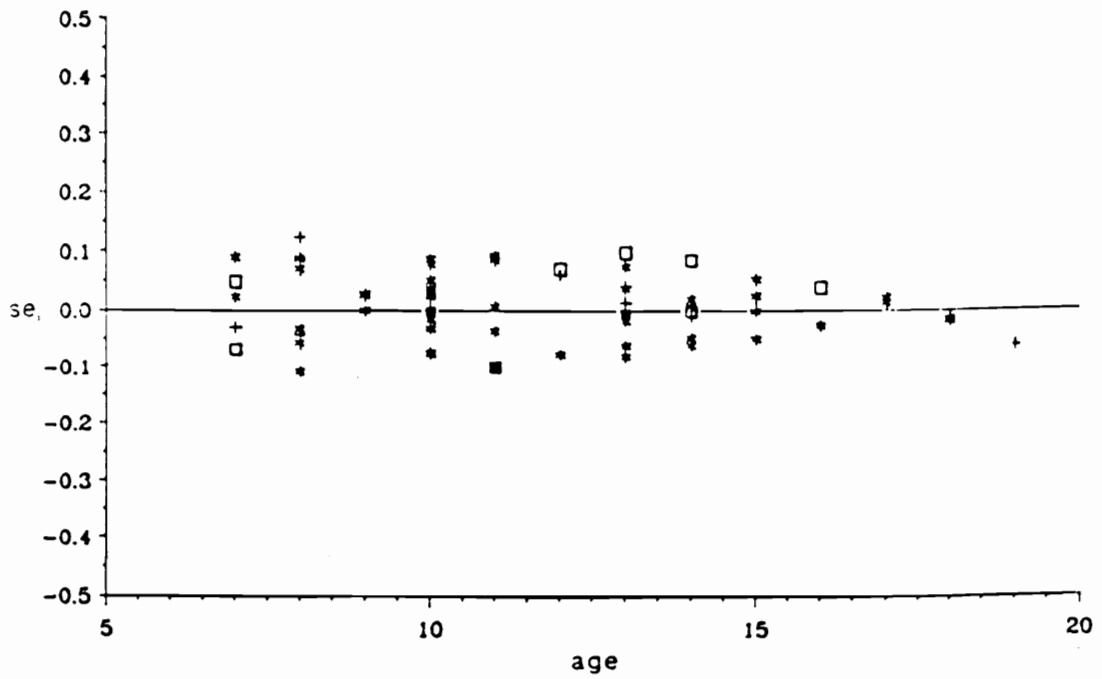
\* : slope site. + : flat site. □ : ridge.

Figure 7.15. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of stem infected loblolly pine trees.



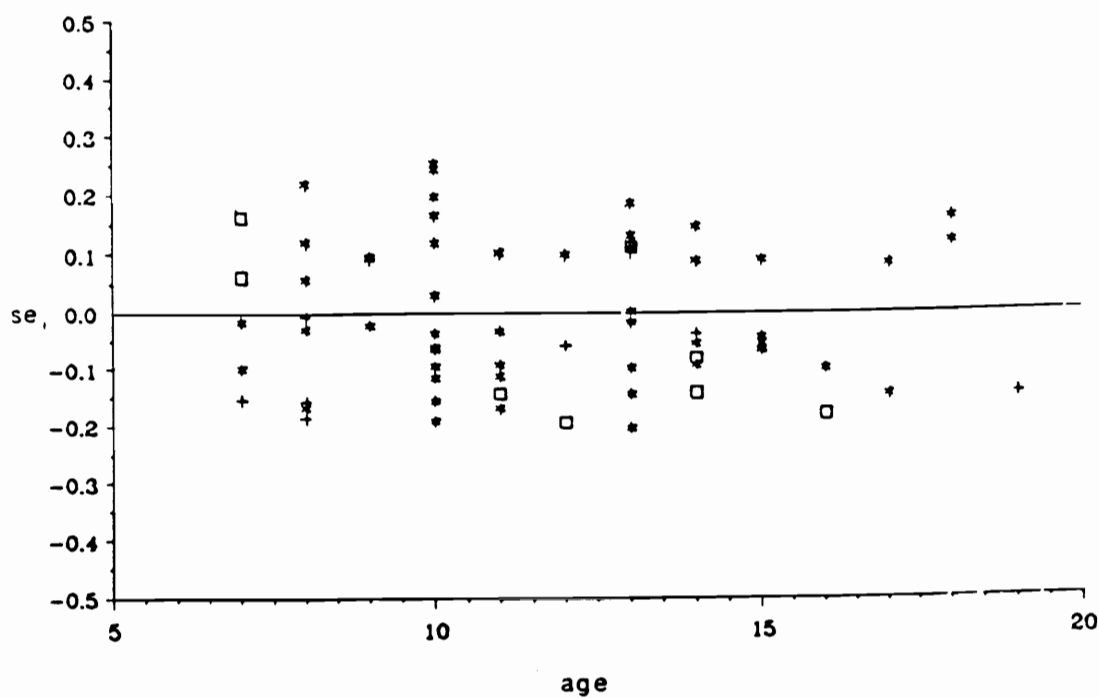
\* : slope site. + : flat site. □ : ridge.

Figure 7.16. Standardized residuals plotted against site index and landform for the OMNL model predicting the proportion of dead loblolly pine trees.



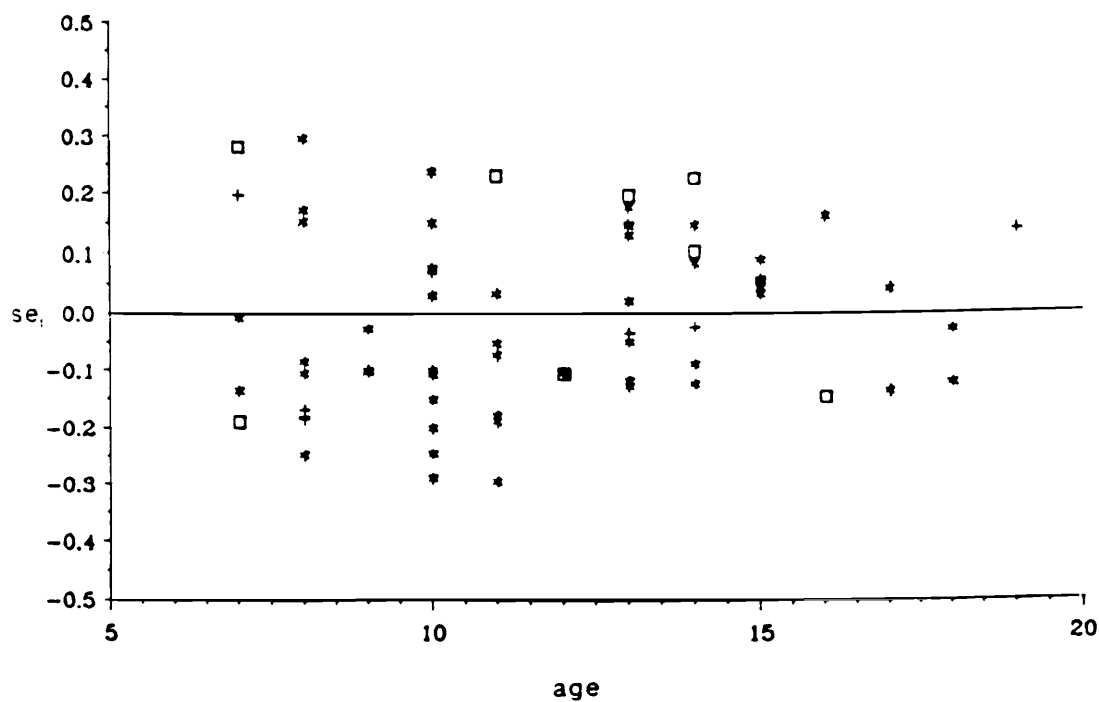
\* : slope site. + : flat site. □ : ridge.

Figure 7.17. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of healthy loblolly pine trees.



\* : slope site. + : flat site. □ : ridge.

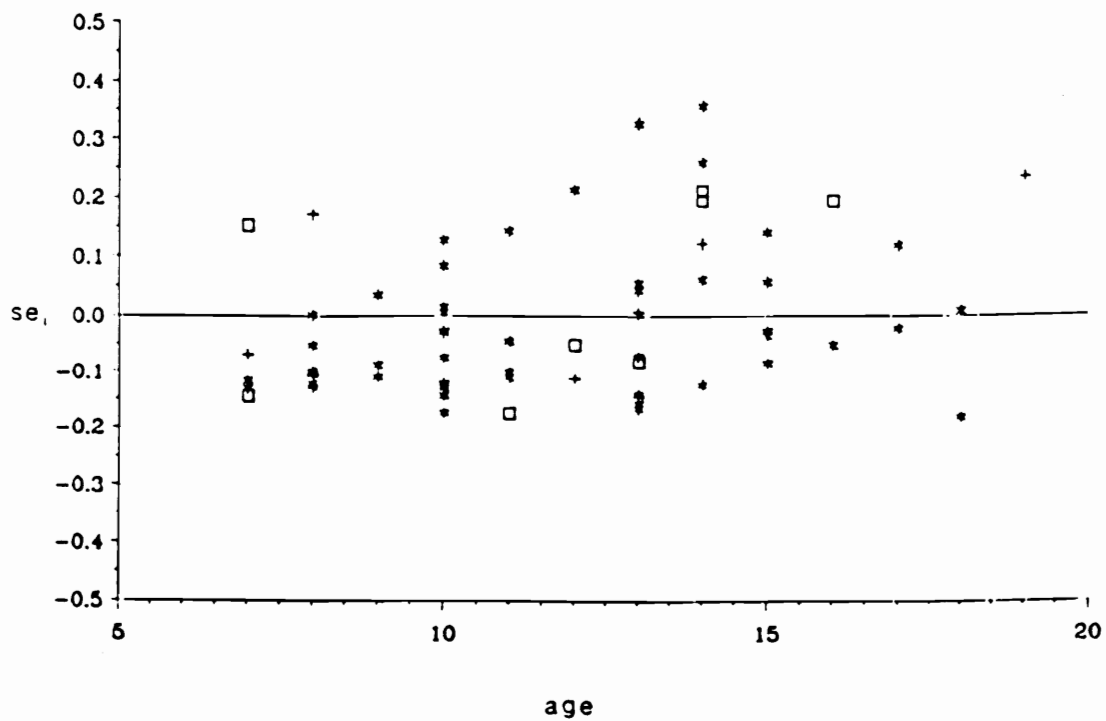
Figure 7.18. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of branch infected loblolly pine trees.



\* : slope site. + : flat site. □ : ridge.

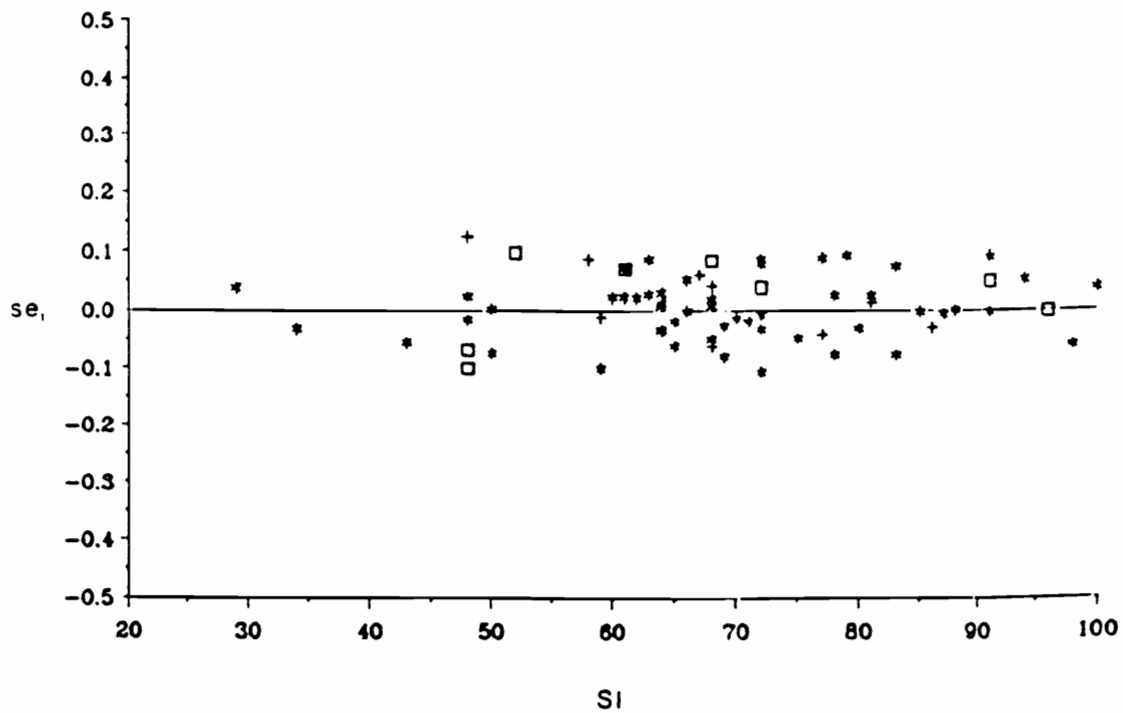
Figure 7.19. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of stem infected loblolly pine trees.





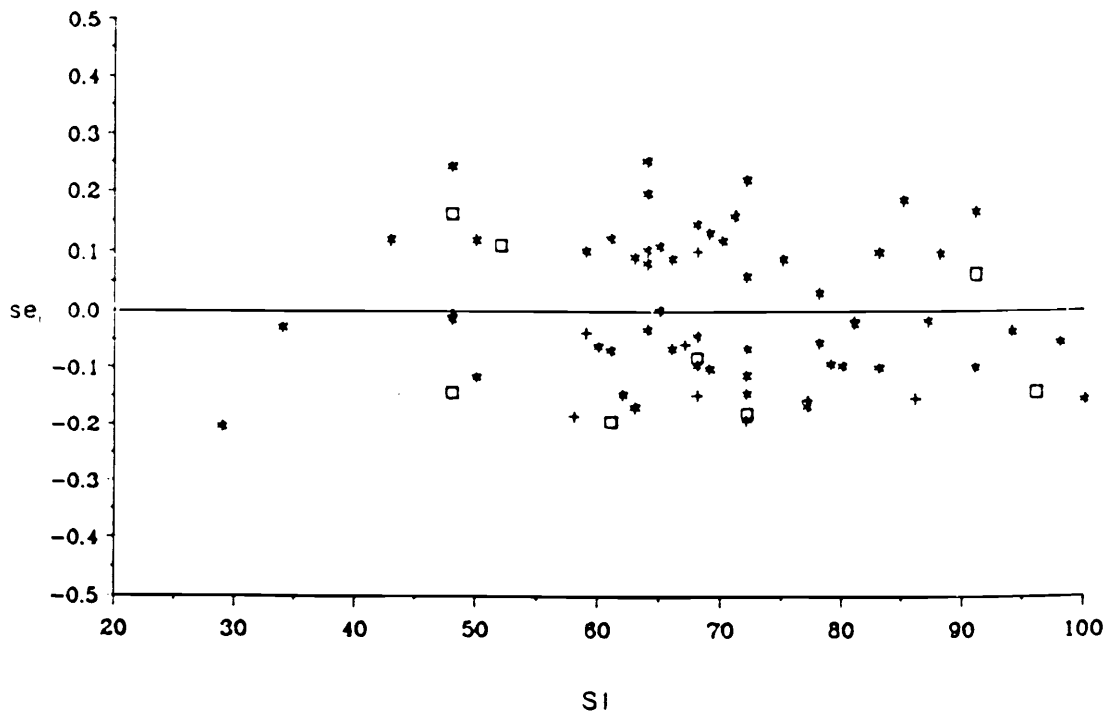
\* : slope site. + : flat site. □ : ridge.

Figure 7.20. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of dead loblolly pine trees.



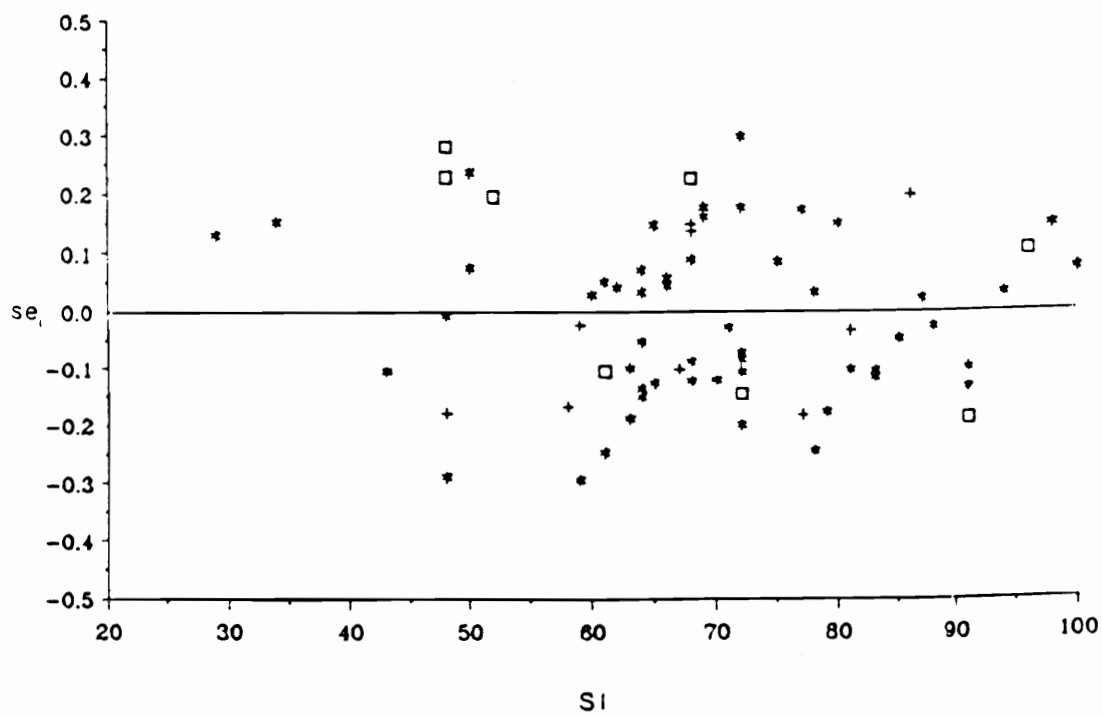
\* : slope site. + : flat site. □ : ridge.

Figure 7.21. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of healthy loblolly pine trees.



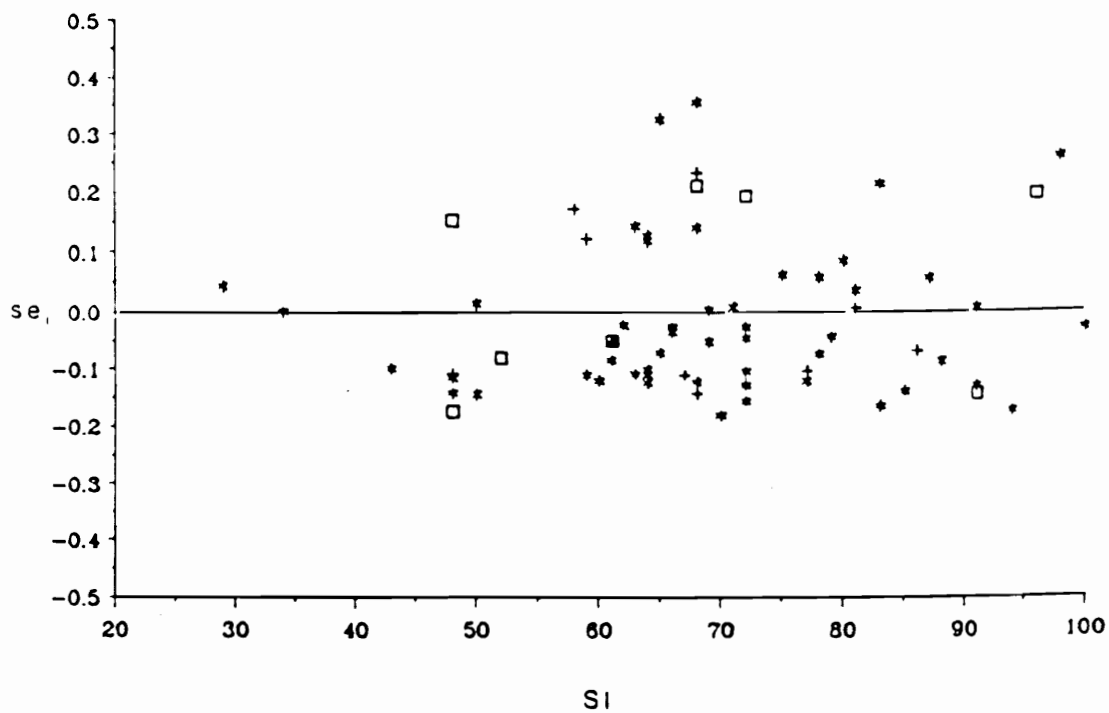
\* : slope site. + : flat site. □ : ridge.

Figure 7.22. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of branch infected loblolly pine trees.



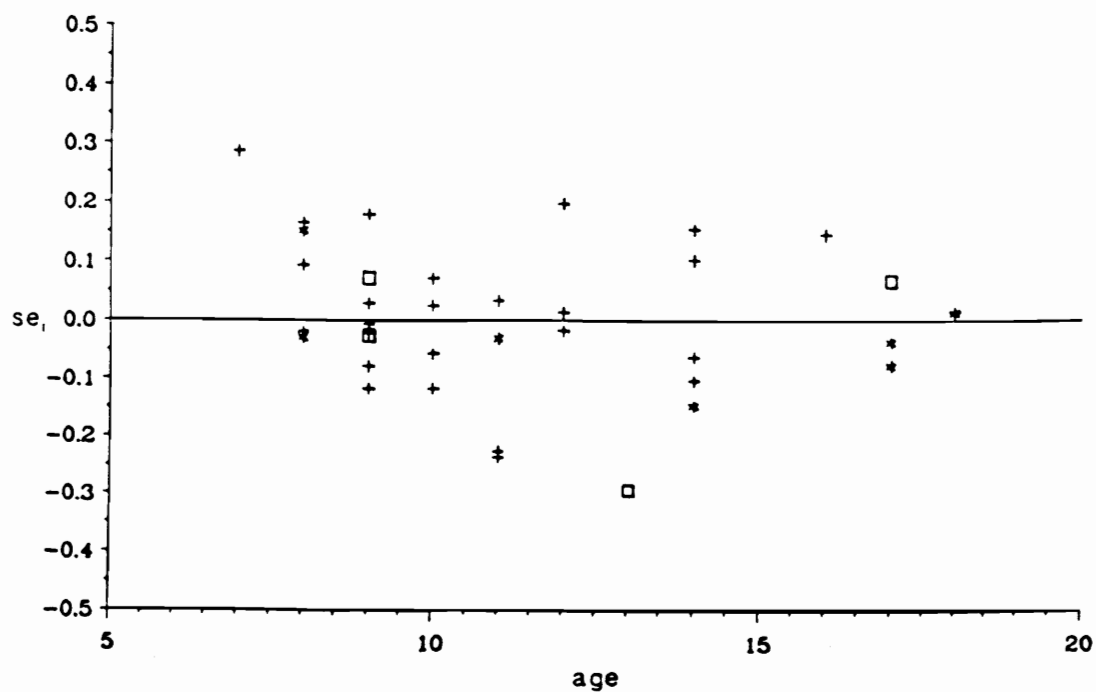
\* : slope site. + : flat site. □ : ridge.

Figure 7.23. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of stem infected loblolly pine trees.



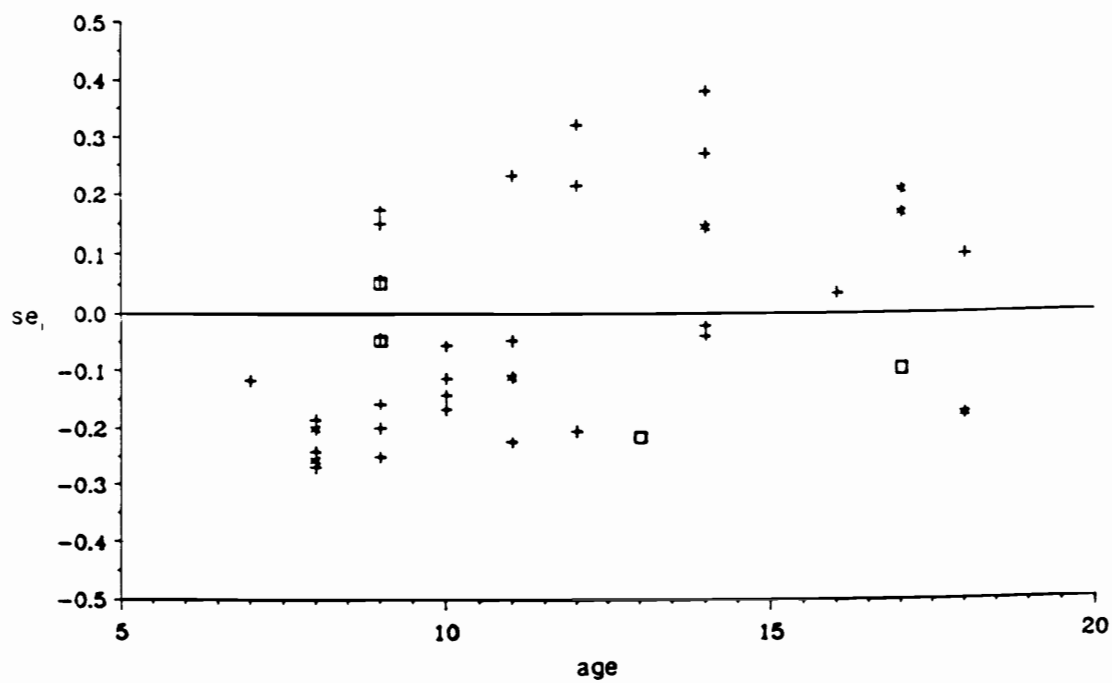
\* : slope site. + : flat site. □ : ridge.

Figure 7.24. Standardized residuals plotted against site index and landform for the UMNL model predicting the proportion of dead loblolly pine trees.



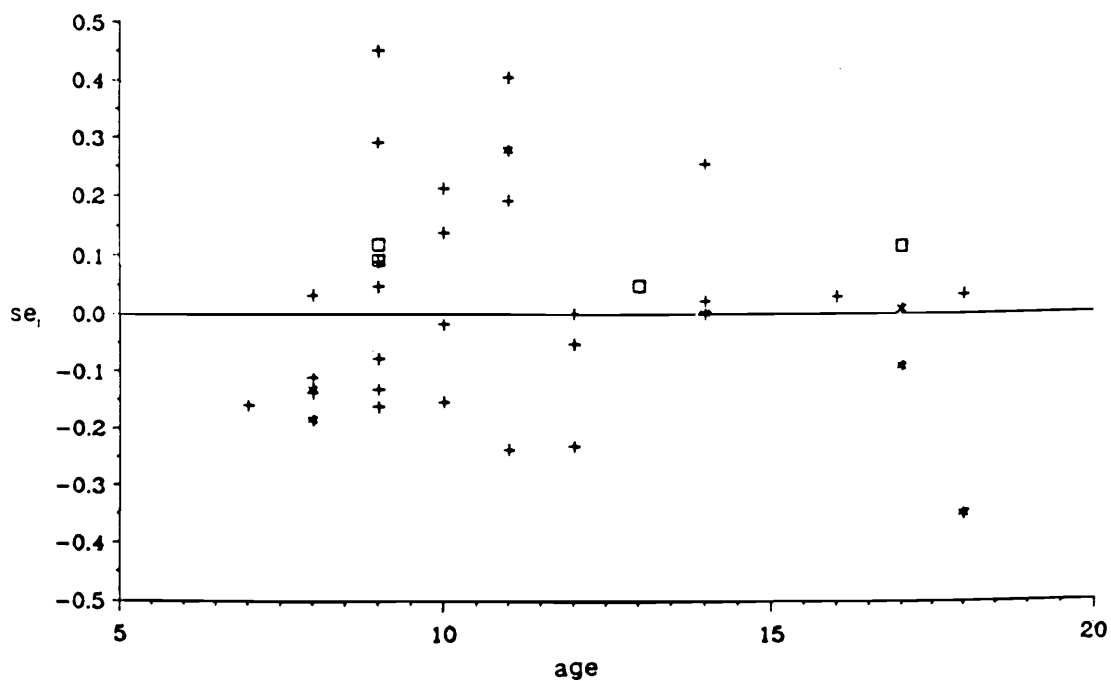
\* : slope site. + : flat site. □ : ridge.

Figure 7.25. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of healthy slash pine trees.



\* : slope site. + : flat site. □ : ridge.

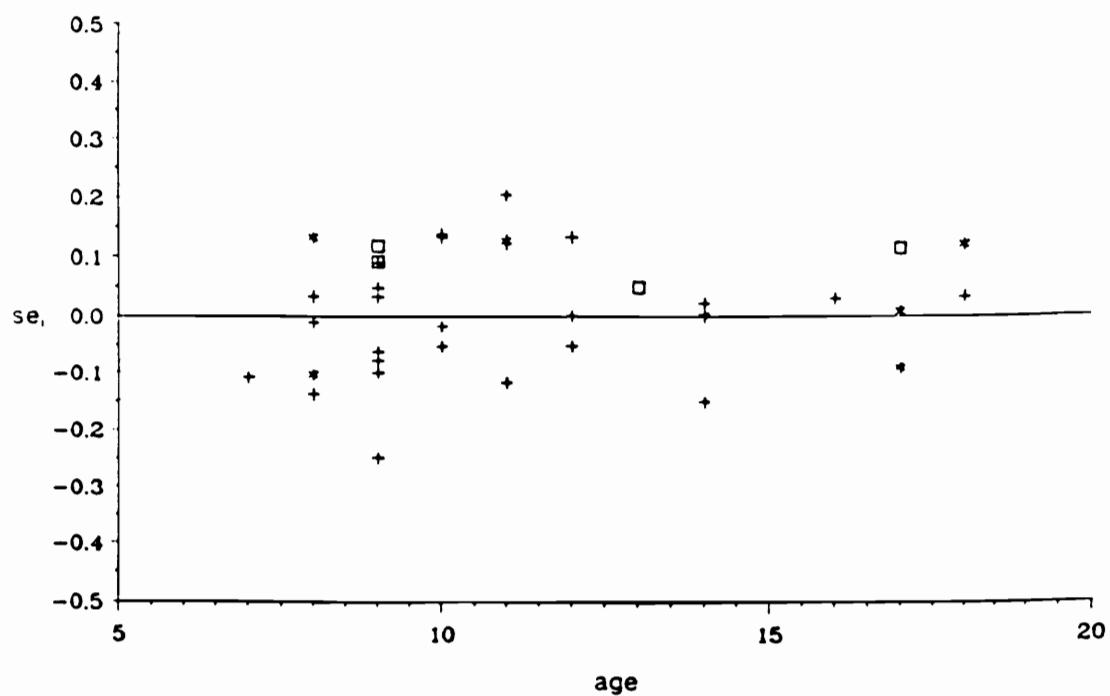
Figure 7.26. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of branch infected slash pine trees.



\* : slope site. + : flat site. □ : ridge.

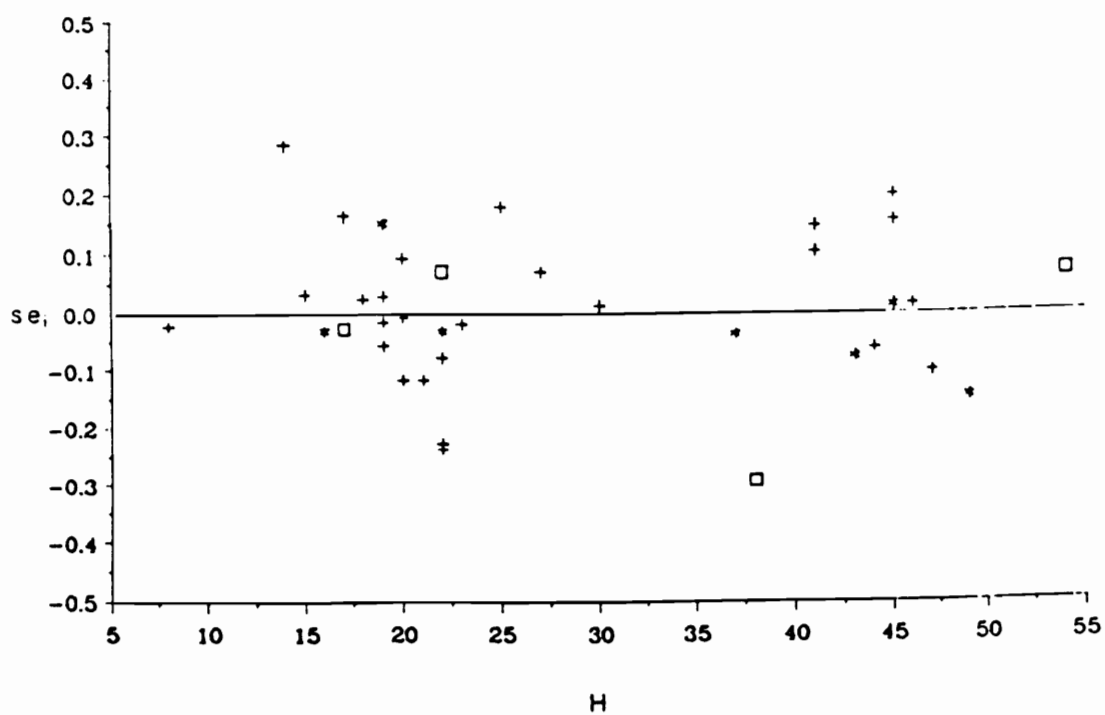
Figure 7.27. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of stem infected slash pine trees.





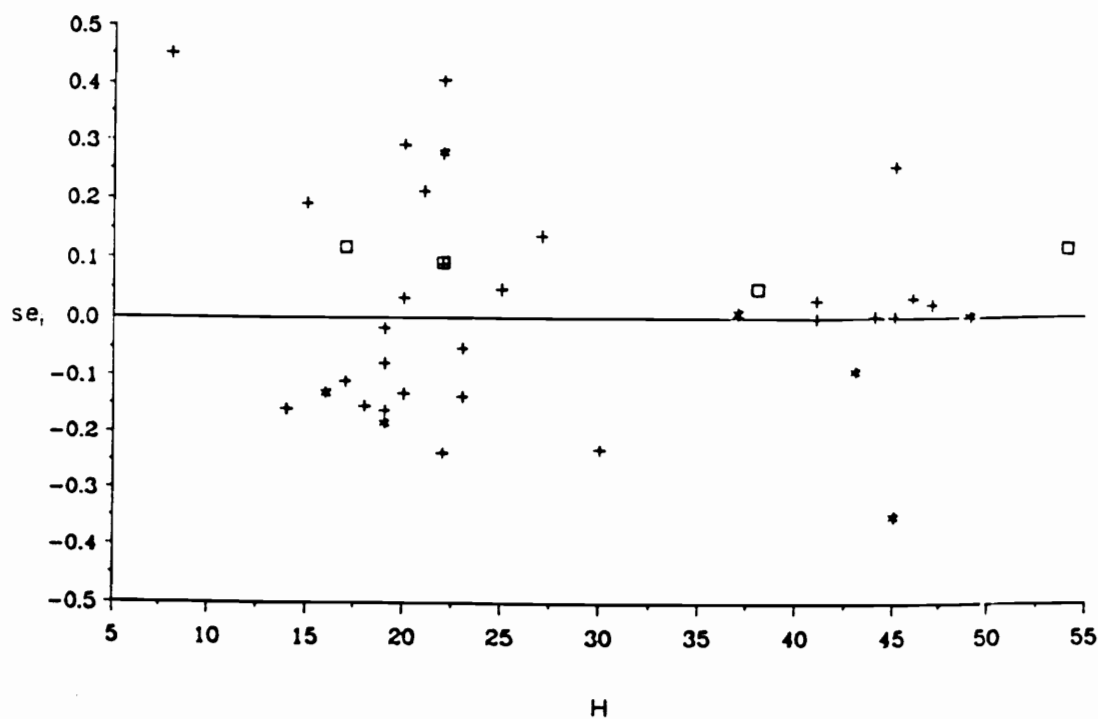
\* : slope site. + : flat site. □ : ridge.

Figure 7.28. Standardized residuals plotted against age and landform for the OMNL model predicting the proportion of dead slash pine trees.



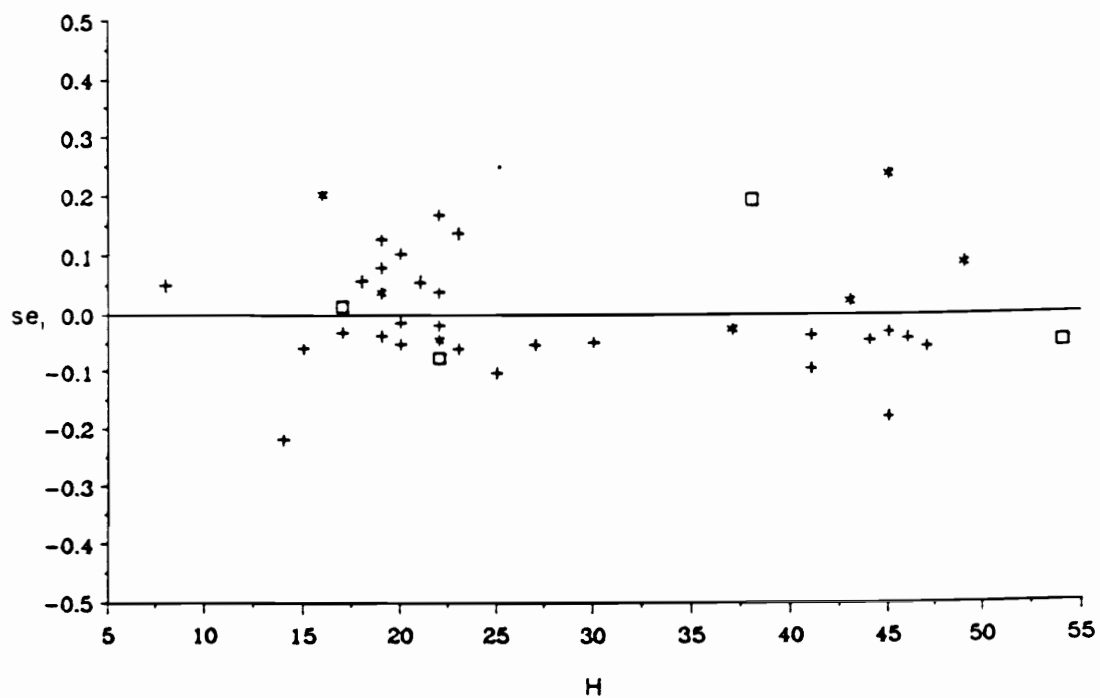
\* : slope site. + : flat site. □ : ridge.

Figure 7.29. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of healthy slash pine trees.



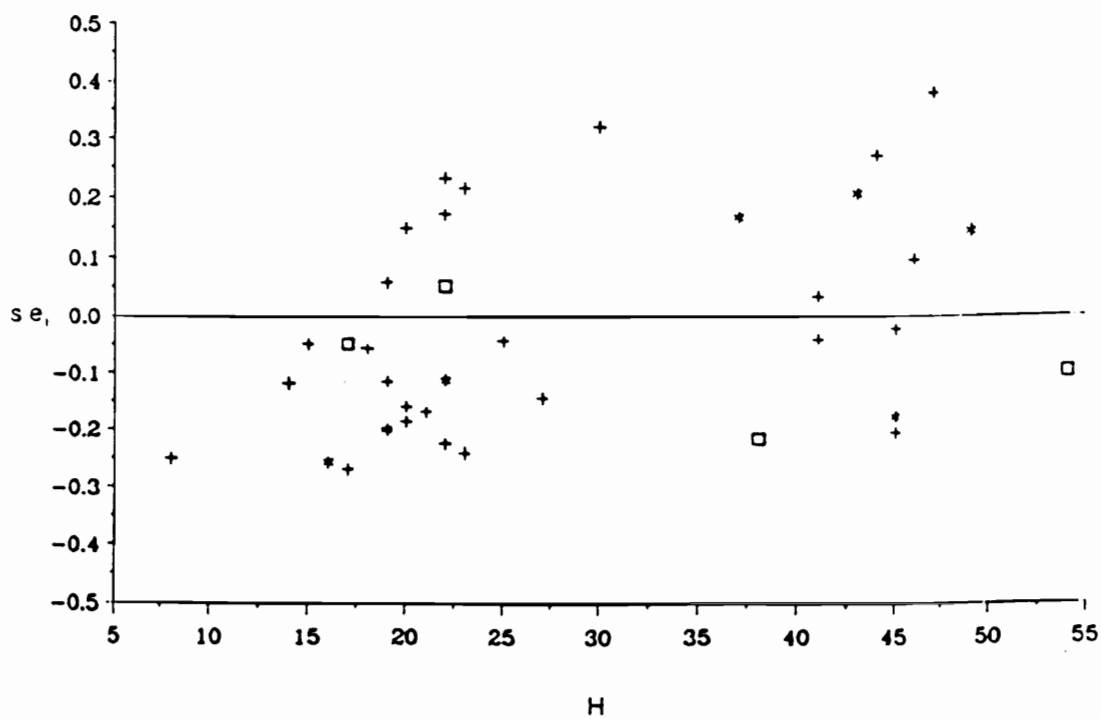
\* : slope site. + : flat site. □ : ridge.

Figure 7.30. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of branch infected slash pine trees.



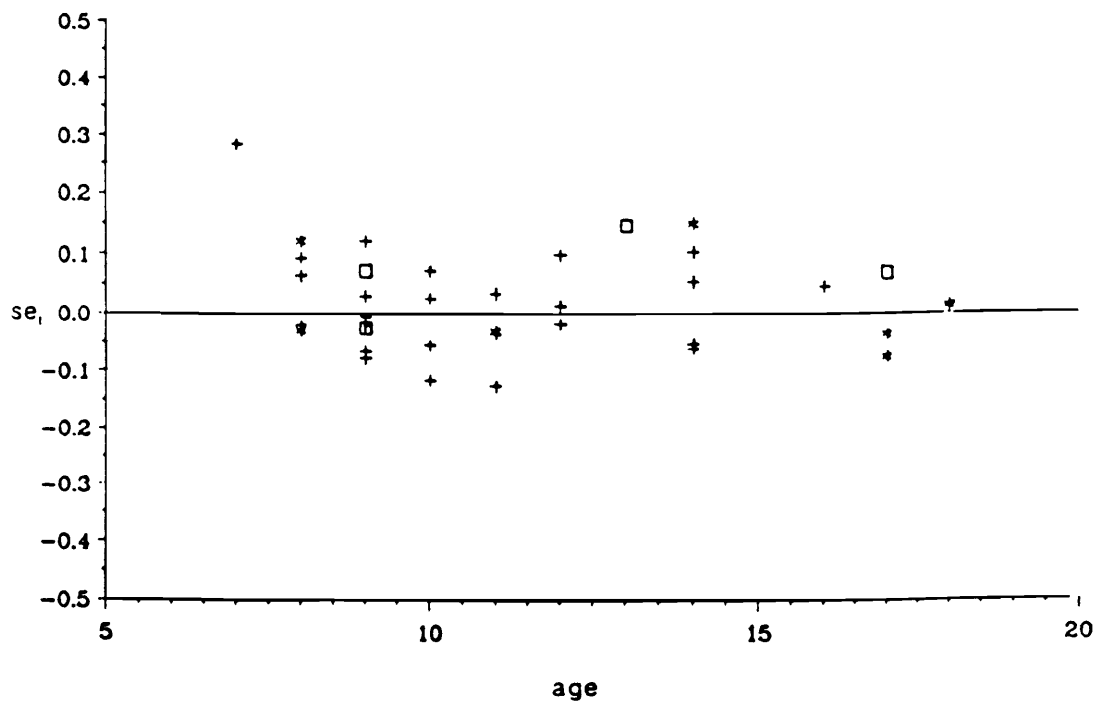
\* : slope site. + : flat site. □ : ridge.

Figure 7.31. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of stem infected slash pine trees.



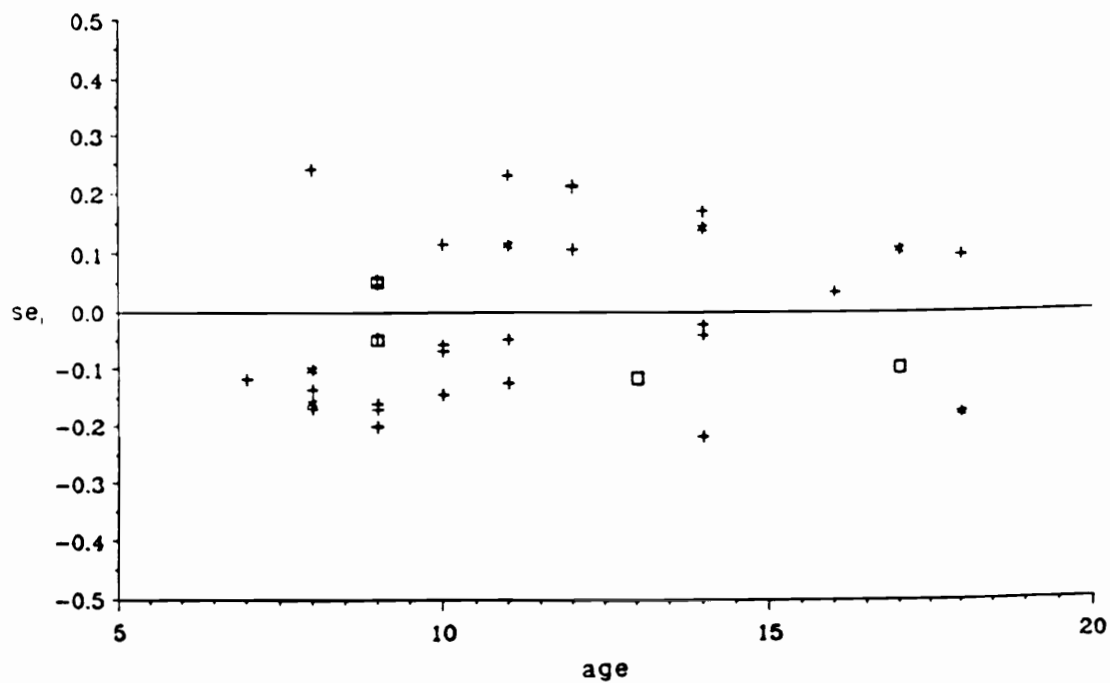
\* : slope site. + : flat site. □ : ridge.

Figure 7.32. Standardized residuals plotted against average height and landform for the OMNL model predicting the proportion of dead slash pine trees.



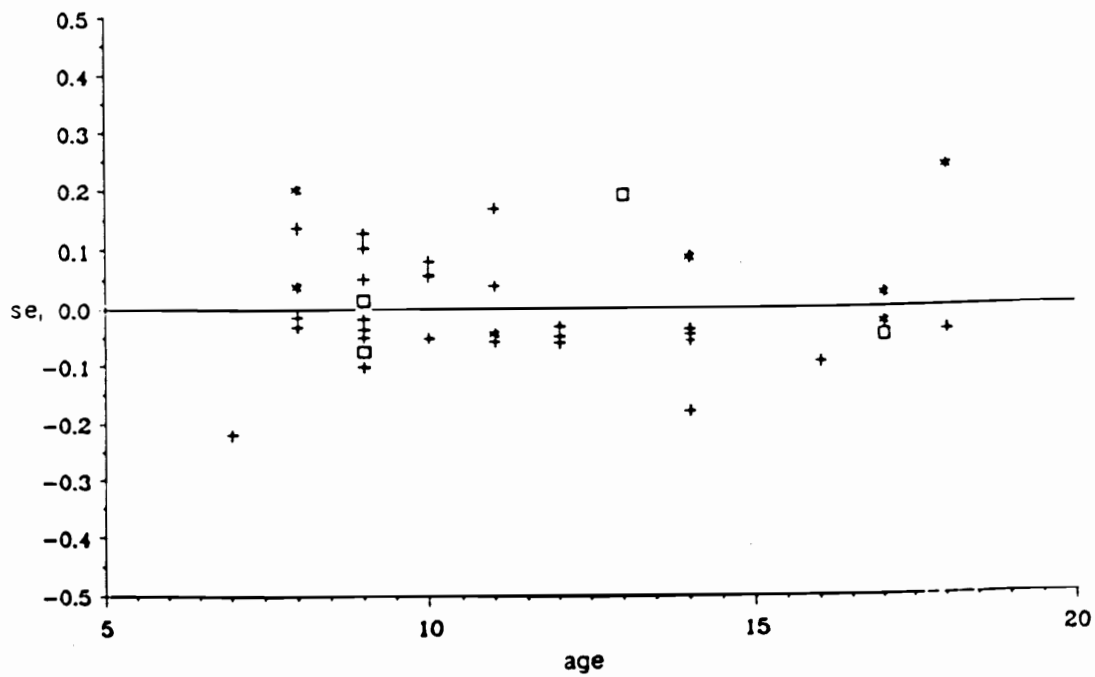
\* : slope site. + : flat site. □ : ridge.

Figure 7.33. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of healthy slash pine trees.



\* : slope site. + : flat site. □ : ridge.

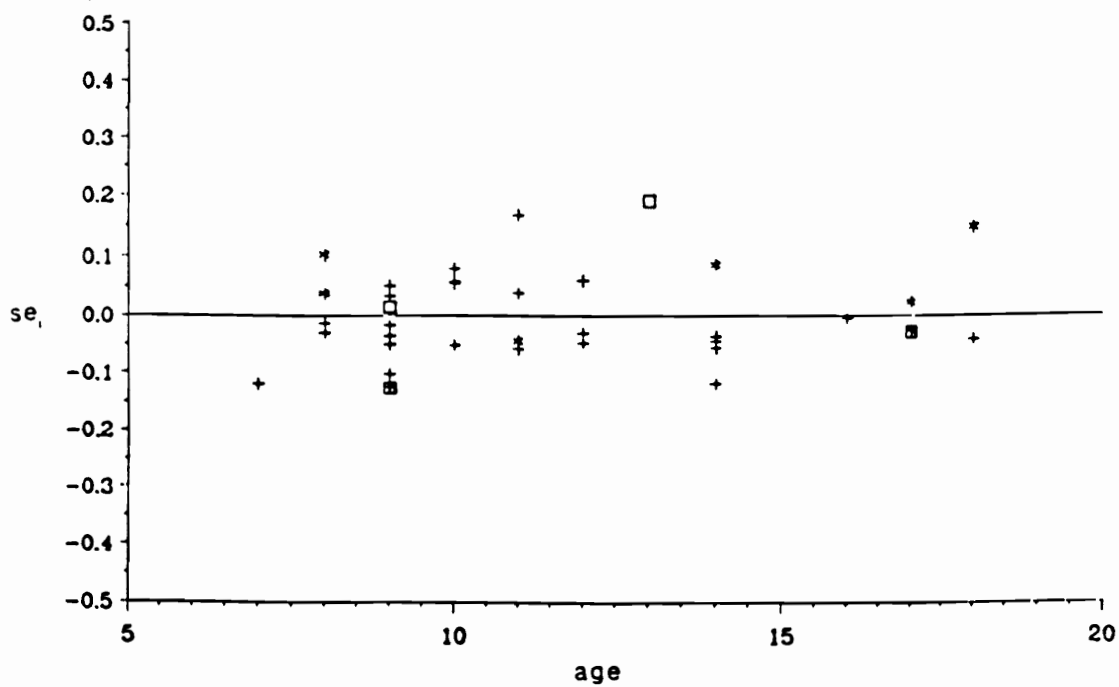
Figure 7.34. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of branch infected slash pine trees.



\* : slope site. + : flat site. □ : ridge.

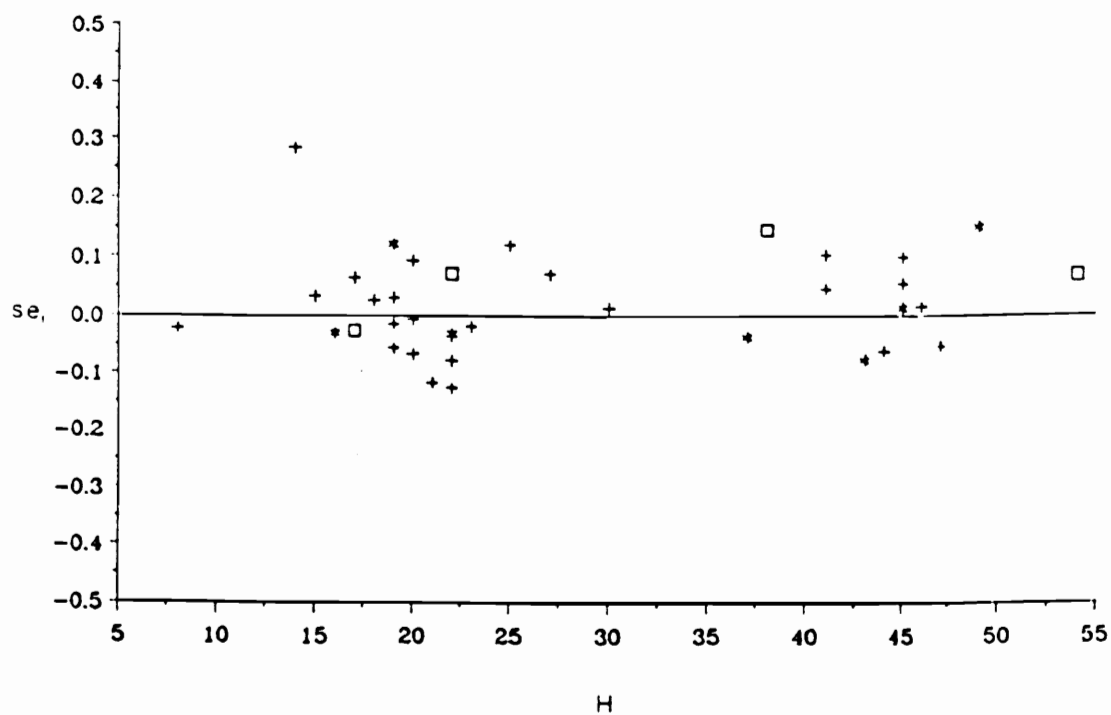
Figure 7.35. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of stem infected slash pine trees.





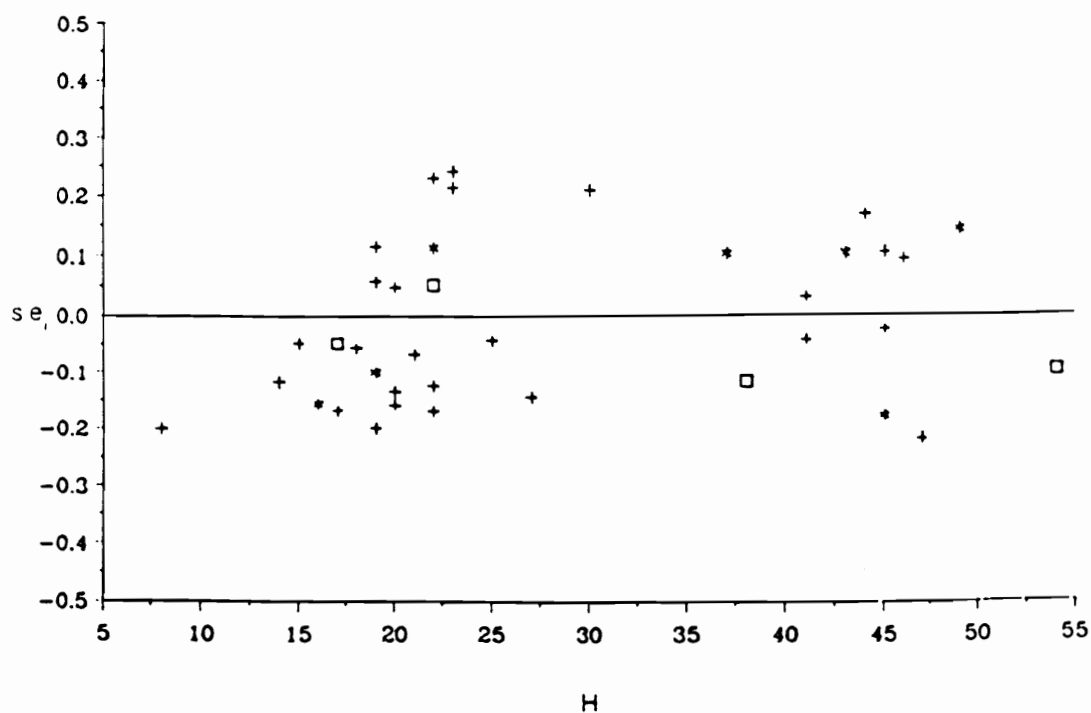
\* : slope site. + : flat site. □ : ridge.

Figure 7.36. Standardized residuals plotted against age and landform for the UMNL model predicting the proportion of dead slash pine trees.



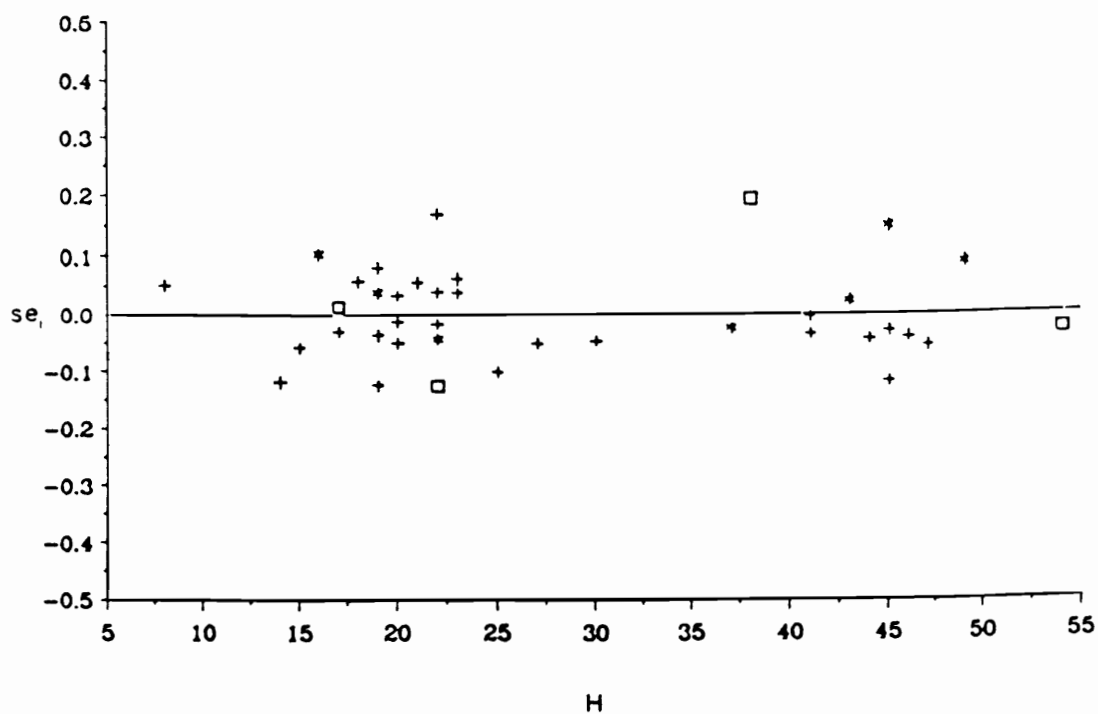
\* : slope site. + : flat site. □ : ridge.

Figure 7.37. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of healthy slash pine trees.



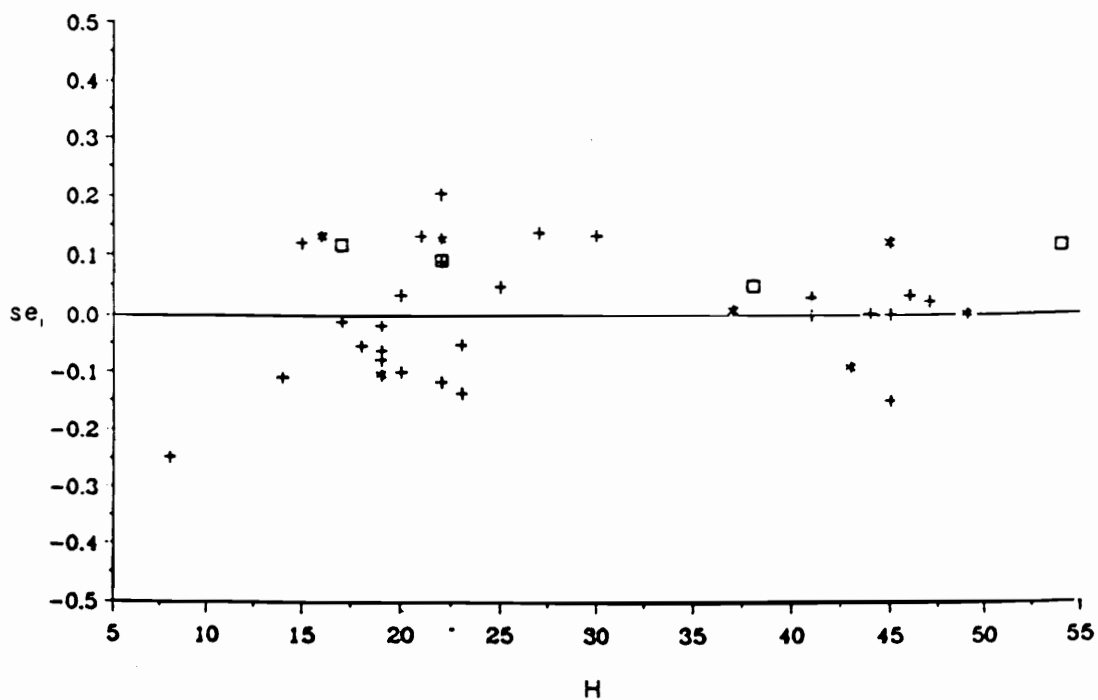
\* : slope site. + : flat site. □ : ridge.

Figure 7.38. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of branch infected slash pine trees.



\* : slope site. + : flat site. □ : ridge.

Figure 7.39. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of stem infected slash pine trees.



\* : slope site. + : flat site. □ : ridge.

Figure 7.40. Standardized residuals plotted against average height and landform for the UMNL model predicting the proportion of dead slash pine trees.

## **Chapter VIII**

### **Summary - Conclusions**

The first objective of this study was to introduce the philosophy of qualitative response models and to present certain dichotomous and polychotomous formulations that appear to be suitable for forestry applications. A substantial part of this study is devoted to the discussion of dichotomous models not only because these models are important on their own, but also because this discussion provides a necessary introduction to the subsequent discussion of multi-response models. In presenting this theory, special attention was paid to the following problems: i) how to motivate a qualitative response model which is theoretically correct and statistically manageable, ii) how to estimate and draw inferences about the parameters of interest, iii) what criteria to use when choosing among competing models and iv) how to identify observations which may have the potential for seriously distorting the fit of a model (outliers, high leverage and high influence observations).

The second objective was to illustrate the use of qualitative response models by considering two, forestry related, case studies. First, to assess the merchantability of loblolly pine trees growing on plantations in southern United States and second, to model the incidence and spread of fusiform rust on loblolly and slash pine plantations in east Texas. In both studies a variety of model forms

have been examined and important factors affecting the responses have been identified. The specific conclusions are discussed in detail at the end of chapters six and seven. From the overall discussion-application of qualitative response models conducted in this study the following conclusions can be drawn:

1) The theory of qualitative response models can be successfully applied in a variety of forestry problems mainly because of the fact that many important variable in forestry are discrete or recorded in a discrete manner. Of particular importance is that these models can be easily incorporated into existing forest growth and yield prediction systems to enhance their performance. Also, this theory may prove useful in the development of new prediction systems. In particular, we have in mind a diameter distribution growth and yield prediction system based on the type I extreme value distribution rather than the Weibull distribution which is commonly used. The theory of multinomial choice models as presented in chapter III provides the theoretical justification for such a system which may be preferable to current systems using the Weibull distribution on empirical only basis.

2) For dichotomous responses, both logit and probit formulations produce similar results. The linear probability formulation is not recommended because of serious weaknesses associated with the specification of this model the most serious of which is that the predicted probabilities or proportions are not restricted to fall within the  $[0,1]$  interval especially when data other than those used to fit the models are used.

3) For polychotomous responses, the use of the unordered multinomial logit (UMNL) model is recommended even when the analyst has substantial evidence that the response variable is ordered. If such is the case then, a well specified unordered model will reveal the suspected ordering through hypothesis testing. The IIA property associated the UMNL model is not a negative factor as long as the alternatives are distinct enough.

4) Further research is required on outlier and influence diagnostics in qualitative response models. Although considerable contributions to this matter have been made by Pregibon (1981) and Cook and Weisberg (1982), these apply only to binary logit models. No such theory has been developed yet for polychotomous response models.

5) The need to develop qualitative response models that account for the correlation among clustered observations is perhaps more immense in forestry than in any other science because the vast majority of forestry data bases consist of clustered data with large cluster sizes. Of particular importance is the investigation of the consequences from the violation of the independence assumption and also the development of efficient computational procedures to estimate the parameters when large clusters of observations are involved. Finally, future research efforts must be devoted to the analysis of correlated polychotomous observations with an emphasis on large cluster sizes.



## Bibliography

- Aitchison, J. and Bennett, J. (1970). Polychotomous Quantal Response by Maximum Indicant. *Biometrika*, **57**, 253-262.
- Aitchison, J. and Silvey, S. (1957). The generalization of Probit Analysis to the Case of Multiple Responses. *Biometrika*, **44**, 131-140.
- Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: *Second International Symposium on Information Theory*. B. N. Petrov and F. Csaki, eds., Budapest, Akademiai Kiado, 267-281.
- Amateis, R.L. and Burkhart, H.E. (1985). Site Index Curves for Loblolly Pine Plantations on Cutover Site-Prepared Lands. *Southern Journal of Applied Forestry*, **9**, 166-169.
- Amemiya, T. (1975). Qualitative Response Models. *Annals of Economic and Social Measurement*, **4**, 363-372.
- Amemiya, T. (1980). Selection of Regressor. *International Economic Review*, **21**, 331-345.
- Amemiya, T. (1981). Qualitative Response Models: A Survey. *Journal of Economic Literature*, **19**, 1483-1536.
- Amemiya, T. (1985). *Advanced Econometrics*. Harvard Univ. Press, Cambridge, Mass.
- Anderson, J.A. and Philips, P.R. (1981). Regression, Discrimination and Measurement Models for Ordered Categorical Variables. *Applied Statistics*, **30**, 22-31.
- Anderson, J.A. (1984). Regression and Ordered Categorical Variables. *Journal of the Royal Statistical Society B*, **46**, 1-30.
- Arvanitis, L.G. and Amateis, R.L. (1978). A Markov Model for Predicting the Development of Slash Pine Plantations Infected by Fusiform Rust. IN: *Statistical Methods, Mathematics and Data Processing*, IUFRO, Subject Group S6.02. Mitteilugen der Forstlichen Versuchs-und Forschungsalstalt, Baden-Wurtemberg. Abt. Biometric und Info. No. 28: 79-99.

- Arrow, K.J. (1951). *Residential Location Markets and Urban Transportation*. Academic Press, New York.
- Ashford, J.R. (1959). An Approach to the Analysis of Data for Semi-Quantal Responses in Biological Response. *Biometrics*, **15**, 573-581.
- Attneave, F. (1959). *Applications of Information Theory to Psychology*. Henry Holt, New York.
- Avery, T.E. and Burkhart, H.E. (1983). *Forest Measurements*. Third Edition, McGraw-Hill, New York.
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Data and Sources of Collinearity*. Wiley, New York.
- Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Predict Travel Demand*. The MIT Press, Cambridge Mass.
- Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, **39**, 357-365.
- Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics*, **7**, 327-339.
- Bliss, C.I. (1934a). The Method of Probits. *Science*, **79**, 38-39.
- Bliss, C.I. (1934b). The Method of Probits - A Correction. *Science*, **79**, 409-410.
- Bliss, C.I. (1935a). The Calculation of the Dosage - Mortality Curve. *Annals of Applied Biology*, **22**, 134-167.
- Bliss, C.I. (1935b). The Comparison of Dosage - Mortality Data. *Annals of Applied Biology*, **22**, 307-333.
- Bliss, C.I. (1937). The Calculation of the Time - Mortality Curve. *Annals of Applied Biology*, **24**, 815-852.
- Bock, R. (1969). Estimating Multinomial Response Relations. In: *Contributions to Statistics and Probability: Essays in memory of S. N. Roy*. R. Bose, ed., University of North Carolina Press.
- Borders, B.E. and Bailey, R.L. (1986). Fusiform Rust Prediction Models for Site-Prepared Slash and Loblolly Pine Plantations in the Southeast. *Southern Journal of Applied Forestry*, **10**, 145-151.
- Burkhart, H.E. (1987). Data Collection and Modelling Approaches for Forest Growth and Yield Prediction. In: *Predicting Forest Growth and Yield: Current Issues, Future Prospects*. B. N. Chappel and D. A. Maguire, eds., Institute of Forest Resources, Contribution 58, University of Washington, Seattle, 3-16.
- Burkhart, H.E. and Bredenkamp, B.V. (1989). Product-Class Proportions for Thinned and Unthinned Loblolly Pine Plantations. Accepted for publication by the *Southern Journal of Applied Forestry*.
- Burkhart, H.E., Cloeren, D.C. and Amateis, R.L. (1985). Yield Relationships in Unthinned Loblolly Pine Plantations on Cutover, Site-Prepared Lands. *Southern Journal of Applied Forestry*, **2**, 84-91.

- Burton, J.D., Shoulders, E. and Snow, G.A. (1985). Incidence and Impact of Fusiform Rust Vary with Silviculture in Slash Pine Plantations. *Forest Science*, 31 , 671-680.
- Chambers, J.M. (1973). Fitting Non-Linear Models: Numerical Techniques. *Biometrika*, 60 , 1-13.
- Clutter, J.L., Fortson, J.C., Pienaar, L.V., Brister, G.H. and Bailey, R.L. (1983) *Timber Management: A Quantitative Approach*. Wiley, New York.
- Connolly, M.A. and Liang, K.Y. (1988). Conditional Logistic Regression Models for Correlated Binary Data. *Biometrika*, 75 , 501-506.
- Cook, R.D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics*, 19 , 15-18.
- Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.
- Cox, D.R. (1970). *The Analysis of Binary Data*. Methuen, London.
- Daganzo, C. (1979). *Multinomial Probit: The Theory and Its Applications to Demand Forecasting*. Academic Press, New York.
- Daganzo, C., Bouthelier, F. and Sheffi, Y. (1977). Multinomial Probit and Qualitative Choice: A Computationally Efficient Algorithm. *Transportation Science*, 11 , 338-358.
- Daniels, R.F., Leuschner, W.A., Zarnoch, S.J., Burkhart, H.E. and Hicks, R.R. (1979). A Method for Estimating the Probability of Southern Pine Beetle Outbreaks. *Forest Science*, 25 , 265-269.
- David, J.M. and Legg, W.E. (1975). An Application of Multivariate Probit Analysis to the Demand of Housing: A Contribution to the Improvement of the Predictive Performance of Demand Theory, Preliminary Results. *Proceedings of the American Statistical Association (Business and Economics Section)*, pp. 295-300.
- Deacon, R. and Shapiro, P. (1975). Private Preference for Collective Goods Revealed through Voting and Referenda. *American Economic Review*, 65 , 943-955.
- Debreu, G. (1960). Review of R.D. Luce Individual Choice Behavior. Occupation Using Multiple Logit Models. *American Economic Review*, 50 , 186-188.
- DeGroot, M.H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.
- Domencich, T. and McFadden, L. (1975). *Urban Travel Demand- A Behavioral Analysis*. North Holland, Amsterdam.
- Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*. Second Edition, Wiley, New York.
- Duncan, G.M. (1980). Formulation and Statistical Analysis of the Mixed, Continuous/Discrete Dependent Variable Model in Classical Production Theory. *Econometrica*, 67 , 761-767.
- Dutt, J. (1976). Numerical Aspects of Multivariate Normal Probabilities in Econometrics. *Annals of Economic and Social Measurement*, 5 .

- Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. Second Edition, Cambridge, The MIT Press.
- Finney, D.J. (1947). *Probit Analysis*. Cambridge University Press.
- Finney, D.J. (1971). *Probit Analysis*. Third Edition, Cambridge University Press.
- Finney, D.J. (1978). *Statistical Method in Biological Assay*. Third Edition, Charles Griffin & Company Ltd.
- Fomby, T.B., Hill, R.C. and Johnson, S.R. (1984). *Advanced Econometric Methods*. Springer-Verlag, New York.
- Froelich, R.C. and Snow, G.A. (1986). Predicting site hazard to Fusiform Rust. *Forest Science*, **23**, 69-77.
- Gaddum, J.H. (1933). Reports on Biological Standards. III. Methods of Biological Assay Depending on a Quantal Response. *Medical Research Council, Special Report Series, no. 183*.
- Gallagher, R. (1968). *Information Theory and Reliable Communication*. Wiley, New York.
- Goldberger, A.S. (1964). *Econometric Theory*. Wiley, New York.
- Goldberger, A.S. (1973). Correlations Between Binary Choices and Probabilistic Predictions. *Journal of the American Statistical Association*, **68**, 84.
- Goldman, S. (1953). *Information Theory*. Prentice-Hall, New York.
- Greenland, S. (1985). An Application of Logistic Models to the Analysis of Ordinal Responses. *Biometric Journal*, **27**, 189-197.
- Griffiths, D.A. (1973). Maximum Likelihood Estimation for the Beta- Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease. *Biometrics*, **29**, 637-648.
- Gumbel, E.J. (1962). Statistical Theory of Extreme Values (main results). In: *Contributions to Order Statistics*. A.E. Sarhan and B.G. Greenberg, eds., Wiley, New York.
- Gurland, J., Lee, T. and Dahm, P. (1960). Polychotomous Quantal Response in Biological Assay. *Biometrics*, **16**, 382-398.
- Hamilton, D.A. Jr. (1984). Sampling and Estimation of Conifer Mortality Using Large-Scale Aerial Photography. *Forest Science*, **30**, 333-342.
- Hamilton, D.A. Jr. (1986). A Logistic Model of Mortality in Thinned and Unthinned Conifer Stands of Northern Idaho. *Forest Science*, **32**, 989-1000.
- Hamilton, D.A. Jr. and Edwards, B.M. (1976). Modeling the Probability of Individual Tree Mortality. USDA Forest Service Research Paper, INT-185, 22p.
- Hauck, W.W. and Donner, A. (1977). Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, **72**, 851-853.
- Hauser, J.R. (1978). Testing the Accuracy, Usefulness and Significance of Probabilistic Choice Models. *Operations Research*, **26**, 406-421.

- Hausman, J.A. and McFadden, D. (1984). Specification Tests for the Multinomial Logit Model. *Econometrica*, **52** , 1219-1240.
- Hausman, J.A. and Wise, D.A. (1978). A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*, **45** , 919-938.
- Hensher, D.A. and Johnson, L.W. (1981). *Applied Discrete Choice Modelling*. Wiley, New York.
- Hoaglin, D.C. and Welsch, R.E. (1978). The Hat Matrix in Regression and ANOVA. *American Statistician*, **32** , 17-22.
- Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.
- Hollis, C.A., Smith, W.H., Schmidt, R.A. and Pritchett, W.L. (1975). Soil and Tissue Nutrients, Soil Drainage, Fertilization and Tree Growth as Related to Fusiform rust Incidence in Slash Pine. *Forest Science*, **22** , 141-148.
- Hunt, E.V. and Lenhart, J.D. (1986). Fusiform rust Trends in East Texas. *Southern Journal of Applied Forestry*, **10** , 215-216.
- Jennings, D.E. (1986). Judging Inference Adequacy in Logistic Regression. *Journal of the American Statistical Association*, **81** , 471-476.
- Johnson, N.L. and Kotz, S. (1970). *Continuous Univariate Distributions, Vol. I*. Wiley, New York.
- Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. and Lee, T.C..L. (1985). *The Theory and Practice of Econometrics*. Second Edition, Wiley, New York
- Judge, G.G., Hill, R.C., Griffiths, W.E., Lutkepohl, H. and Lee, T.C.L. (1988). *Introduction to the Theory and Practice of Econometrics*. Second Edition, Wiley, New York
- Kakwani, N.C. (1967). The Unbiasedness of Zellner's Seemingly Unrelated Equations Estimators. *Journal of the American Statistical Association*, **62** , 141-142.
- Kennard, R.W. and Stone, M. (1969). Computer Aided Design of Experiments. *Technometrics*, **11** , 137-148.
- Kmenta, J. (1986). *Elements of Econometrics*. Second Edition, Macmillan, New York.
- Lave, C.A. (1970). The Demand for Urban Mass Transportation *Review of Economics and Statistics*, **52** , 320-323.
- Lenhart, J.D., Hunt, E.V. Jr. and Blackard, J.A. (1985). Establishment of Permanent Growth and Yield Plots in Loblolly and Slash Pine Plantations in East Texas. In: *Proceedings of the Third Biennial Southern Silviculture Research Conference*. E. Shoulders, Ed., USDA Forest Service General Technical Reports SO-54, 436-437.
- Lerman, S.R. and Manski, C.F. (1981). On the Use of Simulated Frequencies to Approximate Choice Probabilities. In: *Structural Analysis of Discrete Data with Econometric Applications*. C. Manski and D. McFadden, eds., Cambridge, The MIT Press, 305-319.
- Li, M.M. (1977). A Logit Model of Homeownership. *Econometrica*, **45** , 1081-1098.

- Lowell, K.E. and Mitchell, R.J. (1987). Stand Growth Projection: Simultaneous Estimation of Growth and Mortality Using a Single Probabilistic Function. *Canadian Journal of Forestry Research*, **17**, 1466-1470.
- Luce, R. and Suppes, P. (1965). Preference, Utility and Subjective Probability. In: *Handbook of Mathematical Psychology*, Vol. 3. R. Luce, R. Bush and E. Galanter, eds., Wiley, New York.
- Maddala, G.S. (1977). *Econometrics*. McGraw-Hill, New York.
- Magnus, J.R. (1978). Maximum Likelihood Estimation of the GLS Model with Unknown Parameters in the Disturbance Covariance Matrix. *Journal of Econometrics*, **7**, 281-312.
- Manski, C (1981). Structural Models for Discrete Data. *Sociological Methodology*, 58-109.
- Mantel, N. (1966). Models for Complex Contingency Tables and Polychotomous Response Curves. *Biometrics*, **22**, 83-110.
- May, J.T., Rahman, S. and Worst, R.H. (1973). Effects of Site Preparation and Spacing on Planted Slash Pine. *Forest Science*, **22**, 141-148.
- McCullagh, P. (1980). Regression Models for Ordinal Data (with discussion). *Journal of the Royal Statistical Society B*, **42**, 109-142.
- McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In: *Frontiers in Econometrics*. P. Zarembka, ed., New York, Academic Press, 105-142.
- McFadden, D. (1981). Econometric Models of Probabilistic Choice. In: *Structural Analysis of Discrete Data with Econometric Applications*. C. Manski and D. McFadden, eds., Cambridge, The MIT Press, 198-272.
- McFadden, D., Puig, C. and Kirschner, D. (1977). Determinants of the Long-Run Demand for Electricity. *Proceedings of the American Statistical Association (Business and Economics Section)*, pp. 109-117.
- McGillivray, R.G. (1970). Estimating the Linear Probability Function. *Econometrica*, **38**, 775-776.
- McKelvey, R. and Zavoina, W. (1975). A Statistical Model for the Analysis of Ordinal Dependent Variables. *Journal of Mathematical Sociology*, **4**, 103-120.
- Miller, T. (1973). Fusiform Rust in Planted Slash Pines: Influence of Site Preparation and Spacing. *Forest Science*, **18**, 70-75.
- Monserud, R.A. (1976). Simulation of Forest Tree Mortality. *Forest Science*, **22**, 438-444.
- Montgomery, D.C. and Peck, E.A. (1982). *Introduction to Linear Regression Analysis*. Wiley, New York.
- Morrison, D.G. (1972). Upper Bounds for Correlations Between Binary Outcomes and Probabilistic Predictions. *Journal of the American Statistical Association*, **7**, 68-70.
- Myers, R.H. (1986). *Classical and Modern Regression with Applications*. Duxbury Press, Boston.

- Nance, W.L., Froelich, R.C. and Shoulders, E. (1981). Effects of Fusiform Rust on Survival and Structure of Mississippi and Louisiana Slash Pine Plantations. USDA Forest Service Research Paper SO-172.
- Nerlove, M. and Press, S.J. (1973). Univariate and Multivariate Log-linear and Logistic models. Rand Corporation Technical Report R-1306-EDA/NIH, Santa Monica, California.
- Neter, J., Wasserman, W. and Kutner, M.H. (1985). *Applied Linear Statistical Models*. Second Edition, Irwin, Homewood, IL.
- Ochi, Y. and Prentice, R.L. (1984). Likelihood Inference in a Correlated Probit Regression Model. *Biometrika*, **71**, 531-543.
- Paul, S.R. (1979). A Clumped Beta-Binomial Model for the Analysis of Clustered Attribute Data. *Biometrics*, **35**, 821-825.
- Powers, H.R., McClure, J.P., Knight, H.A. and Dutrow, G.F. (1974). Incidence and Financial Impact of Fusiform Rust in the South. *Journal of Forestry*, **72**, 398-401.
- Pratt, J.W. (1981). Concavity of the Log-Likelihood. *Journal of the American Statistical Association*, **76**, 137-159.
- Pregibon, D. (1980). Goodness of Link Tests for Generalized Linear Models. *Applied Statistics*, **29**, 15-24.
- Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics*, **9**, 705-724.
- Press, S.J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart and Winston, New York.
- Qu, Y.S., Williams, G.W., Beck, G.J. and Goormastic, M. (1987). A Generalized Model of Logistic Regression for Correlated Data. *Communications in Statistics*, **A 16**, 3447-3476.
- Rao, C.R. (1965). *Linear Statistical Inference and its Applications*. Wiley, New York.
- Reed, D.D., Burkhart, H.E., Leuschner, W.A. and Hedden, R.L. (1981). A Severity Model for Southern Pine Beetle Infestations. *Forest Science*, **27**, 290-296.
- Rosner, B. (1984). Multivariate Methods in Ophthalmology with Application to Other Paired-Data Situations. *Biometrics*, **40**, 1025-1035.
- Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379-423, 623-656.
- Shannon, C.E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana.
- Schmidt, P. (1976). *Econometrics*. Marcel Dekker, New York.
- Schmidt, P. and Strauss, R.P. (1975). The Predictions of Occupation Using Multiple Logit Models. *International Economic Review*, **16**, 471-486.
- Schmidt, R.A., Squillace, A.E. and Swindel, B.F. (1979). Predicting the Incidence of Fusiform Rust in Five to Ten-Year-Old Slash and Loblolly Pine Plantations. *Southern Journal of Applied Forestry*, **3**, 138-140.

- Schmidt, R.A. and Klapproth, M.C. (1982). Delineation of Fusiform Rust Hazard Based on Estimated Volume Loss as a Guide to Rust Management Decisions in Slash Pine Plantations. *Southern Journal of Applied Forestry*, **6** , 59-63.
- Schmidt, R.A., Miller, T., Holley, R.C., Belanger R.P., and Allen J.E. (1988). Relation of Site Factors to Fusiform Rust Incidence in Young Slash and Loblolly Pine Plantations in the Coastal Plain of Florida and Georgia. *Plant Disease*, **72** , 710-714.
- Sheffi, Y. (1979). Estimating Choice Probabilities among Nested Alternatives. *Transportation Research*, **B13** , 113-205.
- Shoulders E. and Nance, W.L. (1987). Effects of Fusiform Rust on Survival and Structure of Mississippi and Louisiana Loblolly Pine Plantations. USDA Forest Service Research Paper SO-232.
- Snee D. R. (1977). Validation of Regression Models: Methods and Examples. *Technometrics*, **19** , 415-428.
- Steinberg, D. (1987). Interpretation and Diagnostics of the Multinomial and Logistic Regression. In: *Proceedings of the Twelveth Annual SAS User's Group International Conference*, SAS Institute, Cary, N.C.
- Stone, M. (1974). Cross-Validation Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, Series B*, **36** , 111-147.
- Strub, M.R., Green, E.J., Burkhardt, H.E. and Pirie, W.R. (1986). Merchantability of Loblolly Pine-An Application of Nonlinear Regression with a Discrete Dependent Variable. *Forest Science*, **32** , 254-261.
- Stynes, D.J. and Peterson, G.L. (1984). A Review of Logit Models with Implications for Modeling Recreation Choices. *Journal of Leisure Research*, **16** , 295-310.
- Theil, H. (1967). *Economics and Information Theory*. Rand McNally, Chicago and North-Holland, Amsterdam.
- Theil, H. (1969). A Multinomial Extension of the Linear Logit Model. *International Economic Review*, **10** , 251-259.
- Theil, H. (1970). On the Estimations of Relationships Involving Qualitative Variables. *American Journal of Sociology*, **76** , 103-154.
- Theil, H. (1971). *Principles of Econometrics*. Wiley, New York.
- Thomson, G.H. (1919). A Direct Deduction of the Constant Process Used in the Method of Right and Wrong Cases. *Psychology Review*, **26** , 454-464.
- Uhler, R.S. and Cragg, J.G. (1971). The Structure of Asset Portfolios of Households. *Review of Economic Studies*, **38** , 341-357.
- Wakeley, P.C. (1969). Results of Southern Pine Planting Experiments Established in the Middle Twenties. *Journal of Forestry*, **67** , 237-241.
- Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, **54** , 426-482.



- Walker, S.H. and Duncan, D.B. (1967). Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika*, **54** , 167-179.
- Weisberg, S. (1985). *Applied Linear Regression*. Second Edition, Wiley, New York.
- Wells, O.O. and Dinus, R.J. (1978). Early Infection as a Predictor of Mortality Associated with Fusiform Rust of Southern Pines. *Journal of Forestry*, **76** , 8-12.
- Williams, D.A. (1975). The analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity. *Biometrics*, **31** , 949-952.
- Witte, A.D. (1980). Estimating the Economic Model of Crime with Individual Data. *Quarterly Journal of Economics*, **94** , 57-84.
- Wrigley, N. (1982). *Categorical Data Analysis for Geographers and Environmental Scientists*. Longman, London.
- Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, **57** , 348-368.
- Zellner, A. and Lee, T.H. (1965). Joint Estimation of Relationships Involving Discrete Random Variables. *Econometrica*, **32** , 382-394.
- Zutter, B.R., Gjerstad, D.H. and Glover, G.R. (1987). Fusiform Rust Incidence in Loblolly Pine Plantations Following Herbaceous Weed Control. *Forest Science*, **33** , 790-800.

## Vita

The author was born on November 27, 1960 in Thessaloniki, Greece. Upon graduation from high school on June, 1978, he entered the five year program of the Forestry School at the Aristotelean University of Thessaloniki from where he graduated on June, 1983 with a B.Sc. degree in Forestry. On September of the same year he was accepted as a Master's candidate by the State University of New York, College of Environmental Science and Forestry at Syracuse, New York, under the guidance of professor Tiberius Cunia. His thesis, titled "Evaluating the Error of Tree Biomass Regressions by Simulation; Sample Trees Selected by Stratified Cluster Sampling" involved research funded by the Research Foundation of the State University of New York, the United States Department of Agriculture Forest Service and the Department of Energy. After receiving his M.Sc. degree in Resource Management and Policy, he was enrolled as a Ph.D. student in the Department of Forestry at Virginia Polytechnic Institute and State University. Concurrently with his Ph.D program and as a result of a heavy emphasis on formal statistical coursework, the author was awarded a M.Sc. degree in Statistics by the Statistics Department at VPI & SU on May, 1988. During his three years of doctoral studies, he served as Graduate Assistant in the Department of Forestry. At the time of this writing, he is a post-doctorate fellow at the same Department.

